A Simple Algorithm for a Class of Nonsmooth Convex-Concave Saddle-Point Problems

Yoel Drori* Shoham Sabach† Marc Teboulle‡

February 2, 2015

Abstract

We introduce a novel algorithm for solving a class of structured nonsmooth convex-concave saddle-point problems involving a smooth function and a sum of finitely many bilinear terms and nonsmooth functions. The proposed method is simple and proven to globally converges to a saddle-point with an $O(1/\varepsilon)$ efficiency estimate. We demonstrate its usefulness for tackling a broad class of minimization models with a finitely sum of composite nonsmooth functions.

Keywords: Saddle-point problems, nonsmooth convex minimization, iteration complexity.

1 Introduction

In this paper, we consider a class of nonsmooth structured convex-concave saddle-point (SP) problems. By structured we mean that the model consists of a saddle-point function that is a sum of a smooth function (i.e., with Lipschitz continuous gradient), with a finite collection of nonsmooth functions and bilinear terms, see Section 2. This model is very rich and encompasses most convex optimization models arising in a wide array of applications in signal/image processing and machine learning, see for instance the two very recent edited volumes [16, 22] and references therein.

The past and current research activities in the search of methods for solving the alluded class of convex-concave SP problems and their relatives composite nonsmooth minimization problems, have been intensive over the past five decades and has been recently revived due to their relevance in many applications, see, e.g., [6, 7, 9, 10, 11, 17, 23, 24]. The recent trends of research efforts has been focusing on first order methods that allow to adequately handle very large scale problems and produce efficient schemes for modest accuracy requirement $\varepsilon > 0$. There exist already such algorithms which achieve the best known $O(1/\varepsilon)$ efficiency estimate under various modeling/assumptions on the problem structures, and under various degrees of sophistication in their derivations and analysis. Let us briefly mention some of the main underlying approaches and some of their limitations.

A classical way to solve saddle-point problems is via the variational inequality framework [8]. For *smooth* convex-concave SP, extra-gradient based methods which originated from Korpelevitch [10] have been recently shown to exhibit an $O(1/\varepsilon)$ efficiency estimate [13, 2]. However, these extra-gradient type methods cannot in general be applied for nonsmooth SP problems without adequate reformulations or further assumptions, and they also double the amount of computation of projections which often severely affect their performance. Another approach is to exploit the "max-structure" inherently present in an SP formulation and combine smoothing with fast gradient schemes [15, 3]. Such methods require knowledge of the smoothing parameter (which depends on the desired

*School of Mathematical Sciences, Tel-Aviv University, Ramat-Aviv 69978, Israel (dyoel@post.tau.ac.il).

†Corresponding Author. Faculty of Industrial Engineering and Management, Technion, Haifa, Israel (ssabach@ie.technion.ac.il), phone: +972-4-8294442.

‡School of Mathematical Sciences, Tel-Aviv University, Ramat-Aviv 69978, Israel (teboulle@math.tau.ac.il).

accuracy) and some other assumptions, such as compactness. A third approach relies on primal-dual decomposition methods revolving around variants of the well known Alternating Direction of Multipliers (ADM) for which an $O(1/\varepsilon)$ efficiency estimate has been established, see e.g., the very recent study [21] for more details and many relevant references. However, ADM like methods do not easily extend to problems involving the sum of finitely many composite nonsmooth terms, see e.g., [5] where even the convergence become an issue.

Motivated by the above recent developments, we present a novel and simple algorithm which allows to tackle a broader class of SP problems and is proven to globally converge to a saddle-point with efficiency estimate $O(1/\varepsilon)$. By simple, we mean an algorithm which at each iteration utilizes one gradient and one proximal map operation on the given nonsmooth function, assumed to be easy to compute or/and can be efficiently computed. Moreover, no matrix inversion is involved and furthermore we do not rely on nested optimization schemes see Section 2 for details and the proposed algorithm. In Section 3 we derive the promised global nonasymptotic efficiency estimate, and as an easy by-product the sequential convergence is also obtained. Our approach, also allows to efficiently address the important class of structured convex models involving the sum of smooth function with a finite sum of nonsmooth functions composed with linear maps in the objective or in the constraints. In particular, our method avoids the difficult task of computing the proximal map of the (sum of) composition with linear maps, see Section 4. Numerical illustration demonstrating the relevance and potential of the proposed method when compared to recent state-of-the art schemes can be found in the supplementary material `http://www.math.tau.ac.il/~teboulle/pub/orl-papc-numeric.pdf`.

Notation. The set of symmetric $p \times p$ positive (semi)-definite matrices is denoted by $\mathbb{S}^p_{++}$ ($\mathbb{S}^p_{+}$), we also use $M \succ 0$ ($M \succeq 0$). For any $z \in \mathbb{R}^p$ and any $M \in \mathbb{S}^p_{+}$, $\|z\|_M := \langle z, Mz \rangle^{1/2}$. Other standard convex analysis notations can be found in any text, e.g., [18].

