Lifted Primal-Dual Method for Bilinearly Coupled Smooth Minimax Optimization

Kiran Koshy Thekumparampil
University of Illinois at Urbana-Champaign
thekump2@illinois.edu

Niao He
ETH Zürich
niao.he@inf.ethz.ch

Sewoong Oh
University of Washington
sewoong@cs.washington.edu

Abstract

We study the bilinearly coupled minimax problem: $\min_x \max_y f(x) + \langle y, Ax\rangle - h(y)$, where f and h are both strongly convex smooth functions and admit first-order gradient oracles. Surprisingly, no known first-order algorithms have hitherto achieved the lower complexity bound of $\Omega((\sqrt{\frac{L_x}{\mu_x}} + \frac{\|A\|}{\sqrt{\mu_x\mu_y}} + \sqrt{\frac{L_y}{\mu_y}})\log(\frac{1}{\varepsilon}))$ for solving this problem up to an ε primal-dual gap in the general parameter regime, where L_x, L_y, μ_x, μ_y are the corresponding smoothness and strongly convexity constants.

We close this gap by devising the first *optimal* algorithm, the *Lifted Primal-Dual (LPD) method*. Our method lifts the objective into an extended form that allows both the smooth terms and the bilinear term to be handled optimally and seamlessly with the same primal-dual framework. Besides optimality, our method yields a desirably simple *single-loop* algorithm that uses only one gradient oracle call per iteration. Moreover, when f is just convex, the same algorithm applied to a smoothed objective achieves the nearly optimal iteration complexity. We also provide a direct single-loop algorithm, using the LPD method, that achieves the iteration complexity of $\mathcal{O}(\sqrt{\frac{L_x}{\varepsilon}} + \frac{\|A\|}{\sqrt{\mu_y\varepsilon}} + \sqrt{\frac{L_y}{\varepsilon}})$. Numerical experiments on quadratic minimax problems and policy evaluation problems further demonstrate the fast convergence of our algorithm in practice.

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

1 INTRODUCTION

Smooth minimax optimization has gained renewed interest driven by a wide spectrum of applications in machine learning, especially those arising in adversarial training, generative adversarial networks, and reinforcement learning. A plethora of first-order algorithms have been developed in the classical and recent literature, ranging from convex to nonconvex settings, from deterministic to stochastic oracles, from single-loop to multiple-loop schemes. However, our theoretical understanding of the iteration complexity of minimax optimization is far from complete even in the canonical *strongly-convex-strongly-concave (SC-SC)* setting. In particular, the optimal dependence on the condition numbers of different blocks of variables has not been fully characterized.

Consider the smooth convex-concave minimax problem (a.k.a. saddle point problem):

$$\min_x \max_y \ \phi(x, y), \tag{1}$$

where $\phi(x, y)$ is μ_x-strongly convex in x and μ_y-strongly concave in y. Let L_x, L_y, L_{xy} be the corresponding gradient Lipshitz constants with respect to different blocks of variables. To find an ε-approximate saddle point, Zhang et al. (2019) recently showed that any first-order algorithm with the linear span assumption requires at least

$$\Omega\left(\left(\sqrt{\frac{L_x}{\mu_x} + \frac{L_{xy}^2}{\mu_x\mu_y} + \frac{L_y}{\mu_y}}\right)\log\left(\frac{1}{\varepsilon}\right)\right) \tag{2}$$

calls to a gradient oracle for $\phi(x, y)$. Notably, the lower iteration complexity bound applies to even the class of bilinearly coupled quadratic minimax problems, which was used to construct the hard instance.

In the special parameter regime when $L_x = L_y = L_{xy} := L$ and $\mu_x = \mu_y := \mu$, this lower bound is matched by several popular algorithms including the extragradient and optimistic methods (Korpelevich 1976; Nemirovski 2004; Gidel et al. 2018; Mokhtari et al. 2020)

and the accelerated dual extrapolation (Nesterov et al. 2018), with iteration complexity of $O((L/\mu)\log(1/\varepsilon))$.

However, in the general parameter regime, despite several recent attempts (Cohen et al. 2020; Lin et al. 2020; Wang et al. 2020; Zhang et al. 2021b), no known algorithms have yet exactly matched the lower bound. For instance, the algorithm in Lin et al. (2020) achieves an upper complexity bound of $\tilde{O}((\mathcal{L}/\sqrt{\mu_x\mu_y})\log^3(1/\varepsilon))$. One of the best-known results is obtained in Wang et al. (2020), that gives the complexity of $\tilde{O}(\sqrt{(L_x/\mu_x)+(\mathcal{L}\,L_{xy}/\mu_x\mu_y)+(L_y/\mu_y)}\log(1/\varepsilon))$, where $\tilde{O}$ hides a polylogarithmic factor in problem parameters and $\mathcal{L}=\max\{L_x,L_{xy},L_y\}$. These advances all rely on carefully designed multi-loop algorithms.

We close this gap for a class of SC-SC minimax problems with bilinear coupling (Bi-SC-SC). Specifically, we consider problems of the general form:

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}}\,[\phi(x,y)=f(x)+\langle y,Ax\rangle-h(y)], \qquad (3)$$

where $f(x)$ is L_x-smooth and μ_x-strongly convex, $h(y)$ is L_y-smooth and μ_y-strongly convex, $\mathcal{X}$ and $\mathcal{Y}$ are closed convex sets. We assume access to first-order gradient oracles of f and g as well as the matrix A. Note that the lower bound in (2) also holds for this class of problems. This class of problems by itself has found numerous applications in machine learning, as detailed in Section 2.

The main challenge in designing an optimal algorithm is that the objective consists of two different classes of functions: smooth convex terms f and h, and bilinear coupling $\langle y,Ax\rangle$. These two classes are traditionally optimized using conceptually different algorithms. On one hand, accelerated gradient methods (like AGD, Nesterov et al. 2018) are optimal at solving smooth strongly convex problems like $\min_x f(x)$ or $\min_y h(y)$. On the other hand, bilinear problems of the form, $\min_x\max_y\langle y,Ax\rangle$ or the like (with additional proximal-friendly terms), are optimally solved using a seemingly different class of algorithms such as primal-dual methods; see e.g., Chen et al. (1997), Bauschke et al. (2011), Chen et al. (2014), Chambolle et al. (2016), and He et al. (2016), just to name a few. Such a conceptual difference makes it hard to design an algorithm that achieves optimal dependence on the smoothness and strong convexity parameters of each of the three terms in the objective.

1.1 Our Contributions

We introduce a new algorithm that reconciles these different components by lifting the objective to an extended saddle point formulation. Our key idea hinges on the recent interpretation (Lan et al. 2018) of accelerated gradient descent for convex minimization as a variant of primal-dual method for an equivalent minimax problem. Based on the reformulation, we can handle both the smooth terms and bilinear coupling term under the same umbrella of primal-dual method. We make the following key contributions.

- We provide the first optimal algorithm for the class of bilinearly coupled SC-SC minimax problems, called the *lifted primal-dual (LPD) method*, achieving the iteration complexity of $\mathcal{O}\big((\sqrt{L_x/\mu_x}+\|A\|/\sqrt{\mu_x\mu_y}+\sqrt{L_y/\mu_y})\log(1/\varepsilon)\big)$ (Theorem 2), tightly matching the lower bound. The LPD method is also *single-loop*, using only one gradient oracle call per iteration, which is more desirable in practice.
- For bilinearly coupled *convex-strongly-concave* (Bi-C-SC) minimax problems where f is only convex, namely, $\mu_x=0$, we can apply the LPD method to a *smoothed* objective $\phi(x,y)+\lambda\varepsilon\|x\|^2$, which transforms the objective into a SC-SC one (Nesterov 2005). The LPD method is the first to achieve optimal complexity up to logarithmic factors in this setting (Remark 1) as shown in Table 1. However, smoothing might not be desirable in practice (see Section 5). To this end, we design a direct algorithm by selecting appropriate stepsizes in LPD. This achieves an iteration complexity that is suboptimal but the best among those not using smoothing (Theorem 3).

Detailed comparisons with existing algorithms are presented in Table 1.

1.2 Related Work

Below we highlight key distinctions of our work to the most closely related literature. Our list of related work is by no means comprehensive. There exists optimal algorithm for the case when both f and h are just convex ($\mu_x=\mu_y=0$) (Chen et al. 2014, 2017). However, when either of f or h is strongly convex it is not readily clear how to optimally solve the problem.

Bilinear coupling with simple terms. Existing work on bilinearly coupled minimax problems primarily focuses on the case when f and/or h are proximal-friendly, i.e., it is easy to compute the proximal operator. If both f and h are proximal-friendly and strongly convex, then the primal-dual method (Chambolle et al. 2016) and accelerated forward-backward algorithm (Palaniappan et al. 2016) already achieve the optimal rate $\mathcal{O}((\|A\|/\sqrt{\mu_x\mu_y})\log(1/\varepsilon))$ (Xie et al. 2021). If only h is proximal-friendly and f is smooth, but both are strongly convex, Chambolle et al. (2016) provides a linearly convergent but sub-optimal algorithm.

Table 1: Gradient oracle complexity of first-order methods for solving bilinearly-coupled smooth minimax problem upto an ε primal-dual gap. We define $\mathcal{L} := \max(L_x, \|A\|, L_y)$. ♣Each term of this tight lower bound is implicitly implied by existing lower-bounds for special cases of the Bi-C-SC problem (Nesterov et al. 2018; Ouyang et al. 2021).

Method	# Loops	Gradient Complexity
Strongly-Convex–Strongly-Concave (Bi-SC-SC)		
MP/EG, OGDA (Mokhtari et al. 2020)	Single	$\mathcal{O}\big(\frac{L_x+\|A\|+L_y}{\min(\mu_x,\mu_y)}\big)\log\big(\frac{1}{\varepsilon}\big)$
MP Bal. (App. F), MP RL (Cohen et al. 2020)	Single	$\mathcal{O}\big(\frac{L_x}{\mu_x}+\frac{\|A\|}{\sqrt{\mu_x\mu_y}}+\frac{L_y}{\mu_y}\big)\log\big(\frac{1}{\varepsilon}\big)$
Proximal Best Response (Wang et al. 2020)	Multi	$\widetilde{\mathcal{O}}\big(\sqrt{\frac{L_x}{\mu_x}}+\sqrt{\frac{\|A\|\mathcal{L}}{\mu_x\mu_y}}+\sqrt{\frac{L_y}{\mu_y}}\big)\log\big(\frac{1}{\varepsilon}\big)$
DIPPA (Xie et al. 2021)	Multi	$\widetilde{\mathcal{O}}\big((\frac{L_x^2L_y}{\mu_x^2\mu_y})^{\frac{1}{4}}+\frac{\|A\|}{\sqrt{\mu_x\mu_y}}+(\frac{L_y^2L_x}{\mu_y^2\mu_x})^{\frac{1}{4}}\big)\log\big(\frac{1}{\varepsilon}\big)$
Lifted PD (Theorem 2)	Single	$\mathcal{O}\big(\sqrt{\frac{L_x}{\mu_x}}+\frac{\|A\|}{\sqrt{\mu_x\mu_y}}+\sqrt{\frac{L_y}{\mu_y}}\big)\log\big(\frac{1}{\varepsilon}\big)$
Lower bound (Zhang et al. 2019)	N/A	$\Omega\big(\sqrt{\frac{L_x}{\mu_x}}+\frac{\|A\|}{\sqrt{\mu_x\mu_y}}+\sqrt{\frac{L_y}{\mu_y}}\big)\log\big(\frac{1}{\varepsilon}\big)$
Convex–Strongly-Concave (Bi-C-SC)		
MP/EG, OGDA (Mokhtari et al. 2020)	Single	$\mathcal{O}\big(\frac{L_x+\|A\|+L_y}{\varepsilon}\big)$
PDHG-type (Zhao 2019)	Multi	$\mathcal{O}\big(\frac{L_x}{\varepsilon}+\frac{\|A\|}{\sqrt{\mu_y\varepsilon}}+\sqrt{\frac{L_y}{\mu_y}}\log(\frac{1}{\varepsilon})\big)$
DIAG (Thekumparampil et al. 2019)	Multi	$\mathcal{O}\big(\sqrt{\frac{L_y}{\mu_y}}\big(\sqrt{\frac{L_x}{\varepsilon}}+\frac{\|A\|}{\sqrt{\mu_y\varepsilon}}\big)\big)\log^2\big(\frac{1}{\varepsilon}\big)$
Lifted PD (Theorem 3)	Single	$\mathcal{O}\big(\sqrt{\frac{L_x}{\varepsilon}}+\frac{\|A\|}{\sqrt{\mu_y\varepsilon}}+\sqrt{\frac{L_y}{\varepsilon}}\big)$
Lifted PD + Smoothing (Remark 1)	Single	$\mathcal{O}\big(\sqrt{\frac{L_x}{\varepsilon}}+\frac{\|A\|}{\sqrt{\mu_y\varepsilon}}+\sqrt{\frac{L_y}{\mu_y}}\big)\log\big(\frac{1}{\varepsilon}\big)$
Lower bound♣	N/A	$\Omega\big(\sqrt{\frac{L_x}{\varepsilon}}+\frac{\|A\|}{\sqrt{\mu_y\varepsilon}}+\sqrt{\frac{L_y}{\mu_y}}\log(\frac{1}{\varepsilon})\big)$

Our work differs from this line of results as we do not require computing the proximal operators of neither f nor h, but instead only their gradients. One exception is DIPPA, a complex multi-loop algorithm (Xie et al. 2021), which achieves $\mathcal{O}((((L_xL_y/\mu_x\mu_y)(L_x/\mu_x + L_y/\mu_y))^{1/4} + \|A\|/\sqrt{\mu_x\mu_y})\log(1/\varepsilon))$ complexity under the same setting (Bi-SC-SC) as the one we study. Additionally, for the special case of quadratic Bi-SC-SC problem, Wang et al. (2020) provides a recursive multi-loop algorithm which achieves a sub-optimal iteration complexity of $\mathcal{O}((\sqrt{L_x/\mu_x} + \|A\|/\sqrt{\mu_x\mu_y}+\sqrt{L_y/\mu_y})\,(\mathcal{L}/\mu_x\mu_y)^{o(1)}\,\log(1/\varepsilon))$, where $\mathcal{L} = \max(L_x, L_{xy} = \|A\|, L_y)$.

Beyond bilinear coupling. Beyond bilinear coupling, most existing work either treat the objective as a whole or consider special couplings. In the SC-SC setting with a general coupling (1), there are many algorithms which achieve linear convergence; one of the first such algorithm is the Extragradient (EG) method (Korpelevich 1976; Tseng 1995). Here, Gradient Descent Ascent (GDA) achieves an iteration complexity of $\mathcal{O}(\kappa_{\max}^2\log(1/\varepsilon))$ (Facchinei et al. 2007, Chapter 12), while Mirror-Prox (MP/EG) (Nemirovski 2004), Dual Extrapolation (DE) (Nesterov et al. 2006; Nesterov 2007), and Optimistic Gradient Descent Ascent (OGDA) (Daskalakis et al. 2017; Gidel et al. 2018; Mokhtari et al. 2020) achieve an iteration complexity of $\mathcal{O}(\kappa_{\max}\log(1/\varepsilon))$, where $\kappa_{\max} = \mathcal{L}/\min(\mu_x,\mu_y)$. Further, the above complexities can be improved (replacing $\kappa_{\max}$ with $L_x/\mu_x + L_{xy}/\sqrt{\mu_x\mu_y} + L_y/\mu_y$) with proper balancing of distance functions (see Appendix F). This improved complexity can also be attained by a modified MP via relative Lipschitzness (we refer to as MP RL)(Cohen et al. 2020). Two multi-loop algorithms: Minimax-APPA (Lin et al. 2020) and Catalyst-type method (Alkousa et al. 2020), which are based on accelerated minimization methods, achieve the iteration complexities of $\widetilde{\mathcal{O}}((\mathcal{L}/\sqrt{\mu_x\mu_y})\log^3(1/\varepsilon))$ and $\widetilde{\mathcal{O}}(\sqrt{L_y/\mu_y}(\sqrt{L_x/\mu_x} + L_{xy}/\sqrt{\mu_x\mu_y})\log^2(1/\varepsilon))$, respectively. The state-of-the-art complexity for general coupling is achieved by a multi-loop algorithm (Wang et al. 2020), but there is a gap to the known lower bound of (2); see Table 1 and Section 1.

Convex-Strongly-Concave minimax problems. As an example of special couplings, if the coupling is linear only in x but nonlinear in y, and f is proximal-friendly and convex and h is smooth and strongly convex, Juditsky et al. (2011) and Hamedani et al. (2021) achieve $\mathcal{O}((\|A\|/\sqrt{\mu_y\varepsilon}+(L_y/\mu_y))\log(1/\varepsilon))$ and $\mathcal{O}((\|A\|+L_y)/\sqrt{\mu_y\varepsilon})$ complexities, respectively, under our setting. In the same setting as these works, (Zhao 2019) provides a PDHG-type algorithm which works even when the coupling is non-linear instead of our bilinear one $\langle y, Ax\rangle$. This leads to a sub-optimal (in ε) complexity of $\mathcal{O}(L_x/\varepsilon+\|A\|/\sqrt{\mu_y\varepsilon}+\sqrt{L_y/\mu_y}\log(1/\varepsilon))$ for our Bi-C-SC setting. For minimax problems that are not necessarily SC-SC, recent works (Du et al. 2019; Azizian et al. 2020a; Yang et al. 2020a) show that linear convergence can still be achieved under additional assumptions. A recent work (Thekumparampil et al. 2019) discussed general convex-concave minimax problems with one-sided strong convexity (i.e., C-SC setting) and obtained a $\mathcal{O}(1/\sqrt{\varepsilon})$ complexity (see Table 1). In principle, any of the known algorithms (Lin et al. 2020; Mokhtari et al. 2020; Wang et al. 2020; Yang et al. 2020b; Xie et al. 2021) for the Bi-SC-SC setting (3) can be applied to solve Bi-C-SC problem after a smoothing transformation (Nesterov 2005). However, due to their sub-optimality in the original Bi-SC-SC case itself, their complexities for C-SC case are sub-optimal as well. We omit discussions of other minimax optimization settings as they are less relevant.

1.3 Notations

We use $\langle x, y\rangle$ to denote the inner product between vectors x and y, and $\|x\|$ to denote Euclidean norm of x. For a convex set $\mathcal{X}$, $\mathcal{P}_{\mathcal{X}}(\cdot)$ denotes its projection operator. We use the standard big-O $\mathcal{O}$ and Ω notations. *Iteration complexity or (gradient) complexity* of an algorithm is the number of iterations or gradients used by it find an *ε-approximate saddle point* $(\hat{x}, \hat{y})$, which means that its primal-dual gap $\max_y \phi(\hat{x}, y) - \min_x \phi(x, \hat{y}) \le \varepsilon$. Standard definitions of L-smoothness, μ-strong convexity, Fenchel/convex conjugate, Bregman divergence and its distance generating function, and proximal operators are given in Appendix A.

2 PROBLEM AND APPLICATIONS

We are mainly interested in the *bilinearly coupled strongly-convex–strongly-concave (Bi-SC-SC) minimax problem* of the form (3). Throughout, we make the following assumption.

Assumption 1. *f is L_x-smooth and μ_x-strongly convex, and h is L_y-smooth and μ_y-strongly convex on the entire Euclidean space.*

In addition, we assume that sets $\mathcal{X}$ and $\mathcal{Y}$ are closed convex and the projection onto these sets is easily computable. Functions f and h have well defined gradient on $\mathcal{X}$ and $\mathcal{Y}$ and they can be accessed through gradient oracles. Two distinctions that differ from most existing work are (i) no requirement on computing proximal operator of either f or h, and (ii) the linear coupling term. This type of problems find numerous applications in machine learning. Below we list only a few.

