# A Primal-Dual Algorithm for General Convex-Concave Saddle Point Problems

Erfan Yazdandoost Hamedani[*],   Necdet Serhat Aybat[*]

July 24, 2018

## Abstract

In this paper we propose a primal-dual algorithm with a novel momentum term using the partial gradients of the coupling function that can be viewed as a generalization of the method proposed by Chambolle and Pock in 2016 to solve saddle point problems defined by a convex-concave function $\mathcal{L}(x, y) = f(x) + \Phi(x, y) - h(y)$ with a general coupling term $\Phi(x, y)$ that is *not* assumed to be bilinear. Given a saddle point $(x^*, y^*)$, assuming $\nabla_x \Phi(\cdot, y)$ is Lipschitz in $x$ for any fixed $y$, and $\nabla_y \Phi(\cdot, \cdot)$ is Lipschitz, we derive error bounds in terms of $\mathcal{L}(\bar{x}_k, y^*) - \mathcal{L}(x^*, \bar{y}_k)$ for the ergodic sequence $\{\bar{x}_k, \bar{y}_k\}$; in particular, we show $\mathcal{O}(1/k)$ rate when the problem is merely convex in $x$. Furthermore, assuming $\Phi(x, \cdot)$ is linear in $y$ for each fixed $x$ and $f$ is strongly convex, we obtain the ergodic convergence rate of $\mathcal{O}(1/k^2)$ – we are not aware of another single-loop method in the related literature achieving the same rate when $\Phi$ is not bilinear. We tested our method for solving kernel matrix learning problem, and compare it against the Mirror-prox algorithm and interior point methods.

## 1   Introduction

Let $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ and $(\mathcal{Y}, \|\cdot\|_{\mathcal{Y}})$ be finite dimensional, normed vector spaces. In this paper, we study the following saddle point (SP) problem:

$$(P): \quad \min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \mathcal{L}(x, y) \triangleq f(x) + \Phi(x, y) - h(y), \tag{1}$$

[*]Pennsylvania State University, University Park, PA 16803 E-mail: evy5047@psu.edu, nsa10@psu.edu

where $f : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ and $h : \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ are convex functions (possibly nonsmooth) and $\Phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is a differentiable function, convex in $x$ and concave in $y$. Our objective is to design an efficient first-order method to compute a saddle point of the structured convex-concave function $\mathcal{L}$ in (1). The problem $(P)$ covers a broad class of optimization problems, e.g., convex optimization with nonlinear conic constraints which itself includes LP, QP, QCQP, SOCP, and SDP as its subclasses. Indeed, consider

$$\min_{x \in \mathbb{R}^n} \ \rho(x) \triangleq f(x) + g(x) \quad \text{s.t.} \quad G(x) \in -\mathcal{K}, \tag{2}$$

where $\mathcal{K} \subseteq \mathcal{Y}^*$ is a closed convex cone in the dual space $\mathcal{Y}^*$, $f$ is convex (possibly nonsmooth), $g$ is convex with a Lipschitz continuous gradient, $G : \mathcal{X} \to \mathcal{Y}^*$ is a smooth $\mathcal{K}$-convex, Lipschitz function having Lipschitz continuous derivative. Various optimization problems that frequently arise in many important machine learning applications are special cases of the conic problem in (2), e.g., primal or dual formulations of $\ell_1$ or $\ell_2$-norm soft margin SVM, ellipsoidal kernel machines, kernel matrix learning, and etc. Using Lagrangian duality, one can equivalently write (2) as

$$\min_{x \in \mathbb{R}^n} \max_{y \in \mathcal{K}^*} f(x) + g(x) + \langle G(x), y \rangle, \tag{3}$$

which is a special case of (1), i.e., $\Phi(x, y) = g(x) + \langle G(x), y \rangle$ and $h(y) = \mathbb{I}_{\mathcal{K}^*}(y)$ is the indicator function of $\mathcal{K}^*$, where $\mathcal{K}^* \subseteq \mathcal{Y}$ denotes the dual cone of $\mathcal{K}$.

**Related Work.** Constained convex optimization can be viewed as a special of case of SP (1), and recently some first-order methods and their randomized-coordinate variants are proposed to solve $\min \{f(\mathbf{x}) + g(\mathbf{x}) \ : \ A\mathbf{x} = b, \ G_j(\mathbf{x}) \leq 0, \ \forall j \in \{1, 2, ..., m\}\}$. In [21], a level-set method with iteration complexity guarantees is proposed for nonsmooth/smooth and strongly/merely convex settings. Moreover, in [32], a primal-dual method based on linearized augmented Lagrangian method is proposed with a sublinear convergence rate in terms of suboptimality and infeasibility. However, none of these methods can solve the more general SP problem we considered in this paper.

SP problems has become popular in recent years due to their generality and ability to directly solve constrained optimization problems. There has been several studies on proposing first-order primal-dual algorithms for (1) when $\Phi(x, y)$ is bilinear, such as [7, 10, 8, 16, 30, 11], and few others

considered the more general setting similar to this paper [24, 22, 18, 15, 19]. Here, we briefly review some recent work that is closely related to ours. In the rest, suppose (1) has a saddle point $(x^*, y^*)$.

In [7], a special case of (1) with bilinear coupling term is studied:

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \hat{f}(x) + \langle Kx, \ y \rangle - h(y), \tag{4}$$

for some linear operator $K : \mathcal{X} \to \mathcal{Y}^*$, where $\hat{f}$ and $h$ are convex functions with easily computable prox (Moreau) maps [17]. The authors proposed a primal-dual algorithm which guarantees that $\mathcal{L}(\bar{x}_K, y^*) - \mathcal{L}(x^*, \bar{y}_K)$ converges to 0 with $\mathcal{O}(1/K)$ rate when $\hat{f}$ is merely convex and with $\mathcal{O}(1/K^2)$ rate when $\hat{f}$ is strongly convex, where $\{(\bar{x}_k, \bar{y}_k)\}_k$ is a weighted ergodic average sequence. In a follow-up paper, Chambolle and Pock [8] consider an SP problem of the form in (4) such that $\hat{f}$ has a composite convex structure, i.e., $\hat{f}(x) = f(x) + g(x)$ such that $f$ has an easy prox map and $g$ has Lipschitz continuous gradient. It is shown that their previous work can be extended to handle non-linear proximity operators based on Bregman distance functions while guaranteeing the same rate results – see also [9] for an optimal method with $\mathcal{O}(1/K)$ rate to solve *bilinear* SP problems.

In a recent work, He and Monteiro [16] have considered a *bilinear* SP problem from a monotone inclusion perspective. They proposed an accelerated algorithm based on hybrid proximal extragradient (HPE) method, and showed that an $\epsilon$-saddle point $(x_\epsilon, y_\epsilon)$ can be computed within $\mathcal{O}(1/\epsilon)$ iterations. More recently, Kolossoski and Monteiro [19] propose another HPE-type method to solve a more general SP problems as in (1) over *bounded* sets – it is worth emphasizing that for a nonlinearly constrained convex optimization problem, the dual optimal solution set may be unbounded and/or it may not be trivial to get an upper bound on a dual solution. Indeed, the method in [19] is an inexact proximal point method, each prox subinclusion (outer step) is solved using an accelerated gradient method (inner steps). This work generalizes the method in [16] as the new method can deal with SP problems that are not bilinear, and it can use general Bregman distances instead of the Euclidean one.

Nemirovski [22] and Juditsky & Nemirovski [18] studied $\min_{x \in X} \max_{y \in Y} \Phi(x, y)$ as a variational inequality problem, and proposed a prox-type method. Assuming that $X$ and $Y$ are convex compact sets, $\Phi(x, y)$ is differentiable, and $F(x, y) = [\nabla_x \Phi(x, y)^\top, -\nabla_y \Phi(x, y)^\top]^\top$ is Lipschitz with constant $L$, $\mathcal{O}(L/K)$ ergodic convergence rate is shown for Mirror-prox where in each iteration $F$ is computed

*twice* and a projection onto $X \times Y$ is computed with respect to general (Bregman) distance. Moreover, in [18] under the assumption of strong concavity of $\Phi(x, y)$ in $y$ and compactness of $X$, if $\nabla_x \Phi(x, y)$ is independent of $x$, i.e., $L_{xx} = 0$, then convergence rate of $\mathcal{O}(1/K^2)$ is shown for a multi-stage method which repeatedly calls Mirror-Prox in each stage. In a more recent paper, He et al. [15] extend the original prox method to solve the composite version $\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} f(x) + \Phi(x, y) - h(y)$ with the same convergence rate where $f$ and $h$ are possibly nonsmooth convex functions with simple prox maps with respect to a general (Bregman) distance. In those papers, the primal and dual step-sizes are at most $1/L$.

**Application.** From application perspective, there are many real-life problems arising in machine learning, signal processing, image processing, finance, etc. and they can be formulated as a special case of (1). In particular, the following problems arising in machine learning can be efficiently solved using the methodology proposed in this paper: to name a few, **i)** robust classification under Gaussian uncertainty in feature observations leads to SOCP problems [5]; **ii)** distance metric learning formulation proposed in [31] is a convex optimization problem over positive semidefinite matrices subject to nonlinear convex constraints; **iii)** training ellipsoidal kernel machines [28] requires solving nonlinear SDPs; **iv)** learning a kernel matrix for transduction problem can be cast as an SDP or a QCQP [20, 12].

In this paper, following [20], we implemented our method for learning a kernel matrix to predict the labels of partially labeled data sets. To summarize the problem briefly, suppose we are given a set of labeled data points consisting of feature vectors $\{\mathbf{a}_i\}_{i \in \mathcal{S}} \subset \mathbb{R}^m$, corresponding labels $\{b_i\}_{i \in \mathcal{S}} \subset \{-1, +1\}$, and a set of unlabeled test data $\{\mathbf{a}_i\}_{i \in \mathcal{T}} \subset \mathbb{R}^m$. Let $n_{tr} \triangleq |\mathcal{S}|$ and $n_t \triangleq |\mathcal{T}|$ denote the cardinality of the training and test sets, respectively, and define $n \triangleq n_{tr} + n_t$. Consider $M$ different embedding of the data corresponding to kernel functions $k_\ell : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$ for $\ell = 1, ..., M$. Let $K_\ell \in \mathbb{S}_+^n$ be the kernel matrix such that $[K_\ell]_{ij} = k_\ell(\mathbf{a}_i, \mathbf{a}_j)$ for $i, j \in \mathcal{S} \cup \mathcal{T}$ and consider the partition of $K_\ell = \begin{pmatrix} K_\ell^{tr} & K_\ell^{tr,t} \\ K_\ell^{t,tr} & K_\ell^t \end{pmatrix}$ for all $\ell \in \{1, \ldots, M\}$.

The objective is to learn a kernel matrix $K$ belonging to a class of kernel matrices, which is a convex set generated by $\{K_\ell\}_{\ell=1}^M$, such that it minimizes the training error of a kernel SVM as a function of $K$. Skipping the details [20], one can study both $\ell_1$- and $\ell_2$-norm soft margin SVMs by

4

considering the following generic formulation:

$$\min_{\substack{K \in \mathcal{K}, \\ \text{trace}(K)=c}} \max_{\substack{0 \leq \alpha \leq C, \\ \langle \mathbf{b}, \alpha \rangle = 0}} 2\mathbf{e}^\top \alpha - \alpha^\top (G(K^{tr}) + \lambda \mathbf{I})\alpha, \tag{5}$$

where $c, C > 0$ and $\lambda \geq 0$ are model parameters, $\mathbf{b} = [b_i]_{i=1}^{n_{tr}}$ and $G(K^{tr}) \triangleq \mathbf{diag}(\mathbf{b})K^{tr}\,\mathbf{diag}(\mathbf{b})$. Suppose we want to learn a kernel matrix belonging to $\mathcal{K} = \{\sum_{\ell=1}^M \eta_\ell K_\ell : \eta_\ell \geq 0, \ \ell = 1, \ldots, M\}$; clearly, $K \in \mathcal{K}$ implies $K \succeq 0$. For kernel class $\mathcal{K}$, (5) takes the following form:

$$\min_{\substack{\eta: \ \langle \mathbf{r}, \eta \rangle = c, \\ \eta \geq 0}} \max_{\substack{\alpha: \ 0 \leq \alpha \leq C, \\ \langle \mathbf{b}, \alpha \rangle = 0}} 2\mathbf{e}^\top \alpha - \sum_{\ell=1}^M \eta_\ell \alpha^\top G(K_\ell^{tr})\alpha - \lambda \|\alpha\|_2^2, \tag{6}$$

where $\eta = [\eta_\ell]_{\ell=1}^M$ and $\mathbf{r} = [r_\ell]_{\ell=1}^M$ for $r_\ell = \text{trace}(K_\ell)$. Clearly, (6) is a special case of (1). In [20], (6) is equivalently represented as a QCQP and then solved using MOSEK [1], a comercial interior-point method (IPM). Computational complexity of a generic IPM is $\mathcal{O}(Mn_{tr}^3)$ for solving the resulting QCQP [23]. On the other hand, the first-order primal-dual method we proposed in this paper has $\mathcal{O}(Mn_{tr}^2)$ per-iteration complexity. Therefore, when $n_{tr}$ is very large, IPMs are not suitable for solving large-scale problems unless the data matrix has certain sparsity structure; and in practice as $n_{tr}$ grows, the first-order methods with much lower per-iteration complexity will have the advantage over IPMs for computing low to medium level accuracy solutions.

