Accelerated Primal-Dual Methods for Convex-Strongly-Concave Saddle Point Problems

Mohammad Khalafi [1] **Digvijay Boob** [1]

Abstract

We investigate a primal-dual (PD) method for the saddle point problem (SPP) that uses a linear approximation of the primal function instead of the standard proximal step, resulting in a linearized PD (LPD) method. For convex-strongly concave SPP, we observe that the LPD method has a suboptimal dependence on the Lipschitz constant of the primal function. To fix this issue, we combine features of Accelerated Gradient Descent with the LPD method resulting in a single-loop Accelerated Linearized Primal-Dual (ALPD) method. ALPD method achieves the optimal gradient complexity when the SPP has a *semi-linear* coupling function. We also present an inexact ALPD method for SPPs with a general nonlinear coupling function that maintains the optimal gradient evaluations of the primal parts and significantly improves the gradient evaluations of the coupling term compared to the ALPD method. We verify our findings with numerical experiments.

1. Introduction

As a class of optimization problems, the min-max saddle point problem (SPP) has attracted much attention in the optimization and machine learning literature. The SPPs contain many classical problems as a special case. E.g., we can transform convex optimization problems with smooth or nonsmooth objective functions into a min-max saddle point form. One can extend this observation to nonsmooth nonconvex problems relatively easily. Given their strong modeling power, SPPs have extensive applications in (distributionally) robust optimization and adversarial learning.

In this paper, we are interested in the following SPP

$$\mathcal{L}(x, y) := \min_{x\in X} \max_{y\in Y} f(x) + \phi(x, y) - g(y), \qquad (1)$$

where we refer to f, g and ϕ as the primal, dual and coupling functions, respectively.

The broad applicability of the SPP model has resulted in various algorithmic complexity studies in the literature. The major focus was on the computationally tractable *convex-concave* case, i.e., $\mathcal{L}(\cdot, y)$ is convex in x for all $y \in Y$ and $\mathcal{L}(x, \cdot)$ is concave in y for all $x \in X$. In this setting, $\max_{y\in Y} \mathcal{L}(x, y)$ is a nonsmooth function in x. According to Nemirovski & Yudin (1983), subgradient descent for a black-box nonsmooth convex function achieves an ϵ optimality error in $\mathcal{O}(\frac{1}{\epsilon^2})$ subgradient evaluations. In a seminal work, Nesterov (2005) exploited the max-form of the problem to obtain a significantly improved gradient complexity of $\mathcal{O}(\frac{1}{\epsilon})$. This result broke the earlier established complexity lower bounds and is popularly known as *Nesterov's smoothing* technique. Nemirovski (2004) presented an Extragradient method that performs one extra gradient descent-ascent step in each iteration. This method can obtain an ϵ error on the stronger *gap function* criterion (c.f. Definition 2.1) using $\mathcal{O}(\frac{1}{\epsilon})$ gradient evaluations. Subsequently, (Chambolle & Pock, 2011; 2016; Chen et al., 2014) showed primal-dual (PD) type methods which remove the additional gradient descent-ascent step and maintain an $\mathcal{O}(\frac{1}{\epsilon})$ complexity when ϕ is a bilinear coupling. Later, (Hamedani & Aybat, 2021) extended it to the general convex-concave coupling functions.

The PD methods in (Chambolle & Pock, 2011; Hamedani & Aybat, 2021) assume that the proximal operators of f and g are easy to evaluate. For the bilinear coupling term, i.e., $\phi(x, y) = y^\top Ax$, Condat (2013); Vu (2011) introduced LPD method where they used the linear approximation of f in a PD method and proved the convergence of its iterates to saddle point. Chambolle & Pock (2016) considered the same design and showed LPD method has the convergence complexity of $\mathcal{O}(\frac{L_f+\|A\|}{\epsilon})$, where L_f is the Lipschitz constant of ∇f and $\|A\|$ is the operator norm of A. Observing that this dependence is not optimal in L_f, Chen et al. (2014) proposed an accelerated PD method whose complexity is of $\mathcal{O}(\sqrt{\frac{L_f}{\epsilon}} + \frac{\|A\|}{\epsilon})$ which significantly reduces the impact of

[1] Department of Operations Research and Engineering Management, Southern Methodist University, Dallas TX, USA. Correspondence to: Mohammad Khalafi <mohamadk@smu.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

Lipschitz constant L_f on the complexity.

Chambolle & Pock (2011; 2016) also show that when f is strongly convex with modulus $\mu_f > 0$ and the coupling term is bilinear, the LPD method exhibits a much smaller complexity of $\mathcal{O}(\frac{\|A\|}{\sqrt{\mu_f \epsilon}})$, while using the exact proximal operators for f and g. Hamedani & Aybat (2021) extend similar results for *semi-linear* couplings (linear in y only).

However, to our best knowledge, a few works study the impact of linearization of $f(x)$ when $g(y)$ is strongly convex with modulus $\mu_g > 0$. Kovalev et al. (2022), showed linear convergence under a restricted strong concavity-type condition for a bilinear coupling function. Thekumparampil et al. (2019), introduced a three-loop algorithm called *Dual Implicit Accelerated Gradient* (DIAG) where each iteration contains an implicit step in which an AGD is run. Thekumparampil et al. (2022) proposed the first single-loop optimal algorithm called *Lifted Primal-Dual method* for SPPs under strong concavity. However, their analysis heavily relies on the bilinear coupling function and it is unclear whether it can be extended for nonlinear coupling.

The SPPs with strong concavity have a direct application in the *Nesterov's smoothing* framework: a nonsmooth convex function $\max_{y\in Y} f(x) + \phi(x, y)$ can be smoothened by adding a strongly concave regularizer $-g(y)$ resulting in (1). Moreover, using appropriate Y and g, we obtain equivalent formulations of a variety of (smoothened) penalty functions used in constrained optimization. Assuming the exact proximal operator of objective f in such cases is quite impractical. Hence, we need to study methods that can handle linearization. We intend to make contributions to this setting, i.e., $\mu_g > 0$ and f is linearized. See Table 1 for a comparison of our work with the relevant literature.

1. Our first contribution is to observe the subtle but important difference due to linearization. In particular, when f is linearized, the case of $\mu_g > 0$ is qualitatively "harder" than $\mu_f > 0$. Hence, the LPD method exhibits a weaker complexity of $\mathcal{O}(\frac{L_f}{\epsilon} + \frac{\|A\|}{\sqrt{\mu_g \epsilon}})$ (c.f. Theorem 3.1 and 3.2).

2. A careful observation of the above complexity yields that the LPD algorithm is unable to mitigate the impact of the primal Lipschitz constant L_f when $\mu_g > 0$. Hence, we seek an algorithm that can accelerate convergence in the primal. Moreover, we expand the scope of the problem to include the general nonlinear couplings. To address both questions, we imbibe elements of Nesterov's Accelerated Gradient Descent (AGD) in the PD method for general nonlinear couplings, and propose a novel single-loop Accelerated Linearized PD (ALPD) method (see Algorithm 2). We show that (i) for the semi-linear coupling (linear in x-only), the ALPD method exhibits the complexity of $\mathcal{O}(\sqrt{\frac{L_f}{\epsilon}})$ which significantly improves the dependence on L_f compared to the LPD method[1]; (ii) for the general coupling, it exhibits the complexity of $\mathcal{O}(\sqrt{\frac{L_f}{\epsilon}} + \frac{L_{xx}}{\epsilon})$ where L_{xx} is the Lipschitz constant of $\nabla_x \phi(\cdot, y)$.

3. To improve the above complexity in L_{xx}, we propose an Inexact ALPD method. It is a two-loop algorithm that solves a proximal problem using AGD in the inner loop while the outer loop follows a "conceptual" ALPD method. The Inexact ALPD method obtains an ϵ-error in $\mathcal{O}(\sqrt{\frac{L_f}{\epsilon}})$ evaluations of ∇f and $\tilde{\mathcal{O}}(\frac{\sqrt{L_{xx}}}{\epsilon^{3/4}})$ evaluations of $\nabla_x \phi$. Essentially, this method maintains the optimal dependence of the complexity on L_f and improves the dependence on L_{xx}.

4. We verify our findings using numerical experiments on the penalty problems for linear and nonlinear constraints.

1.1. Related works

The SPPs are extensively studied in the literature due to their broad applicability and strong modeling power. Here, we provide a brief review of the most relevant first-order methods that consider the issue of algorithmic complexity for the SPPs.

Classical results: Nesterov (2005) reformulated a deterministic optimization problem into an SPP form and showed the first optimally converging algorithm using the smoothing framework. Subsequently, Nemirovski (2004) showed the optimal convergence of the mirror-prox method (a generalization of the extragradient method (Korpelevich, 1976)) for the variational inequality problem which contains the nonlinear SPP as a special case. Separately, Nesterov (2007) and Tseng (2008) provided two optimally converging algorithms for the SPPs. This approach was further extended by Monteiro & Svaiter (2010) in an HPE framework to relax the bounded domain assumption. Nemirovski et al. (2009) presented a mirror-descent type algorithm for the stochastic SPP. Juditsky et al. (2011) proposed a stochastic version of the mirror-prox method. Chen et al. (2017) incorporated a multi-step acceleration scheme into the stochastic mirror-prox to improve the convergence rate.

Bilinear case: While extragradient (or mirror-prox) required two ∇_x, ∇_y evaluations in each iteration, the primal-dual method of (Chambolle & Pock, 2011) required only one such evaluation per iteration and maintained the same convergence rate. Several variants of this method are proposed in the literature for bilinear couplings. E.g., the linearization of f is presented in (Chambolle & Pock, 2016), optimal accelerated-version is introduced in (Chen et al., 2014), randomized block-coordinate settings are considered in (Dang & Lan, 2014; Zhu & Storkey, 2015; Yu et al., 2015; Zhang & Lin, 2015).

Nonlinear coupling: For the nonlinear coupling term,

[1] See Remark 4.6 for similarity with (Hamedani & Aybat, 2021)

Table 1: Comparison of our work. Gradient complexity is for obtaining an ϵ error in gap function.

	Coupling	Linearizing f	Gradient Complexity	
			$\mu_f > 0$	$\mu_g > 0$
(Chambolle & Pock, 2011)	bilinear	No	$\mathcal{O}(\frac{1}{\sqrt{\epsilon}})$	NA
(Chambolle & Pock, 2016)	bilinear	Yes	$\mathcal{O}(\frac{1}{\sqrt{\epsilon}})$	NA
(Hamedani & Aybat, 2021)	semi-linear	No	$\mathcal{O}(\frac{1}{\sqrt{\epsilon}})$	NA
(Thekumparampil et al., 2022)	bilinear	Yes	NA	$\mathcal{O}(\sqrt{\frac{L_f}{\epsilon}} + \frac{\|A\|}{\sqrt{\mu_g \epsilon}})$
LPD (Algorithm 1)	bilinear	Yes	$\mathcal{O}(\frac{1}{\sqrt{\epsilon}})$	$\mathcal{O}(\frac{L_f}{\epsilon} + \frac{\|A\|}{\sqrt{\mu_g \epsilon}})$
ALPD (Algorithm 2)	semi-linear	Yes	NA	$\mathcal{O}(\sqrt{\frac{L_f + L_{yy}}{\epsilon}} + \frac{L_{xy}}{\sqrt{\mu_g \epsilon}})$
	general			$\mathcal{O}(\sqrt{\frac{L_f + L_{yy}}{\epsilon}} + \frac{L_{xy}}{\sqrt{\mu_g \epsilon}} + \frac{L_{xx}}{\epsilon})$
Inexact ALPD (Algorithm 3)	general	Yes	NA	For $\nabla f, \nabla_y \phi :\ \mathcal{O}(\sqrt{\frac{L_f + L_{yy}}{\epsilon}} + \frac{L_{xy}}{\sqrt{\mu_g \epsilon}})$ For $\nabla_x \phi :\ \mathcal{O}(\frac{\sqrt{L_{xx}\sqrt{L_f + L_{xy}^2/\mu_g}}}{\epsilon^{3/4}} \log(\frac{1}{\epsilon}))$

Hamedani & Aybat (2021) proposed a primal-dual method which can be seen as an extension of the original primal-dual method. Its extension to a randomized block-coordinate version was presented in (Hamedani et al., 2018). Another variation of significant consequence is proposed in (Boob et al., 2022b) for the stochastic smooth/nonsmooth function-constrained optimization.

Strong convexity: To our best knowledge, the existing works look at the strongly convex case ($\mu_f > 0$). For the bilinear couplings, Chambolle & Pock (2011) shows a smaller complexity of $\mathcal{O}(\frac{1}{\sqrt{\epsilon}})$. Hamedani & Aybat (2021) presents the first accelerated convergence result for semi-linear coupling (linear in y-only). Lin et al. (2020) proposed an inexact accelerated proximal point algorithm which has a nested three-loop structure and obtains an optimal complexity up to a $\log^3(\frac{1}{\epsilon})$ factor. The problem of obtaining optimal rates for general nonlinear couplings with single-loop algorithms remains open.

2. Notation and Definitions

We use $\|\cdot\|_q$ and $\|\cdot\|$ to denote ℓ_q-norm and Euclidean norm of any vector, respectively. $\langle \cdot, \cdot \rangle$ stands for the standard inner product of two vectors. For a general function h, ∇h expresses the gradient of h. $\nabla_v h$ implies the partial gradient of h with respect to variable v. We use $[m]$ to denote $\{1, \ldots, m\}$. For a compact set $\mathcal{W}$, we define its diameter $D_{\mathcal{W}} := \max_{w', w \in \mathcal{W}} \|w' - w\|/\sqrt{2}$. We use $z = (x, y)$ as the combined variable defined on the set $X \times Y \equiv Z$. We naturally extend this notation for $\bar{z} = (\bar{x}, \bar{y})$, $z_t = (x_t, y_t)$, $\bar{z}_t = (\bar{x}_t, \bar{y}_t)$ and so on.

Problems setting. In problem (1), $X \subseteq \mathbb{R}^n$ and $Y \subseteq \mathbb{R}^m$ are compact convex sets, $f : X \to \mathbb{R}$ is a convex primal function, $g : Y \to \mathbb{R}$ is a convex dual function and $\phi(x, y) : X \times Y \to \mathbb{R}$ is a convex-concave coupling function, i.e., $\phi(\cdot, y)$ is convex for all $y \in Y$ and $\phi(x, \cdot)$ is concave for all $x \in X$. The *gap function* defined below acts as a measure of convergence.

Definition 2.1. For a point $\bar{z} \in Z$, we define its gap as

$$\text{Gap}(\bar{z}) = \max_{z \in Z} Q(\bar{z}, z).$$

where $Q(\bar{z}, z) := \mathcal{L}(\bar{x}, y) - \mathcal{L}(x, \bar{y})$.

It is easy to see that $\text{Gap}(\bar{z}) \geq 0$ and $z^\star \in Z$ is the saddle point for (1) if and only if $\text{Gap}(z^\star) = 0$. Hence, we can measure the quality of an approximate solution using the Gap function.

Definition 2.2. For $\epsilon > 0$, we say that $\bar{z} \in Z$ is an ϵ-solution of problem (1) if $\text{Gap}(\bar{z}) \leq \epsilon$.