2.1 Quadratic Minimax Problems

Quadratic minimax problems are fundamental problems which arise in numerical analyses (Bai et al. 2003; Benzi et al. 2005; Bai 2009; Wang et al. 2020), optimal control problems (Rockafellar 1987; Liu et al. 2015), and constrained matrix games (Xie et al. 2021). They also appear naturally when solving subspace proximal sub-problems of Sequential Subspace Optimization for quadratic saddle-point problems (Choukroun et al. 2020), and when solving sub-problems of minimax (cubic regularized) Newton method (Huang et al. 2020; Schäfer et al. 2020; Zhang et al. 2020). Here $f(x) = x^T Bx$ and $h(y) = y^\top Cy$ correspond to positive definite (p.d.) matrices $B \succ 0$ and $C \succ 0$. Thus the minimax objective is quadratic in x and y:

$$\phi(x, y) = x^\top Bx + y^\top Ax - y^T Cy. \tag{4}$$

Despite their simplicity, quadratic minimax problems are not trivial to solve (Zhang et al. 2021a). Further, even nonconvex-nonconcave minimax problems can behave like an Bi-SC-SC problem near a strict local saddle point (Azizian et al. 2020b).

2.2 Robust Least Squares

Consider the robust least squares problem (El Ghaoui et al. 1997; Yang et al. 2020a) with a coefficient matrix A and noisy vector y, where y is corrupted by a deterministic perturbation δ of a bounded norm ρ:

$$\min_x \max_{\delta:\|\delta\|\le\rho} \|Ax - y\|^2, \text{ where } \delta = y - y_0.$$

The corresponding penalized version of the objective is a Bi-SC-Concave minimax problem if A is p.d.:

$$\min_x \max_y \phi(x, y) := \|Ax - y\|^2 - \lambda\|y - y_0\|^2.$$

Selecting $\lambda > 1$, we get a Bi-SC-SC problem.

2.3 Policy Evaluation

Bi-SC-SC arise in policy evaluation problem in reinforcement learning (Du et al. 2017, 2019) when finding minimum the mean squared projected Bellman error

(MSPBE). Empirical estimator of minimum MSPBE has the form:

$$\arg\min_{\theta} \frac{1}{2}\|A\theta - b\|^2_{C^{-1}} + \frac{\rho}{2}\|\theta\|^2, \tag{5}$$

where A, b, C are defined as follows. Suppose we have a trace of n tuples of current-state s_t, action a_t, next-state s_{t+1}, and reward r_t under some policy π on some MDP. Then we define $b = 1/n \sum_t^n r_t \phi_t$

$$A = \frac{1}{n}\sum_t^n \phi_t(\phi_t - \gamma\phi_{t+1})^\top, \text{ and } C = \frac{1}{n}\sum_t^n \phi_t\phi_t^\top$$

where ϕ_t is the feature of state s_t, γ is the discount factor. In practice, inverting C can be computationally costly. Therefore, one may resort to solving the following minimax reformulation, eliminating the need for matrix inversion.

$$\min_{\theta}\max_{w} \frac{\rho}{2}\|\theta\|^2 - w^\top A\theta - (\frac{1}{2}\|w\|^2_C - w^\top b) \tag{6}$$

This is a Bi-SC-SC problem if C is positive definite.

Note that, all these problems becomes a Bi-Convex–Strongly-Concave (Bi-C-SC) or Bi-Strongly-Convex–Concave (Bi-SC-C) problem, if the Hessian of convex quadratic of the primal or dual variables, respectively, becomes positive semi-definite.

3 BUILDING BLOCKS

We present known results that serve as the intuition behind the design of Algo. 1. We first revisit the primal-dual method (Chambolle et al. 2016), which is originally designed to solve bilinearly coupled minimax problems with simple terms whose proximal operators are easy to compute. After that we discuss how this method can be used to minimize smooth convex objectives with accelerated convergence (Lan et al. 2018).

3.1 Primal-Dual method

Consider the bilinearly coupled minimax problem:

$$\min_x \max_y \; F(x) + \langle y, Ax\rangle - H(y) \tag{7}$$

with a unique solution $z^* = (x^*, y^*)$. Additionally, let r and s be 1-strongly convex distance generating functions (d.g.f.) that induce Bregman divergences $V^r_{x_0}(x)$ and $V^s_{y_0}(y)$. Then, we assume that F and H are relatively μ_x- and μ_y-strongly convex with respect to $V^r_{x_0}(x)$ and $V^s_{y_0}(y)$, respectively. We also assume the access to their Bregman proximal operators, with respect to the corresponding divergences.

The PD method can be viewed as an approximation of proximal point method (PPM) (Rockafellar 1976). We emphasize this connection as the analyses of our main results closely follow that of PPM (Lemma 1). Readers who are familiar with this connection may skip to after Lemma 1. The PPM updating rule is as follows:

$$(x_{k+1}, y_{k+1}) = \arg\min_x \arg\max_y \left\{\frac{1}{\eta_x}V^r_{x_k}(x) + F(x) + \langle y, Ax\rangle - H(x) - \frac{1}{\eta_y}V^s_{y_k}(y)\right\}.$$

This is equivalent to the implicit update rule

$$\begin{cases} x_{k+1} = \arg\min_x \langle A^\top y_{k+1}, x\rangle + \frac{1}{\eta_x}V^r_{x_k}(x) + F(x) \\ y_{k+1} = \arg\min_y -\langle Ax_{k+1}, y\rangle + \frac{1}{\eta_y}V^s_{y_k}(y) + H(y). \end{cases}$$

This is a conceptual rule and not an implementable one because finding x_{k+1} requires the gradient at y_{k+1} and vice versa. It is easy to prove that iterates of PPM linearly converges to the solution of (7). We provide a proof in Appendix B.1 for completeness.

Lemma 1. *The iterates of the PPM for the problem (7) satisfy* $(\|x^* - x_K\|^2/\eta_x + \|y^* - y_K\|^2/\eta_y) \le 2\exp(-K/(1+\kappa))(V^r_{x_0}(x^*)/\eta_x + V^s_{y_0}(y^*)/\eta_y)$ *for all* $K \ge 0$, *where* $\kappa = 1/\min(\mu_x\eta_x, \mu_y\eta_y)$.

PD method is the following approximation of PPM:

$$\begin{cases} \widetilde{y}_{k+1} = y_k + \theta(y_k - y_{k-1}) \\ x_{k+1} = \arg\min_x \langle A^\top \widetilde{y}_{k+1}, x\rangle + \frac{1}{\eta_x}V^r_{x_k}(x) + F(x) \\ y_{k+1} = \arg\min_y -\langle Ax_{k+1}, y\rangle + \frac{1}{\eta_y}V^s_{y_k}(y) + H(x) \end{cases} \tag{8}$$

where $\theta = 1/\gamma$ and $\gamma \le 1 + \min(\mu_x\eta_x, \mu_y\eta_y)$. Different from PPM, the PD method uses a pseudo-gradient $A^\top\widetilde{y}_{k+1}$ computed at the extrapolated $\widetilde{y}_{k+1}$, instead of the actual gradient $A^\top y_{k+1}$ at y_{k+1}, to update x_{k+1}. This approximation leads to an implementable algorithm with the same linear convergence as PPM.

Theorem 1 (Chambolle et al. 2016)**.** *If* $\sqrt{\mu_x/\mu_y}\eta_x = \sqrt{\mu_y/\mu_x}\eta_y = 1/2\|A\|$ *and* $y_{-1} = y_0$, *then the iterates of the PD update rule (8) satisfy the same conclusion as Lemma 1, with* $\kappa = 2\|A\|/\sqrt{\mu_x\mu_y}$.

For completeness, we provide a proof in Appendix B.2. The PD method obtains the optimal iteration complexity of $\mathcal{O}(\|A\|\log(1/\varepsilon)/\sqrt{\mu_x\mu_y})$ (Xie et al. 2020; Han et al. 2021). We also note that some versions of the PD method can also be interpreted as an *exact* PPM update using the the Bregman divergence corresponding to the bilinear operator A (He et al. 2012).

3.2 Accelerated Convex Minimization

In this section, we illustrate that the PD method can also be deployed to optimally solve strongly convex and smooth minimization problems. Consider the

problem: $\min_x f(x)$, where function f is L-smooth and μ-strongly convex, and the optimal solution is x^*.

First, we reformulate it into the following minimax problem by introducing a dual variable u and *lifting* it into a larger variable space (x, u):

$$\min_x \left[f(x) = \max_u \ \frac{\mu}{2}\|x\|^2 + \langle x, u\rangle - \underline{f}^*(u) \right], \quad (9)$$

where

$$\underline{f}(x) = f(x) - \frac{\mu}{2}\|x\|^2, \quad \underline{f}^*(u) = \max_x \langle u, x\rangle - \underline{f}(x).$$

Here $\underline{f}^*$ is the Fenchel/convex conjugate of $\underline{f}$. Following the definition, we have $\underline{f}(x)$ is $(L-\mu)$-smooth and convex. A proof is provided in Appendix B.3 for completeness. Then its dual $\underline{f}^*$ is $(L-\mu)^{-1}$ strongly convex with respect to Euclidean norm (Beck 2017).

Notice that this new minimax problem is in the form (7), with bilinear coupling matrix $A = \mathbf{I}$, μ-strongly convex function $F(x) = \frac{\mu}{2}\|x\|^2$, and 1-relatively strongly convex function $H(u) = \underline{f}^*(u)$ with respect to the Bregman divergence $V_{u_k}^{\underline{f}^*}(u)$ generated by $\underline{f}^*$ itself.

Then, if we instantiate the PD update rule (8) for this problem, we obtain the updates:

$$\begin{cases} \widetilde{u}_{k+1} = u_k + \theta(u_k - u_{k-1}) \\ x_{k+1} = \arg\min_x \langle \tilde{u}_{k+1}, x\rangle + \frac{\mu}{2}\|x\|^2 + \frac{1}{2\eta_x}\|x - x_k\|^2 \\ u_{k+1} = \arg\min_u -\langle x_{k+1}, u\rangle + \underline{f}^*(u) + \frac{1}{\eta_u} V_{u_k}^{\underline{f}^*}(u) \end{cases} \quad (10)$$

Corollary 1 (of Theorem 1)**.** *Let $u_{-1} = u_0 = \nabla f(x_0)$. Then the iterates of the PD update rule (10) with stepsizes $\eta_x = 1/\sqrt{\mu(L-\mu)}$, $\eta_u = \sqrt{\mu/(L-\mu)}$ and $\theta = (1 + \sqrt{\mu/(L-\mu)})^{-1}$, for problem (9) satisfies $\|x^* - x_K\|^2 \le \mathcal{O}(\exp(-K/2\sqrt{\kappa - 1})\|x^* - x_0\|^2)$ for all $K \ge 0$, where $\kappa = L/\mu$.*

A proof is in Appendix B.4. Note that this matches the optimal convergence rate achieved by accelerated gradient descent (AGD, Nesterov et al. 2018).

Finally, we show that Bregman proximal update rule in (10) admits an elegant implementation based on the gradients, which resembles AGD.

Lemma 2. *For problem (9), iterates x_k of the PD update rule (10) are the same as the iterates x_k of the following update rule when $(u_{-1}, u_0) = (\nabla \underline{f}(\underline{x}_{-1}), \nabla \underline{f}(\underline{x}_0))$.*

$$\begin{cases} \widetilde{\nabla}_{k+1} = \nabla \underline{f}(\underline{x}_k) + \theta(\nabla \underline{f}(\underline{x}_k) - \nabla \underline{f}(\underline{x}_{k-1})) \\ x_{k+1} = (x_k - \eta_x \widetilde{\nabla}_{k+1})/(1 + \eta_x \mu) \\ \underline{x}_{k+1} = (\underline{x}_k + \eta_u x_{k+1})/(1 + \eta_u) \end{cases} \quad (11)$$

A proof is in Appendix B.5. This close connection between the PD method and AGD was first identified in Lan et al. (2018). The above analysis based on the primal-dual interpretation is conceptually much simpler than the more opaque estimate sequence (Nesterov et al. 2018) or Lyapunov-based (Lan 2012) analyses of AGD. Note that the above update rule is slightly different from the one used in Lan et al. (2018). The latter uses an extrapolated primal iterate $\widetilde{x}_{k+1}$ to update dual iterate u_{k+1}, whereas we use an extrapolated dual iterate $\widetilde{u}_{k+1}$ to update the primal iterate x_{k+1}.

4 LIFTED PRIMAL-DUAL METHOD

The previous section indicates that both bilinear minimax problems and smooth strongly convex minimization problems can be optimally solved using the same PD method after appropriate reformulation. Naturally, this suggests that the PD method has the potential to solve the Bi-SC-SC problem of our interest:

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} [\phi(x, y) = f(x) + \langle y, Ax\rangle - h(y)], \quad (12)$$

which consists of a bilinear term $\langle y, Ax\rangle$ and two smooth strongly-convex functions f and h.

Our strategy to solve (12) is to first transform the objective into a form where the proximal operators are easy to compute and then solve this new objective using the PD method. Introducing dual variables u and v for f and h, respectively, the Bi-SC-SC problem can be equivalently reformulated (or *lifted*) as

$$\min_{x\in\mathcal{X},v} \max_{y\in\mathcal{Y},u} \Phi(x, y; u, v), \text{ where} \quad (13)$$

$$\Phi(x, y; u, v) := \left[-\underline{f}^*(u) + \langle u, x\rangle + (\mu_x/2)\|x\|^2 \right] + \langle y, Ax\rangle - \left[(\mu_y/2)\|y\|^2 + \langle v, y\rangle - \underline{h}^*(v) \right], \quad (14)$$

$$\underline{f}^*(u) := \max_x \langle u, x\rangle - [\underline{f} := f(x) - (\mu_x/2)\|x\|^2], \text{ and}$$
$$\underline{h}^*(v) := \max_y \langle v, y\rangle - [\underline{h} := h(x) - (\mu_y/2)\|y\|^2]. \quad (15)$$

By Fenchel duality, it follows that $\phi(x, y) = \min_v \max_u \Phi(x, y; u, v)$ (Lemma 4(c)). Note that both $\underline{f}^*$ and $\underline{h}^*$ are strongly convex. Intriguingly, the first three terms, the middle three terms, and the last three terms in (13) are all of the form (7), and hence amenable to the Primal-Dual approach. To this end, we introduce the following the PD update to each of the four variables with their respective stepsizes, Breg-

Algorithm 1 LPD: Lifted Primal-Dual algorithm

Required: $\mathcal{X}$, $\mathcal{Y}$, (f, L_x, μ_x), $(A, \|A\|)$, (h, L_y, μ_y), K, $\{(\eta_{x,k}, \eta_{y,k}, \eta_{u,k}, \eta_{v,k}, \theta_k)\}_{k=0}^{K-1}$

1 Initialize $(x_{-1}, y_{-1}) = (x_0, y_0) \in \mathcal{X} \times \mathcal{Y}$

2 Set $\underline{f} = f - (\mu_x/2)\|\cdot\|^2$, $\underline{h} = h - (\mu_y/2)\|\cdot\|^2$, $(\underline{x}_{-1}, \underline{y}_{-1}) = (\underline{x}_0, \underline{y}_0) = (x_0, y_0)$

for $0 \le k \le K-1$ **do**

3 $\quad \widetilde{x}_{k+1} = x_k + \theta_k(x_k - x_{k-1})$,
$\quad \widetilde{y}_{k+1} = y_k + \theta_k(y_k - y_{k-1})$,
$\quad \widetilde{\nabla}_{x,k+1} = \nabla \underline{f}(\underline{x}_k) + \theta_k(\nabla \underline{f}(\underline{x}_k) - \nabla \underline{f}(\underline{x}_{k-1}))$,
$\quad \widetilde{\nabla}_{y,k+1} = \nabla \underline{h}(\underline{y}_k) + \theta_k(\nabla \underline{h}(\underline{y}_k) - \nabla \underline{h}(\underline{y}_{k-1}))$

4 $\quad x_{k+1} = \mathcal{P}_{\mathcal{X}}((x_k - \eta_{x,k}(A^\top \widetilde{y}_{k+1} + \widetilde{\nabla}_{x,k+1}))/(1 + \eta_{x,k}\mu_y))$

5 $\quad y_{k+1} = \mathcal{P}_{\mathcal{Y}}((y_k + \eta_{y,k}(A\widetilde{x}_{k+1} - \widetilde{\nabla}_{y,k+1}))/(1 + \eta_{x,k}\mu_y))$

6 $\quad \underline{x}_{k+1} = (\underline{x}_k + \eta_{u,k}\, x_{k+1})/(1 + \eta_{u,k})$,

7 $\quad \underline{y}_{k+1} = (\underline{y}_k + \eta_{v,k}\, y_{k+1})/(1 + \eta_{v,k})$

end

8 **return** $(x_K, y_K, \underline{x}_K, \underline{y}_K)$

man divergences, and extrapolation steps.

$$
\begin{aligned}
&(\widetilde{x}_{k+1}, \widetilde{y}_{k+1}) = (1+\theta)(x_k, y_k) - \theta(x_{k-1}, y_{k-1}) \\
&(\widetilde{u}_{k+1}, \widetilde{v}_{k+1}) = (1+\theta)(u_k, v_k) - \theta(u_{k-1}, v_{k-1}) \\
&x_{k+1} = \arg\min_{x \in \mathcal{X}} \big\langle A^\top \widetilde{y}_{k+1} + \widetilde{u}_{k+1}, x \big\rangle + \\
&\qquad\qquad \|x - x_k\|^2/2\eta_x + \mu_x\|x\|^2/2 \\
&y_{k+1} = \arg\min_{y \in \mathcal{Y}} - \big\langle A^\top \widetilde{x}_{k+1} + \widetilde{v}_{k+1}, y \big\rangle + \qquad (16) \\
&\qquad\qquad \|y - y_k\|^2/2\eta_y + \mu_y\|y\|^2/2 \\
&u_{k+1} = \arg\min_{u} - \langle x_{k+1}, u\rangle + \underline{f}^*(u) + V_{\underline{u}_k}^{\underline{f}^*}(u)/\eta_u \\
&v_{k+1} = \arg\min_{v} - \langle y_{k+1}, v\rangle + \underline{h}^*(v) + V_{\underline{v}_k}^{\underline{h}^*}(v)/\eta_v
\end{aligned}
$$

We show that the above update rule can be easily implemented using Algorithm 1, which we call the *Lifted Primal-Dual (LPD)* method.

Lemma 3. *For problem (13), iterates (x_k, y_k) of the PD update rule (16) is the same as the iterates (x_k, y_k) of Algorithm 1, when $(u_{-1}, u_0) = (\nabla \underline{f}(\underline{x}_{-1}), \nabla \underline{f}(\underline{x}_0))$, $(v_{-1}, v_0) = (\nabla \underline{f}(\underline{y}_{-1}), \nabla \underline{f}(\underline{y}_0))$.*

We omit the proof of the above lemma as it is similar to that of Lemma 2. Note that we update the variables in the order $(x, y) \to (\underline{x}, \underline{y})$, where variables in tuples are simultaneously updated. However, any update ordering can be shown to achieve similar guarantees as we show, by using appropriate extrapolation steps and stepsize choices. We extrapolate all the variables and gradients (in step 3 of Algorithm 1) before the x and y updates (steps 4 and 5 of Algorithm 1) to make our analysis a bit symmetric, hence simpler. However, depending on the order in which we update each of the variables we may not have to extrapolate all the variables. For example, if we update variables in the order $x \to y \to \underline{x} \to \underline{y}$, we only have to use (a) the extrapolated $\widetilde{y}_{k+1}$ and $\widetilde{\nabla}_{x,k+1}$ for updating x, and (b) the extrapolated $\widetilde{\nabla}_{y,k+1}$ for updating y.

5 CONVERGENCE ANALYSIS

Now we provide the main theoretical results.