**Contribution.** In this paper we propose a primal-dual algorithm with a momentum term that can be viewed as a generalization of the method in [8] to solve SP problems with a more general coupling term $\Phi$ that is *not* bilinear. Assuming $\nabla_y \Phi(\cdot, \cdot)$ is Lipschitz and $\nabla_x \Phi(\cdot, y)$ is Lipschitz for any fixed $y$, we derive error bounds in terms of $\mathcal{L}(\bar{x}_K, y^*) - \mathcal{L}(x^*, \bar{y}_K)$ for the ergodic sequence – without requiring primal-dual domains to be bounded; in particular, we show $\mathcal{O}(1/K)$ rate that when the problem is merely convex in $x$ using a constant step-size rule. Furthermore, assuming $\Phi(x, \cdot)$ is linear in $y$ for each fixed $x$ and $f$ is strongly convex, we obtain the ergodic convergence rate of $\mathcal{O}(1/K^2)$ – we are not aware of any other single-loop method with $\mathcal{O}(1/K^2)$ rate when $\Phi$ is not bilinear.

The previous art for solving SP problems in the general setting are the Mirror-Prox algorithm in [22, 15] and HPE-type methods in [16, 19]. All these methods including ours have $\mathcal{O}(1/\epsilon)$ complexity under mere convexity; however, our accelerated primal-dual (APD) method has an

improved $\mathcal{O}(1/\sqrt{\epsilon})$ rate when $f$ is strongly convex. To the best of our knowledge, all the rates derived here are the best rates for our setting. When compared to [19], ours is a simpler one-loop algorithm while HPE [19] is a two-loop method requiring a more stronger oracle for subproblems – see Remark 2.1 and also requiring a bounded domain. Moreover, while convergence to a unique limit point is shown for APD, [19] shows a limit point result (weaker than ours), i.e., any limit point is a saddle point – see end of p.1254 in [19]. The other competitor algorithm, Mirror-Prox, requires computing both the primal and dual gradients twice during each iteration. The proposed APD method only needs to compute them once; thus, saving the computation cost by half yet achieving the same iteration complexity. Moreover, recall that in [15], $\nabla\Phi(\cdot,\cdot)$ is assumed to be Lipschitz with constant $L$ and the method has a primal-dual stepsize less than $1/L$; note that while our assumption on $\Phi$ is weaker, our primal and dual step sizes are longer than $1/L$ – see Remark 2.2 for weakening the assumptions on $\Phi$. Finally, the numerical results clearly demonstrate that APD has roughly the same iteration complexity as proximal Mirror-Prox; but, requires half the computational efforts (reflected by the savings in computation time).

**Organization of the Paper.** In the coming section, we precisely state our assumptions on $\mathcal{L}$ in (1), describe the proposed algorithm, APD, and present convergence guarantees for the APD iterate sequence, which is the main result of this paper. Subsequently, in Section 3, we provide an easy-to-read convergence analysis proving the main result. Next, in Section 4, we apply our APD method to solve the kernel matrix learning problem and numerically compare it with the Mirror-prox method proposed in [15] and off-the-shelf interior point methods. Finally, Section 5 concludes the paper.

## 2    Accelerated Primal-Dual Algorithm

**Definition 1.** *Let $\varphi_{\mathcal{X}} : \mathcal{X} \to \mathbb{R}$ and $\varphi_{\mathcal{Y}} : \mathcal{Y} \to \mathbb{R}$ be differentiable functions on open sets containing* **dom** $f$ *and* **dom** $h$, *respectively. Suppose $\varphi_{\mathcal{X}}$ and $\varphi_{\mathcal{Y}}$ have closed domains and are 1-strongly convex w.r.t. $\|\cdot\|_{\mathcal{X}}$ and $\|\cdot\|_{\mathcal{Y}}$, respectively. Let $\mathbf{D}_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ and $\mathbf{D}_{\mathcal{Y}} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ be Bregman distance functions corresponding to $\varphi_{\mathcal{X}}$ and $\varphi_{\mathcal{Y}}$, i.e., $\mathbf{D}_{\mathcal{X}}(x, \bar{x}) \triangleq \varphi_{\mathcal{X}}(x) - \varphi_{\mathcal{X}}(\bar{x}) - \langle \nabla\varphi_{\mathcal{X}}(\bar{x}), x - \bar{x} \rangle$, and $\mathbf{D}_{\mathcal{Y}}$ has a similar form.*

Clearly, $\mathbf{D}_{\mathcal{X}}(x, \bar{x}) \geq \frac{1}{2} \|x - \bar{x}\|_{\mathcal{X}}^2$ for $x \in \mathcal{X}$ and $\bar{x} \in \mathbf{dom}\, f$, and $\mathbf{D}_{\mathcal{Y}}(y, \bar{y}) \geq \frac{1}{2} \|y - \bar{y}\|_{\mathcal{Y}}^2$ for $y \in \mathcal{Y}$

and $\bar{y} \in \mathbf{dom}\, h$. The dual spaces are denoted by $\mathcal{X}^*$ and $\mathcal{Y}^*$. For $x' \in \mathcal{X}^*$, we define the dual norm $\|x'\|_{\mathcal{X}^*} \triangleq \max\{\langle x', x \rangle : \|x\|_{\mathcal{X}} \le 1\}$, and $\|\cdot\|_{\mathcal{Y}^*}$ is defined similarly.

Recall that if $f$ is convex with modulus $\mu \ge 0$, then for any $x, \bar{x} \in \mathbf{dom}\, f$ and $g \in \partial f(\bar{x})$,

$$f(x) \ge f(\bar{x}) + \langle g,\ x - \bar{x} \rangle + \frac{\mu}{2} \|x - \bar{x}\|_{\mathcal{X}}^2. \tag{7}$$

Note also that (9) and convexity imply that for any $y \in \mathcal{Y}$,

$$0 \le \Phi(x, y) - \Phi(\bar{x}, y) - \langle \nabla_x \Phi(\bar{x}, y),\ x - \bar{x} \rangle \le \frac{L_{xx}}{2} \|x - \bar{x}\|_{\mathcal{X}}^2, \quad \forall x, \bar{x} \in \mathcal{X}. \tag{8}$$

We next state our main assumption and propose the APD algorithm for solving (1), and discuss its convergence properties which is the main result of this paper.

**Assumption 2.1.** *Suppose $\mathbf{D}_{\mathcal{X}}$ and $\mathbf{D}_{\mathcal{Y}}$ be some Bregman distance functions as in Definition 1. In case $f$ is strongly convex, i.e., $\mu > 0$, we fix $\|x\|_{\mathcal{X}} = \sqrt{\langle x, x \rangle}$, and set $\mathbf{D}_{\mathcal{X}}(x, \bar{x}) = \frac{1}{2} \|x - \bar{x}\|_{\mathcal{X}}^2$.*

*Suppose $f$ and $h$ are closed convex functions, and $\Phi$ is a differentiable function such that*

**(i)** *for any fixed $y \in \mathcal{Y}$, $\Phi(x, y)$ is convex and differentiable in $x$, and for some $L_{xx} \ge 0$,*

$$\|\nabla_x \Phi(x, y) - \nabla_x \Phi(\bar{x}, y)\|_{\mathcal{X}^*} \le L_{xx} \|x - \bar{x}\|_{\mathcal{X}}, \quad \forall\, x, \bar{x} \in \mathcal{X}, \tag{9}$$

**(ii)** *for any fixed $x \in \mathcal{X}$, $\Phi(x, y)$ is concave and differentiable in $y$; for some $L_{yx} > 0$ and $L_{yy} \ge 0$,*

$$\|\nabla_y \Phi(x, y) - \nabla_y \Phi(\bar{x}, \bar{y})\|_{\mathcal{Y}^*} \le L_{yy} \|y - \bar{y}\|_{\mathcal{Y}} + L_{yx} \|x - \bar{x}\|_{\mathcal{X}}, \quad \forall\, y, \bar{y} \in \mathcal{Y},\ \forall\, x, \bar{x} \in \mathcal{X}. \tag{10}$$

**(iii)** *For any $\bar{x} \in \mathbf{dom}\, f$, $s \in \mathcal{X}^*$ and $t > 0$, $\operatorname{argmin}_{x \in \mathcal{X}} \{tf(x) + \langle s, x \rangle + \mathbf{D}_{\mathcal{X}}(x, \bar{x})\}$ can be computed efficiently. Similarly, $\operatorname{argmin}_{y \in \mathcal{Y}} \{th(y) + \langle s, y \rangle + \mathbf{D}_{\mathcal{Y}}(y, \bar{y})\}$ is easy to compute for any $\bar{y} \in \mathbf{dom}\, h$, $s \in \mathcal{Y}^*$ and $t > 0$.*

**Remark 2.1.** *The APD subproblem $(S_y)$: $\operatorname{argmin}_{y \in \mathcal{Y}} \{th(y) + \langle s, y \rangle + \mathbf{D}_{\mathcal{Y}}(y, \bar{y})\}$ is a generalization of Moreau map [17]. Compared to our subproblem $(S_y)$, HPE-type method in [19] requires solving $\operatorname{argmax}_{y \in Y} \Phi(\bar{x}, y) - \mathbf{D}_{\mathcal{Y}}(y, \bar{y})/\sigma$ for some given $\bar{x}$ and $\bar{y}$ where $Y \subset \mathcal{Y}$ is a bounded convex set such that $y^* \in Y$. This may not be a trivial operation.*

---

**Algorithm 1** Accelerated Primal-Dual (APD) Algorithm

---
1: **Input:** $\{\tau_k, \sigma_k, \theta_k\}_{k \geq 0}$, $(x_0, y_0) \in \mathcal{X} \times \mathcal{Y}$
2: $(x_{-1}, y_{-1}) \leftarrow (x_0, y_0)$
3: **for** $k \geq 0$ **do**
4: $\quad s_k \leftarrow (1 + \theta_k)\nabla_y \Phi(x_k, y_k) - \theta_k \nabla_y \Phi(x_{k-1}, y_{k-1})$
5: $\quad y_{k+1} \leftarrow \operatorname{argmin}_{y \in \mathcal{Y}} h(y) - \langle s_k,\ y \rangle + \frac{1}{\sigma_k} \mathbf{D}_{\mathcal{Y}}(y, y_k)$
6: $\quad x_{k+1} \leftarrow \operatorname{argmin}_{x \in \mathcal{X}} f(x) + \langle \nabla_x \Phi(x_k, y_{k+1}),\ x \rangle + \frac{1}{\tau_k} \mathbf{D}_{\mathcal{X}}(x, x_k)$
7: **end for**

---

**Theorem 2.1. (Main Result)** *Let $\mathbf{D}_{\mathcal{X}}$ and $\mathbf{D}_{\mathcal{Y}}$ be some Bregman distance functions given in Definition 1. Suppose Assumption 2.1 holds, and $\{x_k, y_k\}_{k \geq 0}$ is generated by Algorithm 1 using a parameter sequence $\{\tau_k, \sigma_k, \theta_k\}_{k \geq 0}$.*

*(**Part I.**) For all $k \geq 0$, let $\theta_k = 1$, $t_k = 1$, $\tau_k = \tau_0 = c_\tau (L_{xx} + L_{yx}^2/\alpha)^{-1}$, $\sigma_k = \sigma_0 = c_\sigma (\alpha + 2L_{yy})^{-1}$ for some $\alpha > 0$ and $c_\tau$, $c_\sigma \in (0,1]$. Then for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$,*

$$\mathcal{L}(\bar{x}_K, y) - \mathcal{L}(x, \bar{y}_K) \leq \frac{1}{K} \Delta(x, y), \quad \Delta(x, y) \triangleq \frac{1}{\tau_0} \mathbf{D}_{\mathcal{X}}(x, x_0) + \frac{1}{\sigma_0} \mathbf{D}_{\mathcal{Y}}(y, y_0), \tag{11}$$

*holds for all $K \geq 1$, where $\bar{x}_K \triangleq \frac{1}{K} \sum_{k=1}^{K} x_k$ and $\bar{y}_K \triangleq \frac{1}{K} \sum_{k=1}^{K} y_k$. Moreover, if a saddle point for (1) exists and $c_\tau$, $c_\sigma \in (0,1)$, then the actual sequence $\{(x_k, y_k)\}_{k \geq 0}$ converges to a saddle point $(x^*, y^*)$ such that $0 \leq \mathcal{L}(\bar{x}_K, y^*) - \mathcal{L}(x^*, \bar{y}_K) \leq \mathcal{O}(1/K)$ and $\mathbf{D}_{\mathcal{Y}}(y^*, y_K) \leq \frac{c_\sigma}{1-c_\sigma} \frac{\Delta(x^*,y^*)}{\alpha + 2L_{yy}}$.*