We call a function $h : H \to \mathbb{R}$ to be strongly-convex with modulus $\mu_h > 0$ if it satisfies $h(x') - h(x) - \langle \nabla h(x), x' - x \rangle \geq \frac{\mu_h}{2} \|x' - x\|^2$ for all $x', x \in H$

Throughout the paper, we make the following assumptions on the general coupling function $\phi(x, y)$:

Assumption 2.3. We assume function $\phi(\cdot, y)$ is L_{xx}-smooth for all $y \in Y$, $\phi(x, \cdot)$ is L_{yy}-smooth for all $x \in X$ and ϕ is L_{xy}-smooth, i.e., ϕ satisfies the following relations, respectively, for all $x, x' \in X,\ y, y' \in Y$:

$$\begin{aligned}
\|\nabla_x \phi(x', y) - \nabla_x \phi(x, y)\| &\leq L_{xx} \|x' - x\|, \\
\|\nabla_y \phi(x, y') - \nabla_y \phi(x, y)\| &\leq L_{yy} \|y' - y\|, \\
\|\nabla_y \phi(x', y) - \nabla_y \phi(x, y)\| &\leq L_{xy} \|x' - x\|.
\end{aligned}$$

If all Lipschitz constants above are positive, then $\phi(x, y)$ is a general nonlinear coupling function. If either $L_{xx} = 0$ or $L_{yy} = 0$, then the coupling function is linear in x or y, respectively. We refer to these cases as the *semi-linear coupling*. $L_{xx} = L_{yy} = 0$ implies a bilinear coupling.

3. Technical overview - The LPD method

For the bilinear SPP, i.e., $\phi(x,y) = y^\top Ax$, most PD methods use computationally expensive proximal operators of f and g. This may be reasonable in some applications where g is a regularizing function. However, that is not the case for f which arises from the primal optimization. To overcome this challenge, the linearized PD method (Chambolle & Pock, 2016) uses a linear approximation $f(x_t)+\langle \nabla f(x_t), x-x_t\rangle$ instead of evaluating a proximal operator. Algorithm 1 illustrates a typical LPD method, where parameters τ_t and η_t denote the step-sizes (or learning rates) in the dual and primal updates, respectively. The *momentum* parameter θ_t is used to generate an extrapolated sequence $\{\tilde{x}_t\}$ which is then used for the accelerated update of the dual y (line 3). On the other hand, the method uses a simple gradient descent step to update x (line 4). The algorithm outputs an ergodic average after K iterations. Chambolle & Pock (2016) showed an accelerated convergence of $\mathcal{O}(\frac{1}{K^2})$ for the strongly convex case ($\mu_f > 0, \mu_g = 0$). However, the strongly concave case ($\mu_f = 0, \mu_g > 0$) is missing. Furthermore, it is important to note that the two cases are not symmetric since we are linearizing the primal function f. A closer inspection shows that the two cases are quantitatively different. Here, we present two contrasting (and hence, somewhat surprising) results for the LPD method for these cases. Theorem 3.1 considers $\mu_f > 0$, and show convergence rate of $\mathcal{O}(\frac{1}{K^2})$ for the LPD method [2]. However, the LPD method does not effectively handle the error caused by the linearization of f when $\mu_g > 0$ (see Theorem 3.2). Below, we state the step-size conditions required for the analysis of the LPD method. See Appendix A for proofs of all results in this section.

Step-size conditions for the LPD method: For $t \ge 2$

$$\gamma_{t+1}(\tfrac{1}{\eta_t} - \mu_f) \le \tfrac{\gamma_t}{\eta_{t-1}}, \tag{2a}$$

$$\tfrac{\gamma_{t+1}}{\tau_t} \le \gamma_t \left(\mu_g + \tfrac{1}{\tau_{t-1}}\right), \tag{2b}$$

$$\theta_{t-1} = \tfrac{\gamma_t}{\gamma_{t+1}}, \tag{2c}$$

$$\theta_{t-1}\|A\|^2 \le (\tfrac{1}{\eta_{t-1}} - L_f)\tfrac{1}{\tau_t}. \tag{2d}$$

Theorem 3.1. *Assume that $\mu_f > 0$, $\mu_g = 0$ and set parameters $\{\gamma_t, \theta_t, \eta_t, \tau_t\}$ as per the following:*

$$\begin{aligned} \gamma_t &= \tfrac{t}{2} + \tfrac{L_f}{\mu_f}, \quad \theta_{t-1} = \tfrac{t/2 + L_f/\mu_f}{(t+1)/2 + L_f/\mu_f}, \\ \tfrac{1}{\eta_t} &= \mu_f \tfrac{t+1}{2} + L_f, \quad \tfrac{1}{\tau_t} = \tfrac{4\|A\|^2}{\mu_f(t+1)/2}. \end{aligned} \tag{3}$$

Then, we have

$$\begin{aligned} Gap(\bar{z}_{K+1}) \le & \tfrac{4}{K(K+3+4L_f/\mu_f)}\big[(1+\tfrac{L_f}{\mu_f})[\tfrac{\mu_f+L_f}{2}\|x-x_1\|^2 \\ & + \tfrac{4\|A\|^2}{2\mu_f}\|y-y_1\|^2]\big]. \end{aligned} \tag{4}$$

Algorithm 1 Linearized PD (LPD) method

```
1: Initialize $\tilde{x}_1 = x_1 \in X,\ y_1 \in Y$
2: for $t = 1, \dots, K$ do
3:     $y_{t+1} \leftarrow \arg\min_{y \in Y} \langle -A\tilde{x}_t, y\rangle + g(y) + \frac{1}{2\tau_t}\|y - y_t\|^2$
4:     $x_{t+1} \leftarrow \arg\min_{x \in X} \langle \nabla f(x_t) + A^\top y_{t+1}, x\rangle + \frac{1}{2\eta_t}\|x - x_t\|^2$
5:     $\tilde{x}_{t+1} \leftarrow x_{t+1} + \theta_t(x_{t+1} - x_t)$
6: end for
7: return $\bar{x}_{K+1} \leftarrow \frac{\sum_{t=1}^K \gamma_{t+1}x_{t+1}}{\sum_{t=1}^K \gamma_{t+1}}, \bar{y}_K \leftarrow \frac{\sum_{t=1}^K \gamma_{t+1}y_{t+1}}{\sum_{t=1}^K \gamma_{t+1}}$
```

It is easy to see that the step-size policy (3) satisfies the conditions in (2). Theorem 3.1 shows $\mathcal{O}(\frac{1}{K^2})$ convergence rate for Algorithm 1. It is also interesting to note that (3) provides an explicit expression of the weights γ_t which results in an explicit bound of $\Theta(K^2)$ on $\sum_{t=1}^K \gamma_t$ for $K \ge 1$. This bound is usually shown implicitly and for only large values of K in (Chambolle & Pock, 2011; 2016; Hamedani & Aybat, 2021). For the semi-linear couplings, a similar explicit policy is used in (Boob et al., 2022b).

In the second case ($\mu_f = 0, \mu_g > 0$), however, a step-size approach similar to (3) is not applicable. The following argument provides a rather mechanical intuition: To have an accelerated convergence rate of $\mathcal{O}(\frac{1}{K^2})$, we need $\Gamma_K := \sum_{t=1}^K \gamma_t = \Omega(K^2)$ and hence γ_t needs to increase linearly in t. In view of $\mu_f = 0$, (2a) requires $\frac{\gamma_{t+1}}{\eta_t}$ to be a decreasing sequence and we get $\frac{\gamma_2}{\eta_1} \ge \frac{\gamma_{K+1}}{\eta_K}$. Simultaneously, to mitigate errors generated by linearization of f, we require $\frac{1}{\eta_K} \ge L_f$ (see (2d)). These two relations and linearly increasing nature of γ_t imply that $\frac{1}{\eta_1} \ge \frac{L_f\gamma_{K+1}}{\gamma_2} = \Omega(L_fK)$. This is problematic since the final convergence error of the LPD method is of $\mathcal{O}(\frac{\gamma_2}{\eta_1\Gamma_K}) = \mathcal{O}(\frac{L_f}{K})$, a weaker convergence compared to $\mathcal{O}(\frac{1}{K^2})$. This is not observed when $\mu_f > 0$ and $\mu_g = 0$. Indeed in (3), we see that both γ_t and $\frac{1}{\eta_t}$ are both increasing in t and still (2a) is satisfied.

The critical issue is that (2a) requires $\{\frac{\gamma_{t+1}}{\eta_t}\}$ to be a decreasing sequence when $\mu_f = 0$. To provide a principled solution to this problem, we modify (2a) to allow $\frac{\gamma_{t+1}}{\eta_t}$ to increase with t by a fixed amount (see (5)). This approach requires a new step-size policy discussed below.

Modified step-size condition for the LPD method: Modify (2a) as follows while keeping (2b)-(2d) unchanged:

$$\tfrac{\gamma_{t+1}}{\eta_t} - \tfrac{\gamma_t}{\eta_{t-1}} \le L_f \tag{5}$$

Theorem 3.2. *Suppose $\mu_g > 0, \mu_f = 0$ and set parameters $\{\gamma_t, \theta_t, \eta_t, \tau_t\}$ as per the following:*

$$\begin{aligned} \gamma_t = t, \quad \tfrac{1}{\tau_t} = \mu_g \tfrac{t}{2}, \\ \tfrac{1}{\eta_t} = \tfrac{2\|A\|^2}{\mu_g(t+1)} + L_f, \quad \theta_{t-1} = \tfrac{t}{t+1}. \end{aligned} \tag{6}$$

[2] Though the result is similar to (Chambolle & Pock, 2016), the step-size policy is significantly different.

Then, we have

$$\mathit{Gap}(\bar{z}_{K+1}) \le \tfrac{2D_x^2\|A\|^2/\mu_g + D_y^2\mu_g}{K^2} + \tfrac{2(K+1)L_fD_x^2}{K^2}. \tag{7}$$

Note that (6) satisfies the modified step-size condition (5) and (2b)-(2d). From the result, it is clear that for the strongly concave SPP ($\mu_g > 0$), the convergence rate of the LPD method is of $\mathcal{O}(\frac{\|A\|^2}{K^2} + \frac{L_f}{K})$ when f is linearized.

This result is in sharp contrast with Theorem 3.1 where the convergence rate is of $\mathcal{O}(\frac{1}{K^2})$. We already provided a mechanical reasoning for the ineffectiveness of the LPD method in reducing the impact of Lipschitz constant L_f. At a broader design level, the algorithm itself is not accelerated in the primal iterate. Indeed, it is simply a gradient descent in the x-update (see Line 4 in Algorithm 1). This was not a problem when f was strongly convex. However, when only the dual is strongly-concave, one needs a stronger acceleration in the primal to mitigate the errors caused by the linearization of f. Hence, the rest of this paper is dedicated to presenting the accelerated linearized PD algorithm and its variant for obtaining more robust convergence results for problem (1) when f is linearized and $\mu_g > 0$.

4. The ALPD method for general ϕ

In addition to the primal acceleration mentioned in the earlier section, we consider two more generalizations: (i) we use the linear approximation for g instead of its proximal operator to allow the use of complex dual functions, (ii) the coupling function ϕ is a general nonlinear function.

To address the issues mentioned in Section 3 in the broader settings above, we present the accelerated linearized primal-dual (ALPD) method (see Algorithm 2). Here, we introduce a new parameter β, which is motivated from a (three-sequence) form of Nesterov's AGD algorithm (Nesterov, 1983). If we set $\beta_t = 1$ in Algorithm 2, then it is easy to see that $\underline{x}_t = x_t$ and $\bar{x}_{t+1} = x_{t+1}$ for all t, and we immediately recover the LPD method for the bilinear coupling $\phi(x, y) = y^\top Ax$. Hence, the ALPD method is a generalization of the LPD method in two senses: (i) using the parameter $\beta_t \ge 1$, we aim to put the AGD framework inside the LPD and reduce the impact of L_f in the complexity, and (ii) using a new sequence $\{v_t\}$ in place of $\{A\tilde{x}_t\}$, we allow for the nonlinear coupling function ϕ.

The following lemma provides a useful recursive relation on the primal-dual gap function of the iterates of Algorithm 2. It is later used for bounding the gap function (see Definition 2.1). See Appendix B for proofs of all results in this section.

Lemma 4.1. *Let $\bar{z}_{t+1} = (\bar{x}_{t+1}, \bar{y}_{t+1})$ then:*

$$\begin{aligned}
&\beta_t Q(\bar{z}_{t+1}, z) - (\beta_t - 1)Q(\bar{z}_t, z)\\
&\le \tfrac{1}{2\eta_t}\big[\|x - x_t\|^2 - \|x - x_{t+1}\|^2\big]\\
&\quad + \big[\big(\tfrac{1}{2\tau_t} - \tfrac{\mu_g}{2}\big)\|y - y_t\|^2 - \tfrac{1}{2\tau_t}\|y - y_{t+1}\|^2\big]\\
&\quad - \big(\tfrac{1}{2\eta_t} - \tfrac{L_f}{2\beta_t} - \tfrac{L_{xx}}{2}\big)\|x_t - x_{t+1}\|^2\\
&\quad - \big(\tfrac{1}{2\tau_t} - \tfrac{L_g}{2}\big)\tfrac{1}{2}\|y_t - y_{t+1}\|^2\\
&\quad + [\phi(x_{t+1}, y) - \phi(x_{t+1}, y_{t+1}) - \langle v_t, y - y_{t+1}\rangle]
\end{aligned} \tag{8}$$

Algorithm 2 Accelerated Linearized PD (ALPD) method

```
Initialize $\bar{x}_1 = x_0 = x_1 \in X, \bar{y}_1 = y_0 = y_1 \in Y$
for $t = 1, \ldots, K$ do
    $\underline{x}_t \leftarrow (1 - \beta_t^{-1})\bar{x}_t + \beta_t^{-1}x_t$
    $v_t \leftarrow (1 + \theta_t)\nabla_y\phi(x_t, y_t) - \theta_t\nabla_y\phi(x_{t-1}, y_{t-1})$
    $y_{t+1} \leftarrow \arg\min_{y\in Y}\langle -v_t + \nabla g(y_t), y\rangle + \frac{1}{2\tau_t}\|y - y_t\|^2$
    $x_{t+1} \leftarrow \arg\min_{x\in X}\langle \nabla f(\underline{x}_t) + \nabla_x\phi(x_t, y_{t+1}), x\rangle + \frac{1}{2\eta_t}\|x - x_t\|^2$
    $\bar{x}_{t+1} = (1 - \beta_t^{-1})\bar{x}_t + \beta_t^{-1}x_{t+1}$
    $\bar{y}_{t+1} = (1 - \beta_t^{-1})\bar{y}_t + \beta_t^{-1}y_{t+1}$
end for
return $\bar{x}_{K+1}, \bar{y}_{K+1}$
```

Lemma 4.2 states a step-size condition for parameters $\{\beta_t, \theta_t, \gamma_t, \tau_t, \eta_t\}$ and provides an upper bound on the $\mathrm{Gap}(\bar{z}_{K+1})$ where $\bar{z}_{K+1}$ is the output of the ALPD method.