Strongly-Convex–Strongly-Concave Case. LPD achieves the optimal iteration complexity for solving Bi-SC-SC problem in (12). Define the following condition numbers: $\kappa_x = L_x/\mu_x$, $\kappa_y = L_y/\mu_y$, $\kappa_{xy} = \|A\|/\sqrt{\mu_x\mu_y}$, and define the meta-condition number: $\kappa = \sqrt{\kappa_x - 1} + 2\kappa_{xy} + \sqrt{\kappa_y - 1}$. Let x^*, y^* be the optimal solution. For any candidate solution $(x, y) \in \mathcal{X} \times \mathcal{Y}$, we measure the suboptimality with,

$$\Delta(x, y) = \kappa_{xy}(\mu_x\|x - x^*\|^2 + \mu_y\|y - y^*\|^2).$$

Theorem 2 (Informal, cf. Corollary 2)**.** *For any $k \ge 0$, set the parameters*

$$
\begin{aligned}
&\gamma = 1 + \kappa^{-1}\,,\ \theta_k = 1/\gamma\,, \\
&\eta_{x,k} = (\sqrt{\kappa_x - 1} + 2\kappa_{xy})^{-1}/\mu_x\,, \\
&\eta_{y,k} = (2\kappa_{xy} + \sqrt{\kappa_y - 1})^{-1}/\mu_y\,, \qquad (17) \\
&\eta_{u,k} = (\sqrt{\kappa_x - 1})^{-1},\ \eta_{v,k} = (\sqrt{\kappa_y - 1})^{-1}\,.
\end{aligned}
$$

Then for any $K > 0$, output of Algorithm 1 satisfies

$$
\begin{aligned}
&\Delta(x^K, y^K) \\
&\le \exp\Big(-\frac{(K-1)}{(\kappa+1)}\Big)\Big(\Big(\frac{1}{\eta_{x,0}} + \frac{L_x - \mu_x}{\eta_{u,0}}\Big)\|x^* - x_0\|^2 + \\
&\Big(\frac{1}{\eta_{y,0}} + \frac{L_y - \mu_y}{\eta_{v,0}}\Big)\|y^* - y_0\|^2\Big)\,.
\end{aligned}
$$

Note that the parameter choices in the above theorem are iteration (k) invariant. The gradient complexity of Algorithm 1 is

$$\mathcal{O}\Big(\Big(\sqrt{\frac{L_x}{\mu_x} - 1} + \frac{\|A\|}{\sqrt{\mu_x\mu_y}} + \sqrt{\frac{L_y}{\mu_y} - 1}\Big)\log\Big(\frac{1}{\varepsilon}\Big)\Big), \quad (18)$$

which is optimal and matches the lower-bound (Zhang et al. 2019) for Bi-SC-SC problem (12) up to logarithmic factors in the problem parameters. The lifting of the objective function allows the PD method to be jointly applied to the smooth convex terms (as illustrated in Section 3.2) and to the bilinear minimax terms (as illustrated in Section 3.1), achieving this optimal rate (Zhang et al. 2019). Comparisons to other algorithms are given in Table 1.

We emphasize that LPD inherits the computational and conceptual simplicity of the PD methods. The former leads to a single-loop algorithm, which is much simpler than other state-of-the-art complex multi-loop methods with sub-optimal guarantees (Lin et al. 2020; Wang et al. 2020; Xie et al. 2021). The latter leads to a more transparent analysis, based on the simple analysis of the PD methods (Theorem 1), which is based on an even simpler analysis of PPM (Lemma 1).

Note that we do not directly adapt the original guarantee of the PD method (Chambolle et al. 2016). Our analysis has to be different since the naive application of the existing algorithm and analysis will depend on an effective strong convexity parameter (in (x,v)) of $\min(\mu_x, 1/(L_y-\mu_y))$, an effective strong concavity parameter (in (y,u)) of $\min(\mu_y, 1/(L_x-\mu_x))$, and a Lipschitz constant which is equal to the largest eigenvalue of the matrix effective coupling matrix $[\mathbf{I}, A; 0, \mathbf{I}]$. This leads to a sub-optimal guarantee. Hence, we propose a different approach with a tighter analysis to achieve the optimal rates.

Convex–Strongly-Concave Case. Consider the Bilinearly-coupled Convex–Strongly-Concave (Bi-C-SC) case, where f is merely convex, i.e. $\mu_x = 0$.

Remark 1 (LPD + Smoothing (Nesterov 2005))**.** *Let $\phi(x,y)$ be the objective of a Bi-C-SC problem. Then we can apply LPD for Bi-SC-SC problems (Theorem 2) to the* smoothed *Bi-SC-SC objective $\phi(x,y) + \lambda\varepsilon\|x\|^2$ for some $\lambda > 0$, and achieve an iteration complexity of $\mathcal{O}(\sqrt{L_x/\varepsilon} + \|A\|/\sqrt{\mu_y\varepsilon} + \sqrt{L_y/\mu_y}) \log(1/\varepsilon)$ for solving the original Bi-C-SC problem.*

The above result is optimal up to logarithmic factors. The first term cannot be improved even for a pure minimization of convex f (Nesterov et al. 2018). Due to a lower-bound of $\Omega(\|A\|/\sqrt{\mu_y\varepsilon})$ for the same problem when $f = 0$ (Ouyang et al. 2021), the second term cannot be improved. The third term cannot be improved even for a pure maximization of strongly-concave h (Nesterov et al. 2018).

However, smoothing might not be desirable in practice, because it requires bounded domains and fixing the final target error ε in advance, and it is hard to tune λ (Nesterov 2005). We therefore design a direct algorithm by customizing the stepsizes of LPD. Let $D_{\mathcal{X}} = \max_{x\in\mathcal{X}\cap\mathrm{dom}(f)} \|x - x_0\|$ and $D_{\mathcal{Y}} = \max_{y\in\mathcal{Y}\cap\mathrm{dom}(h)} \|y - y_0\|$. Note that the min variable solution x^* may not be unique.

Theorem 3 (Informal, cf. Corollary 3)**.** *Let*

$$1/\eta_{x,k} = 1/(k+1)\eta_x\,,\ 1/\eta_x = 2L_x + 16\|A\|/\mu_y\,,$$
$$1/\eta_{y,k} = 1/(k+1)\eta_y + k\mu_y/2\,,\ 1/\eta_y = 2(L_y-\mu_y)\,,$$
$$\eta_{u,k} = \eta_{v,k} = 2/k\,,\ \textit{for all } k \geq 0\,. \tag{19}$$

Then for any $K > 0$, output of Algorithm 1 satisfies (a) if $D_{\mathcal{X}} < \infty$ and $D_{\mathcal{Y}} < \infty$,

$$\max_{y\in\mathcal{Y}} \phi(\overline{x}_K, y) - \min_{x\in\mathcal{X}} \phi(x, \overline{y}_K) \leq \frac{2L_x D_{\mathcal{X}}^2}{K(K+1)} + \frac{16\|A\|^2 D_{\mathcal{X}}^2}{\mu_y K(K+1)} + \frac{2(L_y-\mu_y)D_{\mathcal{Y}}^2}{K(K+1)} \tag{20}$$

where $(\overline{x}_K, \overline{y}_K) := \sum_{k=1}^K \frac{2k}{K(K+1)}(x_k, y_k)$,
(b) even if the feasible set is unbounded,

$$\frac{\mu_y}{4}\|y^* - y_K\|^2 \leq \frac{2L_x\|x^*-x_0\|^2}{K(K+1)} + \frac{16\|A\|^2\|x^*-x_0\|^2}{\mu_y K(K+1)} + \frac{2(L_y-\mu_y)\|y^*-y_0\|^2}{K(K+1)} \tag{21}$$

(c) if $D_{\mathcal{X}} < \infty$, $\phi_p(x) = \max_{y\in\mathcal{Y}} \phi(x,y)$, $\phi_d(x) = \min_{x\in\mathcal{X}} \phi(x,y)$, and we do a warm restart on variable y with $K_0 = \Omega_\varepsilon(1)$ initial additional iterations, then

$$\phi_p(\overline{x}_K) - \phi_p(x^*) \leq (L_x + \frac{10L_y\|A\|^2}{\mu_y^2})\frac{4\|x^*-x_0\|^2}{K(K+1)}\,,\ \textit{and}$$
$$\phi_d(y^*) - \phi_d(\overline{y}_K) \leq (L_x + \frac{8\|A\|^2}{\mu_y})\frac{4D_{\mathcal{X}}^2}{K(K+1)}\,.$$

This implies that, for Bi-C-SC problem, LPD has a gradient complexity of

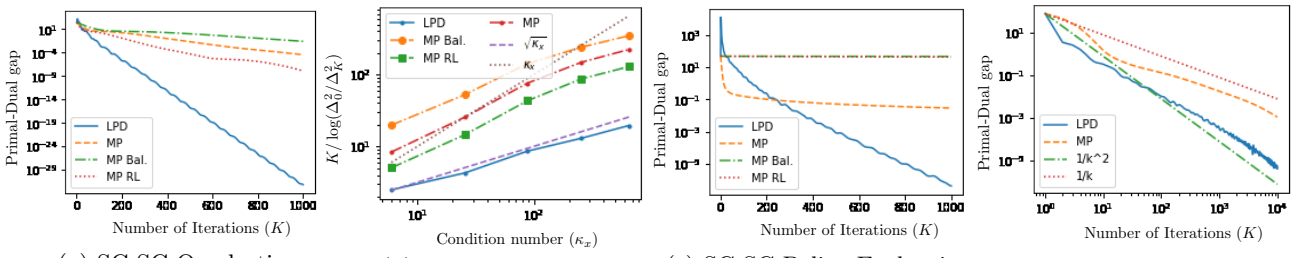
$$\mathcal{O}\Big(\Big(\sqrt{\frac{L_x}{\varepsilon}} + \frac{\|A\|}{\sqrt{\mu_y\varepsilon}} + \sqrt{\frac{L_y-\mu_y}{\varepsilon}}\Big)\Big)\,. \tag{22}$$

The LPD method achieves better complexities than previous single-loop algorithms (Nesterov et al. 2006; Mokhtari et al. 2020), PDHG-type algorithm (Zhao 2019), direct multi-loop algorithm (Thekumparampil et al. 2019), and some smoothing-based multi-loop algorithms (Lin et al. 2020; Wang et al. 2020; Xie et al. 2021) (see Table 1). Earlier single-loop methods such as Chambolle et al. (2016) and Hamedani et al. (2021) achieve $\mathcal{O}(1/K^2)$ rate only under the restriction that $L_x = 0$. This showcases the generality and simplicity of our LPD method, as it is the first single-loop algorithm (to the best of our knowledge) which achieves $\mathcal{O}(1/K^2)$ rate for this problem. It is not known if better rates than in the above theorem are achievable with a direct single-loop algorithm without using the smoothing technique, like in Lifted PD + Smoothing (Remark 1). As discussed after Remark 1, direct algorithms such as the one above are more desirable.

Prox-friendly terms: We point out that LPD can be extended to solve more general (possibly nonsmooth) minmax problems with the same guarantees:

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}}\ F(x) + f(x) + \langle y, Ax\rangle - h(y) - H(y), \tag{23}$$

where F and H are convex and we have access to their proximal operators and f, h satisfy our Assumption 1. We give the details of extension in Appendix B.6.



(a) SC-SC Quadratic (b) SC-SC Quadratic (c) SC-SC Policy Evaluation (d) SC-C Policy Evaluation

Figure 1: LPD method (ours) achieves a faster linear convergence rate than competing algorithms in Strongly-Convex–Strongly-Concave synthetic quadratic minimax (a-b) and policy evaluation (c) problems. LPD method (ours) also achieves a faster $O(1/K^2)$ convergence rate than competing single-loop algorithm in Convex–Strongly-Concave policy evaluation problem (d).

6 EXPERIMENTAL RESULTS

In this section we compare our LPD method with some competing single-loop non-smoothing-based direct algorithms when solving both synthetic and real-world problems. More details of the experiments are provided in Appendix F. First, we compare our LPD method with Mirror Prox (MP) (Mokhtari et al. 2020), Balanced Mirror Prox (MP Bal.) (see Appendix F), and Relative Lipschitzness-based Mirror Prox (MP RL, Cohen et al. 2020) when solving Bi-SC-SC problems. We only compared our (single-loop) algorithm with other single-loop algorithms, because multi-loop algorithms such as in Wang et al. (2020) and Xie et al. (2021) are typically challenging to implement and tune. To the best of our knowledge, there are no publicly available implementations for these algorithms.

Quadratic Problem: First, we consider synthetic quadratic problems of the form (4). We randomly generate the matrices B, A, C in such a way that $\kappa_x = L_x/\mu_x = \kappa_y = L_y/\mu_y$ and $\kappa_{xy} = \|A\|/\sqrt{\mu_x\mu_y} := \sqrt{\kappa_x}$. In Figure 1a, we plot the primal-dual gap against the number of iterations (K) of different algorithms when solving such a problem with $\kappa_x = 256.0$. We see what LPD achieves a faster linear convergence than other methods. In Figure 1b, we plot $K/\log(\Delta_0^2/\Delta_K^2)$ against κ_x where $\Delta_K = \|x_K - x^*\|^2 + \|y_K - y^*\|^2$. We vary κ_x from 5.96 to 656.84. As expected from theory, in this log-log scale plot, slope of the LPD curve is close to 1/2 since $\Delta_K^2 \le \mathcal{O}(\exp(-K/\sqrt{\kappa_x}))$ for LPD, and slope of other algorithms are close to one since $\Delta_K^2 \le \mathcal{O}(\exp(-K/\kappa_x))$ for other algorithms.

Policy Evaluation: Next, we consider policy evaluation problems of the form (6). We consider the same MountainCar (Sutton et al. 2018) reinforcement learning problem used in Du et al. (2017), and use the same copy of policy trace $\{(s_t, a_t, s_{t+1}, r_t)\}_{t=1}^n$ used by Du et al. (2017) to construct the MSPBE minimization problem. We create the feature vectors ϕ_t, by applying PCA to the state vectors s_t to whiten them. This reduces their dimension from 300 to 200. Finally setting $\rho = 1.0$, results in a highly ill-conditioned Bi-SC-SC problem with $\kappa_x = 1.0$, $\kappa_{xy} = 24.35$, and $\kappa_y = 19387.07$. In Figure 1c, we plot the primal-dual gap against the number of iterations (K) of different algorithms when solving this problem. We observe that, our LPD method achieves much faster linear convergence than all other algorithms. Note that MP is better than LPD for small K, because in this regime the $\mathcal{O}(1/K)$ convergence rate of MP dominates its primal-dual gap.

Finally, we compare our LPD method with MP (Mokhtari et al. 2020), when solving a Bi-SC-C problem.

SC-C Policy Evaluation: We consider the same minimum MSPBE estimation problem as above. However we directly use the 300 dimensional state vectors s_t as its feature vector ϕ_t. This results in a Bi-SC-C problem. Note that Bi-SC-C objective is the negative of the objective of a Bi-C-SC problem, which means that we can solve it using LPD with stepsize choice given in Theorem 2. In Figure 1d, we plot the primal-dual gap against the number of iterations (K) of LPD and MP methods when solving this problem. As theory predicts, we observe that the LPD method achieves a much faster $O(1/K^2)$ convergence rate than $O(1/K)$ convergence rate of MP.

7 CONCLUSION

We studied Bi-SC-SC problem and provided an optimal single-loop algorithm: the Lifted Primal-Dual (LPD) method to solve it. The LPD method is designed using simple building blocks of the Primal-Dual method and *lifting*, leading to its generalizability, simplicity, and transparent analysis. Further, we also provide two related algorithms—one optimal (upto logarithmic factors) and another single-loop—to solve Bi-C-SC problem.

Acknowledgement

This work is supported by Google faculty research award and NSF grants CNS-2002664, IIS-1929955, DMS-2134012, CCF-2019844 as a part of NSF Institute for Foundations of Machine Learning (IFML), CNS-2112471 as a part of NSF AI Institute for Future Edge Networks and Distributed Intelligence (AI-EDGE), and CCF-1934986. This work was done prior to the first author joining Amazon and does not relate to his current position there.

References

Korpelevich, Galina M (1976). "The extragradient method for finding saddle points and other problems". In: *Matecon* 12, pp. 747–756.

Rockafellar, R Tyrrell (1976). "Monotone operators and the proximal point algorithm". In: *SIAM journal on control and optimization* 14.5, pp. 877–898.

Rockafellar, R Tyrell (1987). "Linear-quadratic programming and optimal control". In: *SIAM Journal on Control and Optimization* 25.3, pp. 781–814.

Tseng, Paul (1995). "On linear convergence of iterative methods for the variational inequality problem". In: *Journal of Computational and Applied Mathematics* 60.1-2, pp. 237–252.

Chen, George HG and R Tyrrell Rockafellar (1997). "Convergence rates in forward–backward splitting". In: *SIAM Journal on Optimization* 7.2, pp. 421–444.

El Ghaoui, Laurent and Hervé Lebret (1997). "Robust solutions to least-squares problems with uncertain data". In: *SIAM Journal on matrix analysis and applications* 18.4, pp. 1035–1064.

Bai, Zhong-Zhi, Gene H Golub, and Michael K Ng (2003). "Hermitian and skew-Hermitian splitting methods for non-Hermitian positive definite linear systems". In: *SIAM Journal on Matrix Analysis and Applications* 24.3, pp. 603–626.

Nemirovski, Arkadi (2004). "Prox-method with rate of convergence O(1/t) for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems". In: *SIAM Journal on Optimization* 15.1, pp. 229–251.

Benzi, Michele, Gene H Golub, and Jörg Liesen (2005). "Numerical solution of saddle point problems". In: *Acta numerica* 14, pp. 1–137.

Nesterov, Yu (2005). "Smooth minimization of non-smooth functions". In: *Mathematical programming* 103.1, pp. 127–152.

Nesterov, Yurii and Laura Scrimali (2006). "Solving strongly monotone variational and quasi-variational inequalities". In:

Facchinei, Francisco and Jong-Shi Pang (2007). *Finite-dimensional variational inequalities and complementarity problems*. Springer Science & Business Media.

Nesterov, Yurii (2007). "Dual extrapolation and its applications to solving variational inequalities and related problems". In: *Mathematical Programming* 109.2, pp. 319–344.

Bai, Zhong-Zhi (2009). "Optimal parameters in the HSS-like methods for saddle-point problems". In: *Numerical Linear Algebra with Applications* 16.6, pp. 447–479.

Kakade, Sham, Shai Shalev-Shwartz, Ambuj Tewari, et al. (2009). "On the duality of strong convexity and strong smoothness: Learning applications and matrix regularization". In: *Unpublished Manuscript, http://ttic.uchicago.edu/~shai/papers/KakadeShalevTewari09.pdf* 2.1.

Bauschke, Heinz H, Patrick L Combettes, et al. (2011). *Convex analysis and monotone operator theory in Hilbert spaces*. Vol. 408. Springer.

Juditsky, Anatoli, Arkadi Nemirovski, et al. (2011). "First order methods for nonsmooth convex large-scale optimization, ii: utilizing problems structure". In: *Optimization for Machine Learning* 30.9, pp. 149–183.

He, Bingsheng and Xiaoming Yuan (2012). "Convergence analysis of primal-dual algorithms for a saddle-point problem: from contraction perspective". In: *SIAM Journal on Imaging Sciences* 5.1, pp. 119–149.

Lan, Guanghui (2012). "An optimal method for stochastic composite optimization". In: *Mathematical Programming* 133.1, pp. 365–397.

Chen, Yunmei, Guanghui Lan, and Yuyuan Ouyang (2014). "Optimal primal-dual methods for a class of saddle point problems". In: *SIAM Journal on Optimization* 24.4, pp. 1779–1814.

Liu, Qingshan and Jun Wang (2015). "A projection neural network for constrained quadratic minimax optimization". In: *IEEE Transactions on Neural Networks and Learning Systems* 26.11, pp. 2891–2900.

Chambolle, Antonin and Thomas Pock (2016). "On the ergodic convergence rates of a first-order primal–dual algorithm". In: *Mathematical Programming* 159.1, pp. 253–287.

He, Yunlong and Renato DC Monteiro (2016). "An accelerated HPE-type algorithm for a class of composite convex-concave saddle-point problems". In: *SIAM Journal on Optimization* 26.1, pp. 29–56.

Palaniappan, Balamurugan and Francis Bach (2016). "Stochastic variance reduction methods for saddle-point problems". In: *Advances in Neural Information Processing Systems*, pp. 1416–1424.

Beck, Amir (2017). *First-order methods in optimization*. SIAM.

Chen, Yunmei, Guanghui Lan, and Yuyuan Ouyang (2017). "Accelerated schemes for a class of variational inequalities". In: *Mathematical Programming* 165.1, pp. 113–149.