*(**Part II.**) Suppose $\mu > 0$ and $L_{yy} = 0$, in this setting let $\|x\|_{\mathcal{X}} = \sqrt{\langle x, x \rangle}$, and $\mathbf{D}_{\mathcal{X}}(x, \bar{x}) = \frac{1}{2} \|x - \bar{x}\|_{\mathcal{X}}^2$. If $\{\tau_k, \sigma_k, \theta_k\}_{k \geq 0}$ is chosen such that for $k \geq 0$*

$$\theta_{k+1} \leftarrow \frac{1}{\sqrt{1 + \mu\tau_k}}, \quad \tau_{k+1} \leftarrow \theta_{k+1}\tau_k, \quad \sigma_{k+1} \leftarrow \sigma_k/\theta_{k+1}, \tag{12}$$

*where $\tau_0 = c_\tau (L_{xx} + L_{yx}^2/\alpha)^{-1}$, $\sigma_0 = c_\sigma \alpha^{-1}$ for some $\alpha > 0$ and $c_\tau$, $c_\sigma \in (0,1]$, and $\theta_0 = 1$, then for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$,*

$$\mathcal{L}(\bar{x}_K, y) - \mathcal{L}(x, \bar{y}_K) \leq \frac{\sigma_0}{T_K} \Delta(x, y)$$

*holds for all $K \geq 1$, where $\bar{x}_K = \frac{1}{T_K} \sum_{k=0}^{K-1} \sigma_k x_{k+1}$, $\bar{y}_K = \frac{1}{T_K} \sum_{k=0}^{K-1} \sigma_k y_{k+1}$, and $T_K = \sum_{k=0}^{K-1} \sigma_k = \Theta(K^2)$ [1]. Moreover, if a saddle point $(x^*, y^*)$ for (1) exists, then $\{x_k\}_{k \geq 0}$ converges to $x^*$ such that $0 \leq \mathcal{L}(\bar{x}_K, y^*) - \mathcal{L}(x^*, \bar{y}_K) \leq \mathcal{O}(1/K^2)$, $\mathbf{D}_{\mathcal{X}}(x^*, x_K) \leq \frac{\tau_K}{\sigma_K} \sigma_0 \Delta(x^*, y^*) = \mathcal{O}(1/K^2)$, and for $c_\sigma \in (0,1)$ one also has $\mathbf{D}_{\mathcal{Y}}(y^*, y_K) \leq \frac{c_\sigma}{1-c_\sigma} \frac{\Delta(x^*,y^*)}{\alpha}$.*

---

[1] $f(k) = \Theta(k)$ means $f(k) = \mathcal{O}(k)$ and $f(k) = \Omega(k)$.

*Proof.* See Section 3.1 for the proof of the main result. □

**Remark 2.2.** *As in [19], assuming a stronger oracle we can remove $L_{xx}$ assumption in (6), i.e., if we replace Line 6 of APD with $\operatorname{argmin}_x f(x) + \Phi(x, y_{k+1}) + D_{\mathcal{X}}(x, x_k)/\tau_k$, then we can remove assumption in (9); hence, even if $\Phi$ is nonsmooth in $x$, all the rate results in our submission will continue to hold. Let $\Phi(x, y) = x^2 y$ on $x \in [-1, 1]$ and $y \geq 0$; the Lipschitz constant $L$ for $\nabla\Phi$ would not exist in this case, and it is not clear how one can modify the analysis of [15] to deal with problems when $\Phi$ is not jointly differentiable.*

Our method generalizes the primal-dual method proposed by [8] to solve SP problems with coupling term $\Phi$ that is *not* bilinear. Now consider applying our method in their setting, i.e., on (4) with $\hat{f}(x) = f(x) + g(x)$ such that $f$ has an easy prox map and $g$ has a Lipschitz continuous gradient with constant $L$. We can equivalently represent it as (1) by setting $\Phi(x, y) = g(x) + \langle Kx, y \rangle$. Note that $L_{xx} = L$, $L_{yy} = 0$, and $L_{yx} = \|K\|$, where $\|\cdot\|$ denotes the spectral norm; hence, according to Part I of Theorem 2.1 primal-dual steps are constant satisfying $\tau_0 \leq \frac{1}{L+\|K\|}$ and $\sigma_0 \leq \frac{1}{\|K\|}$. This choice of step sizes for our APD algorithm also satisfies the condition $(1/\tau_0 - L) \geq \sigma_0 \|K\|^2$ for Algorithm 1 in [8] to work; and both algorithms generate the same iterate sequence with same error bounds. In case $f$ is strongly convex with modulus $\mu > 0$, when $\{(\tau_k, \sigma_k, \theta_k)\}$ is chosen as described in Part II of Theorem 2.1 with $c = 1$, our APD algorithm and Algorithm 4 in [8] outputs the same iterate sequence with the same error bounds. Therefore, APD algorithm inherits the already established connections of the primal-dual framework in [8] to other well-known methods, e.g., (linearized) ADMM [27, 4] and Arrow-Hurwicz method [2].

## 2.1 Application to Constrained Convex Optimization

As mentioned in Section 1, an important special case of (1) is the convex optimization problem with a nonlinear conic constraint, formulated as in (2). Indeed, (2) can be reformulated as a saddle point problem as shown in (3), which is in the form of (1). Clearly, $L_{yy} = 0$, and $L_{yx} > 0$ exists if $G$ is Lipschitz. Moreover, for any fixed $y \in \mathcal{Y}$, a bound on $L_{xx}$, the Lipschitz constant of $\nabla_x \Phi(x, y)$ as a function of $x$, can be computed as follows:

$$\|\nabla_x \Phi(x, y) - \nabla_x \Phi(\bar{x}, y)\|_{\mathcal{X}^*} \leq \|\nabla g(x) - \nabla g(\bar{x})\|_{\mathcal{X}^*} + \left\|(\nabla G(x) - \nabla G(\bar{x}))^\top y\right\|_{\mathcal{X}^*}$$

$$\leq (L(g) + L(G) \|y\|_{\mathcal{Y}}) \|x - \bar{x}\|_{\mathcal{X}}, \quad \forall \, x, \bar{x} \in \mathcal{X}. \tag{13}$$

In this section, we customize our algorithm and state its convergence result for this problem. For simplifying the notation, we assume that $(\mathcal{X}, \ \|\cdot\|_{\mathcal{X}}) = (\mathbb{R}^n, \ \|\cdot\|)$ and $(\mathcal{Y}, \ \|\cdot\|_{\mathcal{Y}}) = (\mathbb{R}^m, \ \|\cdot\|)$ are Euclidean spaces, i.e., $\|\cdot\|$ denotes the Euclidean norm, and we also assume that a dual optimal solution $y^* \in \mathcal{Y}$ exists. Consider the objective $\rho(x) \triangleq f(x) + g(x)$ in (2), suppose $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is convex (possibly nonsmooth), $g : \mathbb{R}^n \to \mathbb{R}$ is convex with a Lipschitz continuous gradient with constant $L(g)$, and $\mathcal{K} \subset \mathbb{R}^m$ is a closed convex cone. Moreover, $G : \mathbb{R}^n \to \mathbb{R}^m$ is a $\mathcal{K}$-convex [6], Lipschitz continuous function with constant $C(G) > 0$ and it has a Lipschitz continuous Jacobian, denoted by $\nabla G : \mathbb{R}^n \to \mathbb{R}^{m \times n}$, with constant $L(G) \geq 0$. In the rest, let $\mathcal{P}_{\mathcal{K}}(w) \triangleq \operatorname{argmin}_{y \in \mathcal{K}} \|y - w\|$, $d_{\mathcal{K}}(w) \triangleq \|\mathcal{P}_{\mathcal{K}}(w) - w\| = \|\mathcal{P}_{\mathcal{K}^\circ}(w)\|$ where $\mathcal{K}^\circ = -\mathcal{K}^*$ denotes the polar cone of $\mathcal{K}$.

We next consider two scenarios: **i)** a dual bound is known, **ii)** a dual bound is not known.

**i) A dual bound is known.** Let $\mathcal{B} \triangleq \{y \in \mathbb{R}^m \ : \ \|y\| \leq B\}$ for some $B > 0$ such that $\|y^*\| \leq B$ for some dual optimal solution $y^*$. Thus, the Lipschitz constant $L_{xx}$ for (13) can be chosen as $L_{xx} = L(g) + BL(G)$ and we set $h(y) = \mathbb{I}_{\mathcal{K}^* \cap \mathcal{B}}$. Such a bound $B$ can be computed if a slater point for (2) is available. In particular, using the following lemma one can compute a dual bound efficiently.

**Lemma 2.2.** *[3] Let $\bar{x}$ be a Slater point for (2), i.e., $\bar{x} \in \mathbf{relint}(\mathbf{dom}\,\rho)$ such that $G(\bar{x}) \in \mathbf{int}(-\mathcal{K})$, and $q : \mathbb{R}^m \to \mathbb{R} \cup \{-\infty\}$ denote the Lagrangian dual function, i.e.,*

$$q(y) \triangleq \begin{cases} \inf_x \rho(x) + \langle G(x), \ y \rangle, & \text{if } y \in \mathcal{K}^*; \\ -\infty, & \text{o.w.} \end{cases}$$

*For any $\bar{y} \in \mathbf{dom}\,q$, let $Q_{\bar{y}} \triangleq \{y \in \mathbf{dom}\,q : \ q(y) \geq q(\bar{y})\} \subset \mathcal{K}^*$ denote the corresponding superlevel set. Then for all $\bar{y} \in \mathbf{dom}\,q$, the superlevel set $Q_{\bar{y}}$ can be bounded as follows:*

$$\|y\| \leq \frac{\rho(\bar{x}) - q(\bar{y})}{r^*}, \quad \forall y \in Q_{\bar{y}}, \tag{14}$$

*where $0 < r^* \triangleq \min_w \{\langle G(\bar{x}), \ w \rangle : \ \|w\| = 1, \ w \in \mathcal{K}^*\}$. Although this is not a convex problem due to the nonlinear equality constraint, one can upper bound (14) using $0 < \tilde{r} \leq r^*$, which can be efficiently computed by solving a convex problem $\tilde{r} \triangleq \min_w \{\langle G(\bar{x}), \ w \rangle : \ \|w\|_1 = 1, \ w \in \mathcal{K}^*\}$.*

**Corollary 2.2.1.** *Consider the convex optimization problem in (2) and let $\{(x_k, y_k)\}_{k \geq 0}$ be the APD iterate sequence when APD is applied to the following SP problem with $h(y) = \mathbb{I}_{\mathcal{K}^* \cap \mathcal{B}}$,*

$$\min_{x \in \mathbb{R}^n} \max_{y \in \mathbb{R}^m} f(x) + g(x) + \langle G(x), y \rangle - h(y). \tag{15}$$

**i.** *Suppose the objective function in* (2) *is merely convex. For all* $k \geq 0$, *let* $\theta_k = 1$, $\tau_k = \tau_0 = c_\tau \left( L(g) + BL(G) + \frac{1}{\alpha}C(G)^2 \right)^{-1}$ *and* $\sigma_k = \sigma_0 = c_\sigma \alpha^{-1}$ *for some* $\alpha > 0$ *and* $c_\tau, c_\sigma \in (0,1]$, *then*

$$\max \left\{ |\rho(\bar{x}_K) - \rho(x^*)|, \ \|y^*\| \, d_{-\mathcal{K}}(G(\bar{x}_K)) \right\} \leq \frac{1}{K}\Delta(x^*, y^\dagger) = \mathcal{O}(1/K),$$

*where* $y^\dagger = 2\|y^*\| \, \mathcal{P}_{\mathcal{K}^*}(G(\bar{x}_K)) \, \big\|\mathcal{P}_{\mathcal{K}^*}(G(\bar{x}_K))\big\|^{-1}$, $\bar{x}_K$ *is defined in Theorem 2.1 Part I and* $\Delta(x,y)$ *is defined in* (11).

**ii.** *Suppose the objective function in* (2) *is strongly convex with* $\mu > 0$. *For all* $k \geq 0$, *let* $\tau_0 = c_\tau \left( L(g) + BL(G) + \frac{1}{\alpha}C(G)^2 \right)^{-1}$ *and* $\sigma_0 = c_\sigma \alpha^{-1}$ *for some* $\alpha > 0$ *and* $c_\tau, c_\sigma \in (0,1]$. *If* $\{\tau_k, \sigma_k, \theta_k\}_{k \geq 0}$ *sequence is chosen as in* (12), *then*

$$\max \left\{ |\rho(\bar{x}_K) - \rho(x^*)|, \ \|y^*\| \, d_{-\mathcal{K}}(G(\bar{x}_K)) \right\} \leq \frac{\sigma_0}{T_K}\Delta(x^*, y^\ddagger) = \mathcal{O}(1/K^2),$$

$$\mathbf{D}_{\mathcal{X}}(x^*, x_K) \leq \frac{\tau_K}{\sigma_K}\sigma_0\Delta(x^*, y^*) = \mathcal{O}(1/K^2),$$

*where* $y^\ddagger = 2\|y^*\| \, \mathcal{P}_{\mathcal{K}^*}(G(\bar{x}_K)) \, \big\|\mathcal{P}_{\mathcal{K}^*}(G(\bar{x}_K))\big\|^{-1}$, $\bar{x}_K$ *and* $T_K$ *is defined in Theorem 2.1 Part II, and* $\Delta(x,y)$ *is defined in* (11).