Lemma 4.2. *Suppose $\{\beta_t, \theta_t, \gamma_t, \tau_t, \eta_t\}$ satisfy*

$$\begin{aligned}
&\beta_1 = 1, && \beta_{t+1} - 1 = \beta_t\theta_{t+1},\\
&\theta_t = \tfrac{\gamma_{t-1}}{\gamma_t}, && 0 \le \theta_t \le \tfrac{\tau_{t-1}}{\tau_t},\\
&\tfrac{1}{2\eta_t} - \tfrac{L_f}{2\beta_t} - 2L_{xy}^2\tau_t \ge \tfrac{L_{xx}}{2}, &&\\
&\tfrac{1}{4\tau_t} - \tfrac{L_g}{2} - 2L_{yy}^2\tau_t \ge 0, &&
\end{aligned} \tag{9}$$

then, we have

$$\begin{aligned}
&\beta_K\gamma_K Q(\bar{z}_{K+1}, z) \le B_K(z, z_{[K]})\\
&\quad + \gamma_K\langle \nabla_y\phi(x_{K+1}, y_{K+1}) - \nabla_y\phi(x_K, y_K), y - y_{K+1}\rangle\\
&\quad - \gamma_K\left(\tfrac{1}{2\eta_K} - \tfrac{L_f}{2\beta_K} - \tfrac{L_{xx}}{2}\right)\|x_{K+1} - x_K\|^2\\
&\quad - \gamma_K\left(\tfrac{1}{4\tau_K} - \tfrac{L_g}{2}\right)\|y_t - y_{K+1}\|^2,
\end{aligned} \tag{10}$$

where

$$\begin{aligned}
B_K(z, z_{[K]}) := \sum_{t=1}^{K}\{&\tfrac{\gamma_t}{2\eta_t}[\|x - x_t\|^2 - \|x - x_{t+1}\|^2]\\
&+ \gamma_t\big(\tfrac{1}{2\tau_t} - \tfrac{\mu_g}{2}\big)\|y - y_t\|^2 - \tfrac{\gamma_t}{2\tau_t}\|y - y_{t+1}\|^2\}.
\end{aligned}$$

A comparison of the ALPD step-size conditions in (9) with the LPD (in (2)) shows that the impact of L_f can be mitigated using the parameter β_t. Indeed, for the bilinear problems, i.e., $L_{xx} = 0$ and $L_{xy} = \|A\|$, the fifth relation in (9) reveals the necessity of condition $\frac{1}{\eta_t} \ge \frac{L_f}{\beta_t}$ for the ALPD method. Appropriate choice of β_t, (say, increasing with t) may allow us to increase η_t resulting in a stronger learning rate. Besides, (2d) requires $\frac{1}{\eta_t} \ge L_f$ and hence, no scope for improving the learning rate. Theorem 4.3 exhibits a tangible upper bound on the Gap function that explicitly shows the dependence of the convergence rate on β_t.

Theorem 4.3. *In addition to the assumptions in Lemma 4.2, let the following condition hold for $t \geq 2$:*

$$\gamma_t(\tfrac{1}{\tau_t} - \mu_g) \leq \tfrac{\gamma_{t-1}}{\tau_{t-1}}, \quad \tfrac{\gamma_t}{\eta_t} \leq \tfrac{\gamma_{t-1}}{\eta_{t-1}} + \tfrac{L_{xx}}{2} \tag{11}$$

Then, we have

$$Gap(\bar{z}_{K+1}) \leq \big(\tfrac{\gamma_1}{\beta_K\gamma_K\eta_1} + \tfrac{KL_{xx}}{\beta_K\gamma_K}\big)D_X^2 + \tfrac{\gamma_1}{\beta_K\gamma_K\tau_1}D_Y^2. \tag{12}$$

where D_X^2 and D_Y^2 are diameters of set X and Y.

4.1. Step-size policy for the ALPD method

Using the result of Theorem 4.3, we are ready to present step-size policy for the ALPD method. We break our analysis in two cases.

4.1.1. Case 1: Semi-linear coupling with $L_{xx} = 0$

Theorem 4.4. *Assume a semi-linear coupling function $\phi(x, y)$ which is linear in x, i.e., $L_{xx} = 0$ and consider the following choice of parameters for Algorithm 2:*

$$\begin{aligned}
&\gamma_1 = 1, \quad \gamma_t = \tfrac{t+1}{2} + \tfrac{2\sqrt{2}L_{yy}+2L_g}{\mu_g}, \quad t \geq 2,\\
&\theta_t = \tfrac{\gamma_{t-1}}{\gamma_t}, \quad t \geq 2\\
&\beta_1 = 1, \quad \beta_{t+1} = 1 + \theta_{t+1}\beta_t,\\
&\eta_t = \tfrac{t+1}{5L_f+16L_{xy}^2/\mu_g},\\
&\tfrac{1}{\tau_t} = \tfrac{\mu_g t}{2} + 2\sqrt{2}L_{yy} + 2L_g.
\end{aligned} \tag{13}$$

Then we obtain the complexity of $K = O(\sqrt{\frac{L_f+L_{yy}}{\epsilon}} + \frac{L_{xy}}{\sqrt{\mu_g\epsilon}})$ for getting an ϵ-solution of (1).

Proof. Comparing the step-size policy in (13) with conditions in (9) where $L_{xx} = 0$, it is easy to see that the relations $\theta_t = \frac{\gamma_{t-1}}{\gamma_t}$, the recursive relation on β_t and $\frac{1}{4\tau_t} - \frac{L_g}{2} - 2L_{yy}^2\tau_t \geq 0$ are satisfied. Furthermore, since γ_t is increasing and τ_t is decreasing, we have $\theta_t < 1 < \frac{\tau_{t-1}}{\tau_t}$. It is straight-forward to see that $\{\frac{\gamma_t}{\eta_t}\}$ is a decreasing sequence. Besides, by choosing τ_t according to this step-size policy, the first condition in (11) also holds. The proposition below provides a bound on β_t.

Proposition 4.5. *Suppose we set the step-size parameters according to* (13) *then $\beta_{t+1} \in [\frac{t+2}{2}, t+1]$.*

Using the above proposition, we verify the one remaining condition of (9) with $L_{xx} = 0$:

$$\begin{aligned}
&\tfrac{1}{2\eta_t} - \tfrac{L_f}{2\beta_t} - 2L_{xy}^2\tau_t\\
&\geq \tfrac{5L_f}{2(t+1)} - \tfrac{L_f}{2\beta_t} + \tfrac{16L_{xy}^2}{2\mu_g(t+1)} - \tfrac{4L_{xy}^2}{\mu_g t+4\sqrt{2}L_{yy}} \geq 0,
\end{aligned}$$

where the first inequality follows by replacing the values of η_t, τ_t along with the fact that $L_g \geq 0$, and the second inequality holds since $\beta_t \geq \frac{t}{2} \geq \frac{t+1}{5}$ and $\frac{16L_{xy}^2}{2\mu_g(t+1)} - \frac{4L_{xy}^2}{\mu_g t} \geq 0$ for $t \geq 1$.

Using (12) in Theorem 4.3, we obtain the following upper bound on the Gap:

$$\text{Gap}(\bar{z}_{K+1}) \leq \tfrac{1}{\beta_K\gamma_K\eta_1}D_X^2 + \tfrac{1}{\beta_K\gamma_K\tau_1}D_Y^2. \tag{14}$$

Note that since β_t and γ_t are increasing at a linear rate, we obtain the accelerated convergence rate of $O(\frac{L_f+L_{yy}}{K^2} + \frac{L_{xy}^2}{\mu_g K^2})$ which is equivalent to the complexity of $K = O(\sqrt{\frac{L_f+L_{yy}}{\epsilon}} + \frac{L_{xy}}{\sqrt{\mu_g\epsilon}})$ for getting an ϵ-solution. □

Remark 4.6. (Hamedani & Aybat, 2021) is the only known single-loop PD algorithm that shows accelerated convergence when the coupling function is semi-linear with $L_{yy} = 0$ and $\mu_f > 0$. We have a (reflected) result where $\mu_g > 0$ and $L_{xx} = 0$. Even then, (Hamedani & Aybat, 2021) assume f and g have proximal updates. Hence, they do not need any additional acceleration of the ALPD method.

4.1.2. Case 2: nonlinear coupling

Theorem 4.7. *Consider an SPP with a general nonlinear coupling function, i.e., $L_{xx} > 0$. Assuming the step-size policy in* (13) *with the following single change in η_t as*

$$\eta_t = \tfrac{t+1}{5L_f+16L_{xy}^2/\mu_g+(t+1)L_{xx}},$$

we obtain the complexity of $K = \mathcal{O}(\sqrt{\frac{L_f+L_{yy}}{\epsilon}} + \frac{L_{xy}}{\sqrt{\mu_g\epsilon}} + \frac{L_{xx}}{\epsilon})$ for getting an ϵ-solution of (1).

Proof. It is easy to see that the mentioned step-size policy satisfies the condition (9) in Lemma 4.2 (including fifth relation in (9)). Furthermore, (11) is also satisfied. Thus, using Theorem 4.3, we can establish the following upper bound on the Gap function:

$$\text{Gap}(\bar{z}_{K+1}) \leq \big(\tfrac{\gamma_1}{\beta_K\gamma_K\eta_1} + \tfrac{KL_{xx}}{\beta_K\gamma_K}\big)D_X^2 + \tfrac{\gamma_1}{\beta_K\gamma_K\tau_1}D_Y^2.$$

Consequently, the gradient complexity in this case is $K = \mathcal{O}(\sqrt{\frac{L_f+L_{yy}}{\epsilon}} + \frac{L_{xy}}{\sqrt{\mu_g\epsilon}} + \frac{L_{xx}}{\epsilon})$. □

Hence, though we get acceleration in terms of L_f, the convergence rate in terms of L_{xx} is of $\mathcal{O}(\frac{1}{K})$. This is similar to the LPD case where the complexity had a weaker dependence on L_f. ALPD method does accelerate on the primal only term, i.e., L_f. However, accelerating the convergence for the primal coupling term is still difficult. In light of Remark 4.6, accelerating the class of PD methods for nonlinear coupling is a challenging open problem, even without linearization.

Algorithm 3 Inexact ALPD Method

```
1: Initialize $\bar{x}_1 = x_0 = x_1 \in X, \bar{y}_1 = y_0 = y_1 \in Y$
2: for $t = 1, \ldots, K$ do
3:     $\underline{x}_t \leftarrow (1 - \beta_t^{-1})\bar{x}_t + \beta_t^{-1} x_t$
4:     $v_t \leftarrow (1 + \theta_t)\nabla_y \phi(x_t, y_t) - \theta_t \nabla_y \phi(x_{t-1}, y_{t-1})$
5:     $y_{t+1} \leftarrow \arg\min_{y \in Y} \langle -v_t + \nabla g(y_t), y\rangle + \frac{1}{2\tau_t}\|y - y_t\|^2$
6:     $x_{t+1}$ is a $\delta_t$-approximate solution of the problem:
           $\min_{x \in X} \langle \nabla f(\underline{x}_t), x\rangle + \phi(x, y_{t+1}) + \frac{1}{2\eta_t}\|x - x_t\|^2 \quad (15)$
7:     $\bar{x}_{t+1} \leftarrow (1 - \beta_t^{-1})\bar{x}_t + \beta_t^{-1} x_{t+1}$
8:     $\bar{y}_{t+1} \leftarrow (1 - \beta_t^{-1})\bar{y}_t + \beta_t^{-1} y_{t+1}$
9: end for
10: return $\bar{x}_{K+1}, \bar{y}_{K+1}$
```

5. The Inexact ALPD method for general ϕ

This section proposes an Inexact ALPD method to improve the complexity in L_{xx}. The linearization of $\phi(x, y_{t+1})$ in the ALPD method generates errors that depend on L_{xx}. It leads to a slow convergence rate when $L_{xx} > 0$. To fix this issue, we use $\phi(x, y_{t+1})$ instead of its linearization in the x-update (compare line 6 of Algorithms 2 and 3). However, we cannot evaluate the proximal oracle of $\phi(\cdot, y)$ efficiently. To evaluate the truly representative computational effort for this algorithm, we propose an inexact approach in the x-update and perform a detailed analysis of the inner loop to estimate the complexity bounds. The rest of this section is dedicated to the complexity analysis of Algorithm 3 in the outer loop and inner loop.

5.1. Complexity analysis of the Inexact ALPD method

Complexity analysis of the outer loop

Using a proximal oracle of $\phi(\cdot, y_{t+1})$ in the x-update removes the linearization errors that depend on L_{xx}. Hence, the outer loop analysis reduces to Case 4.1.1. Consequently, we have the following theorem.

Theorem 5.1. *Suppose conditions in ((9), (11)) and the step-size policy ((13)) hold, then we have the following upper bound for the Gap function*

$$Gap(\bar{z}_{K+1}) \leq \frac{1}{\beta_K \gamma_K \eta_1} D_X^2 + \frac{1}{\beta_K \gamma_K \tau_1} D_Y^2 + \frac{\sum_{t=1}^K \gamma_t \delta_t}{\beta_K \gamma_K} + \frac{\sum_{t=1}^K \gamma_t \sqrt{4 \frac{1}{\eta_t} \delta_t D_X^2}}{\beta_K \gamma_K}. \tag{16}$$

The above upper bound is similar to (14) with the addition of the last two terms since we are using a δ_t-approximate solution for (15). The detailed proof and analysis of the inexact ALPD method is in Appendix C. We need to manage the error caused by δ_t. Proposition 5.2 provides the required condition to bound such errors.

Proposition 5.2. *Suppose $\delta_t = \frac{1}{t^c}$ and the step-size policy in 13 holds. Then by choosing $c = 3.5$, $\sum_{t=1}^K \gamma_t \delta_t$ and $\sum_{t=1}^K \gamma_t \sqrt{\delta_t / \eta_t}$ are bounded by a constant.*

One important result of Proposition 5.2 is obtaining complexity of ∇f, and $\nabla_y \phi$ as the following corollary

Corollary 5.3. *Suppose Proposition 5.2 holds for δ_t, then we need $\mathcal{O}(\sqrt{\frac{L_f + L_{yy}}{\epsilon}} + \frac{L_{xy}}{\sqrt{\mu_g \epsilon}})$ evaluations in ∇f, and $\nabla_y \phi$ to obtain an ϵ-solution of* (1).

To compute the gradient complexity of $\nabla_x \phi$ and the impact of the above choice of c, we perform the inner loop analysis as below.

Complexity analysis of the inner loop

Theorem 5.4. *The complexity of $\nabla_x \phi$ evaluations is $\mathcal{O}\big(c\sqrt{L_{xx}\sqrt{L_f + L_{xy}^2/\mu_g}} \frac{1}{\epsilon^{3/4}} \log(\frac{1}{\epsilon})\big)$*

Proof. We implement the AGD method ((Nesterov, 2003)) to solve the subproblem (15). Let k_t denote the number of AGD iterations for the t'th iteration in the outer loop. Consequently, the complexity of $\nabla_x \phi$ after K outer loop iterations is $\sum_{t=1}^K k_t$. The number of AGD iterations k_t is directly related to the choice of error δ_t. Nesterov (2003) shows that for a L-smooth and μ-strongly convex function, we need $\mathcal{O}(\sqrt{\frac{L}{\mu}} \log(\frac{1}{\epsilon}))$ AGD iterations to obtain an ϵ error on the optimality. Then, to obtain a δ_t error on the optimality of (15), we need $k_t = \mathcal{O}(\sqrt{L_{xx}\eta_t} \log(\frac{1}{\delta_t}))$ iterations of the AGD method. Here, we used $L = L_{xx}$ and $\mu = \frac{1}{\eta_t}$. Setting δ_t according to Proposition 5.2, we obtain $k_t = \mathcal{O}(\sqrt{L_{xx}\eta_t} \log(t^c))$ or $\mathcal{O}(c\sqrt{L_{xx}\eta_t} \log(t))$. Hence, the total number of iterations of AGD for K outer iterations (i.e., the number of gradients evaluations of $\nabla_x \phi$) is

$$\begin{aligned} \sum_{t=1}^K k_t &= \sum_{t=1}^K c\sqrt{L_{xx}\eta_t} \log(t) \\ &= \sum_{t=1}^K c\sqrt{L_{xx} \frac{t+1}{5L_f + 16L_{xy}^2/\mu_g}} \log(t) \\ &\leq c\sqrt{\frac{L_{xx}}{5L_f + 16L_{xy}^2/\mu_g}} (K+1)^{3/2} \log(K+1). \end{aligned}$$

Using $K = \mathcal{O}(\sqrt{\frac{L_f + L_{xy}^2/\mu_g}{\epsilon}})$, we obtain the result. □

We immediately get the following corollary.