Daskalakis, Constantinos, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng (2017). "Training gans with optimism". In: *arXiv preprint arXiv:1711.00141*.

Du, Simon S, Jianshu Chen, Lihong Li, Lin Xiao, and Dengyong Zhou (2017). "Stochastic variance reduction methods for policy evaluation". In: *International Conference on Machine Learning*. PMLR, pp. 1049–1058.

Gidel, Gauthier, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien (2018). "A Variational Inequality Perspective on Generative Adversarial Networks". In: *International Conference on Learning Representations*.

Lan, Guanghui and Yi Zhou (2018). "An optimal randomized incremental gradient method". In: *Mathematical programming* 171.1, pp. 167–215.

Nesterov, Yurii et al. (2018). *Lectures on convex optimization*. Vol. 137. Springer.

Sutton, Richard S and Andrew G Barto (2018). *Reinforcement learning: An introduction*. MIT press.

Du, Simon S and Wei Hu (2019). "Linear convergence of the primal-dual gradient method for convex-concave saddle point problems without strong convexity". In: *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 196–205.

Thekumparampil, Kiran K, Prateek Jain, Praneeth Netrapalli, and Sewoong Oh (2019). "Efficient algorithms for smooth minimax optimization". In: *Advances in Neural Information Processing Systems*, pp. 12659–12670.

Zhang, Junyu, Mingyi Hong, and Shuzhong Zhang (2019). "On lower iteration complexity bounds for the saddle point problems". In: *arXiv preprint arXiv:1912.07481*.

Zhao, Renbo (2019). "Optimal algorithms for stochastic three-composite convex-concave saddle point problems". In: *arXiv preprint arXiv:1903.01687*.

Alkousa, Mohammad S, Alexander Vladimirovich Gasnikov, Darina Mikhailovna Dvinskikh, Dmitry A Kovalev, and Fedor Sergeevich Stonyakin (2020). "Accelerated methods for saddle-point problem". In: *Computational Mathematics and Mathematical Physics* 60.11, pp. 1787–1809.

Azizian, Waïss, Ioannis Mitliagkas, Simon Lacoste-Julien, and Gauthier Gidel (2020a). "A tight and unified analysis of gradient-based methods for a whole spectrum of differentiable games". In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 2863–2873.

Azizian, Waïss, Damien Scieur, Ioannis Mitliagkas, Simon Lacoste-Julien, and Gauthier Gidel (2020b). "Accelerating smooth games by manipulating spectral shapes". In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 1705–1715.

Choukroun, Yoni, Michael Zibulevsky, and Pavel Kisilev (2020). "Primal-Dual Sequential Subspace Optimization for Saddle-point Problems". In: *arXiv preprint arXiv:2008.09149*.

Cohen, Michael B, Aaron Sidford, and Kevin Tian (2020). "Relative lipschitzness in extragradient methods and a direct recipe for acceleration". In: *arXiv preprint arXiv:2011.06572*.

Huang, Kevin, Junyu Zhang, and Shuzhong Zhang (2020). "Cubic regularized newton method for saddle point models: a global and local convergence analysis". In: *arXiv preprint arXiv:2008.09919*.

Lin, Tianyi, Chi Jin, and Michael I Jordan (2020). "Near-optimal algorithms for minimax optimization". In: *Conference on Learning Theory*. PMLR, pp. 2738–2779.

Mokhtari, Aryan, Asuman Ozdaglar, and Sarath Pattathil (2020). "A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach". In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 1497–1507.

Schäfer, Florian, Anima Anandkumar, and Houman Owhadi (2020). "Competitive Mirror Descent". In: *arXiv preprint arXiv:2006.10179*.

Wang, Yuanhao and Jian Li (2020). "Improved algorithms for convex-concave minimax optimization". In: *arXiv preprint arXiv:2006.06359*.

Xie, Guangzeng, Luo Luo, Yijiang Lian, and Zhihua Zhang (2020). "Lower complexity bounds for finite-sum convex-concave minimax optimization problems". In: *International Conference on Machine Learning*. PMLR, pp. 10504–10513.

Yang, Junchi, Negar Kiyavash, and Niao He (2020a). "Global convergence and variance reduction for a class of nonconvex-nonconcave minimax problems". In: *Advances in Neural Information Processing Systems* 33, pp. 1153–1165.

Yang, Junchi, Siqi Zhang, Negar Kiyavash, and Niao He (2020b). "A catalyst framework for minimax optimization". In: *Advances in Neural Information Processing Systems* 33, pp. 5667–5678.

Zhang, Guojun, Kaiwen Wu, Pascal Poupart, and Yaoliang Yu (2020). "Newton-type methods for minimax optimization". In: *arXiv preprint arXiv:2006.14592*.

Hamedani, Erfan Yazdandoost and Necdet Serhat Aybat (2021). "A Primal-Dual Algorithm with Line Search for General Convex-Concave Saddle Point

Problems". In: *SIAM Journal on Optimization* 31.2, pp. 1299–1329.

Han, Yuze, Guangzeng Xie, and Zhihua Zhang (2021). "Lower complexity bounds of finite-sum optimization problems: The results and construction". In: *arXiv preprint arXiv:2103.08280*.

Ouyang, Yuyuan and Yangyang Xu (2021). "Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems". In: *Mathematical Programming* 185.1, pp. 1–35.

Xie, Guangzeng, Yuze Han, and Zhihua Zhang (2021). "DIPPA: An improved Method for Bilinear Saddle Point Problems". In: *arXiv preprint arXiv:2103.08270*.

Zhang, Guodong, Yuanhao Wang, Laurent Lessard, and Roger Grosse (2021a). "Don't Fix What ain't Broke: Near-optimal Local Convergence of Alternating Gradient Descent-Ascent for Minimax Optimization". In: *arXiv preprint arXiv:2102.09468*.

Zhang, Siqi, Junchi Yang, Cristóbal Guzmán, Negar Kiyavash, and Niao He (2021b). "The complexity of nonconvex-strongly-concave minimax optimization". In: *Uncertainty in Artificial Intelligence*. PMLR, pp. 482–492.

Supplementary Material: Lifted Primal-Dual Method for Bilinearly Coupled Smooth Minimax Optimization

A DEFINITIONS AND STANDARD RESULTS

A.1 Convexity and Smoothness

Definition 1. *We say that a function is μ-strongly convex if*

$$f(\alpha x_1 + (1-\alpha)x_2) \le \alpha f(x_1) + (1-\alpha)f(x_2) - \frac{\mu}{2}\alpha(1-\alpha)\|x_1 - x_2\|^2, \quad \text{for any } \alpha \in [0,1],$$

$$f(x_2) \ge f(x_1) + \langle f'(x_1), x_2 - x_2\rangle + \frac{\mu}{2}\|x_1 - x_2\|^2, \text{ or equivalently}$$

$$\langle f'(x_1)(x_1) - f'(x_1)(x_2), x_1 - x_2\rangle \ge \mu\|x_1 - x_2\|^2$$

for all x_1 and x_2, where at any point x, $f'(x) \in \partial f(x)$ is some sub-gradient of the function in its (Frechet) sub-differential $\partial f(x)$ at that point. Further we say that a function (merely) convex if it is 0-strongly convex.

For a differentiable function f, its gradient at any point x is denoted by $\nabla f(x)$.

Definition 2. *We say that a function is L-smooth if it is differentiable and*

$$f(x_2) \le f(x_1) + \langle \nabla f(x_1), x_2 - x_2\rangle + \frac{L}{2}\|x_2 - x_1\|^2, \text{ or equivalently } \langle \nabla f(x_1) - \nabla f(x_2), x_1 - x_2\rangle \le L\|x_1 - x_2\|^2$$

for all x_1 and x_2, where where at any point x, $\nabla f(x)$ is gradient of the function at that point x.

A.2 Fenchel/Convex Conjugate and Duality

Definition 3. *Let f be a convex function. Then its Fenchel/convex conjugate f^* is defined as $f^*(u) := \max_x \langle u, x\rangle - f(x)$*

Lemma 4 (Kakade et al. 2009; Nesterov et al. 2018)**.** *Fenchel/convex conjugate satisfy the following properties.*

(a) If f is an L-smooth and convex function, then f^* $1/L$-strongly convex.

(b) If f is an L-smooth and convex function, then $\nabla f(x) = \arg\min_x \langle x, u\rangle - f^*(u)$.

(c) If f is a convex function, $(f^*)^*$ is f

(d) If f is an L-smooth and convex function and $u = \nabla f(x)$ then $x = \arg\min_u \langle u, x\rangle - f^*(x) \in \partial(f^*)(u)$.

A.3 Proximal Operator

Definition 4. *For a convex function, F, its proximal operator $\text{prox}_{\eta F}(x)$ (parameterized by some $\eta > 0$) is defined as*

$$\text{prox}_{\eta F}(x) = \arg\min_{\widetilde{x}} F(x) + \frac{1}{2\eta}\|\widetilde{x} - x\|^2 \tag{24}$$

A.4 Bregman Divergence, and Relative Lipschitzness and Relative Convexity

Definition 5. *Let r be a strongly convex function. Then Bregman divergence $V_x^r(\widetilde{x})$ w.r.t. to the distance generating function (d.g.f.) r is defined as the*

$$V_x^r(\widetilde{x}) = r(\widetilde{x}) - r(x) - \langle r'(x), \widetilde{x} - x\rangle \tag{25}$$

where $r'(x) \in \partial r(x)$ is a sub-gradient of r at x.

Lemma 5. *Let r be a σ-strongly convex function. Then Bregman divergence $V_x^r(\widetilde{x})$ w.r.t. to the d.g.f. r satisfies $V_x^r(\widetilde{x}) \geq (\sigma/2)\|\widetilde{x} - x\|^2$.*

Lemma 6. *If f L-smooth convex function, $u = \nabla f(x)$ and $u_0 = \nabla f(x_0)$, then $\frac{1}{2L}\|u - u_0\|^2 \leq V_{u_0}^{f^*}(u) = V_x^f(x_0) \leq L/2\|x - x_0\|^2$*

Proof. By Lemma 4, f^* is $1/L$-strongly convex. Then using Lemma 5 and Lemma 4 we can show that.

$$\begin{aligned}
\frac{1}{2L}\|u - u_0\|^2 \leq V_{u_0}^{f^*}(u) &= f^*(u) - f^*(u_0) - \left\langle f^{*\prime}(u_0), u - u_0 \right\rangle && (26)\\
&= (\langle u, x\rangle - f(x)) - (\langle u_0, x_0\rangle - f(x_0)) - \langle x_0, u - u_0\rangle && (27)\\
&= f(x_0) - f(x) - \langle u, x_0 - x\rangle && (28)\\
&= f(x_0) - f(x) - \langle \nabla f(x), x_0 - x\rangle && (29)\\
&= V_{x_0}^f(x) && (30)\\
&\leq \frac{L}{2}\|x - x_0\|^2 && (31)
\end{aligned}$$

□

Definition 6. *We say that a function is relatively μ-strongly convex w.r.t. to a Bregman divergence V^r (generated by a strongly convex d.g.f. r) if*

$$\begin{aligned}
&f(\alpha x_1 + (1-\alpha)x_2) \leq \alpha f(x_1) + (1-\alpha)f(x_2) - \mu\alpha(1-\alpha)V_{x_1}^r(x_2)\,, \quad \text{for any } \alpha \in [0,1]\,,\\
&f(x_2) \geq f(x_1) + \langle f'(x_1), x_2 - x_2\rangle + \mu V_{x_1}^r(x_2)\,, \text{ or equivalently}\\
&\langle f'(x_1)(x_1) - f'(x_1)(x_2), x_1 - x_2\rangle \geq 2\mu V_{x_1}^r(x_2)
\end{aligned}$$

for all x_1 and x_2, where at any point x, $f'(x) \in \partial f(x)$ is some sub-gradient of the function in its (Frechet) sub-differential $\partial f(x)$ at that point. Further we say that a function (merely) relatively convex w.r.t. to the Bregman divergence V^r if it is relatively 0-strongly convex w.r.t. V^r .

Definition 7. *We say that a convex function, f is relatively smooth w.r.t. to a Bregman divergence V^r (generated by a strongly convex d.g.f. r)*

$$f(x_2) \leq f(x_1) + \langle \nabla f(x_1), x_2 - x_2\rangle + LV_{x_1}^r(x_2)\,, \text{ or equivalently } \langle \nabla f(x_1) - \nabla f(x_2), x_1 - x_2\rangle \leq 2LV_{x_1}^r(x_2)$$

Definition 8. *For a convex function, F, its relative proximal operator $\operatorname{prox}_{\eta F}^r(x)$ (parameterized by some $\eta > 0$) w.r.t. to a Bregman divergence V^r (generated by a strongly convex d.g.f. r) is defined as*

$$\operatorname{prox}_{\eta F}^r(x) = \arg\min_{\widetilde{x}} F(x) + \frac{1}{\eta}V_x^r(\widetilde{x}) \tag{32}$$

A.5 Minimax Problems

Lemma 7. *Let $\phi(x, y)$ be convex-concave objective. Then $\phi(\widetilde{x}, y^*) - \phi(x^*, \widetilde{y}) \geq 0$ for all $(\widetilde{x}, \widetilde{y}) \in \mathcal{X} \times \mathcal{Y}$, if $(x^*, y^*) \in \arg\min_{x\in\mathcal{X}, y\in\mathcal{Y}} \phi(x, y)$.*

Proof. Notice that the LHS above is positive since

$$\begin{aligned}
\phi(\widetilde{x}, y^*) - \phi(x^*, \widetilde{y}) &= (\phi(\widetilde{x}, y^*) - \phi(x^*, y^*)) + (\phi(x^*, y^*) - \phi(x^*, \widetilde{y}))\\
&= (\phi(\widetilde{x}, y^*) - \min_x \phi(x, y^*)) + (\max_y \phi(x^*, y) - \phi(x^*, \widetilde{y}))\\
&\geq 0 && (33)
\end{aligned}$$

□

B SUPPORTING RESULTS

B.1 Proximal Point method: Proof of Lemma 1

Proof. Since F (H) is μ-relatively strong convexity w.r.t. r (s) and (x_{k+1}, y_{k+1}) satisfies PPM rule (8), we can use mirror descent lemma 9 to get

$$\begin{aligned} F(x_{k+1}) - F(x) + \left\langle A^\top y_{k+1}, x_{k+1} - x \right\rangle &\le \frac{1}{\eta_x} V^r_{x_k}(x) - (\frac{1}{\eta_x} + \mu_x) V^r_{x_{k+1}}(x) - \frac{1}{\eta_x} V^r_{x_k}(x_{k+1}) \\ H(y_{k+1}) - H(y) - \langle A x_{k+1}, y_{k+1} - y \rangle &\le \frac{1}{\eta_y} V^s_{y_k}(y) - (\frac{1}{\eta_y} + \mu_y) V^s_{y_{k+1}}(y) - \frac{1}{\eta_y} V^s_{y_k}(y_{k+1}) \end{aligned} \tag{34}$$

Separately, using convexity and concavity of $\langle y, Ax \rangle$ w.r.t. x and y, we get

$$\phi(x_{k+1}, y) - \phi(x, y_{k+1}) \le F(x_{k+1}) - F(x) + \left\langle A^\top y_{k+1}, x_{k+1} - x \right\rangle + \langle A x_{k+1}, y_{k+1} - y \rangle + H(y_{k+1}) - H(x) \tag{35}$$

Summing three equations and setting $(x, y) = (x^*, y^*)$ we get

$$\phi(x_{k+1}, y^*) - \phi(x^*, y_{k+1}) \le \frac{1}{\eta_x} V^r_{x_k}(x^*) - (\frac{1}{\eta_x} + \mu_x) V^r_{x_{k+1}}(x^*) + \frac{1}{\eta_y} V^s_{y_k}(y^*) - (\frac{1}{\eta_y} + \mu_y) V^s_{y_{k+1}}(y^*) \tag{36}$$

Notice that the LHS above is positive by Lemma 7, that is

$$\phi(x_{k+1}, y^*) - \phi(x^*, y_{k+1}) \ge 0 \tag{37}$$

Let us define $\gamma := 1 + \kappa^{-1}$, where we define also $\kappa := 1/\min(\eta_x \mu_x, \eta_y \mu_y)$. Now multiplying both the sides of (36) with γ^k, and using $\gamma \le 1 + \min(\eta_x \mu_x, \eta_y \mu_y)$ we get

$$\begin{aligned} 0 &\le \gamma^k (\phi(x_{k+1}, y^*) - \phi(x^*, y_{k+1})) \\ &\le \frac{\gamma^k}{\eta_x} V^r_{x_k}(x^*) - \frac{\gamma^{k+1}}{\eta_x} V^r_{x_{k+1}}(x^*) + \frac{\gamma^k}{\eta_y} V^s_{y_k}(y^*) - \frac{\gamma^{k+1}}{\eta_y} V^s_{y_{k+1}}(y^*) \,. \end{aligned} \tag{38}$$

Now summing the above equation from $k = 0$ to $k = K - 1$, we get that

$$\frac{\gamma^K}{\eta_x} V^r_{x_K}(x^*) + \frac{\gamma^K}{\eta_y} V^s_{y_K}(y^*) \le \frac{1}{\eta_x} V^r_{x_0}(x^*) + \frac{1}{\eta_y} V^s_{y_0}(y^*) \,. \tag{39}$$

Finally, dividing both sides using $2\gamma^K$ and using the 1-strongly convexity of r and s and Lemma 5(a) we get

$$\frac{1}{\eta_x} \|x^* - x_k\|^2 + \frac{1}{\eta_y} \|y^* - y_k\|^2 \le \frac{2\gamma^{-K}}{\eta_x} V^r_{x_0}(x^*) + \frac{2\gamma^{-K}}{\eta_y} V^s_{y_0}(y^*) \tag{40}$$

Finally we get the desired result using the fact that $\gamma^{-1} = 1/(1 + \kappa^{-1}) = 1 - 1/(1 + \kappa) \le \exp(1/\kappa + 1)$. □

B.2 Primal Dual Method: Proof of Theorem 1

Since the PD method is an approximation of PPM, former's analysis closely follows that of the latter (proof of Lemma 1).