*Proof.* **i.** It is easy to verify that $\langle G(\bar{x}_K), y^\dagger \rangle = 2\|y^*\| \, d_{-\mathcal{K}}(G(\bar{x}_K))$ as for any $w \in \mathbb{R}^m$ we have $w = \mathcal{P}_{-\mathcal{K}}(w) + \mathcal{P}_{\mathcal{K}^*}(w)$ and $\langle \mathcal{P}_{-\mathcal{K}}(w), \ \mathcal{P}_{\mathcal{K}^*}(w) \rangle = 0$. Hence, $\mathcal{L}(\bar{x}_K, y^\dagger) = \rho(\bar{x}_K) + 2\|y^*\|_{\mathcal{Y}} \, d_{-\mathcal{K}}(G(\bar{x}_K))$ since $y^\dagger \in \mathcal{K}^*$. Note that $\rho(x^*) = \mathcal{L}(x^*, y^*) \geq \mathcal{L}(x^*, \bar{y}_K)$. Therefore, (11) implies that

$$\rho(\bar{x}_K) - \rho(x^*) + 2\|y^*\| \, d_{-\mathcal{K}}(G(\bar{x}_K)) \leq \mathcal{L}(\bar{x}_K, y^\dagger) - \mathcal{L}(x^*, \bar{y}_K) \leq \frac{1}{K}\Delta(x^*, y^\dagger). \tag{16}$$

On the other hand, we also have

$$0 \leq \mathcal{L}(\bar{x}_K, y^\dagger) - \mathcal{L}(x^*, y^*) = \rho(\bar{x}_K) - \rho(x^*) + \langle G(\bar{x}_K), y^* \rangle$$

$$\leq \rho(\bar{x}_K) - \rho(x^*) + \|y^*\| \, d_{-\mathcal{K}}(G(\bar{x}_K)), \tag{17}$$

where we used the fact that for any $y \in \mathbb{R}^m$, $\langle y^*, y \rangle \leq \langle y^*, \mathcal{P}_{\mathcal{K}^*}(y) \rangle \leq \|y^*\| \, d_{-\mathcal{K}}(y)$. Combining (16) and (17) gives the result.

**ii.** The proof is similar to part **i.**. $\qquad\qquad\square$

**ii) A dual bound is not known.** Here we consider the situation where the dual bound may not exist or may not be readily available. In this case, one can still implement the algorithm APD provided that the primal step size is initialized sufficiently small.

**Corollary 2.2.2.** *Consider the convex optimization problem in (2) and let $\{(x_k, y_k)\}_{k \geq 0}$ be the APD iterate sequence when APD is applied to (15) with $h(y) = \mathbb{I}_{\mathcal{K}^*}$. The bounds in **i** and **ii** of Corollary 2.2.1 continue to hold when the stepsize sequence is selected as described in Corollary 2.2.1 with a sufficiently small $\tau_0 > 0$. More precisely, $\tau_0 = c_\tau \left( L(g) + BL(G) + \frac{1}{\alpha} C(G)^2 \right)^{-1}$ for any $B > 0$ satisfying*

$$\|y^*\| + \sqrt{\frac{2c_\sigma}{1 - c_\sigma} \left( \frac{L(g) + BL(G) + \frac{1}{\alpha}C(G)^2}{c_\tau \alpha} \, \mathbf{D}_{\mathcal{X}}(x^*, x_0) + \frac{1}{c_\sigma} \mathbf{D}_{\mathcal{Y}}(y^*, y_0) \right)} \leq B. \tag{18}$$

*Proof.* For both cases, we conclude from Theorem 2.1 that $\mathbf{D}_{\mathcal{Y}}(y^*, y_K) \leq \frac{\sigma_0}{1 - c_\sigma} \Delta(x^*, y^*)$. Since $\mathbf{D}_{\mathcal{Y}}(y^*, y_K) \geq \frac{1}{2} \|y_K - y^*\|_{\mathcal{Y}}^2$, the condition in (18) implies that $\|y_K\|_{\mathcal{Y}} \leq B$. Therefore, according to (13), for APD iterate sequence $\{x_k, y_k\}_{k \geq 0}$, one can set $L_{xx} = L(g) + BL(G)$; hence, the rest follows from Theorem 2.1 and Corollary 2.2.1. $\qquad\square$

## 3 Methodology

In this section we provide a general result for the APD algorithm unifying the analysis of both cases described in Theorem 2.1. An easy-to-read convergence analysis is given at the end of this section. Our analysis assume some conditions on $\{(\tau_k, \sigma_k, \theta_k)\}_{k \geq 0}$, stated in Assumption 3.1. Later we discuss that step size rules described in Theorem 2.1 produce step-size sequences satisfying these conditions. We define $0^2/0 = 0$ which may arise when $L_{yy} = 0$.

**Assumption 3.1.** *(**Step-size Condition**) For any $k \geq 0$, the step-sizes $\tau_k, \sigma_k$ and momentum parameter $\theta_k$ satisfy the following conditions: $\theta_0 = t_0 = 1$ and*

$$\frac{1}{\tau_k} \geq L_{xx} + \frac{L_{yx}^2}{\alpha_{k+1}}, \qquad \frac{1}{\sigma_k} \geq \theta_k(\alpha_k + \beta_k) + \frac{L_{yy}^2}{\beta_{k+1}}, \tag{19a}$$

$$t_k \left( \frac{1}{\tau_k} + \mu \right) \geq \frac{t_{k+1}}{\tau_{k+1}}, \qquad \frac{t_k}{\sigma_k} \geq \frac{t_{k+1}}{\sigma_{k+1}}, \qquad \frac{t_k}{t_{k+1}} = \theta_{k+1}, \tag{19b}$$

*for some positive $\{\alpha_k\}_{k \geq 0}$ and nonnegative $\{\beta_k\}_{k \geq 0}$.*

**Theorem 3.1.** *Suppose Assumption 2.1 holds, and $\{x_k, y_k\}_{k \geq 0}$ is generated by APD, stated in Algorithm 1, using a parameter sequence $\{\tau_k, \sigma_k, \theta_k\}_{k \geq 0}$ that satisfies Assumption 3.1. Then for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and $K \geq 1$,*

$$\mathcal{L}(\bar{x}_K, y) - \mathcal{L}(x, \bar{y}_K) \leq \frac{1}{N_K} \Delta(x, y), \tag{20}$$

*where $\Delta(x, y)$ is defined in (11), $N_K = \sum_{k=0}^{K-1} t_k$, $\bar{x}_K = \frac{1}{N_K} \sum_{k=0}^{K-1} t_k x_{k+1}$, and $\bar{y}_K = \frac{1}{N_K} \sum_{k=0}^{K-1} t_k y_{k+1}$. Moreover, for any saddle point $(x^*, y^*)$ of $\mathcal{L}$ in (1),*

$$\frac{1}{\tau_K} \mathbf{D}_{\mathcal{X}}(x^*, x_K) + \big(\frac{1}{\sigma_K} - \theta_K(\alpha_K + \beta_K)\big) \mathbf{D}_{\mathcal{Y}}(y^*, y_K) \leq \frac{1}{t_K} \Delta(x^*, y^*). \tag{21}$$

*Proof.* See Section 6.1 in the appendix. $\qquad\square$

## 3.1 Proof of the main result: Theorem 2.1

Now we are ready to establish the main result in Theorem 2.1. Indeed, we show that there exists $\{\alpha_k, \beta_k\}_{k \geq 0} \subset \mathbb{R}_{++} \times \mathbb{R}_+$ such that the particular step-size instances stated in Theorem 2.1 (constant in Part I and non-constant in Part II) both satisfy the condition (19); hence, the main result will then follow from Theorem 3.1. Consider Part I of Theorem 2.1. For $k \geq 0$, let $\alpha_k = \alpha > 0$ and $\beta_k = L_{yy} \geq 0$. Setting $\theta_k = 1$, $t_k = 1$, $\tau_k = \tau_0$, and $\sigma_k = \sigma_0$ clearly satisfies (19) for any $\mu \geq 0$ and we have $N_K = \sum_{k=0}^{K=1} t_k = K$ which gives us the rate result in Part I. Moreover, (21) implies that $\mathbf{D}_{\mathcal{Y}}(y^*, y_K) \leq \frac{\sigma_0}{1-c_\sigma} \Delta(x^*, y^*)$.

Now suppose $\mu > 0$ and $L_{yy} = 0$. To show Part II of Theorem 2.1, consider $\{\tau_k, \sigma_k, \theta_k\}_{k \geq 0}$ given in Part II. Since $\sigma_{k+1} = \sigma_k / \theta_{k+1}$, we have $\theta_{k+1} = \sigma_k / \sigma_{k+1}$ for $k \geq 0$; therefore, (19) implies that $t_k = \frac{\sigma_k}{\sigma_0}$ for $k \geq 0$. Given $c_\sigma \in (0, 1]$, let $\alpha_k = \frac{c_\sigma}{\theta_k \sigma_k}$ and $\beta_k = 0$ for $k \geq 0$. It is easy to verify that $\{\tau_k, \sigma_k, \theta_k\}_{k \geq 0}$ together with $\{\alpha_k, \beta_k\}_{k \geq 0}$ satisfies (19). Moreover, it can be shown that $\sigma_k = \Theta(k)$, $\tau_k = \Theta(1/k)$, and $\theta_k \in [0, 1)$ – for details see Section 5.2 in [8]. Therefore, we have $N_K = \sum_{k=0}^{K-1} t_k = \Theta(K^2)$ and $\mathcal{O}(1/K^2)$ rate in Part II follows from (20), and finally (21) implies that $\mathbf{D}_{\mathcal{X}}(x^*, x_K) \leq \sigma_0 \frac{\tau_K}{\sigma_K} \Delta(x^*, y^*) = \mathcal{O}(1/K^2)$ and $\mathbf{D}_{\mathcal{Y}}(y^*, y_K) \leq \frac{\sigma_0}{1-c_\sigma} \Delta(x^*, y^*)$.

Next, we show convergence of the iterate sequence to a unique limit point for Part I. The result in Lemma 3.2 will be used to establish the convergence of the primal-dual iterate sequence.

13

**Lemma 3.2.** *[25] Let $\{a_k\}$, $\{b_k\}$, and $\{c_k\}$ be non-negative real sequences such that $a_{k+1} \leq a_k - b_k + c_k$ for all $k \geq 0$, and $\sum_{k=0}^{\infty} c_k < \infty$. Then $a = \lim_{k \to \infty} a_k$ exists, and $\sum_{k=0}^{\infty} b_k < \infty$.*

In the rest of this section, we show the convergence of the iterate sequence for Part I of Theorem 2.1, assuming that for all $k \geq 0$, $\tau_k = \tau_0$ and $\sigma_k = \sigma_0$ such that $c_\tau$, $c_\sigma \in (0,1)$.