Corollary 5.5. *Suppose L_{xx}, L_f, and L_{xy}^2/μ_g are of $\mathcal{O}(L)$, then we need $\tilde{\mathcal{O}}(\frac{L}{\epsilon^{3/4}})$ evaluations of $\nabla_x \phi$ to obtain an ϵ-error in inexact ALPD.*

The following remark is in order.

Remark 5.6. Observe the impact of c on the complexity of $\nabla_x \phi$ is only of a constant factor. This happens since (15) is a strongly convex problem. In view of Corollary 5.5 and

$K = \mathcal{O}(\sqrt{\frac{L_f + L_{xy}^2/\mu_g}{\epsilon}})$, inexact ALPD method exhibits the gradient complexity of $\mathcal{O}(\sqrt{\frac{L_f+L_{yy}}{\epsilon}} + \frac{L_{xy}}{\sqrt{\mu_g \epsilon}})$ for ∇f, and $\nabla_y \phi$. Moreover, its gradient complexity for $\nabla_x \phi$ is of $\tilde{\mathcal{O}}(\frac{L}{\epsilon^{3/4}})$. In comparison, the ALPD method has $\mathcal{O}(\frac{L}{\epsilon})$ gradient complexity for $\nabla_x \phi, \nabla_y \phi$ and ∇f. The Inexact ALPD method improves the complexity in $\nabla_x \phi$, and obtains the optimal complexity in ∇f as well as $\nabla_y \phi$.

6. Numerical Experiments

In this section, we perform numerical experiments to (i) compare the performance of the LPD and ALPD algorithms on the penalty problems with different settings; (ii) evaluate the runtime performance of the step-size policies in Theorem 3.1 and (Chambolle & Pock, 2016); (iii) compare the ALPD and Inexact ALPD on penalty problems for nonlinear constraints. All experiments are performed on 64-bit Windows 10 with Intel i5-9500U @3.00GHz and 16GB RAM.

6.1. ALPD vs. LPD

The ℓ_q-norm penalty problem with linear constraints is

$$\min_{x \in X} f(x) + \rho\|Ax - b\|_q \equiv \min_{x\in X} \max_{\|y\|_p \le 1} f(x) + \rho\langle y, Ax - b\rangle,$$

where ℓ_p-norm is the dual norm of $\|\cdot\|_q$. The equivalence of the dual formulation is well-known where $1/p + 1/q = 1$. We can get a smooth approximation of the nonsmooth penalty term using Nesterov's smoothing technique

$$\min_{x\in X} \max_{\|y\|_p \le 1} \{f(x) + \rho\langle y, Ax - b\rangle - \tfrac{\mu_g}{2}\|y\|^2\}, \qquad (17)$$

where parameter μ_g can be used to calibrate the smoothness of the approximation. We set $f(x) = \frac{1}{2}x^\top Qx + c^\top x$ as a convex quadratic function where $Q \in \mathbb{R}^{n\times n}$ is a randomly generated positive semidefinite matrix and $c \in \mathbb{R}^n$ is a random vector. We also generate matrix $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$ randomly. For these experiments, we set the penalty parameter $\rho = 1$ and $m = n = 100$. Appendix D provides detailed information on the exact functions used for the random number generation. We set $L_f = 200$ since eigenvalues of Q are generated uniformly on $[0, 200]$.

We implement two versions of the ALPD method. The first method is implemented exactly as presented in Algorithm 2. The second method uses a proximal operator of g as follows: line 5 of Algorithm 2 is replaced by $y_{t+1} = \arg\min_{y\in Y}\langle -v_t, y\rangle + g(y) + \frac{1}{2\tau_t}\|y - y_t\|^2$. We make these changes (1) to measure the effect of using linearization in g on the numerical performance of the ALPD method, and (2) to perform a fair comparison with the LPD method as it uses the more advantageous proximal operator of g. We refer to this method as ALPD-prox-g. The step-size policy for this version is similar to (13) with $L_g = 0$ since g is used exactly without linearization. We measure the performance of the algorithms using three metrics: (1) Gap function which is the standard metric used in the convergence analysis, (2) Primal relative error $\|\bar{x}_t - x^*\|/\|x^*\|$, and (3) Dual relative error $\|\bar{y}_t - y^*\|/\|y^*\|$. All algorithms start at the same randomly generated initial point in the domain $X \times Y$. Figure 1 compares the three algorithms in three metrics. Each plot is generated using the average performance of the algorithms on 10 instances of (17) generated independently with identical distribution (i.i.d. instances). We plot the metrics for the last 50 iterations to focus on the major performance differences. Figure 1 shows that when $L_f = 200$ (a large number), the LPD method performs poorly compared to both versions of ALPD. Moreover, ALPD-prox-g gives a slight advantage over ALPD which is expected. Note that in these experiments, we use $p = q = 2$. In Appendix E), we provide a similar comparison for two settings of (17): $q = 1$ and $q = \infty$.

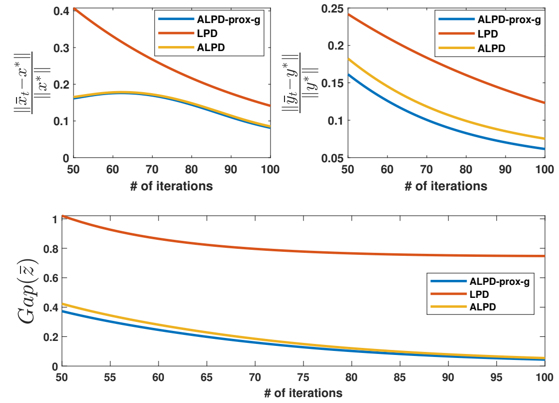

Figure 1: Comparison of the methods in terms of the mean errors in primal (top left), dual (top right), and Gap function (bottom) for 10 i.i.d. instances of 17 with $p = q = 2$.

6.2. ALPD vs. Inexact ALPD

In this subsection, we compare the performances of Algorithms 2 and 3 on the penalty problem with nonlinear constraints. We replace the linear constraints in the previous case with quadratic constraints $\frac{1}{2}x^\top A_j x + b_j^\top x - d_j \le 0$, for all $j \in [m]$ where A_j, b_j and $d_j(> 0)$ are randomly generated as in the previous experiment. The dual form of the penalty functions on nonlinear constraints has $L_{xx} > 0$. As we proved in Section 5.1, when $L_{xx} > 0$, Inexact ALPD is superior to ALPD in terms of gradient complexity. To verify our results, we run 10 i.i.d. instances of the nonlinear penalty problem and plot the Gap function against the average run time of each algorithm. For the ALPD method, we use the step-size policy in Section 4.1.2 and Inexact ALPD method is employed as described in Algorithm 3. Moreover, we implement prox versions of both algorithms where we use the proximal oracle of g instead of linearizing it. We call these versions ALPD-prox-g and Inexact-ALPD-prox-g respectively. Figure 2 illustrates the behavior of these algorithms for 100-dimensional ($n = 100$) penalty problems with 10 non-linear constraints ($m = 10$). We run the ALPD method for 200 iterations and its inexact counterpart for 100 iterations. We can see that Inexact ALPD and Inexact ALPD-prox-g dominate the performance of ALPD and ALPD-prox-g, respectively.

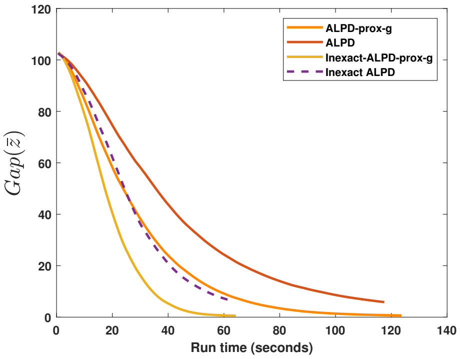

Figure 2: Comparison of the ALPD and inexact ALPD method and their prox-g variants using the Gap function vs run-time (seconds) plot for 10 i.i.d. instances.

6.3. LPD step-size policy comparison

As we mentioned in Section 3, both policies in (3) and (Chambolle & Pock, 2016) give similar convergence rates asymptotically. To make the numerical comparison, we use the SPP in (17) with $\mu_g = 0$. We set μ_f as the minimum eigenvalue of the randomly generated matrix Q. Note that $\mu_f > 0$ almost surely. We run the LPD method for 10 i.i.d. instances of this problem for each step-size policy. See Appendix F for the details of our numerical study. It seems that the LPD method using step-size in (3) performs better than (Chambolle & Pock, 2016). We conjecture the following reason for this deviation in the performance: Chambolle & Pock (2016) show that $\sum_{t=1}^{K}\gamma_t = \Omega(K^2)$ only for large values of K whereas (3) defines $\gamma_t = \Theta(t)$ explicitly and hence $\sum_{t=1}^{K}\gamma_t = \Theta(K^2)$ for all $K \geq 1$. The quadratic growth of $\sum_{t=1}^{K}\gamma_t$ is important to obtain the accelerated $O(\frac{1}{K^2})$ convergence rate. Hence, the step-size policy in (3) seems to be performing well in our experiments.

7. Conclusion

We showed that the standard LPD methods do not mitigate the impact of the linearization of the primal function for convex-strongly-concave SPP. Therefore, we designed the ALPD method which exhibits the optimal complexity for the semi-linear coupling case. For the general nonlinear coupling, we designed a two-loop Inexact ALPD method that maintains the optimal gradient complexity of the primal function and significantly improves the gradient complexity of the coupling function. We verified our findings through numerical experiments.

Acknowledgements

We would like to thank the anonymous reviewers for their helpful comments that enriched the literature review of this paper. DB's research is partially supported by NSF grant CCF 2245705.

References

Boob, D., Deng, Q., and Lan, G. Level constrained first order methods for function constrained optimization. *arXiv preprint arXiv:2205.08011*, 2022a.

Boob, D., Deng, Q., and Lan, G. Stochastic first-order methods for convex and nonconvex functional constrained optimization. *Mathematical Programming*, pp. 1–65, 2022b.

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

Chambolle, A. and Pock, T. On the ergodic convergence rates of a first-order primal–dual algorithm. *Mathematical Programming*, 159(1):253–287, 2016.

Chen, Y., Lan, G., and Ouyang, Y. Optimal primal-dual methods for a class of saddle point problems. *SIAM Journal on Optimization*, 24(4):1779–1814, 2014.

Chen, Y., Lan, G., and Ouyang, Y. Accelerated schemes for a class of variational inequalities. *Mathematical Programming*, 165(1):113–149, 2017.

Condat, L. A primal–dual splitting method for convex optimization involving lipschitzian, proximable and linear composite terms. *Journal of Optimization Theory and Applications*, 158:460–479, 2013.

Dang, C. and Lan, G. Randomized first-order methods for saddle point optimization. *arXiv preprint arXiv:1409.8625*, 2014.

Hamedani, E. Y. and Aybat, N. S. A primal-dual algorithm with line search for general convex-concave saddle point problems. *SIAM Journal on Optimization*, 31(2):1299–1329, 2021.

Hamedani, E. Y., Jalilzadeh, A., Aybat, N. S., and Shanbhag, U. V. Iteration complexity of randomized primal-dual methods for convex-concave saddle point problems. *arXiv preprint arXiv:1806.04118*, 2018.

Juditsky, A., Nemirovski, A. S., and Tauvel, C. Solving variational inequalities with Stochastic Mirror-Prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011. doi: 10.1214/10-SSY011. URL https://hal.archives-ouvertes.fr/hal-00318043.

Korpelevich, G. M. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976.

Kovalev, D., Gasnikov, A., and Richtárik, P. Accelerated primal-dual gradient method for smooth and convex-concave saddle-point problems with bilinear coupling, 2022.

Lin, T., Jin, C., and Jordan, M. I. Near-optimal algorithms for minimax optimization. In *Conference on Learning Theory*, pp. 2738–2779. PMLR, 2020.

Monteiro, R. D. and Svaiter, B. F. On the complexity of the hybrid proximal extragradient method for the iterates and the ergodic mean. *SIAM Journal on Optimization*, 20(6): 2755–2787, 2010.

Nemirovski, A. Prox-method with rate of convergence o(1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004. doi: 10.1137/S1052623403425629. URL https://doi.org/10.1137/S1052623403425629.

Nemirovski, A. and Yudin, D. *Problem complexity and method efficiency in optimization*. Wiley-Interscience publication in Discrete Mathematics. John Wiley, XV, 1983.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009. doi: 10.1137/070704277.

Nesterov, Y. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

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

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

Nesterov, Y. E. A method for solving the convex programming problem with convergence rate $O\left(\frac{1}{k^2}\right)$. In *Dokl. Akad. Nauk SSSR,*, volume 269, pp. 543–547, 1983.

Thekumparampil, K. K., Jain, P., Netrapalli, P., and Oh, S. Efficient algorithms for smooth minimax optimization, 2019.

Thekumparampil, K. K., He, N., and Oh, S. Lifted primal-dual method for bilinearly coupled smooth minimax optimization, 2022.

Tseng, P. On accelerated proximal gradient methods for convex-concave optimization. *submitted to SIAM Journal on Optimization*, 2(3), 2008.

Vu, B. C. A splitting algorithm for dual monotone inclusions involving cocoercive operators. *Advances in Computational Mathematics*, 38:667–681, 2011.

Yu, A. W., Lin, Q., and Yang, T. Doubly stochastic primal-dual coordinate method for regularized empirical risk minimization with factorized data. *CoRR, abs/1508.03390*, 2015.

Zhang, Y. and Lin, X. Stochastic primal-dual coordinate method for regularized empirical risk minimization. In *International Conference on Machine Learning*, pp. 353–361. PMLR, 2015.

Zhu, Z. and Storkey, A. J. Adaptive stochastic primal-dual coordinate descent for separable saddle point problems. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 645–658. Springer, 2015.

Appendix

A. General Analysis of Algorithm 1 (LPD)

In this section, we state some technical results that are ultimately used for obtaining (2a)-(2d) and Theorems 3.1 and 3.2. First, let us state two important lemmas that are utilized in the rest of the discussion especially when we want to construct relations related to optimality points.

Lemma A.1. *Let $x^\star$ be a δ-approximate solution of problem $\min_{x\in X}\{h(x)+\frac{\lambda}{2}\|x-\hat{x}\|^2\}$ where $h(x)$ is a convex function. Then,*

$$h(x^\star)-h(x)\leq\tfrac{\lambda}{2}\left[\|x-\hat{x}\|^2-\|x^\star-x\|^2-\|x^\star-\hat{x}\|^2\right]+\delta+\sqrt{2\lambda\delta}\|x^\star-x\|. \tag{18}$$

This lemma is known as "Three-point" lemma and also can be stated for a strongly-convex function h with modulus μ_h as below

$$h(x^\star)-h(x)\leq\tfrac{\lambda}{2}\left[\|x-\hat{x}\|^2-\|x^\star-x\|^2\ -\|x^\star-\hat{x}\|^2\right]-\tfrac{\mu_h}{2}\|x^\star-x\|^2+\delta+\sqrt{2\lambda\delta}\|x^\star-x\|. \tag{19}$$

Note that the proof can be found in Lemma 7 of Boob et al. (2022a).