Proof. Since F (H) is μ-relatively strong convexity w.r.t. r (s) and (x_{k+1}, y_{k+1}) satisfies PPM rule (8), we can use mirror descent lemma 9 to get

$$\begin{aligned} F(x_{k+1}) - F(x) + \left\langle A^\top \widetilde{y}_{k+1}, x_{k+1} - x \right\rangle &\le \frac{1}{\eta_x} V^r_{x_k}(x) - (\frac{1}{\eta_x} + \mu_x) V^r_{x_{k+1}}(x) - \frac{1}{\eta_x} V^r_{x_k}(x_{k+1}) \\ H(y_{k+1}) - H(y) - \langle A x_{k+1}, y_{k+1} - y \rangle &\le \frac{1}{\eta_y} V^s_{y_k}(y) - (\frac{1}{\eta_y} + \mu_y) V^s_{y_{k+1}}(y) - \frac{1}{\eta_y} V^s_{y_k}(y_{k+1}) \end{aligned} \tag{41}$$

Separately, using convexity and concavity of $\langle y, Ax \rangle$ w.r.t. x and y, we get

$$\phi(x_{k+1}, y) - \phi(x, y_{k+1}) \le F(x_{k+1}) - F(x) + \left\langle A^\top y_{k+1}, x_{k+1} - x \right\rangle + \langle A x_{k+1}, y_{k+1} - y \rangle + H(y_{k+1}) - H(x) \tag{42}$$

Summing above three equations and setting $(x, y) = (x^*, y^*)$ we get

$$\phi(x_{k+1}, y^*) - \phi(x^*, y_{k+1}) \leq \frac{1}{\eta_x} V^r_{x_k}(x^*) - (\frac{1}{\eta_x} + \mu_x) V^r_{x_{k+1}}(x^*) + \frac{1}{\eta_y} V^s_{y_k}(y^*) - (\frac{1}{\eta_y} + \mu_y) V^s_{y_{k+1}}(y^*) + \left\langle A^\top(y_{k+1} - \widetilde{y}_{k+1}), x_{k+1} - x^* \right\rangle - \frac{1}{\eta_x} V^r_{x_k}(x_{k+1}) - \frac{1}{\eta_v} V^s_{y_k}(y_{k+1}) \tag{43}$$

We can further expand out the last four term in the above inequality as follows. Using $\widetilde{y}_{k+1} = y_k + \theta(y_k - y_{k-1})$ (equation (8)) and Cauchy-Schwarz inequality we get

$$\begin{aligned}\left\langle A^\top(y_{k+1} - \widetilde{y}_{k+1}), x_{k+1} - x^* \right\rangle &= \theta \left\langle y_{k-1} - y_k, A(x_k - x^*) \right\rangle - \left\langle y_k - y_{k+1}, A(x_{k+1} - x^*) \right\rangle + \\ &\quad \theta \left\langle y_{k-1} - y_k, A(x_{k+1} - x_k) \right\rangle \\ &\leq \theta \left\langle y_{k-1} - y_k, A(x_k - x^*) \right\rangle - \left\langle y_k - y_{k+1}, A(x_{k+1} - x^*) \right\rangle + \\ &\quad \frac{\theta \|A\| \alpha_y}{2} \|y_{k-1} - y_k\|^2 + \frac{\theta \|A\|}{2\alpha_y} \|x_{k+1} - x_k\|^2 \end{aligned} \tag{44}$$

for some $\alpha_y = \sqrt{\mu_x/\mu_y}$. Using Lemma 5(a) and 1-strong convexity of $\underline{f}^*$ we get that

$$-\frac{1}{\eta_x} V^r_{x_k}(x_{k+1}) \leq -\frac{1}{2\eta_x} \|x_{k+1} - x_k\|^2 \tag{45}$$

$$-\frac{1}{\eta_y} V^s_{y_k}(y_{k+1}) \leq -\frac{1}{2\eta_y} \|y_{k+1} - y_k\|^2 \tag{46}$$

Summing equations (44), (45) and (46), and using $\frac{\alpha_x \|A\|}{2} = \frac{\|A\|}{2}\sqrt{\frac{\mu_x}{\mu_y}} \leq \|A\|\sqrt{\frac{\mu_x}{\mu_y}} = \frac{1}{2\eta_y}$ and $\theta\frac{\|A\|}{2\alpha_y} = \theta\frac{\|A\|}{2}\sqrt{\frac{\mu_y}{\mu_x}} \leq \|A\|\sqrt{\frac{\mu_y}{\mu_x}} = \frac{1}{2\eta_x}$ we get

$$\begin{aligned}&\left\langle A^\top(y_{k+1} - \widetilde{y}_{k+1}), x_{k+1} - x^* \right\rangle - \frac{1}{\eta_x} V^r_{x_k}(x_{k+1}) - \frac{1}{\eta_v} V^s_{y_k}(y_{k+1}) \\ &\leq \theta \left\langle y_{k-1} - y_k, A(x_k - x^*) \right\rangle - \left\langle y_k - y_{k+1}, A(x_{k+1} - x^*) \right\rangle + \\ &\quad - (\frac{1}{2\eta_x} - \theta\frac{\|A\|}{2\alpha_y}) \|x_{k+1} - x_k\|^2 + \theta\frac{\|A\|\alpha_y}{2} \|y_k - y_{k-1}\|^2 - \frac{1}{2\eta_y} \|y_{k+1} - y_k\|^2 \\ &\leq \theta \left\langle y_{k-1} - y_k, A(x_k - x^*) \right\rangle - \left\langle y_k - y_{k+1}, A(x_{k+1} - x^*) \right\rangle + \\ &\quad \theta\frac{\|A\|\alpha_y}{2} \|y_k - y_{k-1}\|^2 - \frac{\|A\|\alpha_y}{2} \|y_{k+1} - y_k\|^2 \end{aligned} \tag{47}$$

Notice that the LHS of (43) is positive by Lemma 7, that is $\phi(x_{k+1}, y^*) - \phi(x^*, y_{k+1}) \geq 0$. Summing equations (43) and (47),and using the above fact we get

$$\begin{aligned} 0 &\leq \phi(x_{k+1}, y^*) - \phi(x^*, y_{k+1}) \\ &\leq \frac{1}{\eta_x} V^r_{x_k}(x^*) - (\frac{1}{\eta_x} + \mu_x) V^r_{x_{k+1}}(x^*) + \frac{1}{\eta_y} V^s_{y_k}(y) - (\frac{1}{\eta_y} + \mu_y) V^s_{y_{k+1}}(y) + \\ &\quad \theta \left\langle y_{k-1} - y_k, A(x_k - x^*) \right\rangle - \left\langle y_k - y_{k+1}, A(x_{k+1} - x^*) \right\rangle + \\ &\quad \theta\frac{\|A\|\alpha_y}{2} \|y_k - y_{k-1}\|^2 - \frac{\|A\|\alpha_y}{2} \|y_{k+1} - y_k\|^2 \end{aligned} \tag{48}$$

Let us define $\gamma := 1 + \kappa^{-1}$, where we define also $\kappa := 1/\min(\eta_x\mu_x, \eta_y\mu_y) = 2\|A\|/\sqrt{\mu_x\mu_y}$. Now multiplying both the sides of (48) with γ^k, and using $\gamma \leq 1 + \min(\eta_x\mu_x, \eta_y\mu_y)$ and $\theta = 1/\gamma$ we get

$$\begin{aligned} 0 \leq &\frac{\gamma^k}{\eta_x} V^r_{x_k}(x^*) - \frac{\gamma^{k+1}}{\eta_x} V^r_{x_{k+1}}(x^*) + \frac{\gamma^k}{\eta_y} V^s_{y_k}(y^*) - \frac{\gamma^{k+1}}{\eta_y} V^s_{y_{k+1}}(y^*) \\ &\gamma^{k-1} \left\langle y_{k-1} - y_k, A(x_k - x^*) \right\rangle - \gamma^k \left\langle y_k - y_{k+1}, A(x_{k+1} - x^*) \right\rangle + \\ &\gamma^{k-1}\frac{\|A\|\alpha_y}{2} \|y_k - y_{k-1}\|^2 - \gamma^k\frac{\|A\|\alpha_y}{2} \|y_{k+1} - y_k\|^2 \end{aligned} \tag{49}$$

Now summing the above equation from $k=0$ to $k=K-1$ and using $y_{-1}=y_0$, we get that

$$\frac{\gamma^K}{\eta_x}V^r_{x_K}(x^*)+\frac{\gamma^K}{\eta_y}V^s_{y_K}(y^*)\le\frac{1}{\eta_x}V^r_{x_0}(x^*)+\frac{1}{\eta_y}V^s_{y_0}(y^*)+\\ \gamma^{K-1}\langle y_{K-1}-y_K,A(x_K-x^*)\rangle+\gamma^{K-1}\frac{\|A\|\alpha_y}{2}\|y_K-y_{K-1}\|^2 \tag{50}$$

Using Cauchy-Schwarz inequality, $\theta\frac{\|A\|}{2\alpha_y}=\theta\frac{\|A\|}{2}\sqrt{\frac{\mu_y}{\mu_x}}\le\|A\|\sqrt{\frac{\mu_y}{\mu_x}}=\frac{1}{2\eta_x}$, Lemma 5(a) we can show that

$$\begin{aligned}&\gamma^{K-1}\langle y_{K-1}-y_K,A(x_K-x^*)\rangle-\gamma^{K-1}\frac{\|A\|\alpha_y}{2}\|y_{k+1}-y_k\|^2\\&\le\gamma^{K-1}\frac{\|A\|}{2\alpha_y}\|x_K-x^*\|^2+\gamma^{K-1}\frac{\|A\|\alpha_y}{2}\|y_K-y_{K-1}\|^2-\gamma^{K-1}\frac{\|A\|\alpha_y}{2}\|y_K-y_{K-1}\|^2\\&\le\gamma^{K-1}\frac{\|A\|}{2\alpha_y}\|x_K-x^*\|^2\\&\le\gamma^K\frac{1}{4\eta_y}\|x_K-x^*\|^2\\&\le\gamma^K\frac{1}{2\eta_y}V^r x_K(x^*)\end{aligned} \tag{51}$$

Summing equations (51) and (52), and dividing both sides of the resulting equation using $2\gamma^K$ and using the 1-strongly convexity of r and s and Lemma 5(a) we get

$$\frac{1}{2\eta_x}\|x^*-x_k\|^2+\frac{1}{\eta_y}\|y^*-y_k\|^2\le\frac{2\gamma^{-K}}{\eta_x}V^r_{x_0}(x^*)+\frac{2\gamma^{-K}}{\eta_y}V^r_{y_0}(y^*) \tag{52}$$

Finally we get the desired result by using the choice $\gamma=1+\min(\eta_x\mu_x,\eta_y\mu_y)=1+\kappa^{-1}$, which implies that $\gamma^{-1}=1/(1+\kappa^{-1})=1-1/(1+\kappa)\le\exp(1/\kappa+1)$. □

B.3 Proof of Lemma 8

Lemma 8. *If f is μ-strongly convex and L-smooth, then $\underline{f}=f-\mu\|\cdot\|^2/2$ is convex and $(L-\mu)$-smooth.*

Proof. It can be easily proved by noticing that

$$\begin{aligned}\langle\nabla\underline{f}(x)-\nabla\underline{f}(\widetilde{x}),x-\widetilde{x}\rangle&=\langle\nabla f(x)-\nabla f(\widetilde{x}),x-\widetilde{x}\rangle+\langle-\mu x+\mu\widetilde{x},x-\widetilde{x}\rangle\\&\le(L-\mu)\|x-\widetilde{x}\|^2\end{aligned} \tag{53}$$

Similarly we can also easily show that $\langle\nabla\underline{f}(x)-\nabla\underline{f}(\widetilde{x}),x-\widetilde{x}\rangle\ge0$ □

B.4 Proof of Corollary 1

We omit the proof of Corollary 1 since it is very similar to that of Theorem 1. Only additional step is to upper-bound $V^{f^*}_{u_0}(u^*)$ by $(L_x-\mu_x)\|x_0-x^*\|^2/2$ using $u^*=\nabla\underline{f}(x^*)$ and $u_0=\nabla\underline{f}(x_0)$ and Lemma 6.

B.5 Proof of Lemma 2

Proof. We want to prove that x_k iterates of (8) (repeated below)

$$\begin{cases}\widetilde{u}_{k+1}=u_k+\theta(u_k-u_{k-1})\\x_{k+1}=\arg\min_x\langle\tilde{u}_{k+1},x\rangle+\frac{\mu}{2}\|x\|^2+\frac{1}{2\eta_x}\|x-x_k\|^2\\u_{k+1}=\arg\min_u-\langle x_{k+1},u\rangle+\underline{f}^*(u)+\frac{1}{\eta_u}V^{\underline{f}^*}_{u_k}(u)\end{cases} \tag{54}$$

and (11) (repeated below)

$$\begin{cases} \widetilde{\nabla}_{k+1} = \nabla \underline{f}(\underline{x}_k) + \theta(\nabla \underline{f}(\underline{x}_k) - \nabla \underline{f}(\underline{x}_{k-1})) \\ x_{k+1} = (x_k - \eta_x \widetilde{\nabla}_{k+1})/(1+\eta_x\mu) \\ \underline{x}_{k+1} = (\underline{x}_k + \eta_u x_{k+1})/(1+\eta_u) \end{cases} \tag{55}$$

are equivalent under the given condition. For this we will prove a stronger condition which additionally states that $u_k = \nabla \underline{f}(\underline{x}_k)$ for all $k = -1, 0, 1, \ldots$.

We prove this by induction. Let us initialize both the updates using x_0. For the base case it is easy to see that when $u_{-1} = u_0 = \nabla f(\underline{x}_{-1}) = \nabla f(\underline{x}_0) = \nabla f(x_0)$.

Let $u_{\widetilde{k}-1} = \nabla f(\underline{x}_{\widetilde{k}-1})$ and $u_{\widetilde{k}} = \nabla f(\underline{x}_{\widetilde{k}})$, and $x_{\widetilde{k}}$ iterates of the both the rules match for $\widetilde{k} = 0, 1, \ldots, k$. Then clearly, $\widetilde{u}_{k+1} = \widetilde{\nabla}_{k+1}$. This implies that x_{k+1} iterates are the same for both the rules.

Next we will prove that $u_{k+1} = \nabla f(\underline{x}_{k+1})$. Note that $u_{k+1} = \arg\min_u -\langle x_{k+1}, u\rangle + \underline{f}^*(u) + \frac{1}{\eta_u} V_{u_k}^{\underline{f}^*}(u)$. However, $V_{u_k}^{\underline{f}^*}(u) = \underline{f}^*(u) - \underline{f}^*(u_k) - \langle (\underline{f}^*)'(u_k), u - u_k\rangle$ is not defined unless we fix a sub-gradient $(\underline{f}^*)'(u_k) \in \partial \underline{f}^*(u_k)$ at u_k. For making the rules equivalent we set $(\underline{f}^*)'(u_k) = \underline{x}_k$. Note that $\underline{x}_k \in \partial \underline{f}^*(u_k)$ since $u_k = \nabla \underline{f}(\underline{x}_k)$ (Lemma 4(d)). Then $u_{k+1} = \nabla f(\underline{x}_{k+1})$, since

$$\begin{aligned} u_{k+1} &= \arg\min_u -\langle x_{k+1}, u\rangle + \underline{f}^*(u) + \frac{1}{\eta_u} V_{u_k}^{\underline{f}^*}(u) \\ &= \arg\min_u -\langle \eta_u x_{k+1} + (\underline{f}^*)'(u_k), u\rangle + (1+\eta_u)\underline{f}^*(u) \\ &= \arg\min_u -\langle \eta_u x_{k+1} + \underline{x}_k, u\rangle + (1+\eta_u)\underline{f}^*(u) \end{aligned} \tag{56}$$

and by Lemma 4(b) $u_{k+1} = \nabla f(\underline{x}_{k+1})$ is a valid and only solution (because of strong convexity of $\underline{f}^*$) to the above optimization, where $\underline{x}_{k+1} = (\underline{x}_k + \eta_u x_{k+1})/(1+\eta_u)$. Hence, we prove the equivalence between the rules by induction. □

B.6 Extension of LPD to a problem with additional proximal-friendly terms

LPD can be extended to solve more general (possibly nonsmooth) minimax problems with the same guarantees:

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \; F(x) + f(x) + \langle y, Ax\rangle - h(y) - H(x), \tag{57}$$

where F and H are convex (and possibly non-smooth) and we have access to their proximal operators and f, h satisfy Assumption 1. The only change we need to make is to replace the x_{k+1} and y_{k+1} update steps in Algorithm 2 with

$$\begin{aligned} x_{k+1} = \arg\min_{x\in\mathcal{X}} &\left\langle A^\top \widetilde{y}_{k+1} + \widetilde{u}_{k+1}, x\right\rangle + \\ &\|x - x_k\|^2/2\eta_x + \mu_x\|x\|^2/2 + F(x) \\ y_{k+1} = \arg\min_{y\in\mathcal{Y}} &-\left\langle A^\top \widetilde{x}_{k+1} + \widetilde{v}_{k+1}, y\right\rangle + \\ &\|y - y_k\|^2/2\eta_y + \mu_y\|y\|^2/2 + H(x)\,. \end{aligned} \tag{58}$$

Then the same guarantees as Corollaries 2 and 3 holds for this update. We omit the analysis since it is similar to the proof of Theorem 4.

B.7 Mirror-Descent lemma

Lemma 9 (Nesterov et al. 2018)**.** *Let r be strongly convex, F be μ-(relatively) strongly w.r.t. to r, and*

$$x_{k+1} = \arg\min_x \langle g, x\rangle + F(x) + \frac{1}{\eta} V_{x_k}^r(x) \tag{59}$$

then

$$\langle g, x_{k+1} - x\rangle + F(x_{k+1}) - F(x) \le \frac{1}{\eta} V_{x_k}^r(x) - (\frac{1}{\eta} + \mu) V_{x_{k+1}}^r(x) - \frac{1}{\eta} V_{x_k}^r(x_{k+1}) \tag{60}$$

Algorithm 2 O-LPD: Original Lifted Primal-Dual algorithm

Required: $\mathcal{X}$, $\mathcal{Y}$, (f, L_x, μ_x), $(A, \|A\|)$, (h, L_y, μ_y), K, $\{(\eta_{x,k}, \eta_{y,k}, \eta_{u,k}, \eta_{v,k}, \theta_k)\}_{k=0}^{K-1}$

Initialize $(x_{-1}, y_{-1}) = (x_0, y_0) \in \mathcal{X} \times \mathcal{Y}$

Set $\underline{f} = f - (\mu_x/2)\|\cdot\|^2$, $\underline{h} = h - (\mu_y/2)\|\cdot\|^2$, $(\underline{x}_{-1}, \underline{y}_{-1}) = (\underline{x}_0, \underline{y}_0) = (x_0, y_0)$

for $0 \le k \le K-1$ **do**

$\widetilde{x}_{k+1} = x_k + \theta_k(x_k - x_{k-1})$, $\widetilde{y}_{k+1} = y_k + \theta_k(y_k - y_{k-1})$,
$\widetilde{u}_{k+1} = u_k + \theta_k(u_k - u_{k-1})$, $\widetilde{v}_{k+1}) = v_k + \theta_k(v_k - v_{k-1})$

$x_{k+1} = \arg\min_{x\in\mathcal{X}} \left\langle A^\top \widetilde{y}_{k+1} + \widetilde{u}_{k+1}, x\right\rangle + \frac{1}{2\eta_{x,k}}\|x - x_k\|^2 + \frac{\mu_x}{2}\|x\|^2$

$y_{k+1} = \arg\min_{y\in\mathcal{Y}} -\langle A\widetilde{x}_{k+1} - \widetilde{v}_{k+1}, y\rangle + \frac{1}{2\eta_{y,k}}\|y - y_k\|^2 + \frac{\mu_y}{2}\|y\|^2$

$u_{k+1} = \arg\min_u -\langle x_{k+1}, u\rangle + \underline{f}^*(u) + V_{u_k}^{f^*}(u)/\eta_u$

$v_{k+1} = \arg\min_v -\langle y_{k+1}, v\rangle + \underline{h}^*(v) + V_{v_k}^{h^*}(v)/\eta_v$

end

return (x_K, y_K, u_K, v_K)

C ALGORITHM FOR BILINERALY-COUPLED SMOOTH MINIMAX PROBLEM

First we will prove a general result for Bilineraly-coupled smooth minimax problem. Then we specialize it to the Bi-SC-SC and Bi-C-SC cases.

As mentioned in the main text we first apply the follow reformulation to (12).

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} [g(x,y) = f(x) + \langle y, Ax\rangle - h(y)] \tag{61}$$

$$= \min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}}\min_{v}\max_{u}[g(x,y;u,v) = -\underline{f}^*(u) + \langle u, x\rangle + \frac{\mu_x}{2}\|x\|^2 + \langle y, Ax\rangle - \frac{\mu_y}{2}\|y\|^2 - \langle v, y\rangle + \underline{h}^*(v)] \tag{62}$$

where

$$\underline{f}^*(u) := \max_{x\in\mathcal{X}\cap\mathrm{dom}(f)} \langle u, x\rangle - [\underline{f} := f(x) - (\mu_x/2)\|x\|^2] \tag{63}$$

$$\underline{h}^*(v) := \max_{y\in\mathcal{Y}\cap\mathrm{dom}(h)} \langle v, y\rangle - [\underline{h} := h(x) - (\mu_y/2)\|y\|^2] \tag{64}$$

Note that by Lemma 8, $\underline{f}$ is convex and $(L_x - \mu_x)$-smooth, and $\underline{h}$ is convex and $(L_y - \mu_y)$-smooth. Then by Lemma 4(a) $\underline{f}^*$ is $1/(L_x - \mu_x)$-strongly convex, and $\underline{h}^*$ is $1/(L_y - \mu_y)$-strongly convex.

Instead of analyzing the Algorithm 1, we analyze the original update rule (16) (Algorithm 2) which is a conceptually easier implementation of LPD. By the following lemma we show that Algorithm 1 and Algorithm 2 of these are equivalent, when initialized appropriately.