Suppose $t_k = 1$, $\theta_k = 1$, $\alpha_k = \alpha$, $\beta_k = L_{yy}$ for all $k \geq 0$ for some $\alpha > 0$ such that the conditions in (19a) hold. More precisely,

$$\frac{1}{\tau_0} \geq \delta + L_{xx} + \frac{L_{yx}^2}{\alpha}, \qquad \frac{1}{\sigma_0} \geq \delta + \alpha + 2L_{yy}, \tag{22}$$

where $\delta = \min\left\{ \left(\frac{1}{c_\tau} - 1\right)(L_{xx} + L_{yx}^2/\alpha), \left(\frac{1}{c_\sigma} - 1\right)(\alpha + 2L_{yy}) \right\} > 0$. Recall that we define $0^2/0 = 0$. Suppose $z^{\#} = (x^{\#}, y^{\#})$ is a saddle point $\mathcal{L}$ in (1). Let $x = x^{\#}$ and $y = y^{\#}$ in (34), then the following simple observations can be made using (22):

$$Q_k(z^{\#}) \geq \frac{1}{\tau_0} \mathbf{D}_{\mathcal{X}}(x^{\#}, x_k) + (\frac{1}{\sigma_0} - (\alpha + \beta))\mathbf{D}_{\mathcal{Y}}(y^{\#}, y_k)$$

$$\geq \delta \mathbf{D}_{\mathcal{X}}(x^{\#}, x_k) + \delta \mathbf{D}_{\mathcal{Y}}(y^{\#}, y_k) \geq 0, \tag{23a}$$

$$R_{k+1}(z^{\#}) \geq Q_{k+1}(z^{\#}) + \delta \mathbf{D}_{\mathcal{X}}(x_{k+1}, x_k) + \delta \mathbf{D}_{\mathcal{Y}}(y_{k+1}, y_k). \tag{23b}$$

Using $\mathcal{L}(x_{k+1}, y^{\#}) - \mathcal{L}(x^{\#}, y_{k+1}) \geq 0$, the inequalities (34a) and (23b) lead to

$$0 \leq Q_k(z^{\#}) - Q_{k+1}(z^{\#}) - \delta(\mathbf{D}_{\mathcal{X}}(x_{k+1}, x_k) + \mathbf{D}_{\mathcal{Y}}(y_{k+1}, y_k)). \tag{24}$$

Let $a_k = Q_k(z^{\#})$, $b_k = \delta(\mathbf{D}_{\mathcal{X}}(x_{k+1}, x_k) + \mathbf{D}_{\mathcal{Y}}(y_{k+1}, y_k))$, and $c_k = 0$ for $k \geq 0$, then Lemma 3.2 implies that $a \triangleq \lim_{k \to \infty} a_k$ exist. Therefore, (23a) implies that $\{z_k\}$ is a bounded sequence, where $z_k \triangleq (x_k, y_k)$; hence, it has a convergent subsequence $z_{k_n} \to z^*$ as $n \to \infty$ for some $z^* \in \mathcal{X} \times \mathcal{Y}$ where $z^* = (x^*, y^*)$. Lemma 3.2 also implies that $\sum_{k=0}^{\infty} b_k < \infty$; hence, for any $\epsilon > 0$ there exists $N_1$ such that for any $n \geq N_1$, $\max\{\|z_{k_n} - z_{k_n-1}\|, \|z_{k_n} - z_{k_n+1}\|\} < \frac{\epsilon}{2}$. Convergence of $\{z_{k_n}\}$ sequence also implies that there exists $N_2$ such that for any $n \geq N_2$, $\|z_{k_n} - z^*\| < \frac{\epsilon}{2}$. Therefore, letting $N \triangleq \max\{N_1, N_2\}$ we conclude that $\|z_{k_n \pm 1} - z^*\| < \epsilon$, i.e., $z_{k_n \pm 1} \to z^*$ as $n \to \infty$.

Now we show that $z^*$ is indeed a saddle point of (1) by considering the optimality conditions for Line 5 and Line 6 of the APD Algorithm. In particular, for all $n \in \mathbb{Z}_+$, we have $u_n \in$

$\partial f(x_{k_n+1})$ and $v_n \in \partial h(y_{k_n+1})$ where $u_n \triangleq \frac{1}{\tau}\left(\nabla\psi_x(x_{k_n}) - \nabla\psi_x(x_{k_n+1})\right) - \nabla_x\Phi(x_{k_n}, y_{k_n+1})$ and $v_n \triangleq \frac{1}{\sigma}\left(\nabla\psi_y(y_{k_n}) - \nabla\psi_y(y_{k_n+1})\right) + s_{k_n}$. Since $\nabla\psi_x$ and $\nabla\psi_y$ are continuously differentiable on $\mathbf{dom}\, f$ and $\mathbf{dom}\, h$, respectively, it follows from Theorem 24.4 in [26] that $\partial f(x^*) \ni \lim_{n\to\infty} u_n = -\nabla_x\Phi(x^*, y^*)$, $\partial h(y^*) \ni \lim_{n\to\infty} v_n = \nabla_y\Phi(x^*, y^*)$, which implies that $z^*$ is a saddle point of (1).

Finally, since (23) and (24) are true for any saddle point $z^\#$, letting $z^\# = z^*$ and invoking Lemma 3.2 again, one can conclude that $w^* \triangleq \lim_{k\to\infty} w_k \geq 0$ exists, where $w_k \triangleq Q_k(z^*)$, i.e.,

$$w_k = \frac{1}{\tau_0}\mathbf{D}_{\mathcal{X}}(x^*, x_k) + \frac{1}{\sigma_0}\mathbf{D}_{\mathcal{Y}}(y^*, y_k) + \langle q_k, y_k - y^* \rangle + \frac{L_{yx}^2}{\alpha}\mathbf{D}_{\mathcal{X}}(x_k, x_{k-1}) + \frac{L_{yy}^2}{\beta}\mathbf{D}_{\mathcal{Y}}(y_k, y_{k-1}).$$

On the other hand, from $z_{k_n} \to z^*$ and $z_{k_n-1} \to z^*$, we have that $\lim_{n\to\infty} w_{k_n} = 0$; henceforth, $w^* = \lim_{k\to\infty} w_k = \lim_{n\to\infty} w_{k_n} = 0$, and (23a) evaluated at $z^*$ implies that $z_k \to z^*$. $\qquad\square$

# 4 Numerical Experiments

In this section, we test the implementation of our method for solving the kernel matrix learning problem discussed in Section 1 for classification. In particular, given a set of kernel matrices $\{K_\ell\}_{\ell=1}^M \subset \mathbb{S}_+^n$, consider the problem in (6). When $\lambda > 0$ and $C = \infty$, the objective is to find a kernel matrix $K^* \in \mathcal{K} \triangleq \{\sum_{\ell=1}^M \eta_\ell K_\ell : \eta \geq \mathbf{0}\}$ that achieves the best training error for $\ell_2$-norm soft margin SVM, and when $\lambda = 0$ and $C > 0$, the objective is to find a kernel matrix $K^* \in \mathcal{K}$ that gives the best performance for $\ell_1$-norm soft margin SVM. Once $(\alpha^*, \eta^*)$, a saddle point for (6), is computed, one can construct $K^* = \sum_{\ell=1}^M \eta_\ell^* K_\ell$ and predict unlabeled data in the test set using the model $\mathcal{M} : \mathbb{R}^m \to \{-1, +1\}$ such that the predicted label of $\mathbf{a}_i$ is $\mathcal{M}(\mathbf{a}_i) = \mathrm{sign}\left(\sum_{j\in\mathcal{S}} b_j \alpha_j^* K_{ji}^* + \gamma^*\right)$, for all $i \in \mathcal{T}$ where for $\ell_1$ soft margin SVM, $\gamma^* = b_{i^*} - \sum_{j\in\mathcal{S}} b_j \alpha_j^* K_{ji^*}^*$ for some $i^* \in \mathcal{S}$ such that $\alpha_{i^*}^* \in (0, C)$, and for $\ell_2$ soft margin SVM, $\gamma^* = b_{i^*}(1 - \lambda\alpha_{i^*}^*) - \sum_{j\in\mathcal{S}} b_j \alpha_j^* K_{ji^*}^*$ for some $i^* \in \mathcal{S}$ such that $\alpha_{i^*}^* > 0$. Note that (6) is a special case of (1) for $f$, $\Phi$ and $h$ chosen as follows: let $y_\ell \triangleq \frac{\eta_\ell r_\ell}{c}$ for each $\ell$ and define $h(y) = \mathbb{I}_\Delta(y)$ where $y = [y_\ell]_{\ell=1}^M \in \mathbb{R}^M$ and $\Delta$ is an $M$-dimensional unit simplex; $\Phi(x, y) = -2\mathbf{e}^\top x + \sum_{\ell=1}^M \frac{c}{r_\ell} y_\ell x^\top G(K_\ell^{tr}) x + \lambda \|x\|_2^2$, and $f(x) = \mathbb{I}_X(x)$ where $X = \{x \in \mathbb{R}^{n_{tr}} : 0 \leq x \leq C, \ \langle \mathbf{b}, x \rangle = 0\}$. In the rest, we consider the following equivalent

reformulation of $\ell_2$-norm soft margin problem:

$$\min_{\substack{x \geq 0 \\ \langle \mathbf{b}, x \rangle = 0}} \max_{y \in \Delta} -2x^\top \mathbf{e} + \sum_{\ell=1}^{M} \frac{c}{r_\ell} \, y_\ell \, x^\top G(K_\ell^{tr})x + \lambda \, \|x\|_2^2. \tag{25}$$

Additional numerical experiments for $\ell_1$-norm Soft Margin SVM is given in Section 6.2 in the appendix.

## 4.1 APD vs Mirror-prox for soft-margin SVMs

In this experiment we compared our method against **Mirror-prox**, the primal-dual algorithm proposed by He et al. [15]. We used four different data sets available in UCI repository: `Ionosphere` (351 observations, 33 features), `Sonar` (208 observations, 60 features), `Heart` (270 observations, 13 features) and `Breast-Cancer` (608 observations, 9 features) with three given kernel functions ($M = 3$); polynomial kernel function $k_1(\mathbf{a}, \bar{\mathbf{a}}) = (1 + \mathbf{a}^\top \bar{\mathbf{a}})^2$, Gaussian kernel function $k_2(\mathbf{a}, \bar{\mathbf{a}}) = \exp(-0.5(\mathbf{a} - \bar{\mathbf{a}})^\top (\mathbf{a} - \bar{\mathbf{a}})/0.1)$, and linear kernel function $k_3(\mathbf{a}, \bar{\mathbf{a}}) = \mathbf{a}^\top \bar{\mathbf{a}}$ to compute $K_1, K_2, K_3$ respectively. All the data sets are normalized such that each feature column is mean-centered and divided by its standard deviation. For $\ell_2$-norm soft margin we set $\lambda = 1$, for $\ell_1$-norm soft margin SVM we set $C = 1$ and for both SVMs $c = \sum_{\ell=1}^{3} r_\ell$, where $r_\ell = \text{trace}(K_\ell)$ for $\ell = 1, 2, 3$. The kernel matrices are normalized as in [20]; thus, $\mathbf{diag}(K_\ell) = \mathbf{1}$ and $r_\ell = n_{tr} + n_t$ for each $\ell$.

We tested four different implementations of the APD algorithm: we will refer to the constant step version of APD, stated in Part I of the main result in Theorem 2.1, as **APD1**; and we refer to the adaptive step version of APD, stated in Part II of the main result, as **APD2**. Finally, we also implemented a variant of **APD2** with periodic restarts, and we call it **APD2-restart**. The **APD2-restart** method is implemented simply by restarting the algorithm periodically after every 500 iterations and using the most current iterate as the initial solution for the next call of **APD2**. All the algorithms are initialized from $x_0 = \mathbf{0}$ and $y_0 = \frac{1}{M}\mathbf{1}$.

The results reported are the average values over 10 random replications. In each replication, %80 of the dataset is selected uniformly at random, and used for training; the rest of data set (%20) is reserved as test data to calculate the test set accuracy (TSA), which is defined as the fraction of the correctly labeled data in the test set. All experiments are performed on a machine running 64-bit Windows 10 with Intel i7-8650U @2.11GHz and 16GB RAM. The algorithms are

compared in terms of relative error for the function value $(|\mathcal{L}(x_k, y_k) - \mathcal{L}^*|/|\mathcal{L}^*|)$ and for the solution $(\|x_k - x^*\|_2 / \|x^*\|_2)$, where $(x^*, y^*)$ denotes a saddle point for the problem of interest, i.e., (37) or (25), and $\mathcal{L}^* \triangleq \mathcal{L}(x^*, y^*)$. To compute $(x^*, y^*)$, we called MOSEK through CVX [13].