Lemma A.2. *For point $z_{t+1}=(x_{t+1},y_{t+1})\in Z$ in Algorithm 1, the primal-dual gap function is upper bounded as follows*

$$Q(z_{t+1},z)\leq\tfrac{L_f}{2}\|x_{t+1}-x_t\|^2-\tfrac{\mu_f}{2}\|x-x_t\|^2+\langle\nabla f(x_t),x_{t+1}-x\rangle+[g(y_{t+1})-g(y)]+\langle Ax_{t+1},y\rangle-\langle Ax,y_{t+1}\rangle. \tag{20}$$

Proof. Since f is L_f-smooth, we have

$$\begin{aligned}f(x_{t+1})&\leq f(x_t)+\langle\nabla f(x_t),x_{t+1}-x_t\rangle+\tfrac{L_f}{2}\|x_{t+1}-x_t\|^2\\&=f(x_t)+\langle\nabla f(x_t),x_{t+1}-x\rangle+\langle\nabla f(x_t),x-x_t\rangle+\tfrac{L_f}{2}\|x_{t+1}-x_t\|^2\\&\leq f(x)+\langle\nabla f(x_t),x_{t+1}-x\rangle+\tfrac{L_f}{2}\|x_{t+1}-x_t\|^2-\tfrac{\mu_f}{2}\|x-x_t\|^2.\end{aligned}$$

Adding $[g(y_{t+1})-g(y)]$, and $[\langle Ax_{t+1},y\rangle-\langle Ax,y_{t+1}\rangle]$ to the both sides leads to the (20). □

We can elaborate on the upper bound by using the optimality conditions of y_{t+1} and x_{t+1} respectively. The following theorem illustrates a useful upper bound for the weighted gap function for the LPD method.

Theorem A.3. *if for $t\geq 2$*

$$\gamma_{t+1}(\tfrac{1}{\eta_t}-\mu_f)\leq\tfrac{\gamma_t}{\eta_{t-1}}, \tag{21a}$$

$$\tfrac{\gamma_{t+1}}{\tau_t}\leq\gamma_t\left(\mu_g+\tfrac{1}{\tau_{t-1}}\right), \tag{21b}$$

$$\theta_{t-1}=\tfrac{\gamma_t}{\gamma_{t+1}}, \tag{21c}$$

$$\theta_{t-1}\|A\|^2\leq(\tfrac{1}{\eta_{t-1}}-L_f)\tfrac{1}{\tau_t}. \tag{21d}$$

then

$$\begin{aligned}\sum_{t=1}^{K}\gamma_{t+1}Q(z_{t+1},z)\leq\ &\tfrac{\gamma_2}{2}(\tfrac{1}{\eta_1}-\mu_f)\|x-x_1\|^2-\tfrac{\gamma_{K+1}}{2\eta_K}\|x-x_{K+1}\|^2+\tfrac{\gamma_2}{2\tau_1}\|y-y_1\|^2\\&-\tfrac{\gamma_{K+1}}{2}\left(\mu_g+\tfrac{1}{\tau_K}-\tfrac{\|A\|^2}{\frac{1}{\eta_K}-L_f}\right)\|y-y_{K+1}\|^2,\end{aligned} \tag{22}$$

and at optimality

$$\tfrac{\gamma_{K+1}}{2}\left(\mu_g+\tfrac{1}{\tau_K}-\tfrac{\|A\|^2}{\frac{1}{\eta_K}-L_f}\right)\|y-y_{K+1}\|^2\leq\tfrac{\gamma_2}{2\eta_1}\|x^\star-x_1\|^2+\tfrac{\gamma_2}{2\tau_1}\|y^\star-y_1\|^2.$$

Proof. Using the optimality of y_{t+1} and Lemma A.1 where $\delta=0$ (note y_{t+1} is an exact solution.), we have

$$g(y_{t+1})-g(y)\leq\tfrac{1}{2\tau_t}\left(\|y-y_t\|^2-\|y_{t+1}-y_t\|^2\right)-(\tfrac{1}{2\tau_t}+\tfrac{\mu_g}{2})\|y-y_{t+1}\|^2+\langle A\tilde{x_t},y_{t+1}-y\rangle. \tag{23}$$

Also, from the optimality of x_{t+1}, we have the following

$$\langle \nabla f(x_t), x_{t+1}-x\rangle \le \tfrac{1}{2\eta_t}\left(\|x-x_t\|^2 - \|x_{t+1}-x_t\|^2\right) - \tfrac{1}{2\eta_t}\|x-x_{t+1}\|^2 - \langle A(x_{t+1}-x), y_{t+1}\rangle. \tag{24}$$

From (20), (23) and (24), one can reconstruct the following upper bound on the gap function at one iteration

$$\begin{aligned} Q(z_{t+1},z) \le{}& \Big((\tfrac{1}{2\eta_t} - \tfrac{\mu_f}{2})\|x-x_t\|^2 - \tfrac{1}{2\eta_t}\|x-x_{t+1}\|^2\Big) - (\tfrac{1}{2\eta_t} - \tfrac{L_f}{2})\|x_{t+1}-x_t\|^2 + \tfrac{1}{2\tau_t}\left(\|y-y_t\|^2 - \|y_{t+1}-y_t\|^2\right) \\ & - (\tfrac{1}{2\tau_t} + \tfrac{\mu_g}{2})\|y-y_{t+1}\|^2 - \langle A(x_{t+1}-x), y_{t+1}\rangle + \langle A\tilde{x_t}, y_{t+1}-y\rangle + \langle Ax_{t+1}, y\rangle - \langle Ax, y_{t+1}\rangle. \end{aligned} \tag{25}$$

We can simplify the upper bound with respect to the inner products.

$$\begin{aligned} &- \langle A(x_{t+1}-x), y_{t+1}\rangle + \langle A\tilde{x_t}, y_{t+1}-y\rangle + \langle Ax_{t+1}, y\rangle - \langle Ax, y_{t+1}\rangle \\ &= -\langle A(x_{t+1}-x), y_{t+1}\rangle + \langle A(x_t + \theta_{t-1}(x_t - x_{t-1})), y_{t+1}-y\rangle + \langle Ax_{t+1}, y\rangle - \langle Ax, y_{t+1}\rangle \\ &= -\langle Ax_{t+1}, y_{t+1}\rangle + \langle Ax, y_{t+1}\rangle + \langle Ax_t, y_{t+1}\rangle - \langle Ax_t, y\rangle \\ &\quad + \theta_{t-1}\langle A(x_t - x_{t-1}), y_{t+1}-y\rangle + \langle Ax_{t+1}, y\rangle - \langle Ax, y_{t+1}\rangle \\ &= -\langle Ax_{t+1}, y_{t+1}\rangle + \langle Ax_t, y_{t+1}\rangle - \langle Ax_t, y\rangle + \theta_{t-1}\langle A(x_t - x_{t-1}), y_{t+1}-y\rangle + \langle Ax_{t+1}, y\rangle \\ &= -\langle A(x_{t+1}-x_t), y_{t+1}-y\rangle + \theta_{t-1}\langle A(x_t - x_{t-1}), y_{t+1}-y\rangle. \end{aligned}$$

Also, we can write the above expression as follows

$$\begin{aligned} &- \langle A(x_{t+1}-x_t), y_{t+1}-y\rangle + \theta_{t-1}\langle A(x_t - x_{t-1}), y_{t+1}-y\rangle \\ &= -[\langle A(x_{t+1}-x_t), y_{t+1}-y\rangle - \theta_{t-1}\langle A(x_t - x_{t-1}), y_t - y\rangle + \theta_{t-1}\langle A(x_t - x_{t-1}), y_t - y_{t+1}\rangle]. \end{aligned}$$

From (25), we can rewrite the upper bound for gap function as follows

$$\begin{aligned} Q(z_{t+1},z) \le{}& (\tfrac{1}{2\eta_t} - \tfrac{\mu_f}{2})\|x-x_t\|^2 - \tfrac{1}{2\eta_t}\|x-x_{t+1}\|^2 + \tfrac{1}{2\tau_t}\|y-y_t\|^2 - (\tfrac{1}{2\tau_t} + \tfrac{\mu_g}{2})\|y-y_{t+1}\|^2 \\ & - \langle A(x_{t+1}-x_t), y_{t+1}-y\rangle + \theta_{t-1}\langle A(x_t - x_{t-1}), y_t - y\rangle \\ & - \tfrac{1}{2\tau_t}\|y_{t+1}-y_t\|^2 - (\tfrac{1}{2\eta_t} - \tfrac{L_f}{2})\|x_{t+1}-x_t\|^2 - \theta_{t-1}\langle A(x_t - x_{t-1}), y_t - y_{t+1}\rangle. \end{aligned} \tag{26}$$

Hence, multiplying both sides by γ_{t+1} and summing up till K gives us an upper bound for the average gap function for LPD. We have

$$\begin{aligned} \sum_{t=1}^{K}\gamma_{t+1}Q(z_{t+1},z) \le{}& \sum_{t=1}^{K}\gamma_{t+1}[(\tfrac{1}{2\eta_t} - \tfrac{\mu_f}{2})\|x-x_t\|^2 - \tfrac{1}{2\eta_t}\|x-x_{t+1}\|^2 + \tfrac{1}{2\tau_t}\|y-y_t\|^2 - (\tfrac{1}{2\tau_t} + \tfrac{\mu_g}{2})\|y-y_{t+1}\|^2] \\ & - \sum_{t=1}^{K}\gamma_{t+1}[\langle A(x_{t+1}-x_t), y_{t+1}-y\rangle + \theta_{t-1}\langle A(x_t - x_{t-1}), y_t - y\rangle] \\ & - \sum_{t=1}^{K}\gamma_{t+1}[\tfrac{1}{2\tau_t}\|y_{t+1}-y_t\|^2 + (\tfrac{1}{2\eta_t} - \tfrac{L_f}{2})\|x_{t+1}-x_t\|^2 - \theta_{t-1}\langle A(x_t - x_{t-1}), y_t - y_{t+1}\rangle]. \end{aligned} \tag{27}$$

To simplify each summation in (27), let us start with the first one

$$\sum_{t=1}^{K}\gamma_{t+1}[(\tfrac{1}{2\eta_t} - \tfrac{\mu_f}{2})\|x-x_t\|^2 - \tfrac{1}{2\eta_t}\|x-x_{t+1}\|^2 + \tfrac{1}{2\tau_t}\|y-y_t\|^2 - (\tfrac{1}{2\tau_t} + \tfrac{\mu_g}{2})\|y-y_{t+1}\|^2] \tag{28}$$

If we assume that for each $t \ge 2$, we have (21a) and (21b), the above summation (28) is upper bounded by

$$\le \gamma_2(\tfrac{1}{2\eta_1} - \tfrac{\mu_f}{2})\|x-x_1\|^2 - \gamma_{K+1}\tfrac{1}{2\eta_K}\|x-x_{K+1}\|^2 + \gamma_2\tfrac{1}{2\tau_1}\|y-y_1\|^2 - \gamma_{K+1}(\tfrac{1}{2\tau_K} + \tfrac{\mu_g}{2})\|y-y_{K+1}\|^2. \tag{29}$$

For the second summation by assuming (21c) for $t \ge 2$, we have

$$\begin{aligned} &- \sum_{t=1}^{K}\gamma_{t+1}[\langle A(x_{t+1}-x_t), y_{t+1}-y\rangle + \theta_{t-1}\langle A(x_t - x_{t-1}), y_t - y\rangle] \\ &\le \gamma_{t+1}\|A\|\|x_{K+1}-x_K\|\|y_{K+1}-y\|. \end{aligned} \tag{30}$$

For the third summation in (27), by assuming condition (21d), we have

$$\begin{aligned}
&-\sum_{t=1}^{K}\gamma_{t+1}\big[\tfrac{1}{2\tau_t}\|y_{t+1}-y_t\|^2+(\tfrac{1}{2\eta_t}-\tfrac{L_f}{2})\|x_{t+1}-x_t\|^2-\theta_{t-1}\langle\, A(x_t-x_{t-1}),y_t-y_{t+1}\rangle\big]\\
&\qquad\le-\sum_{t=2}^{K}\big[\gamma_{t+1}\tfrac{1}{2\tau_t}\|y_{t+1}-y_t\|^2+\gamma_t(\tfrac{1}{2\eta_t}-\tfrac{L_f}{2})\|x_t-x_{t-1}\|^2\\
&\qquad\qquad-\gamma_{t+1}\theta_{t-1}\|A\|\|x_t-x_{t-1}\|\|y_{t+1}-y_t\|\big]-\gamma_{K+1}(\tfrac{1}{2\eta_K}-\tfrac{L_f}{2})\|x_{K+1}-x_K\|^2\\
&\qquad\le-\gamma_{K+1}(\tfrac{1}{2\eta_K}-\tfrac{L_f}{2})\|x_{K+1}-x_K\|^2
\end{aligned}\tag{31}$$

Therefore, from (29), (30) and (31), one can reestablish (27) as

$$\begin{aligned}
\sum_{t=1}^{K}\gamma_{t+1}Q(z_{t+1},z)&\le\gamma_2(\tfrac{1}{2\eta_1}-\tfrac{\mu_f}{2})\|x-x_1\|^2-\gamma_{K+1}\tfrac{1}{2\eta_K}\|x-x_{K+1}\|^2+\gamma_2\tfrac{1}{2\tau_1}\|y-y_1\|^2\\
&\qquad-\gamma_{K+1}(\tfrac{1}{2\tau_K}+\tfrac{\mu_g}{2})\|y-y_{K+1}\|^2-\gamma_{K+1}(\tfrac{1}{2\eta_K}-\tfrac{L_f}{2})\|x_{K+1}-x_K\|^2\\
&\qquad+\gamma_{K+1}\|A\|\|x_{K+1}-x_K\|\|y_{K+1}-y\|.
\end{aligned}\tag{32}$$

Note that $\sum_{t=1}^{K}\gamma_{t+1}Q(z_{t+1},z)$ can be rewritten since

$$\begin{aligned}
&-(\tfrac{1}{2\tau_K}+\tfrac{\mu_g}{2})\|y-y_{K+1}\|^2-(\tfrac{1}{2\eta_K}-\tfrac{L_f}{2})\|x_{K+1}-x_K\|^2+\|A\|\|x_{K+1}-x_K\|\|y-y_{K+1}\|\\
&\qquad\qquad\le-\left((\tfrac{1}{\tau_K}+\mu_g)-\frac{\|A\|^2}{\frac{1}{\eta_K}-L_f}\right)\tfrac{1}{2}\|y-y_{K+1}\|^2.
\end{aligned}$$

Note the above relation holds since

$$-(\tfrac{1}{2\eta_K}-\tfrac{L_f}{2})\|x_{K+1}-x_K\|^2+\|A\|\|x_{K+1}-x_K\|\|y-y_{K+1}\|\le\frac{\|A\|^2}{\frac{1}{\eta_K}-L_f}\tfrac{1}{2}\|y-y_{K+1}\|^2.$$

Thus

$$\begin{aligned}
\sum_{t=1}^{K}\gamma_{t+1}Q(z_{t+1},z)&\le\tfrac{\gamma_2}{2}(\tfrac{1}{\eta_1}-\mu_f)\|x-x_1\|^2-\tfrac{\gamma_{K+1}}{2\eta_K}\|x-x_{K+1}\|^2+\tfrac{\gamma_2}{2\tau_1}\|y-y_1\|^2\\
&\qquad-\tfrac{\gamma_{K+1}}{2}\left(\mu_g+\tfrac{1}{\tau_K}-\frac{\|A\|^2}{\frac{1}{\eta_K}-L_f}\right)\|y-y_{K+1}\|^2.
\end{aligned}\tag{33}$$

Also, at $z=z^\star$, since the gap function is non-negative, we have

$$\tfrac{\gamma_{K+1}}{2}\left(\mu_g+\tfrac{1}{\tau_K}-\frac{\|A\|^2}{\frac{1}{\eta_K}-L_f}\right)\|y-y_{K+1}\|^2\le\tfrac{\gamma_2}{2}(\tfrac{1}{\eta_1}-\mu_f)\|x-x_1\|^2-\tfrac{\gamma_{K+1}}{2\eta_K}\|x-x_{K+1}\|^2+\tfrac{\gamma_2}{2\tau_1}\|y-y_1\|^2.\tag{34}$$

□

As a consequence of Theorem A.3 and the convexity of Gap function, one can conclude the following

$$\begin{aligned}
\mathrm{Gap}(\bar z_{K+1})\le\frac{1}{\sum_{t=1}^{K}\gamma_{t+1}}\Big[&\tfrac{\gamma_2}{2}(\tfrac{1}{\eta_1}-\mu_f)\|x-x_1\|^2-\tfrac{\gamma_{K+1}}{2\eta_K}\|x-x_{K+1}\|^2+\tfrac{\gamma_2}{2\tau_1}\|y-y_1\|^2\\
&-\tfrac{\gamma_{K+1}}{2}\left(\mu_g+\tfrac{1}{\tau_K}-\frac{\|A\|^2}{\frac{1}{\eta_K}-L_f}\right)\|y-y_{K+1}\|^2\Big].
\end{aligned}\tag{35}$$

Where $\bar z_{K+1}=\frac{\sum_{t=1}^{K}\gamma_{t+1}z_{t+1}}{\sum_{t=1}^{K}\gamma_{t+1}}$.