Lemma 10 (Same as Lemma 3)**.** *Let us initialize Algorithm 1 with* $(\underline{x}_{-1}, \underline{x}_0, \underline{y}_{-1}, \underline{y}_0) = (x_0, x_0, y_0, y_0)$*, Algorithm 2 with* $(u_{-1}, u_0, v_{-1}, v_0) = (\nabla\underline{f}(\underline{x}_{-1}), \nabla\underline{f}(\underline{x}_0), \nabla\underline{h}(\underline{x}_{-1}), \nabla\underline{h}(\underline{x}_0))$*, and both the algorithms with the same* (x_0, y_0)*. Then for problem (61), iterates* (x_k, y_k) *of the Algorithm 1 and Algorithm 2 are the same.*

Proof. We omit the proof since we can easily prove it using the same techniques as used in the proof of Lemma 2. □

We prove the follow Theorem for characterizing the output of Algorithm 2.

Theorem 4. *Let there exists positive numbers* λ_k, $\alpha_{x,k}$, $\alpha_{y,k}$, $\alpha_{u,k}$, $\alpha_{v,k}$ *for all* $k = -1, 0, 1\ldots$, *such that*

$\lambda_{k-1} = \theta_k \lambda_k$,

$$\theta_k\left(\frac{\|A\|}{\alpha_{y,k}} + \frac{1}{\alpha_{u,k}}\right) + \|A\|\alpha_{x,k+1} \le \frac{1}{\eta_{x,k}}\,, \quad \alpha_u \le \frac{1}{\eta_{u,k}(L_x - \mu_x)}\,, \tag{65}$$

$$\theta_k\left(\frac{\|A\|}{\alpha_x} + \frac{1}{\alpha_v}\right) + \|A\|\alpha_{y,k+1} \le \frac{1}{\eta_{y,k}}\,, \quad \alpha_v \le \frac{1}{\eta_{v,k}(L_y - \mu_y)} \tag{66}$$

$$\frac{\lambda_{k+1}}{\lambda_k} \le \min\left(\frac{\eta_{x,k+1}(1+\eta_{x,k}\mu_x)}{\eta_{x,k}}\,, \frac{\eta_{y,k+1}(1+\eta_{y,k}\mu_y)}{\eta_{x,k}}\,, \frac{\eta_{u,k+1}(1+\eta_{u,k})}{\eta_{x,k}}\,, \frac{\eta_{v,k+1}(1+\eta_{v,k})}{\eta_{x,k}}\right) \tag{67}$$

for all $k = 0, 1, \ldots$. Then the following is true for any $K = 1, 2, \ldots$, $x \in \mathcal{X}$, $y \in \mathcal{Y}$, u, v

$$\begin{aligned}
&\sum_{k=0}^{K-1} \lambda_k[\Phi(x_{k+1}, y; u, v_{k+1}) - \Phi(x, y_{k+1}; u_{k+1}, v)] \\
&\quad\le \frac{\lambda_0}{2\eta_{x,0}}\|x - x_0\|^2 - \frac{\lambda_K\|A\|\alpha_{x,K+1}}{2}\|x - x_K\|^2 + \frac{\lambda_0}{2\eta_{y,0}}\|y - y_0\|^2 - \frac{\lambda_K\|A\|\alpha_{y,K+1}}{2}\|y - y_K\|^2 + \\
&\quad\quad \frac{\lambda_0}{\eta_{u,0}} V_{u_0}^{\underline{f}^*}(u) - \frac{\lambda_K}{\eta_{u,K}} V_{u_K}^{\underline{f}^*}(u) + \frac{\lambda_0}{\eta_{v,0}} V_{v_0}^{\underline{h}^*}(v) - \frac{\lambda_K}{\eta_{v,K}} V_{v_K}^{\underline{h}^*}(v)
\end{aligned} \tag{68}$$

Note that proof of Theorem 4 closely follows the steps used in the proof of Lemma 1.

Proof. Let $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Using Steps 4 and 5 (Algorithm 2) and Lemma 9 twice—once with $g = A^\top \widetilde{y}_{k+1} + \widetilde{u}_{k+1}$, $F = (\mu_x/2)\|\cdot\|^2 + F$ and $r = \|\cdot\|^2/2$, and second time with $g = -A\widetilde{x}_{k+1} + \widetilde{v}_{k+1}$, $F = (\mu_y/2)\|\cdot\|^2 + H$ and $r = \|\cdot\|^2/2$—we get

$$\begin{aligned}
&\langle A^\top \widetilde{y}_{k+1} + \widetilde{u}_{k+1}, x_{k+1} - x\rangle + \frac{\mu_x}{2}(\|x_{k+1}\|^2 - \|x\|^2) + F(x_{k+1}) - F(x) \\
&\qquad\le \frac{1}{2\eta_{x,k}}(\|x - x_k\|^2 - \|x - x_{k+1}\|^2 - \|x_{k+1} - x_k\|^2) - \frac{\mu_x}{2}\|x - x_{k+1}\|^2
\end{aligned} \tag{69}$$

$$\begin{aligned}
&\langle -A\widetilde{x}_{k+1} + \widetilde{v}_{k+1}, y_{k+1} - x\rangle + \frac{\mu_y}{2}(\|y_{k+1}\|^2 - \|y\|^2) + H(y_{k+1}) - H(y) \\
&\qquad\le \frac{1}{2\eta_{y,k}}(\|y - y_k\|^2 - \|y - y_{k+1}\|^2 - \|y_{k+1} - y_k\|^2) - \frac{\mu_y}{2}\|y - y_{k+1}\|^2
\end{aligned} \tag{70}$$

Note that $\underline{f}^*$ and $\underline{h}^*$ are 1-strong convex w.r.t themselves. Again using Step 6 (Algorithm 2) and Lemma 9 twice—once with $g = -x_{k+1}$, $F = \underline{f}^*$ and $r = \underline{f}^*$, and second time time with $g = -y_{k+1}$, $F = \underline{h}^*$ and $r = \underline{h}^*$—we get

$$\langle -x_{k+1}, u_{k+1} - u\rangle + \underline{f}^*(u_{k+1}) - \underline{f}^*(u) \le \frac{1}{\eta_{u,k}}(V_{u_k}^{\underline{f}^*}(u) - V_{u_{k+1}}^{\underline{f}^*}(u) - V_{u_k}^{\underline{f}^*}(u_{k+1})) - V_{u_{k+1}}^{\underline{f}^*}(u) \tag{71}$$

$$\langle y_{k+1}, v_{k+1} - v\rangle + \underline{h}^*(v_{k+1}) - \underline{h}^*(v) \le \frac{1}{\eta_{v,k}}(V_{v_k}^{\underline{h}^*}(v) - V_{v_{k+1}}^{\underline{h}^*}(v) - V_{v_k}^{\underline{h}^*}(v_{k+1})) - V_{v_{k+1}}^{\underline{h}^*}(v) \tag{72}$$

Adding the above four equations and using the definition $\mathrm{gap}_{z,w}(z_{k+1}, w_{k+1}) = \Phi(x_{k+1}, y; u, v_{k+1}) - \Phi(x, y_{k+1}; u_{k+1}, v)$, where $z = (x, y)$ and $w = (u, v)$

$$\begin{aligned}
\mathrm{gap}_{z,w}(z_{k+1}, w_{k+1}) &= \Phi(x_{k+1}, y; u, v_{k+1}) - \Phi(x, y_{k+1}, u_{k+1}, v) \\
&\le \frac{1}{2\eta_{x,k}}\|x - x_k\|^2 - (\frac{1}{2\eta_{x,k}} + \frac{\mu_x}{2})\|x - x_{k+1}\|^2 - \frac{1}{2\eta_{x,k}}\|x_{k+1} - x_k\|^2 + \\
&\quad \frac{1}{2\eta_{y,k}}\|y - y_k\|^2 - (\frac{1}{2\eta_{y,k}} + \frac{\mu_y}{2})\|y - y_{k+1}\|^2 - \frac{1}{2\eta_{y,k}}\|y_{k+1} - y_k\|^2 + \\
&\quad \frac{1}{\eta_{u,k}} V_{u_k}^{\underline{f}^*}(u) - (\frac{1}{\eta_{u,k}} + 1) V_{u_{k+1}}^{\underline{f}^*}(u) - \frac{1}{\eta_{u,k}} V_{u_k}^{\underline{f}^*}(u_{k+1}) + \\
&\quad \frac{1}{\eta_{v,k}}(V_{v_k}^{\underline{h}^*}(v) - (\frac{1}{\eta_{v,k}} + 1) V_{v_{k+1}}^{\underline{h}^*}(v) - \frac{1}{\eta_{v,k}} V_{v_k}^{\underline{h}^*}(v_{k+1}) + \\
&\quad \langle y_{k+1} - \widetilde{y}_{k+1}, A(x_{k+1} - x)\rangle + -\langle y_{k+1} - y, A(x_{k+1} - \widetilde{x}_{k+1})\rangle + \\
&\quad \langle u_{k+1} - \widetilde{u}_{k+1}, x_{k+1} - x\rangle + \langle v_{k+1} - \widetilde{v}_{k+1}, y_{k+1} - y\rangle
\end{aligned} \tag{73}$$

We can further expand out the last four term in the above inequality as follows. Using Step 3 (Algorithm 2) and Cauchy-Schwarz inequality we get

$$\begin{aligned}\langle y_{k+1}-\widetilde{y}_{k+1}, A(x_{k+1}-x)\rangle &= \theta_k \langle y_{k-1}-y_k, A(x_k-x)\rangle - \langle y_k-y_{k+1}, A(x_{k+1}-x)\rangle \ + \\ &\quad \theta_k \langle y_{k-1}-y_k, A(x_{k+1}-x_k)\rangle \\ &\le \theta_k \langle y_{k-1}-y_k, A(x_k-x)\rangle - \langle y_k-y_{k+1}, A(x_{k+1}-x)\rangle \ + \\ &\quad \frac{\theta_k\|A\|\alpha_{y,k}}{2}\|y_{k-1}-y_k\|^2 + \frac{\theta_k\|A\|}{2\alpha_{y,k}}\|x_{k+1}-x_k\|^2 \end{aligned} \tag{74}$$

for some $\alpha_{y,k}\ge 0$. Similarly we can show that

$$\begin{aligned}-\langle y_{k+1}-y, A(x_{k+1}-\widetilde{x}_{k+1})\rangle &\le -\theta_k \langle y_k-y, A(x_{k-1}-x_k)\rangle + \langle y_{k+1}-y, A(x_k-x_{k+1})\rangle \ + \\ &\quad \frac{\theta_k\|A\|\alpha_{x,k}}{2}\|x_{k-1}-x_k\|^2 + \frac{\theta_k\|A\|}{2\alpha_{x,k}}\|y_{k+1}-y_k\|^2 \end{aligned} \tag{75}$$

$$\begin{aligned}\langle u_{k+1}-\widetilde{u}_{k+1}, x_{k+1}-x\rangle &\le \theta_k \langle u_{k-1}-u_k, x_k-x\rangle - \langle u_k-u_{k+1}, x_{k+1}-x\rangle \ + \\ &\quad \frac{\theta_k\alpha_{u,k}}{2}\|u_{k-1}-u_k\|^2 + \frac{\theta_k}{2\alpha_{u,k}}\|x_{k+1}-x_k\|^2 \end{aligned} \tag{76}$$

$$\begin{aligned}\langle v_{k+1}-\widetilde{v}_{k+1}, y_{k+1}-y\rangle &\le \theta_k \langle v_{k-1}-v_k, y_k-y\rangle - \langle v_k-v_{k+1}, y_{k+1}-y\rangle \ + \\ &\quad \frac{\theta_k\alpha_{v,k}}{2}\|v_{k-1}-v_k\|^2 + \frac{\theta_k}{2\alpha_{v,k}}\|y_{k+1}-y_k\|^2 \end{aligned} \tag{77}$$

for some $\alpha_{x,k}\ge 0$, $\alpha_{u,k}\ge 0$, and $\alpha_{v,k}\ge 0$. Using Lemma 5(a) and $1/(L_x-\mu_x)$- and $1/(L_y-\mu_y)$-strong convexity of $\underline{f}^*$ and $\underline{h}^*$, respectively we get that

$$-\frac{1}{\eta_{u,k}}V_{u_k}^{\underline{f}^*}(u_{k+1}) \le -\frac{1}{2\eta_{u,k}(L_x-\mu_x)}\|u_{k+1}-u_k\|^2 \tag{78}$$

$$-\frac{1}{\eta_{v,k}}V_{v_k}^{\underline{h}^*}(v_{k+1}) \le -\frac{1}{2\eta_{v,k}(L_y-\mu_y)}\|v_{k+1}-v_k\|^2 \tag{79}$$

Summing equations (73), (74), (75), (76), (77), (78), and (79) up we get

$$\begin{aligned}\mathrm{gap}_{z,w}(z_{k+1},w_{k+1}) \ \le\ & \frac{1}{2\eta_{x,k}}\|x-x_k\|^2 - (\frac{1}{2\eta_{x,k}}+\frac{\mu_x}{2})\|x-x_{k+1}\|^2 + \frac{1}{2\eta_{y,k}}\|y-y_k\|^2 - (\frac{1}{2\eta_{y,k}}+\frac{\mu_y}{2})\|y-y_{k+1}\|^2 \ + \\ & \frac{1}{\eta_{u,k}}V_{u_k}^{\underline{f}^*}(u) - (\frac{1}{\eta_{u,k}}+1)V_{u_{k+1}}^{\underline{f}^*}(u) + \frac{1}{\eta_{v,k}}V_{v_k}^{\underline{h}^*}(v) - (\frac{1}{\eta_{v,k}}+1)V_{v_{k+1}}^{\underline{h}^*}(v) \ + \\ & \theta_k \langle y_{k-1}-y_k, A(x_k-x)\rangle - \langle y_k-y_{k+1}, A(x_{k+1}-x)\rangle \ + \\ & -\theta_k \langle y_k-y, A(x_{k-1}-x_k)\rangle + \langle y_{k+1}-y, A(x_k-x_{k+1})\rangle \ + \\ & \theta_k \langle u_{k-1}-u_k, x_k-x\rangle - \langle u_k-u_{k+1}, x_{k+1}-x\rangle \ + \\ & \theta_k \langle v_{k-1}-v_k, y_k-y\rangle - \langle v_k-v_{k+1}, y_{k+1}-y\rangle \ + \\ & \theta_k\frac{\|A\|\alpha_{x,k}}{2}\|x_k-x_{k-1}\|^2 - (\frac{1}{2\eta_{x,k}} - \theta_k(\frac{\|A\|}{2\alpha_{y,k}}+\frac{1}{2\alpha_{u,k}}))\|x_{k+1}-x_k\|^2 \ + \\ & \theta_k\frac{\|A\|\alpha_{y,k}}{2}\|y_k-y_{k-1}\|^2 - (\frac{1}{2\eta_{y,k}} - \theta_k(\frac{\|A\|}{2\alpha_{x,k}}+\frac{1}{2\alpha_{v,k}}))\|y_{k+1}-y_k\|^2 \ + \\ & \theta_k\frac{\alpha_{u,k}}{2}\|u_k-u_{k-1}\|^2 - \frac{1}{2\eta_{u,k}(L_x-\mu_x)}\|u_{k+1}-u_k\|^2 \ + \\ & \theta_k\frac{\alpha_{v,k}}{2}\|v_k-v_{k-1}\|^2 - \frac{1}{2\eta_{v,k}(L_y-\mu_y)}\|v_{k+1}-v_k\|^2 \end{aligned} \tag{80}$$

Assuming $\theta_k(\frac{\|A\|}{2\alpha_{y,k}}+\frac{1}{2\alpha_{u,k}})+\frac{\|A\|\alpha_{x,k+1}}{2}\le\frac{1}{2\eta_{x,k}}$, $\theta_k(\frac{\|A\|}{2\alpha_{x,k}}+\frac{1}{2\alpha_{v,k}})+\frac{\|A\|\alpha_{y,k+1}}{2}\le\frac{1}{2\eta_{y,k}}$, $\alpha_{u,k}\le\frac{1}{\eta_{u,k+1}(L_x-\mu_x)}$, and

$\alpha_{v,k} \leq \frac{1}{\eta_{v,k+1}(L_y-\mu_y)}$ we get

$$\begin{aligned}
\text{gap}_{z,w}(z_{k+1}, w_{k+1}) \leq\ & \frac{1}{2\eta_{x,k}}\|x-x_k\|^2 - (\frac{1}{2\eta_{x,k}}+\frac{\mu_x}{2})\|x-x_{k+1}\|^2 + \frac{1}{2\eta_{y,k}}\|y-y_k\|^2 - (\frac{1}{2\eta_{y,k}}+\frac{\mu_y}{2})\|y-y_{k+1}\|^2 + \\
& \frac{1}{\eta_{u,k}}V_{u_k}^{f^*}(u) - (\frac{1}{\eta_{u,k}}+1)V_{u_{k+1}}^{f^*}(u) + \frac{1}{\eta_{v,k}}V_{v_k}^{h^*}(v) - (\frac{1}{\eta_{v,k}}+1)V_{v_{k+1}}^{h^*}(v) + \\
& \theta_k\langle y_{k-1}-y_k, A(x_k-x)\rangle - \langle y_k - y_{k+1}, A(x_{k+1}-x)\rangle + \\
& -\theta_k\langle y_k - y, A(x_{k-1}-x_k)\rangle + \langle y_{k+1}-y, A(x_k-x_{k+1})\rangle + \\
& \theta_k\langle u_{k-1}-u_k, x_k-x\rangle - \langle u_k-u_{k+1}, x_{k+1}-x\rangle + \\
& \theta_k\langle v_{k-1}-v_k, y_k-y\rangle - \langle v_k-v_{k+1}, y_{k+1}-y\rangle + \\
& \theta_k\frac{\|A\|\alpha_{x,k}}{2}\|x_k-x_{k-1}\|^2 - \frac{\|A\|\alpha_{x,k+1}}{2}\|x_{k+1}-x_k\|^2 + \\
& \theta_k\frac{\|A\|\alpha_{y,k}}{2}\|y_k-y_{k-1}\|^2 - \frac{\|A\|\alpha_{y,k+1}}{2}\|y_{k+1}-y_k\|^2 + \\
& \theta_k\frac{\alpha_{u,k}}{2}\|u_k-u_{k-1}\|^2 - \frac{\alpha_{u,k+1}}{2}\|u_{k+1}-u_k\|^2 + \\
& \theta_k\frac{\alpha_{v,k}}{2}\|v_k-v_{k-1}\|^2 - \frac{\alpha_{v,k+1}}{2}\|v_{k+1}-v_k\|^2
\end{aligned} \tag{81}$$