### 4.1.1 $\ell_2$-norm Soft Margin SVM

Since (25) is strongly convex in $x$ and linear in $y$, we implement both **APD2** and **APD2-restart** methods in addition to **APD1**. Due to strong convexity, $\|x^*\|_2$ can be bounded depending on $\lambda > 0$ and the Lipschitz constants can be computed similarly as in $\ell_1$-norm soft margin problem. In these experiments on $\ell_2$ soft margin problems, **APD2-restart** outperformed all other methods on all four data sets. In particular, **APD1**, **APD2**, **APD2-restart** and **Mirror-prox** are compared in terms of relative errors for function value and for solution in Figures 1 and 2 respectively. Given an accuracy level, both **APD1** and **APD2-restart** can compute a solution with a given accuracy requiring much fewer iterations than **Mirror-prox** needs. In addition, we observed that for fixed number of iterations the run time for **Mirror-prox** is almost twice the run time for any APD implementation – see Table 1, and for runtime comparison on larger size problems, see Section 4.2. Interpreting the results in Figures 1 and 2, and computational time, we conclude that APD implementations can compute a solution with a given accuracy in a significantly lower time than **Mirror-prox** requires. For instance, consider the results for Sonar data set in Figure 1, to compute a solution with $|\mathcal{L}(x_k, y_k) - \mathcal{L}^*|/|\mathcal{L}^*| < 10^{-6}$, **APD2-restart** requires 1000 iterations, on the other hand, **Mirror-prox** needs around 2000 iterations; hence, **APD2-restart** can compute it in 1/4 of the run time for **Mirror-prox**. This effect is more apparent when these methods are compared on larger scale problems, e.g., see Figure 3b.

| Method | Iteration #<br>Data Set | K=1000<br>Time | Rel.<br>error | TSA | K=1500<br>Time | Rel.<br>error | TSA | K=2000<br>Time | Rel.<br>error | TSA | K=2500<br>Time | Rel.<br>error | TSA |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **APD1** | Ionosphere | 0.29 | 6.2e-07 | 93.9 | 0.45 | 1.6e-06 | 93.9 | 0.60 | 1.6e-06 | 93.9 | 0.75 | 1.6e-06 | 93.9 |
| | Sonar | 0.15 | 8.3e-05 | 82.6 | 0.23 | 1.3e-06 | 82.4 | 0.31 | 2.3e-08 | 82.1 | 0.40 | 3.6e-10 | 82.1 |
| | Heart | 0.26 | 3.0e-11 | 83.0 | 0.38 | 3.0e-11 | 83.0 | 0.50 | 3.0e-11 | 83.0 | 0.62 | 3.0e-11 | 83.0 |
| | Breast-Cancer | 0.76 | 7.5e-05 | 96.9 | 1.22 | 4.4e-06 | 96.9 | 1.65 | 4.4e-07 | 96.9 | 2.08 | 5.5e-08 | 96.9 |
| **APD2** | Ionosphere | 0.32 | 1.6e-06 | 93.9 | 0.48 | 1.6e-06 | 93.9 | 0.65 | 1.6e-06 | 93.9 | 0.82 | 1.6e-06 | 93.9 |
| | Sonar | 0.15 | 4.1e-06 | 82.4 | 0.23 | 2.0e-07 | 82.1 | 0.31 | 9.5e-09 | 82.1 | 0.38 | 9.4e-10 | 82.1 |
| | Heart | 0.23 | 4.5e-11 | 83.0 | 0.34 | 3.3e-11 | 83.0 | 0.45 | 3.1e-11 | 83.0 | 0.57 | 3.1e-11 | 83.0 |
| | Breast-Cancer | 0.82 | 4.9e-06 | 96.9 | 1.23 | 7.9e-07 | 96.9 | 1.66 | 2.4e-07 | 96.9 | 2.08 | 9.3e-08 | 96.9 |
| **APD2-restart** | Ionosphere | 0.33 | 1.6e-06 | 93.9 | 0.49 | 1.6e-06 | 93.9 | 0.64 | 1.6e-06 | 93.9 | 0.80 | 1.6e-06 | 93.9 |
| | Sonar | 0.15 | 1.0e-06 | 82.1 | 0.23 | 2.1e-08 | 82.1 | 0.32 | 6.5e-11 | 82.1 | 0.40 | 9.9e-12 | 82.1 |
| | Heart | 0.23 | 3.0e-11 | 83.0 | 0.35 | 3.0e-11 | 83.0 | 0.47 | 3.0e-11 | 83.0 | 0.58 | 3.0e-11 | 83.0 |
| | Breast-Cancer | 0.80 | 6.9e-07 | 96.9 | 1.23 | 1.7e-08 | 96.9 | 1.67 | 5.7e-10 | 96.9 | 2.10 | 7.2e-11 | 96.9 |
| **Mirror-prox** | Ionosphere | 0.60 | 1.6e-06 | 93.9 | 0.89 | 1.6e-06 | 93.9 | 1.20 | 1.6e-06 | 93.9 | 1.51 | 1.6e-06 | 93.9 |
| | Sonar | 0.31 | 5.5e-04 | 82.4 | 0.45 | 1.4e-05 | 82.4 | 0.61 | 6.9e-07 | 82.1 | 0.76 | 2.1e-08 | 82.1 |
| | Heart | 0.45 | 3.0e-11 | 83.0 | 0.68 | 3.0e-11 | 83.0 | 0.91 | 3.0e-11 | 83.0 | 1.14 | 3.0e-11 | 83.0 |
| | Breast-Cancer | 1.48 | 2.0e-04 | 96.9 | 2.35 | 1.4e-05 | 96.9 | 3.23 | 1.6e-06 | 96.9 | 4.07 | 2.5e-07 | 96.9 |

Table 1: $\ell_2$-norm soft margin: runtime (sec), relative error for $\mathcal{L}(x_k, y_k)$ and TSA (%) for **APD1**, **APD2**, **APD2-restart** and **Mirror-prox** at iteration $K \in \{1000, 1500, 2000, 2500\}$.
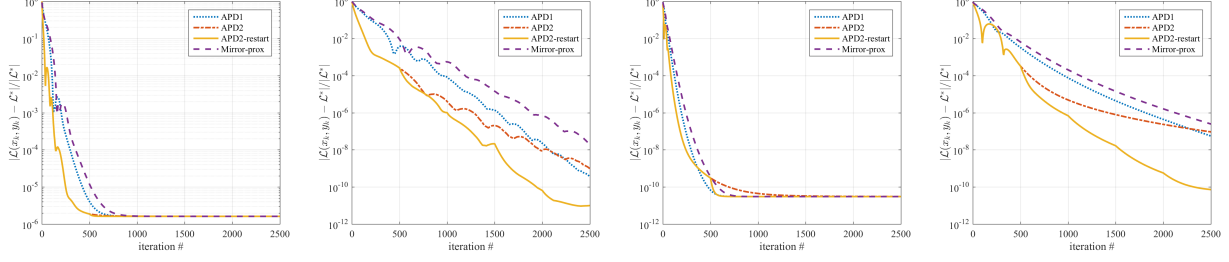
Figure 1: $\ell_2$-norm soft margin: Comparison of **APD1**, **APD2**, **APD2-restart** and **Mirror-prox** in terms of rel. error for $\mathcal{L}(x_k, y_k)$. From left to right: `Ionosphere`, `Sonar`, `Heart`, `Breast-Cancer`.
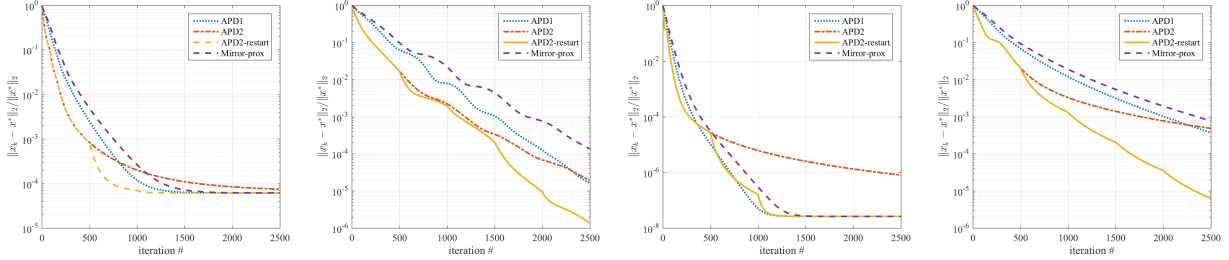


Figure 2: $\ell_2$-norm soft margin: Comparison of **APD1**, **APD2**, **APD2-restart** and **Mirror-prox** in terms of rel. error for $x_k$. From left to right: `Ionosphere`, `Sonar`, `Heart`, `Breast-Cancer`.

## 4.2 APD vs off-the-shelf interior point methods

In this section we compare time complexity of our methods against widely-used, open-source interior point method (IPM) solvers **Sedumi** v4.0 and **SDPT3** v1.3. Here we used two data sets: `Breast-Cancer` available in UCI repository and `SIDO0` [14] (6339 observations, 4932 features – we used half of the observations in data set). The goal is to investigate how the quality of iterates, measured in terms of relative solution error $\|x_k - x^*\|_2 / \|x^*\|_2$, changes as run time progresses. Let $x^*$ be the solution of problem (25) which we computed using MOSEK in CVX with the best accuracy option. To compare **APD1**, **APD2**, **APD2-restart** and **Mirror-prox** against the interior point methods **Sedumi** and **SDPT3**, the problem in (25) is first solved by **Sedumi** and **SDPT3** using their default setting. Let $t_1$ and $t_2$ denote the run time of **Sedumi** and **SDPT3** in seconds, respectively. Next, the primal-dual methods **APD1**, **APD2**, **APD2-restart** and **Mirror-prox** were run with the same settings as in Section 4.1.1 for $\max\{t_1, t_2\}$ seconds. The mean of relative solution error of the iterates over 10 replications are plotted against time (seconds) for each of these methods in Figure 3. **Sedumi** and **SDPT3** are second-order methods and have much better theoretical convergence rates compared to the first-order primal dual methods proposed in this paper. However, for large-scale machine learning problems with dense data, the work per iteration

of an IPM is significantly more than that of a first-order method; hence, if the objective is to attain low-to-medium level accuracy solutions, then first-order methods are better suited for this task, e.g., in Figure 3b, **APD2-restart** iterates are more accurate than **Sedumi** and **SDPT3** for the first 2000 and 1000 seconds respectively.
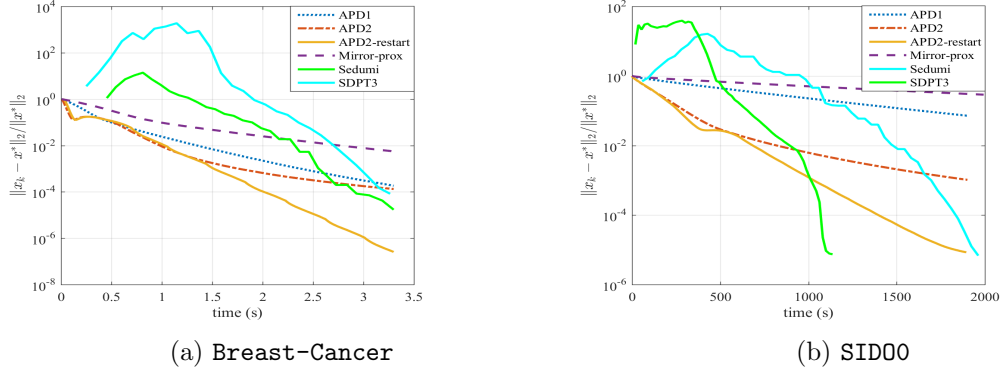


(a) `Breast-Cancer`

(b) `SIDO0`

Figure 3: Run time comparison of **APD1**, **APD2**, **APD2-restart**, **Mirror-prox**, **Sedumi**, and **SDPT3** in terms of relative error for $x_k$ on $\ell_2$ soft margin problems.

## 5 Conclusions

We proposed a primal-dual algorithm with a novel momentum term based on gradient extrapolation to solve saddle point problems defined by a convex-concave function $\mathcal{L}(x, y) = f(x) + \Phi(x, y) - h(y)$ with a general coupling term $\Phi(x, y)$ that is *not* assumed to be bilinear. Assuming $\nabla_y \Phi(\cdot, \cdot)$ is Lipschitz and $\nabla_x \Phi(\cdot, y)$ is Lipschitz for any fixed $y$, we derive error bounds in terms of $\mathcal{L}(\bar{x}_K, y^*) - \mathcal{L}(x^*, \bar{y}_K)$ for the ergodic sequence – without requiring primal-dual domains to be bounded; in particular, we show $\mathcal{O}(1/K)$ rate when the problem is merely convex in $x$ using a constant step-size rule. Furthermore, assuming $\Phi(x, \cdot)$ is linear in $y$ for each fixed $x$ and $f$ is strongly convex, we obtain the ergodic convergence rate of $\mathcal{O}(1/K^2)$. To best of our knowledge, this is the first time $\mathcal{O}(1/K^2)$ rate is shown for a single-loop method for solving a saddle point problem when $\Phi$ is not bilinear. Our new method captures a wide range of optimization problems, especially, those that can be cast as an SDP or a QCQP, or more generally conic convex problems with nonlinear constraints. It has been illustrated in the numerical experiments that the APD method can compete against second-order methods when we aim at a low to medium level accuracy. Moreover, APD with single-loop has roughly the same iteration complexity as proximal Mirror-Prox with double-loop iterations; on the other hand, it requires half the computational efforts needed by Mirror-prox.

# References

[1] M. ApS, *The MOSEK optimization toolbox for MATLAB manual. Version 8.1.*, 2017.

[2] K. J. Arrow, L. Hurwicz, H. Uzawa, and H. B. Chenery, *Studies in linear and non-linear programming*, (1958).

[3] N. S. Aybat and E. Y. Hamedani, *A distributed admm-like method for resource sharing under conic constraints over time-varying networks*, arXiv preprint arXiv:1611.07393, (2016).

[4] N. S. Aybat, Z. Wang, T. Lin, and S. Ma, *Distributed linearized alternating direction method of multipliers for composite convex consensus optimization*, IEEE Transactions on Automatic Control, 63 (2018), pp. 5–20.

[5] C. Bhattacharyya, P. K. Shivaswamy, and A. J. Smola, *A second order cone programming formulation for classifying missing data*, in Advances in neural information processing systems, 2005, pp. 153–160.

[6] S. Boyd and L. Vandenberghe, *Convex optimization*, Cambridge university press, 2004.

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

[8] ——, *On the ergodic convergence rates of a first-order primal–dual algorithm*, Mathematical Programming, 159 (2016), pp. 253–287.

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

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

[11] S. S. Du and W. Hu, *Linear convergence of the primal-dual gradient method for convex-concave saddle point problems without strong convexity*, arXiv preprint arXiv:1802.01504, (2018).