A.1. Proof of Theorem 3.1

Proof. As one can observe, the mentioned values as step-size policy parameters satisfy the required conditions (21a)-(21d). Additionally, from (22) we know that

$$\text{Gap}(\bar{z}_{K+1}) \leq \tfrac{1}{\sum_{t=1}^{K} \gamma_{t+1}} \big[\gamma_2 \tfrac{1}{2\eta_1} \|x - x_1\|^2 + \gamma_2 \tfrac{1}{2\tau_1} \|y - y_1\|^2\big],$$

By considering mentioned values in (3) for the parameters, the upper bound is

$$\text{Gap}(\bar{z}_{K+1}) \leq \tfrac{1}{\sum_{t=1}^{K} \frac{t+1}{2} + \frac{L_f}{\mu_f}} \big[(1 + \tfrac{L_f}{\mu_f}) \tfrac{\mu_f + L_f}{2} \|x - x_1\|^2 + (1 + \tfrac{L_f}{\mu_f}) \tfrac{4\|A\|^2}{2\mu_f} \|y - y_1\|^2\big].$$

Thus

$$\text{Gap}(\bar{z}_{t+1}) \leq \tfrac{4}{k(k+3+\frac{4L_f}{\mu_f})} \big[(1 + \tfrac{L_f}{\mu_f})[\tfrac{\mu_f + L_f}{2} \|x - x_1\|^2 + \tfrac{4\|A\|^2}{2\mu_f} \|y - y_1\|^2]\big].$$

□

A.2. Proof of Theorem 3.2

Proof. First, note that the chosen values in (6) for the algorithm parameters hold the conditions (21b)-(21d) and (5). From the upper bound defined for the weighted gap function in (27) we know

$$\begin{aligned} \sum_{t=1}^{K} \gamma_{t+1} Q(z_{t+1}, z) \leq & \sum_{t=1}^{K} \gamma_{t+1} [(\tfrac{1}{2\eta_t} - \tfrac{\mu_f}{2}) \|x - x_t\|^2 - \tfrac{1}{2\eta_t} \|x - x_{t+1}\|^2 + \tfrac{1}{2\tau_t} \|y - y_t\|^2 - (\tfrac{1}{2\tau_t} + \tfrac{\mu_g}{2}) \|y - y_{t+1}\|^2] \\ & - \sum_{t=1}^{K} \gamma_{t+1} [\langle A(x_{t+1} - x_t), y_{t+1} - y \rangle + \theta_{t-1} \langle A(x_t - x_{t-1}), y_t - y \rangle] \\ & - \sum_{t=1}^{K} \gamma_{t+1} [\tfrac{1}{2\tau_t} \|y_{t+1} - y_t\|^2 + (\tfrac{1}{2\eta_t} - \tfrac{L_f}{2}) \|x_{t+1} - x_t\|^2 - \theta_{t-1} \langle A(x_t - x_{t-1}), y_t - y_{t+1} \rangle]. \end{aligned} \tag{36}$$

One can rewrite $\sum_{t=1}^{K} \gamma_{t+1} [(\frac{1}{2\eta_t} - \frac{\mu_f}{2}) \|x - x_t\|^2 - \frac{1}{2\eta_t} \|x - x_{t+1}\|^2]$ as following

$$\begin{aligned} & = \gamma_2 \tfrac{\|x - x_1\|^2}{2\eta_1} + \sum_{t=2}^{K} (\tfrac{\gamma_{t+1}}{\eta_t} - \tfrac{\gamma_t}{\gamma_{t-1}}) \tfrac{\|x - x_t\|^2}{2}, \\ \text{From (5)} \; & \leq \tfrac{\gamma_2}{\eta_1} D_X^2 + (K-1) L_f D_X^2. \end{aligned} \tag{37}$$

Using the similar procedure we used in proving (22), and by the fact we showed in (35), Gap function at $\bar{z}_{K+1}$ has the following upper bound

$$\begin{aligned} \text{Gap}(\bar{z}_{K+1}) \leq & \tfrac{1}{\sum_{t=1}^{K} t+1} \big(\tfrac{\gamma_2}{\eta_1} D_X^2 + (K-1) L_f D_X^2 + \tfrac{\gamma_2}{\tau_1} D_Y^2\big) \\ = & \tfrac{1}{\sum_{t=1}^{K} t+1} \big(2 \tfrac{\|A\|^2}{\mu_g} D_X^2 + 2 L_f D_X^2 + (K-1) L_f D_X^2 + \mu_g D_Y^2\big). \end{aligned}$$

Then

$$\begin{aligned} \text{Gap}(\bar{z}_{K+1}, z) \leq & \tfrac{2}{K^2} \left(\tfrac{2\|A\|^2}{\mu_g} D_X^2 + (K+1) L_f D_X^2 + \mu_g D_Y^2\right) \\ = & \tfrac{4 D_X^2 \|A\|^2 / \mu_g + D_Y^2 \mu_g}{K^2} + \tfrac{2(K+1) L_f D_X^2}{K^2}. \end{aligned}$$

□

B. General Analysis of Algorithm 2 (ALPD)

In this part, we focus on the proofs of the statements we mentioned in Algorithm 2. Moreover, we present a new proposition (Proposition B.1) which is crucial in convergence analysis.

B.1. Proof of Proposition 4.5

Proof. The approach we use here is induction. First, observe that $\beta_1 = 1 \in [\frac{1}{2}, 1]$. Now let us assume Proposition 4.5 is true for β_t which means $\frac{t+1}{2} \leq \beta_t \leq t$. Let us first verify the lower bound.

Induction hypothesis ($\beta_t \leq t$): By using step-size policy for β_{t+1} in (13) ($\beta_{t+1} = 1 + \theta_{t+1}\beta_t$), and the fact that $\theta_{t+1} \leq 1$, one can conclude that $\beta_{t+1} \leq t+1$.

Induction hypothesis ($\beta_t \geq \frac{t+1}{2}$): Using the similar assumptions for verifying the upper bound, we have

$$\begin{aligned}\beta_{t+1} &= 1 + \theta_{t+1}\beta_t \\ &\geq 1 + \tfrac{t}{t+1}\tfrac{t+1}{2}\end{aligned}$$

then $\beta_{t+1} \geq 1 + \frac{t}{2} = \frac{t+2}{2}$. Hence we proved that $\beta_{t+1} \in [\frac{t+2}{2}, t+1]$. □

B.2. Statement and proof of Proposition B.1

Proposition B.1 captures the impact of introducing $\{\beta_t\}_{t\geq 1}$ on errors incurred by linearizing f in more detail.

Proposition B.1. *Let $\beta_t \geq 1$ then for all $z \in Z$, we have*

$$\begin{aligned}\beta_t Q(\bar{z}_{t+1}, z) - (\beta_t - 1)Q(\bar{z}_t, z) \leq &\langle \nabla f(\underline{x}_t), x_{t+1} - x\rangle + \tfrac{L_f}{2\beta_t}\|x_{t+1} - x_t\|^2 \\ &+ [g(y_{t+1}) - g(y)] + [\phi(x_{t+1}, y) - \phi(x, y_{t+1})].\end{aligned} \tag{38}$$

Proof. From Algorithm 2, one can say $\bar{x}_{t+1} - \underline{x}_t = \beta_t^{-1}\big(x_{t+1} - x_t\big)$. Using this observation and convexity of f, we have

$$\begin{aligned}\beta_t f(\bar{x}_{t+1}) &\leq \beta_t f(\underline{x}_t) + \beta_t\langle \nabla f(\underline{x}_t), \bar{x}_{t+1} - \underline{x}_t\rangle + \tfrac{\beta_t L_f}{2}\|\bar{x}_{t+1} - \underline{x}_t\|^2 \\ &= \beta_t f(\underline{x}_t) + \beta_t\langle \nabla f(\underline{x}_t), \bar{x}_{t+1} - \underline{x}_t\rangle + \tfrac{L_f}{2\beta_t}\|x_{t+1} - x_t\|^2 \\ &= \beta_t f(\underline{x}_t) + (\beta_t - 1)\langle \nabla f(\underline{x}_t), \bar{x}_t - \underline{x}_t\rangle + \langle \nabla f(\underline{x}_t), x_{t+1} - \underline{x}_t\rangle + \tfrac{L_f}{2\beta_t}\|x_{t+1} - x_t\|^2 \\ &= (\beta_t - 1)\big[f(\underline{x}_t) + \langle \nabla f(\underline{x}_t), \bar{x}_t - \underline{x}_t\rangle\big] + \big[f(\underline{x}_t) + \langle \nabla f(\underline{x}_t), x - \underline{x}_t\rangle\big] \\ &\quad + \langle \nabla f(\underline{x}_t), x_{t+1} - x\rangle + \tfrac{L_f}{2\beta_t}\|x_{t+1} - x_t\|^2 \\ &\leq (\beta_t - 1)f(\bar{x}_t) + f(x) + \langle \nabla f(\underline{x}_t), x_{t+1} - x\rangle + \tfrac{L_f}{2\beta_t}\|x_{t+1} - x_t\|^2.\end{aligned}$$

Moreover, by convexity of g and definition of $\bar{y}_{t+1}$, we have

$$\begin{aligned}\beta_t g(\bar{y}_{t+1}) - \beta_t g(y) \leq &(\beta_t - 1)g(\bar{y}_t) + g(y_{t+1}) - \beta_t g(y) \\ &= (\beta_t - 1)[g(\bar{y}_t) - \beta_t g(y)] + g(y_{t+1}) - g(y).\end{aligned} \tag{39}$$

Also, for the coupling function, we have

$$\begin{aligned}&\beta_t[\phi(\bar{x}_{t+1}, y) - \phi(x, \bar{y}_{t+1})] - (\beta_t - 1)[\phi(\bar{x}_t, y) - \phi(x, \bar{y}_t)] \\ &= [\beta_t\phi(\bar{x}_{t+1}, y) - (\beta_t - 1)\phi(\bar{x}_t, y)] + [-\beta_t\phi(x, \bar{y}_{t+1}) + (\beta_t - 1)\phi(x, \bar{y}_t)].\end{aligned} \tag{40}$$

For the first piece in the right hand side of the above inequality, we have

$$\begin{aligned}\beta_t\phi(\bar{x}_{t+1}, y) - (\beta_t - 1)\phi(\bar{x}_t, y) &\leq \phi(\beta_t\bar{x}_{t+1} - (\beta_t - 1)\bar{x}_t, y) \\ &= \phi(x_{t+1}, y).\end{aligned}$$

Note that the above inequality is based on definition of x_{t+1} in Algorithm 2 and convexity of $\phi(\cdot, y)$ for all $y \in Y$. Similarly the second piece of (40) can be upper bounded as follows

$$-\beta_t\phi(x, \bar{y}_{t+1}) + (\beta_t - 1)\phi(x, \bar{y}_t) \leq -\phi(x, y_{t+1}),$$

From the definition of primal-dual gap function and the mentioned upper bounds for each terms, one can construct the following inequality

$$\begin{aligned}\beta_t Q(\bar{z}_{t+1}, z) - (\beta_t - 1)Q(\bar{z}_t, z) \leq &\langle \nabla f(\underline{x}_t), x_{t+1} - x\rangle + \tfrac{L_f}{2\beta_t}\|x_{t+1} - x_t\|^2 \\ &+ [g(y_{t+1}) - g(y)] + \phi(x_{t+1}, y) - \phi(x, y_{t+1}).\end{aligned}$$

□

B.3. Proof of Lemma 4.1

Proof. Using the optimality of y_{t+1} and from LemmaA.1 for $\delta = 0$, we have

$$\langle \nabla g(y_t), y_{t+1} - y\rangle \leq \tfrac{1}{2\tau_t}\left[\|y - y_t\|^2 - \|y - y_{t+1}\|^2 - \|y_t - y_{t+1}\|^2\right] - \langle v_t, y - y_{t+1}\rangle. \tag{41}$$

Note that

$$\langle \nabla g(y_t), y_{t+1} - y\rangle = \langle \nabla g(y_t), y_{t+1} - y_t\rangle + \langle \nabla g(y_t), y_t - y\rangle.$$

From strong-convexity and smoothness of g, we know that

$$\langle \nabla g(y_t), y_{t+1} - y_t\rangle \geq g(y_{t+1}) - g(y_t) - \tfrac{L_g}{2}\|y_t - y_{t+1}\|^2,$$

and

$$\langle \nabla g(y_t), y_t - y\rangle \geq g(y_t) - g(y) + \tfrac{\mu_g}{2}\|y - y_t\|^2.$$

Adding these two inequities and with (41), we can obtain an upper bound on $g(y_{t+1}) - g(y)$

$$g(y_{t+1}) - g(y) \leq \tfrac{1}{2}\big(\tfrac{1}{\tau_t} - \mu_g\big)\|y - y_t\|^2 - \tfrac{1}{2}\big(\tfrac{1}{\tau_t} - L_g\big)\|y_t - y_{t+1}\|^2 - \tfrac{1}{2\tau_t}\|y - y_{t+1}\|^2 - \langle v_t, y - y_{t+1}\rangle. \tag{42}$$

Also, from the optimality of x_{t+1}, we have

$$\langle \nabla f(\underline{x}_t), x_{t+1} - x\rangle \leq \tfrac{1}{2\eta_t}[\|x - x_t\|^2 - \|x_{t+1} - x_t\|^2 - \|x_{t+1} - x\|^2] - \langle \nabla_x \phi(x_t, y_{t+1}), x_{t+1} - x\rangle. \tag{43}$$

From Proposition B.1, (42)and (43), one can reconstruct the following upper bound for the gap function at one single iteration

$$\begin{aligned}\beta_t Q(\bar{z}_{t+1}, z) - (\beta_t - 1)Q(\bar{z}_t, z) \leq{}& \tfrac{1}{2\eta_t}\|x - x_t\|^2 - \tfrac{1}{2\eta_t}\|x - x_{t+1}\|^2 - \big(\tfrac{1}{2\eta_t} - \tfrac{L_f}{2\beta_t}\big)\|x_t - x_{t+1}\|^2\\ &- \tfrac{1}{2\tau_t}\|y - y_{t+1}\|^2 - \big(\tfrac{1}{2\tau_t} - \tfrac{L_g}{2}\big)\|y_t - y_{t+1}\|^2 - \langle v_t, y - y_{t+1}\rangle\\ &- \langle \nabla_x \phi(x_t, y_{t+1}), x_{t+1} - x\rangle + \phi(x_{t+1}, y) - \phi(x, y_{t+1}) + \big(\tfrac{1}{2\tau_t} - \tfrac{\mu_g}{2}\big)\|y - y_t\|^2.\end{aligned}$$