Multiplying both sides with λ_k, and using $\theta_k\lambda_k = \lambda_{k-1}$ and

$$\frac{\lambda_{k+1}}{\lambda_k} \leq \min\left(\frac{\eta_{x,k+1}(1+\eta_{x,k}\mu_x)}{\eta_{x,k}},\ \frac{\eta_{y,k+1}(1+\eta_{y,k}\mu_y)}{\eta_{x,k}},\ \frac{\eta_{u,k+1}(1+\eta_{u,k})}{\eta_{x,k}},\ \frac{\eta_{v,k+1}(1+\eta_{v,k})}{\eta_{x,k}}\right) \tag{82}$$

we get

$$\begin{aligned}
\lambda_k\text{gap}_{z,w}(z_{k+1}, w_{k+1}) \leq\ & \frac{\lambda_k}{2\eta_{x,k+1}}\|x-x_k\|^2 - \frac{\lambda_{k+1}}{2\eta_{x,k+1}}\|x-x_{k+1}\|^2 + \frac{\lambda_k}{2\eta_{y,k}}\|y-y_k\|^2 - \frac{\lambda_{k+1}}{2\eta_{y,k+1}}\|y-y_{k+1}\|^2 + \\
& \frac{\lambda_k}{\eta_{u,k}}V_{u_k}^{f^*}(u) - \frac{\lambda_{k+1}}{\eta_{u,k+1}}V_{u_{k+1}}^{f^*}(u) + \frac{\lambda_k}{\eta_{v,k}}V_{v_k}^{h^*}(v) - \frac{\lambda_{k+1}}{\eta_{v,k+1}}V_{v_{k+1}}^{h^*}(v) + \\
& \lambda_{k-1}\langle y_{k-1}-y_k, A(x_k-x)\rangle - \lambda_k\langle y_k-y_{k+1}, A(x_{k+1}-x)\rangle + \\
& -\lambda_{k-1}\langle y_k-y, A(x_{k-1}-x_k)\rangle + \lambda_k\langle y_{k+1}-y, A(x_k-x_{k+1})\rangle + \\
& \lambda_{k-1}\langle u_{k-1}-u_k, x_k-x\rangle - \lambda_k\langle u_k-u_{k+1}, x_{k+1}-x\rangle + \\
& \lambda_{k-1}\langle v_{k-1}-v_k, y_k-y\rangle - \lambda_k\langle v_k-v_{k+1}, y_{k+1}-y\rangle + \\
& \lambda_{k-1}\frac{\|A\|\alpha_{x,k}}{2}\|x_k-x_{k-1}\|^2 - \lambda_k\frac{\|A\|\alpha_{x,k+1}}{2}\|x_{k+1}-x_k\|^2 + \\
& \lambda_{k-1}\frac{\|A\|\alpha_{y,k}}{2}\|y_k-y_{k-1}\|^2 - \lambda_k\frac{\|A\|\alpha_{y,k+1}}{2}\|y_{k+1}-y_k\|^2 + \\
& \lambda_{k-1}\frac{\alpha_{u,k}}{2}\|u_k-u_{k-1}\|^2 - \lambda_k\frac{\alpha_{u,k+1}}{2}\|u_{k+1}-u_k\|^2 + \\
& \lambda_{k-1}\frac{\alpha_{v,k}}{2}\|v_k-v_{k-1}\|^2 - \lambda_k\frac{\alpha_{v,k+1}}{2}\|v_{k+1}-v_k\|^2
\end{aligned} \tag{83}$$

Summing the iterations of the above inequality for $k = 0, \ldots, K-1$ and without loss of generality setting $\lambda_{-1} = 0$, or $x_{-1} = x_0$, $y_{-1} = y_0$, $u_{-1} = u_0$, and $v_{-1} = v_0$ we get

$$\begin{aligned}
\sum_{k=0}^{K-1}\lambda_k\text{gap}_{z,w}(z_{k+1}, w_{k+1}) \leq\ & \frac{\lambda_0}{2\eta_{x,0}}\|x-x_0\|^2 - \frac{\lambda_K}{2\eta_{x,K}}\|x-x_K\|^2 + \frac{\lambda_0}{2\eta_{y,0}}\|y-y_0\|^2 - \frac{\lambda_K}{2\eta_{y,K}}\|y-y_K\|^2 + \\
& \frac{\lambda_0}{\eta_{u,0}}V_{u_0}^{f^*}(u) - \frac{\lambda_K}{\eta_{u,K}}V_{u_K}^{f^*}(u) + \frac{\lambda_0}{\eta_{v,0}}V_{v_0}^{h^*}(v) - \frac{\lambda_K}{\eta_{v,K}}V_{v_K}^{h^*}(v) + \\
& -\lambda_{K-1}\langle y_{K-1}-y_K, A(x_K-x)\rangle + \lambda_{K-1}\langle y_K-y, A(x_{K-1}-x_K)\rangle + \\
& -\lambda_{K-1}\langle u_{K-1}-u_K, x_K-x\rangle - \lambda_{K-1}\langle v_{K-1}-v_K, y_K-y\rangle + \\
& -\lambda_{K-1}\frac{\|A\|\alpha_{x,K}}{2}\|x_K-x_{K-1}\|^2 - \lambda_{K-1}\frac{\|A\|\alpha_{y,K}}{2}\|y_K-y_{K-1}\|^2 + \\
& -\lambda_{K-1}\frac{\alpha_{u,K}}{2}\|u_K-u_{K-1}\|^2 - \lambda_{K-1}\frac{\alpha_{v,K}}{2}\|v_K-v_{K-1}\|^2
\end{aligned} \tag{84}$$

Using Cauchy-Schwarz inequality we can show that

$$-\lambda_{K-1}\langle y_{K-1}-y_K, A(x_K-x)\rangle \le \frac{\lambda_{K-1}\|A\|\alpha_{y,K}}{2}\|y_{K-1}-y_K\|^2+\frac{\lambda_{K-1}\|A\|}{2\alpha_{y,K}}\|x_K-x\|^2 \tag{85}$$

$$\lambda_{K-1}\langle y_K-y, A(x_{K-1}-x_K)\rangle \le \frac{\lambda_{K-1}\|A\|\alpha_{x,K}}{2}\|x_{K-1}-x_K\|^2+\frac{\lambda_{K-1}\|A\|}{2\alpha_{x,K}}\|y_K-y\|^2 \tag{86}$$

$$-\lambda_{K-1}\langle u_{K-1}-u_K, x_K-x\rangle \le \frac{\lambda_{K-1}\alpha_{u,K}}{2}\|u_{K-1}-u_K\|^2+\frac{\lambda_{K-1}\|A\|}{2\alpha_{u,K}}\|x_K-x\|^2 \tag{87}$$

$$-\lambda_{K-1}\langle v_{K-1}-v_K, y_K-y\rangle \le \frac{\lambda_{K-1}\alpha_{v,K}}{2}\|v_{K-1}-v_K\|^2+\frac{\lambda_{K-1}\|A\|}{2\alpha_{v,K}}\|y_K-y\|^2 \tag{88}$$

Summing equations (84), (85), (86), (87), and (88) and then using $\theta_K\lambda_K=\lambda_{K-1}$, $\theta_K(\frac{\|A\|}{2\alpha_{y,K}}+\frac{1}{2\alpha_{u,K}})+\frac{\|A\|\alpha_{x,K+1}}{2}\le\frac{1}{2\eta_{x,K}}$, and $\theta_K(\frac{\|A\|}{2\alpha_{x,K}}+\frac{1}{2\alpha_{v,K}})+\frac{\|A\|\alpha_{y,K+1}}{2}\le\frac{1}{2\eta_{y,K}}$, we get

$$\begin{aligned}\sum_{k=0}^{K-1}\lambda_k \mathrm{gap}_{z,w}(z_{k+1},w_{k+1}) \le\ & \frac{\lambda_0}{2\eta_{x,0}}\|x-x_0\|^2-\frac{\lambda_K\|A\|\alpha_{x,K+1}}{2}\|x-x_K\|^2+\\ & \frac{\lambda_0}{2\eta_{y,0}}\|y-y_0\|^2-\frac{\lambda_K\|A\|\alpha_{y,K+1}}{2}\|y-y_K\|^2+\\ & \frac{\lambda_0}{\eta_{u,0}}V_{u_0}^{f^*}(u)-\frac{\lambda_K}{\eta_{u,K}}V_{u_K}^{f^*}(u)+\frac{\lambda_0}{\eta_{v,0}}V_{v_0}^{h^*}(v)-\frac{\lambda_K}{\eta_{v,K}}V_{v_K}^{h^*}(v)\end{aligned} \tag{89}$$

□

D GUARANTEE FOR Bi-SC-SC PROBLEM

In this section we provide a guarantee for the output of Algorithm 1 in the Bi-SC-SC setting. We do this by specializing Theorem 4 to this case.

Corollary 2 (Formal version of Theorem 2)**.** *Let $\overline{x}_0=\overline{x}_{-1}=x_0$ and $\overline{y}_0=\overline{y}_{-1}=y_0$. Additionally assume that $\eta_{x,k}=\eta_x$, $\eta_{y,k}=\eta_y$, $\eta_{u,k}=\eta_u$, $\eta_{v,k}=\eta_v$, and $\theta_k=\theta$ for all $k=0,1,\ldots$. If we set*

$$\kappa=\sqrt{\frac{L_x}{\mu_x}-1}+\frac{2\|A\|}{\sqrt{\mu_x\mu_y}}+\sqrt{\frac{L_y}{\mu_y}-1}\ ,\ \text{and} \tag{90}$$

$$\eta_x=\frac{1}{\mu_x}\Big(\sqrt{\frac{L_x}{\mu_x}-1}+\frac{2\|A\|}{\sqrt{\mu_x\mu_y}}\Big)^{-1},\ \eta_y=\frac{1}{\mu_y}\Big(\sqrt{\frac{L_y}{\mu_y}-1}+\frac{2\|A\|}{\sqrt{\mu_x\mu_y}}\Big)^{-1},\ \eta_u=\Big(\sqrt{\frac{L_x}{\mu_x}-1}\Big)^{-1},\ \eta_v=\Big(\sqrt{\frac{L_y}{\mu_y}-1}\Big)^{-1} \tag{91}$$

then for any $K>0$, we can show that

$$\begin{aligned}&\frac{\|A\|}{\sqrt{\mu_x\mu_y}}\Big(\frac{\mu_x}{2}\|x^*-x_K\|^2+\frac{\mu_y}{2}\|y^*-y_K\|^2\Big)+\\ &\qquad\le \exp\Big(-\frac{(K-1)}{(\kappa+1)}\Big)\Big(\Big(\frac{1}{2\eta_x}+\frac{L_x-\mu_x}{2\eta_u}\Big)\|x^*-x_0\|^2+\Big(\frac{1}{2\eta_y}+\frac{L_y-\mu_y}{2\eta_v}\Big)\|y^*-y_0\|^2\Big)\end{aligned} \tag{92}$$

Proof. We will first verify the parameter choices satisfies the required conditions of Theorem 4 for some choice of λ_k, $\alpha_{x,k}$, $\alpha_{y,k}$, $\alpha_{u,k}$, $\alpha_{v,k}$ for $k=-1,0,1,\ldots$.

Let $\lambda_k=\gamma^k$ and $\theta_k=1/\gamma$ where

$$\gamma=1+\kappa^{-1}\ ,\quad \kappa=\sqrt{\frac{L_x}{\mu_x}-1}+\frac{2\|A\|}{\sqrt{\mu_x\mu_y}}+\sqrt{\frac{L_y}{\mu_y}-1} \tag{93}$$

Clearly $\lambda_{k-1} = \theta\lambda_k$. Next we will verify (67) which simplifies to the

$$\gamma \leq 1 + \min(\eta_x\mu_x, \eta_y\mu_y, \eta_u, \eta_v) \tag{94}$$

under our choice of λ_k and k invariant stepsize choices. It is easy to see that

$$\gamma = 1 + \Big(\sqrt{\frac{L_x-\mu_x}{\mu_x}} + \frac{2\|A\|}{\sqrt{\mu_x\mu_y}} + \sqrt{\frac{L_y-\mu_y}{\mu_y}}\Big)^{-1} \leq 1 + \Big(\sqrt{\frac{L_x-\mu_x}{\mu_x}} + \frac{2\|A\|}{\sqrt{\mu_x\mu_y}}\Big)^{-1} = 1 + \mu_x\eta_x \tag{95}$$

Similarly we can also show that $\gamma \leq 1 + \min(\eta_y\mu_y, \eta_u, \eta_v)$.

Let $\alpha_{x,k}$, $\alpha_{y,k}$, $\alpha_{u,k}$, $\alpha_{v,k}$ be invariant to k and $\alpha_{x,k} = \sqrt{\frac{\mu_x}{\mu_y}}$, $\alpha_{y,k} = \sqrt{\frac{\mu_y}{\mu_x}}$, $\alpha_{u,k} = \frac{1}{\sqrt{(L_x-\mu_x)\mu_x}}$, $\alpha_{v,k} = \frac{1}{\sqrt{(L_y-\mu_x)\mu_y}}$ for all $k = 0, 1, \ldots$. Next we verify conditions (65) and (66). We can show that

$$\alpha_{u,k} = \frac{1}{\sqrt{(L_x-\mu_x)\mu_x}} \leq \frac{1}{\sqrt{(L_x-\mu_x)\mu_x}} = \frac{1}{\eta_{u,k}(L_x-\mu_x)} \tag{96}$$

and

$$\theta_k\Big(\frac{\|A\|}{\alpha_{y,k}} + \frac{1}{\alpha_{u,k}}\Big) + \|A\|\alpha_{x,k+1} = \frac{\mu_x}{\gamma}\Big(\frac{\|A\|}{\sqrt{\mu_x\mu_y}} + \sqrt{\frac{L_x}{\mu_x}-1}\Big) + \mu_x\frac{\|A\|}{\sqrt{\mu_x\mu_y}} < \mu_x\Big(\sqrt{\frac{L_x}{\mu_x}-1} + \frac{2\|A\|}{\sqrt{\mu_x\mu_y}}\Big) = \frac{1}{\eta_{x,k}} \tag{97}$$

Similar we can also show that,

$$\theta_k\Big(\frac{\|A\|}{\alpha_{x,k}} + \frac{1}{\alpha_{v,k}}\Big) + \|A\|\alpha_{y,k+1} \leq \frac{1}{\eta_{y,k}}\,, \quad \alpha_{v,k} \leq \frac{1}{\eta_{v,k}(L_y-\mu_y)} \tag{98}$$

Then according to Theorem 4, for any $K \geq 0$, $x \in \mathcal{X}$, $y \in \mathcal{Y}$, u, v

$$\begin{aligned}
&\sum_{k=0}^{K-1} \gamma^k[\Phi(x_{k+1}, y; u, v_{k+1}) - \Phi(x, y_{k+1}; u_{k+1}, v)] \\
&\quad\leq \frac{1}{2\eta_x}\|x-x_0\|^2 - \sqrt{\frac{\mu_x}{\mu_y}}\frac{\gamma^K\|A\|}{2}\|x-x_K\|^2 + \frac{1}{2\eta_y}\|y-y_0\|^2 - \sqrt{\frac{\mu_y}{\mu_x}}\frac{\gamma^K\|A\|}{2}\|y-y_K\|^2 + \\
&\quad\frac{1}{\eta_u}V_{u_0}^{\underline{f}^*}(u) - \frac{\gamma^K}{\eta_u}V_{u_K}^{\underline{f}^*}(u) + \frac{1}{\eta_v}V_{v_0}^{\underline{h}^*}(v) - \frac{\gamma^K}{\eta_v}V_{v_K}^{\underline{h}^*}(v)
\end{aligned} \tag{99}$$

Setting $x = x^*$, $y = y^*$, $u = u^* = \nabla\underline{f}(x^*) = \arg\min_u \langle x, u\rangle - \underline{f}^*(u)$, $v = v^* = \nabla\underline{h}(y^*) = \arg\min_v \langle y, v\rangle - \underline{h}^*(v)$ in (99) we get

$$\begin{aligned}
&\sum_{k=0}^{K-1} \gamma^k[\Phi(x_{k+1}, y^*; u^*, v_{k+1}) - \Phi(x^*, y_{k+1}; u_{k+1}, v^*)] \\
&\quad\leq \frac{1}{2\eta_x}\|x^*-x_0\|^2 - \sqrt{\frac{\mu_x}{\mu_y}}\frac{\gamma^K\|A\|}{2}\|x^*-x_K\|^2 + \frac{1}{2\eta_y}\|y^*-y_0\|^2 - \sqrt{\frac{\mu_y}{\mu_x}}\frac{\gamma^K\|A\|}{2}\|y^*-y_K\|^2 + \\
&\quad\frac{1}{\eta_u}V_{u_0}^{\underline{f}^*}(u^*) - \frac{\gamma^K}{\eta_u}V_{u_K}^{\underline{f}^*}(u^*) + \frac{1}{\eta_v}V_{v_0}^{\underline{h}^*}(v^*) - \frac{\gamma^K}{\eta_v}V_{v_K}^{\underline{h}^*}(v^*)
\end{aligned} \tag{100}$$

Notice that the LHS above is positive, since by Lemma 7, $\Phi(x_{k+1}, y^*; u, v_{k+1}) - \Phi(x^*, y_{k+1}; u_{k+1}, v) \geq 0$ for all $k = 0, 1, \ldots$. Then using this fact and Lemma 6 four times, we get that

$$\begin{aligned}
&\frac{\|A\|}{\sqrt{\mu_x\mu_y}}\Big(\frac{\mu_x}{2}\|x^*-x_K\|^2 + \frac{\mu_y}{2}\|y^*-y_K\|^2\Big) + \frac{1}{2\eta_u}\frac{\|\nabla\underline{f}(x^*) - \nabla\underline{f}(\underline{x}_K)\|^2}{(L_x-\mu_x)} + \frac{1}{2\eta_v}\frac{\|\nabla\underline{h}(y^*) - \nabla\underline{h}(\underline{y}_K)\|^2}{(L_y-\mu_y)} \\
&\qquad\leq \gamma^{-K}\Big(\Big(\frac{1}{2\eta_x} + \frac{L_x-\mu_x}{2\eta_u}\Big)\|x^*-x_0\|^2 + \Big(\frac{1}{2\eta_y} + \frac{L_y-\mu_y}{2\eta_v}\Big)\|y^*-y_0\|^2\Big)
\end{aligned} \tag{101}$$

Using $1 - x \leq \exp(-x)$ we get

$$\gamma^{-K} = \Big(\frac{1}{1+\kappa^{-1}}\Big)^K \leq \Big(1 - \frac{1}{\kappa+1}\Big)^K \leq \exp\Big(-\frac{K}{\kappa+1}\Big) \tag{102}$$

Combining above two inequality gives us the desired result. □

E GUARANTEE FOR Bi-C-SC PROBLEM

In this section we provide a guarantee for the output of Algorithm 1 in the Bi-C-SC setting. We do this by specializing Theorem 4 to this case.

Corollary 3 (Formal version of Theorem 3)**.** *Let* $\overline{x}_0 = \overline{x}_{-1} = x_0$ *and* $\overline{y}_0 = \overline{y}_{-1} = y_0$ *and*

$$\frac{1}{\eta_{x,k}} = \frac{1}{(k+1)\eta_x}\,,\ \frac{1}{\eta_x} = 2L_x + \frac{16\|A\|^2}{\mu_y}\,,\ \frac{1}{\eta_{y,k}} = \frac{1}{(k+1)\eta_y} + \frac{k\mu_y}{2}\,,\ \frac{1}{\eta_y} = 2(L_y-\mu_y)\,,\ \eta_{u,k} = \frac{2}{k}\,,\ \eta_{v,k} = \frac{2}{k}\,. \tag{103}$$

Let $D_{\mathcal{X}} = \max_{x\in\mathcal{X}\cap\mathrm{dom}(f)} \|x - x_0\|$ *and* $D_{\mathcal{Y}} = \max_{y\in\mathcal{Y}\cap\mathrm{dom}(h)} \|y - y_0\|$*. Then for any* $K > 0$*,*
(a) if $D_{\mathcal{X}} < \infty$ *and* $D_{\mathcal{Y}} < \infty$*,*

$$\max_{y\in\mathcal{Y}} \phi(\overline{x}_K, y) - \min_{x\in\mathcal{X}} \phi(x, \overline{y}_K) \le \frac{2L_x}{K(K+1)}D_{\mathcal{X}}^2 + \frac{16\|A\|^2}{\mu_y K(K+1)}D_{\mathcal{X}}^2 + \frac{2(L_y-\mu_y)}{K(K+1)}D_{\mathcal{Y}}^2 \tag{104}$$

where $(\overline{x}_K, \overline{y}_K) := \frac{2}{K(K+1)}\sum_{k=1}^K k(x_k, y_k) = (\underline{x}_K, \underline{y}_K)$*.*
(b) even if the domain is unbounded we can show that

$$\frac{\mu_y}{4}\|y^* - y_K\|^2 \le \frac{4L_x}{K(K+1)}\|x^* - x_0\|^2 + \frac{16\|A\|^2}{\mu_y K(K+1)}\|x^* - x_0\|^2 + \frac{4(L_y-\mu_y)}{K(K+1)}\|y^* - y_0\|^2 \tag{105}$$

where $(\underline{x}_K, \underline{y}_K) = \frac{2}{K(K+1)}\sum_{k=1}^K k\cdot(x_k, y_k) = (\overline{y}_K, \overline{y}_K)$*.*
(c) if $\phi_p(x) = \max_{y\in\mathcal{Y}} \phi(x,y)$*, and we do a warm restart on variable* y *using* $K_0^p = \Omega_\varepsilon(1)$ *initial additional iterations of the same algorithm, then*

$$\phi_p(\overline{x}_K) - \phi_p(x^*) \le \Big(\frac{4L_x}{K(K+1)} + \frac{32\|A\|^2}{\mu_y K(K+1)} + \frac{(L_y-\mu_y)}{\mu_y}\frac{8\|A\|^2}{\mu_y K(K+1)}\Big)\|x^* - x_0\|^2\,. \tag{106}$$

(d) if $D_{\mathcal{X}} < \infty$ *and* $\phi_d(x) = \min_{x\in\mathcal{X}} \phi(x,y)$*, and we do a warm restart on variable* y *with* $K_0^d = \Omega_\varepsilon(1)$ *initial additional iterations of the same algorithm, then*

$$\phi_d(y^*) - \phi_d(\overline{y}_K) \le \frac{4L_x}{K(K+1)}D_{\mathcal{X}}^2 + \frac{32\|A\|^2}{\mu_y K(K+1)}D_{\mathcal{X}}^2\,. \tag{107}$$

Proof. We will first verify the parameter choices satisfies the required conditions of Theorem 4 for some choice of λ_k, $\alpha_{x,k}$, $\alpha_{y,k}$, $\alpha_{u,k}$, $\alpha_{v,k}$ for $k = -1, 0, 1, \ldots$.