[12] M. Gönen and E. Alpaydin, *Multiple kernel learning algorithms*, Journal of machine learning research, 12 (2011), pp. 2211–2268.

[13] M. Grant, S. Boyd, and Y. Ye, *Cvx: Matlab software for disciplined convex programming*, 2008.

[14] I. Guyon, *Sido: A phamacology dataset*, URL: http://www. causality. inf. ethz. ch/data/SIDO. html, (2008).

[15] N. He, A. Juditsky, and A. Nemirovski, *Mirror prox algorithm for multi-term composite minimization and semi-separable problems*, Computational Optimization and Applications, 61 (2015), pp. 275–319.

[16] Y. He and R. D. Monteiro, *An accelerated hpe-type algorithm for a class of composite convex-concave saddle-point problems*, SIAM Journal on Optimization, 26 (2016), pp. 29–56.

[17] J.-B. Hiriart-Urruty and C. Lemaréchal, *Fundamentals of convex analysis*, Springer Science & Business Media, 2012.

[18] A. Juditsky and A. Nemirovski, *First order methods for nonsmooth convex large-scale optimization, ii: utilizing problems structure*, Optimization for Machine Learning, (2011), pp. 149–183.

[19] O. Kolossoski and R. Monteiro, *An accelerated non-euclidean hybrid proximal extragradient-type algorithm for convex–concave saddle-point problems*, Optimization Methods and Software, 32 (2017), pp. 1244–1272.

[20] G. R. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. I. Jordan, *Learning the kernel matrix with semidefinite programming*, Journal of Machine learning research, 5 (2004), pp. 27–72.

[21] Q. Lin, S. Nadarajah, and N. Soheili, *A level-set method for convex optimization with a feasible solution path*, http://www.optimization-online.org/DB_HTML/2017/10/6274.html, (2017).

[22] A. Nemirovski, *Prox-method with rate of convergence $\mathcal{O}(1/t)$ for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems*, SIAM Journal on Optimization, 15 (2004), pp. 229–251.

[23] Y. Nesterov and A. Nemirovskii, *Interior-point polynomial algorithms in convex programming*, vol. 13, Siam, 1994.

[24] B. Palaniappan and F. Bach, *Stochastic variance reduction methods for saddle-point problems*, in Advances in Neural Information Processing Systems, 2016, pp. 1416–1424.

[25] H. Robbins and D. Siegmund, *Optimizing methods in statistics (Proc. Sympos., Ohio State Univ., Columbus, Ohio, 1971)*, New York: Academic Press, 1971, ch. A convergence theorem for non negative almost supermartingales and some applications, pp. 233 – 257.

[26] R. T. Rockafellar, *Convex analysis*, Princeton university press, 1997.

[27] R. Shefi and M. Teboulle, *Rate of convergence analysis of decomposition methods based on the proximal method of multipliers for convex minimization*, SIAM Journal on Optimization, 24 (2014), pp. 269–297.

[28] P. K. SHIVASWAMY AND T. JEBARA, *Ellipsoidal machines*, in Artificial Intelligence and Statistics, 2007, pp. 484–491.

[29] P. TSENG, *On accelerated proximal gradient methods for convex-concave optimization.* Available at `http://www.mit.edu/~dimitrib/PTseng/papers/apgm.pdf`, 2008.

[30] J. WANG AND L. XIAO, *Exploiting strong convexity from data with primal-dual first-order algorithms*, arXiv preprint arXiv:1703.02624, (2017).

[31] E. P. XING, M. I. JORDAN, S. J. RUSSELL, AND A. Y. NG, *Distance metric learning with application to clustering with side-information*, in Advances in neural information processing systems, 2003, pp. 521–528.

[32] Y. XU, *First-order methods for constrained convex programming based on linearized augmented lagrangian function*, arXiv preprint arXiv:1711.08020, (2017).

# 6    Appendix

**Lemma 6.1.** *Let $\mathcal{X}$ be a finite dimensional normed vector space with norm $\|.\|_{\mathcal{X}}$, $f : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ be a closed convex function with convexity modulus $\mu \geq 0$ w.r.t. $\|.\|_{\mathcal{X}}$, and $\mathbf{D} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ be a Bregman distance function corresponding to a strictly convex function $\phi : \mathcal{X} \to \mathbb{R}$ that is differentiable on an open set containing $\mathbf{dom}\, f$. Given $\bar{x} \in \mathbf{dom}\, f$ and $t > 0$, let*

$$x^+ = \operatorname*{argmin}_{x \in \mathcal{X}} f(x) + t\mathbf{D}(x, \bar{x}). \tag{26}$$

*Then for all $x \in \mathcal{X}$, the following inequality holds:*

$$f(x) + t\mathbf{D}(x, \bar{x}) \geq f(x^+) + t\mathbf{D}(x^+, \bar{x}) + t\mathbf{D}(x, x^+) + \frac{\mu}{2} \left\| x - x^+ \right\|_{\mathcal{X}}^2. \tag{27}$$

*Proof.* This result is a trivial extension of Property 1 in [29]. The first-order optimality condition for (26) implies that $0 \in \partial f(x^+) + t\nabla_x \mathbf{D}(x^+, \bar{x})$ – here $\nabla_x \mathbf{D}$ denotes the partial gradient with respect to the first argument. Note that for any $x \in \mathbf{dom}\, f$, we have $\nabla_x \mathbf{D}(x, \bar{x}) = \nabla \phi(x) - \nabla \phi(\bar{x})$. Hence, $t(\nabla \phi(\bar{x}) - \nabla \phi(x^+)) \in \partial f(x^+)$. Using the convexity inequality for $f$, we get

$$f(x) \geq f(x^+) + t \left\langle \nabla \phi(\bar{x}) - \nabla \phi(x^+),\ x - x^+ \right\rangle + \frac{\mu}{2} \left\| x - x^+ \right\|_{\mathcal{X}}^2.$$

The result in (27) immediately follows from this inequality. $\qquad \square$

## 6.1    Proof of Theorem 3.1

For $k \geq 0$, the optimality conditions for the $y$- and $x$-subproblems in Line 5 and Line 6 of APD algorithm imply two inequalities that hold for any $y \in \mathcal{Y}$ and $x \in \mathcal{X}$:

$$h(y_{k+1}) - \langle s_k,\ y_{k+1} - y \rangle \leq h(y) + \frac{1}{\sigma_k} \Big[ \mathbf{D}_{\mathcal{Y}}(y, y_k) - \mathbf{D}_{\mathcal{Y}}(y, y_{k+1}) - \mathbf{D}_{\mathcal{Y}}(y_{k+1}, y_k) \Big], \tag{28}$$

$$f(x_{k+1}) + \langle \nabla_x \Phi(x_k, y_{k+1}), x_{k+1} - x \rangle + \frac{\mu}{2} \| x - x_{k+1} \|_{\mathcal{X}}^2$$
$$\leq f(x) + \frac{1}{\tau_k} \Big[ \mathbf{D}_{\mathcal{X}}(x, x_k) - \mathbf{D}_{\mathcal{X}}(x, x_{k+1}) - \mathbf{D}_{\mathcal{X}}(x_{k+1}, x_k) \Big]. \tag{29}$$

These inequalities follow from the definition of a Bregman function – see Lemma 6.1 for details.

Let $A_{k+1} \triangleq \frac{1}{\sigma_k} \mathbf{D}_{\mathcal{Y}}(y, y_k) - \frac{1}{\sigma_k} \mathbf{D}_{\mathcal{Y}}(y, y_{k+1}) - \frac{1}{\sigma_k} \mathbf{D}_{\mathcal{Y}}(y_{k+1}, y_k)$ and $B_{k+1} \triangleq \frac{1}{\tau_k} \mathbf{D}_{\mathcal{X}}(x, x_k) - \frac{1}{\tau_k} \mathbf{D}_{\mathcal{X}}(x, x_{k+1})$
$- \frac{1}{\tau_k} \mathbf{D}_{\mathcal{X}}(x_{k+1}, x_k) - \frac{\mu}{2} \|x - x_{k+1}\|_{\mathcal{X}}^2$ for $k \geq 0$. Using convexity of $\Phi(x, y_{k+1})$ in $x$ and the bound in (8), the following inequality follows from (29),

$$f(x_{k+1}) + \Phi(x_{k+1}, y_{k+1}) \leq f(x) + \Phi(x, y_{k+1}) + \frac{L_{xx}}{2} \|x_{k+1} - x_k\|_{\mathcal{X}}^2 + B_{k+1}. \tag{30}$$

For $k \geq 0$, let $q_k \triangleq \nabla_y \Phi(x_k, y_k) - \nabla_y \Phi(x_{k-1}, y_{k-1})$, which implies that $s_k = \nabla_y \Phi(x_k, y_k) + \theta_k q_k$ – see Line 4 of APD algorithm. Now summing (28) and (30) and rearranging the terms lead to

$$\mathcal{L}(x_{k+1}, y) - \mathcal{L}(x, y_{k+1})$$
$$\leq \Phi(x_{k+1}, y) - \Phi(x_{k+1}, y_{k+1}) + \langle s_k, y_{k+1} - y \rangle + \frac{L_{xx}}{2} \|x_{k+1} - x_k\|_{\mathcal{X}}^2 + A_{k+1} + B_{k+1}$$
$$\leq - \langle q_{k+1}, y_{k+1} - y \rangle + \theta_k \langle q_k, y_{k+1} - y \rangle + \frac{L_{xx}}{2} \|x_{k+1} - x_k\|_{\mathcal{X}}^2 + A_{k+1} + B_{k+1} \tag{31}$$

for $k \geq 0$, where in the last inequality we use the concavity of $\Phi(x_{k+1}, y)$ in $y$. We can rewrite the bound in (31) as

$$\mathcal{L}(x_{k+1}, y) - \mathcal{L}(x, y_{k+1}) \leq$$
$$\left[ \frac{1}{\tau_k} \mathbf{D}_{\mathcal{X}}(x, x_k) + \frac{1}{\sigma_k} \mathbf{D}_{\mathcal{Y}}(y, y_k) + \theta_k \langle q_k, y_k - y \rangle \right] - \left[ \frac{1}{\tau_k} \mathbf{D}_{\mathcal{X}}(x, x_{k+1}) + \frac{\mu}{2} \|x - x_{k+1}\|_{\mathcal{X}}^2 \right.$$
$$+ \frac{1}{\sigma_k} \mathbf{D}_{\mathcal{Y}}(y, y_{k+1}) + \langle q_{k+1}, y_{k+1} - y \rangle + \frac{1}{\tau_k} \mathbf{D}_{\mathcal{X}}(x_{k+1}, x_k) + \frac{1}{\sigma_k} \mathbf{D}_{\mathcal{Y}}(y_{k+1}, y_k)$$
$$\left. - \frac{L_{xx}}{2} \|x_k - x_{k+1}\|_{\mathcal{X}}^2 \right] + \theta_k \underbrace{\langle q_k, y_{k+1} - y_k \rangle}_{(*)}. \tag{32}$$

The term $(*)$ in (32) can be bounded using the fact that for any $y \in \mathcal{Y}$, $y' \in \mathcal{Y}^*$, and $\eta > 0$, we have $|\langle y', y \rangle| \leq \frac{\eta}{2} \|y\|_{\mathcal{Y}}^2 + \frac{1}{2\eta} \|y'\|_{\mathcal{Y}^*}^2$. Indeed, one can bound $\langle q_k, y - y_k \rangle$ for any given $y \in \mathcal{Y}$ as follows: after adding and subtracting $\nabla_y \Phi(x_{k-1}, y_k)$ to $q_k$, using the previous inequality twice together with (10) and inequalities $\mathbf{D}_{\mathcal{X}}(x, \bar{x}) \geq \frac{1}{2} \|x - \bar{x}\|_{\mathcal{X}}^2$ and $\mathbf{D}_{\mathcal{Y}}(y, \bar{y}) \geq \frac{1}{2} \|y - \bar{y}\|_{\mathcal{Y}}^2$, one can obtain:

$$\langle q_k, y - y_k \rangle \leq \alpha_k \mathbf{D}_{\mathcal{Y}}(y, y_k) + \frac{L_{yx}^2}{\alpha_k} \mathbf{D}_{\mathcal{X}}(x_k, x_{k-1}) \tag{33a}$$
$$+ \beta_k \mathbf{D}_{\mathcal{Y}}(y, y_k) + \frac{L_{yy}^2}{\beta_k} \mathbf{D}_{\mathcal{Y}}(y_k, y_{k-1}), \tag{33b}$$

which is true for any $\alpha_k, \beta_k > 0$. Moreover, if $L_{yy} = 0$, then the bound in (33a) is valid and the term in (33b) becomes 0; hence, (33) holds for any $\alpha_k > 0$ and $\beta_k = 0$ when $L_{yy} = 0$.