Now, let us add and subtract $\phi(x_{t+1}, y_{t+1})$ to the right hand side of above inequality, then

$$\begin{aligned}\beta_t Q(\bar{z}_{t+1}, z) - (\beta_t - 1)Q(\bar{z}_t, z) \leq{}& \tfrac{1}{2\eta_t}\|x - x_t\|^2 - \tfrac{1}{2\eta_t}\|x - x_{t+1}\|^2 - \big(\tfrac{1}{2\eta_t} - \tfrac{L_f}{2\beta_t}\big)\|x_t - x_{t+1}\|^2\\ &- \tfrac{1}{2\tau_t}\|y - y_{t+1}\|^2 - \big(\tfrac{1}{2\tau_t} - \tfrac{L_g}{2}\big)\|y_t - y_{t+1}\|^2 - \langle v_t, y - y_{t+1}\rangle\\ &- \langle \nabla_x \phi(x_t, y_{t+1}), x_{t+1} - x\rangle + \phi(x_{t+1}, y) - \phi(x_{t+1}, y_{t+1})\\ &+ \phi(x_{t+1}, y_{t+1}) - \phi(x, y_{t+1}) + \big(\tfrac{1}{2\tau_t} - \tfrac{\mu_g}{2}\big)\|y - y_t\|^2.\end{aligned}$$

By the L_{xx} of $\phi(\cdot, y)$ for all $y \in Y$ of ϕ one can say that

$$\begin{aligned}&- \langle \nabla_x \phi(x_t, y_{t+1}), x_{t+1} - x\rangle + \phi(x_{t+1}, y_{t+1}) - \phi(x, y_{t+1})\\ &\leq \tfrac{L_{xx}}{2}\|x_t - x_{t+1}\|^2.\end{aligned} \tag{44}$$

Based on these last two inequalities, one can immediately conclude (8). □

B.4. Proof of Lemma 4.2

Proof. From Lemma 4.1, and concavity of ϕ in x, we have

$$\phi(x_{t+1}, y) - \phi(x_{t+1}, y_{t+1}) \leq \langle \nabla_y \phi(x_{t+1}, y_{t+1}), y - y_{t+1}\rangle.$$

Therefore, by the definition of v_t in Algorithm 2, (8) and above inequality, we have

$$\begin{aligned}\beta_t Q(\bar{z}_{t+1}, z) - (\beta_t - 1)Q(\bar{z}_t, z) \leq{}& \tfrac{1}{2\eta_t}\|x - x_t\|^2 - \tfrac{1}{2\eta_t}\|x - x_{t+1}\|^2 + \big(\tfrac{1}{2\tau_t} - \tfrac{\mu_g}{2}\big)\|y - y_t\|^2 - \tfrac{1}{2\tau_t}\|y - y_{t+1}\|^2\\ &- \big(\tfrac{1}{2\eta_t} - \tfrac{L_f}{2\beta_t} - \tfrac{L_{xx}}{2}\big)\|x_t - x_{t+1}\|^2 - \big(\tfrac{1}{2\tau_t} - \tfrac{L_g}{2}\big)\|y_t - y_{t+1}\|^2\\ &+ \langle \nabla_y \phi(x_{t+1}, y_{t+1}) - \nabla_y \phi(x_t, y_t), y - y_{t+1}\rangle\\ &- \theta_t\langle \nabla_y \phi(x_t, y_t) - \nabla_y \phi(x_{t-1}, y_{t-1}), y - y_{t+1}\rangle.\end{aligned} \tag{45}$$

Notice that

$$\begin{aligned}
&- \theta_t\langle \nabla_y\phi(x_t, y_t) - \nabla_y\phi(x_{t-1}, y_{t-1}), y - y_{t+1}\rangle \\
&= -\theta_t\langle \nabla_y\phi(x_t, y_t) - \nabla_y\phi(x_{t-1}, y_{t-1}), y - y_t\rangle - \theta_t\langle \nabla_y\phi(x_t, y_t) - \nabla_y\phi(x_{t-1}, y_{t-1}), y_t - y_{t+1}\rangle.
\end{aligned}$$

Hence, the previous inequality can be written as

$$\begin{aligned}
\beta_t Q(\bar{z}_{t+1}, z) - (\beta_t - 1)Q(\bar{z}_t, z) \le\ & \tfrac{1}{2\eta_t}\|x - x_t\|^2 - \tfrac{1}{2\eta_t}\|x - x_{t+1}\|^2 - \big(\tfrac{1}{2\eta_t} - \tfrac{L_f}{2\beta_t} - \tfrac{L_{xx}}{2}\big)\|x_t - x_{t+1}\|^2 \\
&+ \big(\tfrac{1}{2\tau_t} - \tfrac{\mu_g}{2}\big)\|y - y_t\|^2 - \tfrac{1}{2\tau_t}\|y_{-}y_{t+1}\|^2 - \big(\tfrac{1}{2\tau_t} - \tfrac{L_g}{2}\big)\tfrac{1}{2}\|y_t - y_{t+1}\|^2 \\
&+ \langle \nabla_y\phi(x_{t+1}, y_{t+1}) - \nabla_y\phi(x_t, y_t), y - y_{t+1}\rangle \\
&- \theta_t\langle \nabla_y\phi(x_t, y_t) - \nabla_y\phi(x_{t-1}, y_{t-1}), y - y_t\rangle \\
&- \theta_t\langle \nabla_y\phi(x_t, y_t) - \nabla_y\phi(x_{t-1}, y_{t-1}), y_t - y_{t+1}\rangle.
\end{aligned}$$

Now, by multiplying both sides by γ_t and letting $\theta_t = \frac{\gamma_{t-1}}{\gamma_t}, t \ge 2$, we have

$$\begin{aligned}
\beta_t\gamma_t Q(\bar{z}_{t+1}, z) - (\beta_t - 1)\gamma_t Q(\bar{z}_t, z) \le\ & \tfrac{\gamma_t}{2\eta_t}\|x - x_t\|^2 - \tfrac{\gamma_t}{2\eta_t}\|x - x_{t+1}\|^2 - \gamma_t\big(\tfrac{1}{2\eta_t} - \tfrac{L_f}{2\beta_t} - \tfrac{L_{xx}}{2}\big)\|x_t - x_{t+1}\|^2 \\
&+ \gamma_t\big(\tfrac{1}{2\tau_t} - \tfrac{\mu_g}{2}\big)\|y - y_t\|^2 - \tfrac{\gamma_t}{2\tau_t}\|y_{-}y_{t+1}\|^2 - \gamma_t\big(\tfrac{1}{2\tau_t} - \tfrac{L_g}{2}\big)\tfrac{1}{2}\|y_t - y_{t+1}\|^2 \\
&+ \gamma_t\langle \nabla_y\phi(x_{t+1}, y_{t+1}) - \nabla_y\phi(x_t, y_t), y - y_{t+1}\rangle \\
&- \gamma_{t-1}\langle \nabla_y\phi(x_t, y_t) - \nabla_y\phi(x_{t-1}, y_{t-1}), y - y_t\rangle \\
&- \gamma_{t-1}\langle \nabla_y\phi(x_t, y_t) - \nabla_y\phi(x_{t-1}, y_{t-1}), y_t - y_{t+1}\rangle.
\end{aligned} \tag{46}$$

The last inner product can be written as follows

$$\begin{aligned}
&- \gamma_{t-1}\langle \nabla_y\phi(x_t, y_t) - \nabla_y\phi(x_{t-1}, y_{t-1}), y_t - y_{t+1}\rangle \\
&= -\gamma_{t-1}\langle \nabla_y\phi(x_t, y_t) - \nabla_y\phi(x_{t-1}, y_t), y_t - y_{t+1}\rangle - \gamma_{t-1}\langle \nabla_y\phi(x_{t-1}, y_t) - \nabla_y\phi(x_{t-1}, y_{t-1}), y_t - y_{t+1}\rangle \\
&\le \gamma_{t-1}\|\nabla_y\phi(x_t, y_t) - \nabla_y\phi(x_{t-1}, y_t)\|\|y_t - y_{t+1}\| + \gamma_{t-1}\|\nabla_y\phi(x_{t-1}, y_t) - \nabla_y\phi(x_{t-1}, y_{t-1})\|\|y_t - y_{t+1}\| \\
&\le L_{xy}\gamma_{t-1}\|x_t - x_{t-1}\|\|y_t - y_{t+1}\| + L_{yy}\gamma_{t-1}\|y_t - y_{t-1}\|\|y_t - y_{t+1}\|.
\end{aligned}$$

Since $0 \le \theta_t \le \frac{\tau_{t-1}}{\tau_t}$ for each of norm multiplication, we have

$$\begin{aligned}
&L_{xy}\gamma_{t-1}\|x_t - x_{t-1}\|\|y_t - y_{t+1}\| \\
&\le \tfrac{4L_{xy}^2\gamma_{t-1}^2\tau_t}{2\gamma_t}\|x_t - x_{t-1}\|^2 + \tfrac{\gamma_t}{8\tau_t}\|y_t - y_{t+1}\|^2 \\
&\le \tfrac{4L_{xy}^2\gamma_{t-1}\tau_{t-1}}{2}\|x_t - x_{t-1}\|^2 + \tfrac{\gamma_t}{8\tau_t}\|y_t - y_{t+1}\|^2.
\end{aligned}$$

Similarly

$$\begin{aligned}
&L_{yy}\gamma_{t-1}\|y_t - y_{t-1}\|\|y_t - y_{t+1}\| \\
&\le \tfrac{4L_{yy}^2\gamma_{t-1}^2\tau_t}{2\gamma_t}\|y_t - y_{t-1}\|^2 + \tfrac{\gamma_t}{8\tau_t}\|y_t - y_{t+1}\|^2 \\
&\le \tfrac{4L_{yy}^2\gamma_{t-1}\tau_{t-1}}{2}\|y_t - y_{t-1}\|^2 + \tfrac{\gamma_t}{8\tau_t}\|y_t - y_{t+1}\|^2.
\end{aligned}$$

Using these results and combining it with (46) and $\beta_{t+1} - 1 = \beta_t\theta_{t+1}$, we have

$$\begin{aligned}
\beta_t\gamma_t Q(\bar{z}_{t+1}, z) - (\beta_t - 1)\gamma_t Q(\bar{z}_t, z) \le\ & \tfrac{\gamma_t}{2\eta_t}\|x - x_t\|^2 - \tfrac{\gamma_t}{2\eta_t}\|x - x_{t+1}\|^2 + \gamma_t\big(\tfrac{1}{2\tau_t} - \tfrac{\mu_g}{2}\big)\|y - y_t\|^2 - \tfrac{\gamma_t}{2\tau_t}\|y - y_{t+1}\|^2 \\
&+ \gamma_t\langle \nabla_y\phi(x_{t+1}, y_{t+1}) - \nabla_y\phi(x_t, y_t), y - y_{t+1}\rangle \\
&- \gamma_{t-1}\langle \nabla_y\phi(x_t, y_t) - \nabla_y\phi(x_{t-1}, y_{t-1}), y - y_t\rangle \\
&- \gamma_t\big(\tfrac{1}{2\eta_t} - \tfrac{L_f}{2\beta_t} - \tfrac{L_{xx}}{2}\big)\|x_t - x_{t+1}\|^2 + \tfrac{4L_{xy}^2\gamma_{t-1}\tau_{t-1}}{2}\|x_t - x_{t-1}\|^2 \\
&- \gamma_t\big(\tfrac{1}{4\tau_t} - \tfrac{L_g}{2}\big)\|y_t - y_{t+1}\|^2 + \tfrac{4L_{yy}^2\gamma_{t-1}\tau_{t-1}}{2}\|y_t - y_{t-1}\|^2.
\end{aligned} \tag{47}$$

Applying (47) inductively and letting $x_0 = x_1, \beta_1 = 1$, we conclude that

$$\begin{aligned}
\beta_K\gamma_K Q(\bar z_{K+1}, z) \le{}& B_K(z, z_{[K]}) + \gamma_K\langle \nabla_y\phi(x_{K+1}, y_{K+1}) - \nabla_y\phi(x_K, y_K), y - y_{K+1}\rangle\\
&- \gamma_K\big(\tfrac{1}{2\eta_K} - \tfrac{L_f}{2\beta_K} - \tfrac{L_{xx}}{2}\big)\|x_K - x_{K+1}\|^2 - \sum_{t=1}^{K-1}\gamma_t\big(\tfrac{1}{2\eta_t} - \tfrac{L_f}{2\beta_t} - \tfrac{L_{xx}}{2} - 2L_{xy}^2\tau_t\big)\|x_t - x_{t+1}\|^2\\
&- \gamma_K\big(\tfrac{1}{4\tau_K} - \tfrac{L_g}{2}\big)\|y_K - y_{K+1}\|^2 - \sum_{t=1}^{K-1}\gamma_t\big(\tfrac{1}{4\tau_t} - \tfrac{L_g}{2} - 2L_{yy}^2\tau_t\big)\|y_t - y_{t+1}\|^2.
\end{aligned}$$

By assuming conditions in (9), one can observe that Lemma 4.2 holds. □

B.5. Proof of Theorem 4.3

Proof.

$$\begin{aligned}
B_K(z, z_{[K]}) ={}& \tfrac{\gamma_1}{\eta_1}\tfrac{1}{2}\|x - x_1\|^2 - \sum_{t=1}^{K-1}\left(\tfrac{\gamma_t}{\eta_t} - \tfrac{\gamma_{t+1}}{\eta_{t+1}}\right)\tfrac{1}{2}\|x - x_{t+1}\|^2 - \tfrac{\gamma_K}{2\eta_K}\|x - x_{K+1}\|^2\\
&+ \tfrac{\gamma_1}{2}(\tfrac{1}{\tau_1} - \mu_g)\|y - y_1\|^2 - \tfrac{\gamma_K}{2\tau_K}\|y - y_{K+1}\|^2 - \sum_{t=1}^{K-1}\left(\tfrac{\gamma_t}{\tau_t} - \gamma_{t+1}(\tfrac{1}{\tau_{t+1}} - \mu_g)\right)\tfrac{1}{2}\|y - y_{t+1}\|^2\\
\le{}& \big(\tfrac{\gamma_1}{\eta_1} + KL_{xx}\big)D_X^2 + \tfrac{\gamma_1}{\tau_1}D_Y^2 - \tfrac{\gamma_K}{2\tau_K}\|y - y_{K+1}\|^2,
\end{aligned}$$

where the second last inequality stems from the new condition (11) and the assumption that $\frac{\gamma_t}{\eta_t} \le \frac{\gamma_{t-1}}{\eta_{t-1}} + \frac{L_{xx}}{2}$.