Let $\lambda_k = (k+1)$ and $\theta_k = k/(k+1)$. Clearly $\lambda_{k-1} = \theta\lambda_k$. Next we will verify (67) which simplifies to the

$$\frac{k+2}{\eta_{x,k+1}} \le \frac{k+1}{\eta_{x,k}}\,,\ \frac{k+2}{\eta_{y,k+1}} \le \frac{(k+1)}{\eta_{y,k}} + (k+1)\mu_y\,,\ \frac{k+2}{\eta_{u,k+1}} \le \frac{(k+1)}{\eta_{u,k}} + (k+1)\,,\ \text{and}\ \frac{k+2}{\eta_{v,k+1}} \le \frac{(k+1)}{\eta_{v,k}} + (k+1) \tag{108}$$

under our choice of λ_k and $\mu_x = 0$. It is easy to verify that

$$\begin{aligned}
\frac{k+1}{\eta_{x,k}} = \frac{1}{\eta_x} \ge \frac{1}{\eta_x} &= \frac{k+2}{\eta_{x,k+1}} \\
\frac{(k+1)}{\eta_{y,k}} + (k+1)\mu_y = \frac{1}{\eta_y} + \frac{k(k+1)\mu_y}{2} + (k+1)\mu_y \ge \frac{1}{\eta_y} + \frac{(k+1)(k+2)\mu_y}{2} &= \frac{k+2}{\eta_{y,k+1}} \\
\frac{(k+1)}{\eta_{u,k}} + (k+1) = \frac{k(k+1)}{2} + (k+1) \ge \frac{(k+1)(k+2}{2} &= \frac{k+2}{\eta_{u,k+1}} \\
\frac{(k+1)}{\eta_{v,k}} + (k+1) = \frac{k(k+1)}{2} + (k+1) \ge \frac{(k+1)(k+2}{2} &= \frac{k+2}{\eta_{v,k+1}}
\end{aligned}$$

Let $\alpha_{x,k} = \frac{4\|A\|}{(k+1)\mu_y}$, $\alpha_{y,k} = \frac{k\mu_y}{4\|A\|}$, $\alpha_{u,k} = \frac{k}{2L_x}$, $\alpha_{v,k} = \frac{k}{2(L_y-\mu_y)}$ for all $k = 0, 1, \ldots$.

Next we verify conditions (65) and (66). We can show that

$$\alpha_{u,k} \le \frac{k}{2L_x} \le \frac{k}{2L_x} = \frac{1}{\eta_{u,k}L_x} \tag{109}$$

and

$$\theta_k(\frac{\|A\|}{\alpha_{y,k}} + \frac{1}{\alpha_{u,k}}) + \|A\|\alpha_{x,k+1} = \frac{k}{k+1}(\frac{4\|A\|^2}{k\mu_y} + \frac{2L_x}{k}) + \frac{4\|A\|^2}{(k+2)\mu_y} \le \frac{2L_x}{k+1} + \frac{16\|A\|^2}{\mu_y(k+1)} = \frac{1}{\eta_{x,k}} \tag{110}$$

Similar we can also show that,

$$\alpha_{v,k} \le \frac{k}{2(L_y-\mu_y)} \le \frac{k}{2(L_y-\mu_y)} = \frac{1}{\eta_{v,k}(L_y-\mu_y)} \tag{111}$$

and

$$\theta_k(\frac{\|A\|}{\alpha_{x,k}} + \frac{1}{\alpha_{v,k}}) + \|A\|\alpha_{y,k+1} = \frac{k}{k+1}(\frac{(k+1)\mu_y}{4} + \frac{2(L_x-\mu_y)}{k}) + \frac{k\mu_y}{4\|A\|} \le \frac{2(L_y-\mu_y)}{(k+1)} + \frac{k\mu_y}{2} = \frac{1}{\eta_{y,k}} \tag{112}$$

Then according to Theorem 4, for any $K \ge 0$, $x \in \mathcal{X}$, $y \in \mathcal{Y}$, u, v

$$\begin{aligned}&\sum_{k=0}^{K-1}(k+1)[\Phi(x_{k+1},y;u,v_{k+1}) - \Phi(x,y_{k+1};u_{k+1},v)]\\ &\qquad\le (L_x + \frac{8\|A\|^2}{\mu_y})\|x-x_0\|^2 - \frac{(K+1)16\|A\|^2}{2\mu_y(K+2)}\|x-x_K\|^2 + (L_y-\mu_y)\|y-y_0\|^2 - (K+1)^2\frac{\mu_y}{8}\|y-y_K\|^2\end{aligned} \tag{113}$$

(a) We define that $(\overline{x}_K,\overline{y}_K;\overline{u}_K,\overline{v}_K) = (\sum_{k=1}^K(k+1))^{-1}\sum_{k=1}^K(k+1)(x_k,y_k;u_k,v_k)$. Then $(\overline{x}_K,\overline{y}_K) = (\underline{x}_K,\underline{y}_K)$. Then $\overline{x}_K = \underline{x}_K$ can be shown as follows

$$\begin{aligned}\underline{x}_K = \frac{\underline{x}_{K-1}+\eta_x' x_K}{(1+\eta_x')} &= \frac{K-1}{K+1}\underline{x}_{K-1} + \frac{2}{K+1}x_K\\ &= \frac{(K-2)(K-1)}{K(K+1)}\underline{x}_{K-2} + \frac{2(K-1)}{K(K+1)}x_{K-1} + \frac{2K}{K(K+1)}x_K\\ &= \frac{(K-3)(K-2)}{K(K+1)}\underline{x}_{K-3} + \frac{2(K-2)}{K(K+1)}\underline{x}_{K-2} + \frac{2(K-1)}{K(K+1)}x_{K-1} + \frac{2(K)}{K(K+1)}x_K\\ &\vdots \qquad (114)\\ &= \frac{2}{K(K+1)}\sum_{k=1}^K kx_k = \overline{x}_K \qquad (115)\end{aligned}$$

Similarly, we can prove that $\overline{y}_K = \underline{y}_K$. Then we can lower-bound the LHS of the (113) using Jensen's inequality, convexity of $\Phi(\cdot,y;u,\cdot)$, and concavity of $\Phi(x,\cdot;\cdot,v)$ as follows.

$$(\sum_{k=0}^{K-1}(k+1))[\Phi(\overline{x}_K,y;u,\overline{v}_K) - \Phi(x,\overline{y}_K;\overline{u}_K,v)] \le \sum_{k=0}^{K-1}(k+1)[\Phi(x_{k+1},y;u,v_{k+1}) - \Phi(x,y_{k+1};u_{k+1},v)] \tag{116}$$

Notice that by Lemma 4(a), $\nabla f(x) = \arg\min_u \langle x,u\rangle - f^*(u)$ and $\nabla\underline{h}(x) = \arg\min_v \langle y,v\rangle - \underline{h}^*(v)$. Thus we have

$$\begin{aligned}\phi(\overline{x}_K,y) - \phi(x,\overline{y}_K) &= \min_v\max_u \Phi(\overline{x}_K,y;\nabla f(\overline{x}_K),v) - \Phi(x,\overline{y}_K;u,\nabla\underline{h}(\overline{y}_K))\\ &\le \Phi(\overline{x}_K,y;\nabla f(\overline{x}_K),\overline{v}_K) - \Phi(x,\overline{y}_K;\overline{u}_K,\nabla\underline{h}(\overline{y}_K))\end{aligned} \tag{117}$$

Therefore summing equations (113) and (116), then setting $u = \nabla f(\overline{x}_K)$, $v = \nabla\underline{h}(\overline{y}_K) = \nabla h(\overline{y}_K) - \mu_y\overline{y}_K$ and using (117) we get

$$\phi(\overline{x}_K,y) - \phi(x,\overline{y}_K) \le \frac{2L_x}{K(K+1)}\|x-x_0\|^2 + \frac{16\|A\|^2}{\mu_y K(K+1)})\|x-x_0\|^2 + \frac{2(L_y-\mu_y)}{K(K+1)}\|y-y_0\|^2 \tag{118}$$

Finally maximizing both sides over $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ we get

$$\max_{y\in\mathcal{Y}} \phi(\overline{x}_K, y) - \min_{x\in\mathcal{X}} \phi(x, \overline{y}_K) \leq \frac{2L_x}{K(K+1)} D_{\mathcal{X}}^2 + \frac{16\|A\|^2}{\mu_y K(K+1)} D_{\mathcal{X}}^2 + \frac{2(L_y-\mu_y)}{K(K+1)} D_{\mathcal{Y}}^2 \tag{119}$$

(b) Setting $x = x^*$, $y = y^*$, $u = u^* = \nabla f(x^*) = \arg\min_u \langle x, u\rangle - f^*(u)$, $v = v^* = \nabla \underline{h}(y^*) = \arg\min_v \langle y, v\rangle - \underline{h}^*(v)$ in (113) we get

$$\begin{aligned}
&\sum_{k=0}^{K-1} \gamma^k [\Phi(x_{k+1}, y^*; u^*, v_{k+1}) - \Phi(x^*, y_{k+1}; u_{k+1}, v^*)] \\
&\quad \leq (L_x + \frac{8\|A\|^2}{\mu_y})\|x^* - x_0\|^2 + (L_y - \mu_y)\|y^* - y_0\|^2 - (K+1)^2 \frac{\mu_y}{2}\|y^* - y_K\|^2
\end{aligned} \tag{120}$$

Notice that the LHS above is positive, since by Lemma 7, $\Phi(x_{k+1}, y^*; u, v_{k+1}) - \Phi(x^*, y_{k+1}; u_{k+1}, v) \geq 0$ for all $k = 0, 1, \ldots$. Then using this fact we get that

$$\frac{\mu_y}{4}\|y^* - y_K\|^2 \leq (\frac{2L_x}{(K+1)^2} + \frac{16\|A\|^2}{\mu_y (K+1)^2})\|x^* - x_0\|^2 + \frac{2(L_y-\mu_y)}{(K+1)^2}\|y^* - y_0\|^2 \tag{121}$$

(c) Let $\widehat{y}(x) = \arg\max_y \phi(x, y)$, then $\widehat{y}(x)$ is a $\|A\|/\mu_y$-Lipschitz continuous in x (Nesterov 2005). Then we can show that

$$\begin{aligned}
\|\widehat{y}(x) - y_0\|^2 &\leq 2\|\widehat{y}(x) - y^*\|^2 + 2\|y^* - y_0\|^2 \\
&\leq 2\|\widehat{y}(x) - \widehat{y}(x^*)\|^2 + 2\|y^* - y_0\|^2 \\
&\leq 2\frac{\|A\|^2}{\mu_y^2}\|x - x^*\|^2 + 2\|y^* - y_0\|^2
\end{aligned} \tag{122}$$

Then using the above inequality and (118) we get

$$\begin{aligned}
\phi_p(\overline{x}_K) - \phi_p(x^*) &= \max_{y\in\mathcal{Y}} \phi(\overline{x}_K, y) - \max_{y\in\mathcal{Y}} \phi(x^*, y) \\
&\leq \phi(\overline{x}_K, \widehat{y}(x)) - \phi(x^*, \overline{y}_K) \\
&\leq \frac{2L_x}{K(K+1)}\|x^* - x_0\|^2 + \frac{16\|A\|^2}{\mu_y K(K+1)}\|x^* - x_0\|^2 + \frac{2(L_y-\mu_y)}{K(K+1)}\|\widehat{y}(x) - y_0\|^2 \\
&\leq (\frac{2L_x}{K(K+1)} + \frac{16\|A\|^2}{\mu_y K(K+1)} + \frac{(L_y-\mu_y)}{\mu_y}\frac{4\|A\|^2}{\mu_y K(K+1)})\|x^* - x_0\|^2 + \frac{4(L_y-\mu_y)}{K(K+1)}\|y^* - y_0\|^2
\end{aligned} \tag{123}$$

From the above inequality it is clear that

$$\phi_p(\overline{x}_K) - \phi_p(x^*) \leq (\frac{4L_x}{K(K+1)} + \frac{32\|A\|^2}{\mu_y K(K+1)} + \frac{(L_y-\mu_y)}{\mu_y}\frac{8\|A\|^2}{\mu_y K(K+1)})\|x^* - x_0\|^2 \tag{124}$$

if

$$\|y^* - y_0\|^2 \leq (\frac{L_x}{2(L_y-\mu_y)} + \frac{4\|A\|^2}{\mu_y(L_y-\mu_y)} + \frac{(L_y-\mu_y)}{\mu_y}\frac{\|A\|^2}{\mu_y(L_y-\mu_y)})\|x^* - x_0\|^2 . \tag{125}$$

Because of (121), we can find a y_0 satisfying the above inquality by running our algorithm from from (x_0, y_0) for

$$\begin{aligned}
K_0^p \geq \Omega(\sqrt{4(L_y-\mu_y)/\mu_y} \times \sqrt{(2L_x + \frac{16\|A\|^2}{\mu_y})\|x^* - x_0\|^2 + 2(L_y-\mu_y)\|y^* - y_0\|^2} \times \\
\sqrt{1 \Big/ ((\frac{L_x}{2} + \frac{4\|A\|^2}{\mu_y} + \frac{(L_y-\mu_y)}{\mu_y}\frac{\|A\|^2}{\mu_y})\|x^* - x_0\|^2)}
\end{aligned} \tag{126}$$

iterations.

$$\|y^* - y_{K_0^p}\|^2 \leq \frac{4}{\mu_y}\left(\frac{2L_x}{(K+1)^2} + \frac{16\|A\|^2}{\mu_y(K+1)^2}\right)\|x^* - x_0\|^2 + \frac{4}{\mu_y}\frac{2(L_y-\mu_y)}{(K+1)^2}\|y^* - y_0\|^2$$
$$\leq \left(\frac{L_x}{2(L_y-\mu_y)} + \frac{4\|A\|^2}{\mu_y(L_y-\mu_y)} + \frac{(L_y-\mu_y)}{\mu_y}\frac{\|A\|^2}{\mu_y(L_y-\mu_y)}\right)\|x^* - x_0\|^2 \tag{127}$$

Similarly using the above inequality and (118) we get

$$\phi_d(y^*) - \phi_d(\overline{y}_K) = \min_{x\in\mathcal{X}}\phi(x, y^*) - \min_{x\in\mathcal{X}}\phi(x, \overline{y}_K)$$
$$\leq \phi(\overline{x}_K, y^*) - \min_{x\in\mathcal{X}}\phi(x, \overline{y}_K)$$
$$\leq \frac{2L_x}{K(K+1)}D_{\mathcal{X}}^2 + \frac{16\|A\|^2}{\mu_y K(K+1)}D_{\mathcal{X}}^2 + \frac{2(L_y-\mu_y)}{K(K+1)}\|y^* - y_0\|^2 \tag{128}$$

From the above inequality it is clear that

$$\phi_p(\overline{x}_K) - \phi_p(x^*) \leq \frac{4L_x}{K(K+1)}D_{\mathcal{X}}^2 + \frac{32\|A\|^2}{\mu_y K(K+1)}D_{\mathcal{X}}^2 \tag{129}$$

if

$$\|y^* - y_0\|^2 \leq \left(\frac{L_x}{(L_y-\mu_y)} + \frac{8\|A\|^2}{\mu_y(L_y-\mu_y)}\right)D_{\mathcal{X}}^2\,. \tag{130}$$

Because of (121), we can find a y_0 satisfying the above inquality by running our algorithm from from (x_0, y_0) for

$$K_0^d \geq \Omega\left(\sqrt{4(L_y-\mu_y)/\mu_y} \times \sqrt{\left(2L_x + \frac{16\|A\|^2}{\mu_y}\right)\|x^* - x_0\|^2 + 2(L_y-\mu_y)\|y^* - y_0\|^2} \times \right.$$
$$\left. \sqrt{1\Big/\left(\left(L_x + \frac{8\|A\|^2}{\mu_y}\right)D_{\mathcal{X}}^2\right)}\right) \tag{131}$$

iterations.

$$\|y^* - y_{K_0^d}\|^2 \leq \frac{4}{\mu_y}\left(\frac{2L_x}{(K+1)^2} + \frac{16\|A\|^2}{\mu_y(K+1)^2}\right)\|x^* - x_0\|^2 + \frac{4}{\mu_y}\frac{2(L_y-\mu_y)}{(K+1)^2}\|y^* - y_0\|^2$$
$$\leq \left(\frac{L_x}{(L_y-\mu_y)} + \frac{8\|A\|^2}{\mu_y(L_y-\mu_y)}\right)D_{\mathcal{X}}^2 \tag{132}$$

□

F BALANCED MIRROR-PROX AND ADDITIONAL EXPERIMENTAL DETAILS FOR SECTION 6

For all the experiments we used the theory specified stepsize choices. Balanced Mirror Prox (which we shorten as MP Bal.) is variant of the standard Mirror-Prox algorithm (folklore). For implementing MP Bal. first we normalize the distance functions so that objective becomes 1-strongly convex in both the min variable x and the max variable y. This modifies Lipschitz constants of the gradients as $L_x \leftarrow L_x/\mu_x$, $L_{xy} \leftarrow L_{xy}/\sqrt{\mu_x\mu_y} = \|A\|/\sqrt{\mu_x\mu_y}$, $L_y \leftarrow L_y/\mu_y$. Finally, in this modified geometry (distance metrics), we run the standard MP with the stepsize $1/\max(L_x, L_{xy}, L_y)$. Since we modified the Lipschitz constants of the gradients this leads to a iteration complexity of $\mathcal{O}(\sqrt{\frac{L_x}{\mu_x}} + \frac{\|A\|}{\sqrt{\mu_x\mu_y}} + \sqrt{\frac{L_y}{\mu_y}})\log(\frac{1}{\varepsilon})$. This result was also mentioned as a known folklore in Appendix C of (Cohen et al. 2020).

For experiments using quadratic minimax problems we use $d = 5$ and we generate B, A, C as follows. Let $\Lambda = \mathrm{diag}(r^0, r^1, \ldots, r^{d-1})$. Then $A = Q^{(A,2)}\Lambda(Q^{(A,1)})^\top$, $B = \widetilde{B}^\top\widetilde{B}$, $\widetilde{B} = Q^{(B,1)}\Lambda(Q^{(B,2)})^\top$, $C = \widetilde{C}^\top\widetilde{C}$, $\widetilde{C} = Q^{(C,1)}\Lambda(Q^{(C,2)})^\top$, where $Q^{(A,1)}$, $Q^{(A,2)}$, $Q^{(B,1)}$, $Q^{(B,2)}$, $Q^{(C,1)}$, $Q^{(C,2)}$ are i.i.d. $d \times d$ orthonormal matrices which are generated uniformly at random. For Figure 1a we set $r = 2.0$, and for Figure 1b we vary r using the values $\{1.25, 1.5, 1.75, 2.0, 2.25\}$.