Let $z = (x, y)$, using (33) with $y = y_{k+1}$ within (32), we obtain the following inequality for $k \geq 0$:

$$\mathcal{L}(x_{k+1}, y) - \mathcal{L}(x, y_{k+1}) \leq Q_k(z) - R_{k+1}(z), \tag{34a}$$

$$Q_k(z) \triangleq \left[ \frac{1}{\tau_k} \mathbf{D}_\mathcal{X}(x, x_k) + \frac{1}{\sigma_k} \mathbf{D}_\mathcal{Y}(y, y_k) + \theta_k \langle q_k, y_k - y \rangle + \frac{\theta_k L_{yx}^2}{\alpha_k} \mathbf{D}_\mathcal{X}(x_k, x_{k-1}) \right.$$

$$\left. + \frac{\theta_k L_{yy}^2}{\beta_k} \mathbf{D}_\mathcal{Y}(y_k, y_{k-1}) \right] \tag{34b}$$

$$R_{k+1}(z) \triangleq \left[ \frac{1}{\tau_k} \mathbf{D}_\mathcal{X}(x, x_{k+1}) + \frac{\mu}{2} \|x - x_{k+1}\|_\mathcal{X}^2 + \frac{1}{\sigma_k} \mathbf{D}_\mathcal{Y}(y, y_{k+1}) + \langle q_{k+1}, y_{k+1} - y \rangle \right.$$

$$\left. + \left( \frac{1}{\tau_k} - L_{xx} \right) \mathbf{D}_\mathcal{X}(x_{k+1}, x_k) + \left( \frac{1}{\sigma_k} - \theta_k(\alpha_k + \beta_k) \right) \mathbf{D}_\mathcal{Y}(y_{k+1}, y_k) \right]. \tag{34c}$$

All the derivations until here, including (34), hold for any Bregman distance function $\mathbf{D}_\mathcal{X}$. Recall that if $\mu > 0$, then we set $\mathbf{D}_\mathcal{X}(x, \bar{x}) = \frac{1}{2} \|x - \bar{x}\|_\mathcal{X}^2$ and $\|x\|_\mathcal{X} = \sqrt{\langle x, x \rangle}$. Now, multiplying both sides by $t_k$, summing over $k = 0$ to $K - 1$, and then using Jensens's inequality, we obtain

$$N_K(\mathcal{L}(\bar{x}_K, y) - \mathcal{L}(x, \bar{y}_K)) \leq \sum_{k=0}^{K-1} t_k (Q_k(z) - R_{k+1}(z)) \leq Q_0(z) - t_{K-1} P_{K-1}(z),$$

where the last inequality follows from the step-size conditions in (19), which imply that $t_{k+1} Q_{k+1}(z) - t_k R_{k+1}(z) \leq 0$ for $k = 0$ to $K - 2$. Therefore, we get

$$N_K(\mathcal{L}(\bar{x}_K, y) - \mathcal{L}(x, \bar{y}_K)) \leq$$

$$\frac{1}{\tau_0} \mathbf{D}_\mathcal{X}(x, x_0) + \frac{1}{\sigma_0} \mathbf{D}_\mathcal{Y}(y, y_0) + t_K \theta_K \overbrace{\langle q_K, y - y_K \rangle}^{(**)} - t_K \left[ \frac{1}{\tau_K} \mathbf{D}_\mathcal{X}(x, x_K) \right.$$

$$\left. + \frac{1}{\sigma_K} \mathbf{D}_\mathcal{Y}(y, y_K) + \theta_K \frac{L_{yx}^2}{\alpha_K} \mathbf{D}_\mathcal{X}(x_K, x_{K-1}) + \theta_K \frac{L_{yy}^2}{\beta_K} \mathbf{D}_\mathcal{Y}(y_K, y_{K-1}) \right], \tag{35}$$

where we used the fact that initializing $x_0 = x_{-1}$ and $y_0 = y_{-1}$ implies $q_0 = \mathbf{0}$ and $\mathbf{D}_\mathcal{X}(x_0, x_{-1}) = \mathbf{D}_\mathcal{X}(y_0, y_{-1}) = 0$. Consider (33) with $k = K$, then we get an upper bound on $(**)$. Now we can

combine this upper bound with (35) to get

$$N_K(\mathcal{L}(\bar{x}_K, y) - \mathcal{L}(x, \bar{y}_K)) \leq$$
$$\Delta(x, y) - t_K \Big[ \frac{1}{\tau_K} \mathbf{D}_{\mathcal{X}}(x, x_K) + \Big( \frac{1}{\sigma_K} - \theta_K(\alpha_K + \beta_K) \Big) \mathbf{D}_{\mathcal{Y}}(y, y_K) \Big], \tag{36}$$

where in the last inequality we used the step-size condition (19a) for $k = K$. The result in (20) can be obtained by dropping the second term in (36), which is non-negative. Finally, suppose $(x^*, y^*)$ is a saddle point of (1), the result in (21) can be obtained immediately by letting $x = x^*$ and $y = y^*$ in (36) and using the fact that $\mathcal{L}(\bar{x}_K, y^*) - \mathcal{L}(x^*, \bar{y}_K) \geq 0$. $\qquad\square$

## 6.2 Additional numerical experiments: $\ell_1$-norm Soft Margin SVM

Consider the following equivalent reformulation of $\ell_1$-norm soft margin problem:

$$\min_{\substack{0 \leq x \leq C \\ \langle \mathbf{b}, x \rangle = 0}} \max_{y \in \Delta} -2x^\top \mathbf{e} + \sum_{\ell=1}^{M} \frac{c}{r_\ell} \, y_\ell \, x^\top G(K_\ell^{tr}) x, \tag{37}$$

Note that (37) is merely convex in $x$ and linear in $y$; therefore, we only used **APD1**. Let $\|\cdot\|$ denote the spectral norm; the Lipschitz constants defined in (9) and (10) can be set as $L_{xx} \triangleq 6 \max_{\ell=1,2,3}\{\|G(K_\ell)\|\}$, $L_{yy} = 0$, $L_{yx} \triangleq 6\sqrt{3}C \max_{\ell=1,2,3}\{\|G(K_\ell)\|\}$. Recall that the stepsize of **Mirror-prox** is determined by Lipschitz constant $L$ of $\nabla\Phi$, and which can be set as $L = \sqrt{L_{xx}^2 + L_{xy}^2 + L_{yx}^2 + L_{yy}^2}$, where $L_{xy}$ is defined similarly as $L_{yx}$ in (10), and for (37) one can take $L_{xy} = L_{yx}$. Except for `Sonar` dataset, the Lipschitz constants are multiplied by 0.1.

In these experiments on $\ell_1$ soft margin problems, **APD1** outperformed **Mirror-prox** on all four data sets. In particular, **APD1** and **Mirror-prox** are compared in terms of relative errors for function value and for solution in Figures 4 and 5 respectively. In these figures, relative errorr are plotted against the number of iterations. In this experiment we observed that for fixed number of iterations $K$, the run time for **Mirror-prox** is at least twice of **APD1** run time – while APD requires one primal-dual prox operation, **Mirror-prox** needs *two* primal-dual prox operations at each iteration. It is worth mentioning that for `Ionosphere` and `Sonar` data sets, 1500 iterations of **APD1** took roughly the same time as MOSEK required to solve the problem, and within 1500 iterations **APD1** was able to generate a decent approximate solution with relative error less than

$10^{-4}$ and with a high TSA value; on the other hand, **Mirror-prox** was not able to produce such good quality solutions in a similar amount of time. The average TSA of the optimal solution $(x^*, y^*)$, computed by MOSEK, are 93.81, 84.76, 84.07, 96.79 percent for `Ionosphere`, `Sonar`, `Heart` and `Breast-Cancer` data sets, respectively. Note that TSA is not necessarily increasing in the number of iterations $K$, e.g., **APD1** iterates at $K = 1000$ and $K = 2000$ have 85.95% and 84.76% TSA values, respectively, for `Sonar` data set – note the optimal solution $(x^*, y^*)$ has 84.76% TSA – see Table 2 for details. This is a well-known phenomenon and is related to over fitting; in particular, the model's ability to generalize can weaken as it begins to overfit the training data.

| Iteration # | | K=1000 | | | K=1500 | | | K=2000 | | | K=2500 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | Data Set | Time | Rel. subopt. | TSA | Time | Rel. error | TSA | Time | Rel. error | TSA | Time | Rel. error | TSA |
| **APD1** | Ionosphere | 0.34 | 5.6e-05 | 93.80 | 0.50 | 9.3e-06 | 93.80 | 0.67 | 1.6e-06 | 93.80 | 0.84 | 3.6e-07 | 93.80 |
| | Sonar | 0.16 | 4.6e-04 | 85.95 | 0.24 | 4.1e-05 | 84.52 | 0.32 | 2.1e-06 | 84.76 | 0.41 | 9.7e-08 | 84.76 |
| | Heart | 0.31 | 1.1e-06 | 83.89 | 0.46 | 3.6e-07 | 83.89 | 0.60 | 1.1e-07 | 83.89 | 0.75 | 3.6e-08 | 83.88 |
| | Breast-cancer | 0.98 | 5.5e-03 | 96.86 | 1.59 | 1.0e-03 | 96.79 | 2.19 | 2.2e-04 | 96.72 | 2.79 | 6.3e-05 | 96.71 |
| **Mirror-** | Ionosphere | 0.84 | 1.3e-04 | 93.80 | 1.30 | 2.6e-05 | 93.80 | 1.80 | 6.3e-06 | 93.80 | 2.28 | 1.5e-06 | 93.80 |
| **prox** | Sonar | 0.47 | 4.3e-03 | 85.24 | 0.71 | 3.4e-04 | 85.48 | 0.96 | 2.9e-05 | 84.52 | 1.24 | 2.9e-06 | 84.76 |
| | Heart | 0.75 | 1.9e-06 | 83.89 | 1.16 | 7.5e-07 | 83.89 | 1.60 | 2.9e-07 | 83.89 | 2.04 | 1.2e-07 | 83.88 |
| | Breast-cancer | 2.56 | 1.1e-02 | 96.86 | 4.24 | 2.6e-03 | 96.86 | 5.91 | 6.8e-04 | 96.72 | 7.54 | 2.0e-04 | 96.71 |

Table 2: $\ell_1$-norm soft margin: runtime (sec), relative error for $\mathcal{L}(x_k, y_k)$ and TSA (%) for **APD1** and **Mirror-prox** at iteration $K \in \{1000, 1500, 2000, 2500\}$.
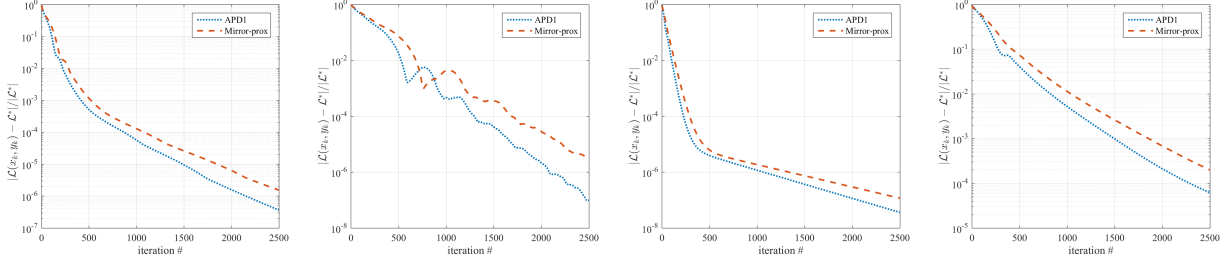


Figure 4: $\ell_1$-norm soft margin: **APD1** vs **Mirror-prox** in terms of relative error for $\mathcal{L}(x_k, y_k)$. The plots from left to right correspond to `Ionosphere`, `Sonar`, `Heart`, `Breast-Cancer`.
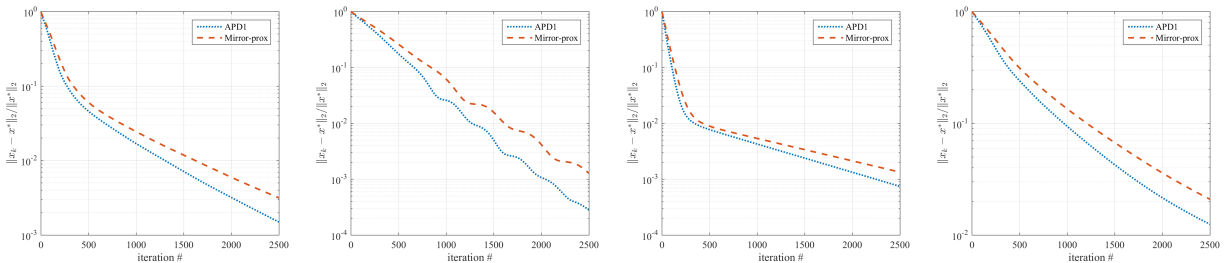


Figure 5: $\ell_1$-norm soft margin: **APD1** vs **Mirror-prox** in terms of relative error for $x_k$. The plots from left to right correspond to: `Ionosphere`, `Sonar`, `Heart`, `Breast-Cancer`.