Moreover, $\gamma_K\langle \nabla_y\phi(x_{K+1}, y_{K+1}) - \nabla_y\phi(x_K, y_K), y - y_{K+1}\rangle$ can be bounded as follows

$$\begin{aligned}
&\gamma_K\langle \nabla_y\phi(x_{K+1}, y_{K+1}) - \nabla_y\phi(x_K, y_K), y - y_{K+1}\rangle\\
&= \gamma_K\langle \nabla_y\phi(x_{K+1}, y_{K+1}) - \nabla_y\phi(x_{K+1}, y_K), y - y_{K+1}\rangle + \gamma_K\langle \nabla_y\phi(x_{t+1}, y_t) - \nabla_y\phi(x_K, y_K), y - y_{K+1}\rangle\\
&\le \gamma_K L_{yy}\|y_K - y_{K+1}\|\|y - y_{K+1}\| + \gamma_K L_{xy}\|x_K - x_{K+1}\|\|y - y_{K+1}\|\\
&\le \tfrac{2L_{yy}^2\gamma_K^2\tau_K}{2\gamma_K}\|y_K - y_{K+1}\|^2 + \tfrac{\gamma_K}{4\tau_K}\|y - y_{K+1}\|^2 + \tfrac{2L_{xy}^2\gamma_K^2\tau_K}{2\gamma_K}\|x_K - x_{K+1}\|^2 + \tfrac{\gamma_K}{4\tau_K}\|y - y_{K+1}\|^2\\
&\le \tfrac{2L_{yy}^2\gamma_K\tau_K}{2}\|y_K - y_{K+1}\|^2 + \tfrac{\gamma_K}{4\tau_K}\|y - y_{K+1}\|^2 + \tfrac{2L_{xy}^2\gamma_K\tau_K}{2}\|x_K - x_{K+1}\|^2 + \tfrac{\gamma_K}{4\tau_K}\|y - y_{K+1}\|^2.
\end{aligned}$$

Then from Lemma 4.2, we have

$$\begin{aligned}
\beta_K\gamma_K Q(\bar z_{K+1}, z) \le{}& B_K(z, z_{[K]}) + \gamma_K\langle \nabla_y\phi(x_{K+1}, y_{K+1}) - \nabla_y\phi(x_K, y_K), y - y_{K+1}\rangle\\
&- \gamma_K\left(\tfrac{1}{2\eta_K} - \tfrac{L_f}{2\beta_K} - \tfrac{L_{xx}}{2}\right)\|x_{K+1} - x_K\|^2 - \gamma_K\left(\tfrac{1}{4\tau_K} - \tfrac{L_g}{2}\right)\|y_K - y_{K+1}\|^2\\
\le{}& \big(\tfrac{\gamma_1}{\eta_1} + tL_{xx}\big)D_X^2 + \tfrac{\gamma_1}{\tau_1}D_Y^2 - \tfrac{\gamma_K}{2\tau_K}\|y - y_{K+1}\|^2 + \tfrac{\gamma_K}{2\tau_K}\|y - y_{K+1}\|^2\\
&- \gamma_K\left(\tfrac{1}{2\eta_K} - \tfrac{L_f}{2\beta_K} - \tfrac{L_{xx}}{2} - L_{xy}^2\tau_K\right)\|x_{K+1} - x_K\|^2 - \gamma_K\big(\tfrac{1}{4\tau_K} - \tfrac{L_g}{2} - L_{yy}^2\tau_K\big)\|y_K - y_{K+1}\|^2.
\end{aligned}$$

From the conditions of Lemma 4.2, we have

$$\beta_K\gamma_K Q(\bar z_{K+1}, z) \le \big(\tfrac{\gamma_1}{\eta_1} + KL_{xx}\big)D_X^2 + \tfrac{\gamma_1}{\tau_1}D_Y^2. \tag{48}$$

Dividing both sides by $\beta_K\gamma_K$ will give us (12). □

C. General Analysis of Algorithm 3 (Inexact ALPD)

We provide this section to highlight the similarities and important differences between ALPD and inexact ALPD algorithms in a mathematical setting. Lemma C.1 shows how the dependence on L_{xx} is alleviated in this approach.

Lemma C.1. *let $\bar z_{t+1} = (\bar x_{t+1}, \bar y_{t+1})$ and if*

$$\begin{aligned}
\beta_t Q(\bar z_{t+1}, z) - (\beta_t - 1)Q(\bar z_t, z) \le{}& \tfrac{1}{2\eta_t}\|x - x_t\|^2 - \tfrac{1}{2\eta_t}\|x - x_{t+1}\|^2 - \big(\tfrac{1}{2\eta_t} - \tfrac{L_f}{2\beta_t}\big)\|x_t - x_{t+1}\|^2\\
&+ \big(\tfrac{1}{2\tau_t} - \tfrac{\mu_g}{2}\big)\|y - y_t\|^2 - \tfrac{1}{2\tau_t}\|y_- y_{t+1}\|^2 - \big(\tfrac{1}{2\tau_t}\tfrac{L_g}{2}\big)\tfrac{1}{2}\|y_t - y_{t+1}\|^2\\
&- \langle v_t, y - y_{t+1}\rangle + \phi(x_{t+1}, y) - \phi(x_{t+1}, y_{t+1}) + \delta_t + \sqrt{2\tfrac{1}{\eta_t}\delta_t}\|x_{t+1} - x\|^2.
\end{aligned} \tag{49}$$

where δ_t denotes to using a δ_t-approximate inexact method in primal.

Proof. The approach we use to prove Lemma C.1 is similar to one we used in Lemma 4.1. The only difference is rooted using the inexact method to find an δ_t-approximate solution for primal which is mentioned below

From the optimality of x_{t+1} using Lemma A.1, we have the following

$$\begin{aligned}\langle \nabla f(\underline{x}_t), x_{t+1} - x\rangle \leq \tfrac{1}{2\eta_t}\|x - x_t\|^2 - \tfrac{1}{2\eta_t}\|x_{t+1} - x_t\|^2 - \tfrac{1}{2\eta_t}\|x_{t+1} - x\|^2 \\ - \phi(x_{t+1}, y_{t+1}) + \phi(x, y_{t+1}) + \delta_t + \sqrt{2\tfrac{1}{\eta_t}\delta_t}\|x_{t+1} - x\|^2.\end{aligned} \tag{50}$$

Above inequality leads to the following change in (44) such that instead of using linear approximation of ϕ in x, we use the exact coupling function. Particularly, (44) changes as

$$-\phi(x_{t+1}, y_{t+1}) + \phi(x, y_{t+1}) + \phi(x_{t+1}, y_{t+1}) - \phi(x, y_{t+1}) = 0. \tag{51}$$

Observe that unlike the case in (44), we do not have any dependence on L_{xx}. □

Lemma C.2. *Suppose these conditions hold*

$$\begin{aligned}&\beta_1 = 1, \quad \beta_{t+1} - 1 = \beta_t\theta_{t+1}, \\ &0 \leq \theta_t \leq \tfrac{\tau_{t-1}}{\tau_t} \quad \tfrac{\gamma_t}{\eta_t} \leq \tfrac{\gamma_{t-1}}{\eta_{t-1}}, \\ &\gamma_1 = 1, \quad \theta_t = \tfrac{\gamma_{t-1}}{\gamma_t}, \quad \tfrac{1}{2\eta_t} - \tfrac{L_f}{2\beta_t} - 2L_{xy}^2\tau_t \geq 0, \\ &\tfrac{1}{4\tau_t} - \tfrac{L_g}{2} - 2L_{yy}^2\tau_t \geq 0.\end{aligned} \tag{52}$$

Then, the following inequality holds

$$\begin{aligned}\beta_K\gamma_K Q(\bar{z}_{K+1}, z) \leq B_K(z, z_{[K]}) + \gamma_K\langle \nabla_y\phi(x_{K+1}, y_{K+1}) - \nabla_y\phi(x_K, y_K), y - y_{K+1}\rangle \\ + \sum_{t=1}^{K}\gamma_t\sqrt{4\tfrac{1}{\eta_t}\delta_t}D_X^2 - \gamma_K\left(\tfrac{1}{2\eta_K} - \tfrac{L_f}{2\beta_K}\right)\|x_{K+1} - x_K\|^2 \\ - \gamma_K\left(\tfrac{1}{4\tau_K} - \tfrac{L_g}{2}\right)\|y_t - y_{K+1}\|^2 + \sum_{t=1}^{K}\gamma_t\delta_t,\end{aligned} \tag{53}$$

where $B_K(z, z_{[K]})$ is the following

$$B_K(z, z_{[K]}) = \sum_{t=1}^{K}\{\tfrac{\gamma_t}{2\eta_t}[\|x - x_t\|^2 - \|x - x_{t+1}\|^2] + \gamma_t\big(\tfrac{1}{2\tau_t} - \tfrac{\mu_g}{2}\big)\|y - y_t\|^2 - \tfrac{\gamma_t}{2\tau_t}\|y - y_{t+1}\|^2\}.$$

Proof. The line of proof we follow in this lemma is the same as we used in proving Lemma 4.2. The only difference in this case is having additional terms in the upper bound which are caused by using a δ_t-approximate solution in x. These additional terms translate into $\sqrt{4\frac{1}{\eta_t}\delta_t}D_X^2$ and $\sum_{t=1}^{K}\gamma_t\delta_t$. □

C.1. Proof of Theorem 5.1

Proof. As we mentioned earlier, the upper bound for the Gap function can be obtained similar to the Section 4.1.1 where L_{xx} is zero. Moreover, the errors caused by minimization step in (15) are added as $\frac{\sum_{t=1}^{K}\gamma_t\delta_t}{\beta_K\gamma_K} + \frac{\sum_{t=1}^{K}\gamma_t\sqrt{4\frac{1}{\eta_t}\delta_t}D_X^2}{\beta_K\gamma_K}$ □

C.2. Proof of Proposition 5.2

Proof. Suppose $\delta_t = \frac{1}{t^c}$ and t is a linearly increasing sequence. Moreover step-size policy in 13 implies sequences γ_t and $\frac{1}{\eta_t}$ are linearly increasing and decreasing respectively. Since $c = 3.5$, two summations $\sum_{t=1}^{K}\gamma_t\delta_t$ and $\sum_{t=1}^{K}\gamma_t\sqrt{\delta_t/\eta_t}$ become in order of $\mathcal{O}(1)$ and bounded by a constant. □

D. Detailed process of problem generation in Section 6

D.1. Process of problem generation in Subsection 6.1

We take the primal objective function $f(x)$ as a quadratic function of the form below

$$f(x) = \tfrac{1}{2}x^\top Qx + c^\top x, \tag{54}$$

where $Q \in \mathbb{R}^{n\times n}$ is a positive semidefinite matrix and $c \in \mathbb{R}^n$ is a random vector with elements drawn from the standard normal distribution. We set $Q = \Lambda^\top D\Lambda$ where $\Lambda \in \mathbb{R}^{n\times n}$ is a random orthonormal matrix and $D \in \mathbb{R}^{n\times n}_+$ is a diagonal matrix whose elements are drawn from a uniform distribution between 0 and 200. To generate the orthonormal matrix Λ, first, we generate a random matrix $\bar{\Lambda}$ whose elements are drawn from the standard normal distribution. Then, we use the MATLAB function orth($\bar{\Lambda}$) to return an orthonormal basis for the range of $\bar{\Lambda}$. For generating the constraint set, we sample the elements of $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$ from a uniform distribution between 0 and 1. In this paper, we take $n = m = 100$ for each problem instance.

For the quadratic constraints, we generate randomized positive semidefinite matrices $A_j,\ j \in [m]$ in a similar fashion as matrix Q. Also, $d_j, j \in [m]$ are uniformly generated in $[0, 1]$. We keep d_j's positive to maintain the feasibility of quadratic constraints (0 is always a feasible solution). For this case, we set $m = 10$

D.2. Process of problem generation in Subsection 6.3

The strongly-convex concave SPP is defined as below

$$\mathcal{L}(x, y) := \min_{x\in X}\max_{y\in Y}\{f(x) + \langle y, Ax - b\rangle\}. \tag{55}$$

Where the primal objective function $f(x)$ is defined as (54) and we generate data for this problem similar to the previous section.

E. Comparison of ALPD and LPD on penalty problems with different norms

In this section, we compare the performance of penalty problems where the norms are not Euclidean anymore. The instances are created similarly to Section D. Figures 3 and 4 show the performances of both versions of ALPD and LPD in terms of gap function for the problem (17) when $q = \infty, p = 1$ and $q = 1, p = \infty$ respectively. To make a better comparison, we set L_f to a sufficiently large number ($L_f \approx 200$) and plot the last 50 iterates of algorithms. Similar to the penalty problem with the Euclidean norm, ALPD has a better performance.

F. Comparing step-size policy for two LPD algorithms

Figure 5 compares the convergence rate of the Gap function between those two step-size policies for 10 i.i.d runs with 200 iterations in each run. Note that the value of L_f is controlled so that 200 iterations of LPD for each problem instance give a satisfactory convergence result. As one can see, our step-size policy has an advantage in terms of having faster convergence.

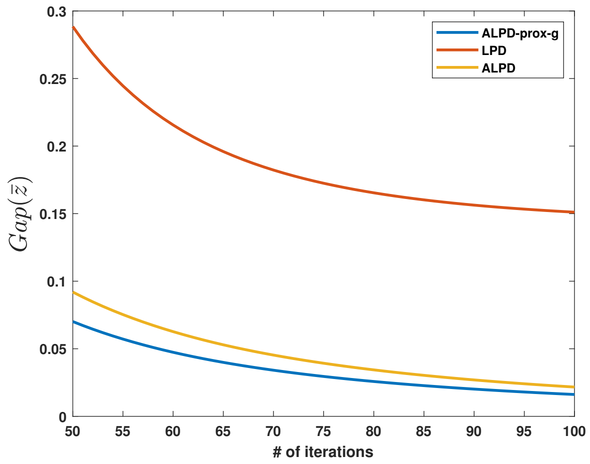

Figure 3: Comparison of the methods in terms of Gap function for 10 i.i.d. replications with 100 iterations in each replication for l^{∞}-norm penalty problem.

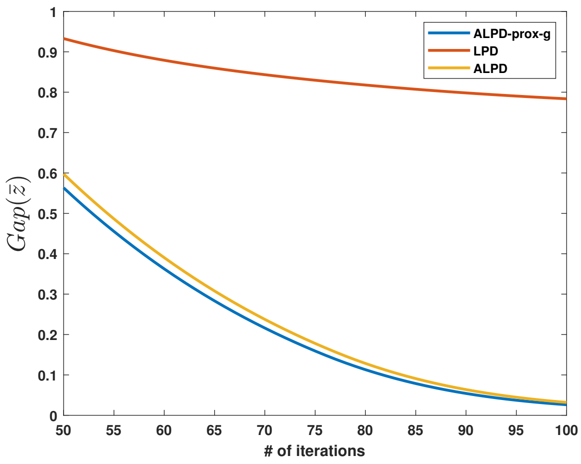

Figure 4: Comparison of the methods in terms of Gap function for 10 i.i.d replications with 100 iterations in each replication for l^{1}-norm penalty problem.

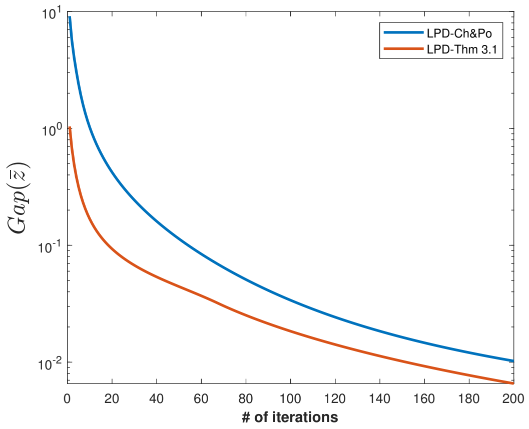

Figure 5: Comparison between the step-size policies of (3) (LPD-Thm 3.1) and Chambolle & Pock (2016) (LPD-Ch&Po) for 10 i.i.d. problem instances. Both policies start from the same initial point. Note that LPD only records $\{\bar{x}_{t+1}\}_{t\geq 1}$.