2 The Saddle-Point Model and The Algorithm

We begin by describing the setting of the nonsmooth structured convex-concave saddle-point problem of interest.

2.1 The Saddle-Point Problem

We consider convex-concave saddle-point problems of the form

$$\text{(M)} \qquad \min_{u \in \mathbb{R}^n} \max_{v \in \mathbb{R}^d} \left\{ K\left(u, v\right) := f\left(u\right) + \langle u, \mathcal{A}v \rangle - g\left(v\right) \right\},$$

where f and g are convex functions and $\mathcal{A}$ is a linear map such that

(i) $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function which is continuously differentiable and its gradient ∇f is Lipschitz continuous with constant L_f, i.e., for all $u_1, u_2 \in \mathbb{R}^n$, we have

$$\left\| \nabla f\left(u_1\right) - \nabla f\left(u_2\right) \right\| \leq L_f \left\| u_1 - u_2 \right\|.$$

(ii) $g_i : \mathbb{R}^{d_i} \to (-\infty, +\infty]$, $i = 1, 2, \ldots, m$, is a proper, lower semicontinuous (lsc) and convex function (possibly nonsmooth). With $v_i \in \mathbb{R}^{d_i}$, we define $v := (v_1, v_2, \ldots, v_m) \in \mathbb{R}^d$ where $d = \sum_{i=1}^m d_i$ and we let $g : \mathbb{R}^d \to (-\infty, +\infty]$ be the proper, lsc and convex function defined by

$$g\left(v\right) := \sum_{i=1}^{m} g_i\left(v\right).$$

(iii) $A_i : \mathbb{R}^{d_i} \to \mathbb{R}^n$, $i = 1, 2, \ldots, m$, is a linear map and we let $\mathcal{A} : \mathbb{R}^d \to \mathbb{R}^n$ be the linear map defined by $\mathcal{A}v = \sum_{i=1}^m A_i v_i$.

Note that our model's formulation does not include constraint on the variable u (constraints on the variable v are built-in thanks to the fact that g is extended valued). However, as we shall see, the choice of our model (M) is not accidental. It will allow not only to easily handle constraints on u, but will also to offer much flexibility in tackling various type of composite optimization models which arise in many important applications (see Section 4).

2.2 The Standing Assumption

Throughout this paper, our standing assumption is that the convex-concave function $K(\cdot, \cdot)$ has a saddle-point, i.e., there exists $(u^*, v^*) \in \mathbb{R}^n \times \mathbb{R}^d$ such that

$$K(u^*, v) \leq K(u^*, v^*) \leq K(u, v^*), \quad \forall\, u \in \mathbb{R}^n,\ v \in \mathbb{R}^d.$$

The existence of a saddle-point corresponds to zero duality gap for the induced optimization problems

$$(P) \qquad \inf_{u \in \mathbb{R}^n} \left[r(u) = \sup_{v \in \mathbb{R}^d} K(u, v) \right] \qquad \text{and} \qquad (D) \qquad \sup_{v \in \mathbb{R}^d} \left[q(v) = \inf_{u \in \mathbb{R}^n} K(u, v) \right].$$

One always has $\inf_{u \in \mathbb{R}^n} r(u) \geq \sup_{v \in \mathbb{R}^d} q(v)$, i.e., weak duality. In addition, (u^*, v^*) is a saddle-point of K if and only if u^* is an optimal solution of the primal problem (P), v^* is an optimal solution of the dual problem (D), and $\inf_{u \in \mathbb{R}^n} \sup_{v \in \mathbb{R}^d} K(u, v) = \sup_{v \in \mathbb{R}^d} \inf_{u \in \mathbb{R}^n} K(u, v) = K(u^*, v^*)$, where $K(u^*, v^*)$ is the saddle-point value. For standard qualification conditions which warrant this equality, see e.g., [1, Chapter 5] and [20, Chapter 11].

2.3 The Algorithm

Before we state our algorithm we need to recall the definition of the Moreau proximal map [12] and to introduce some convenient notations.

Let $h : \mathbb{R}^p \to (-\infty, \infty]$ be a proper, lsc and convex function. For any $x \in \mathbb{R}^p$ and $M \in \mathbb{S}^p_{++}$, the proximal map associated with h is uniquely defined by:

$$\operatorname{prox}^h_M(x) := \operatorname{argmin}_y \left\{ h(y) + \frac{1}{2} \|y - x\|^2_M \right\}. \tag{2.1}$$

When $M = \mu^{-1} I_p$, $\mu > 0$, we simply use the notation $\operatorname{prox}^h_\mu(\cdot)$. We also recall the fundamental Moreau proximal identity [12], that is, for any $z \in \mathbb{R}^p$:

$$\operatorname{prox}^h_M(z) + M \operatorname{prox}^{h^*}_{M^{-1}}\left(M^{-1} z\right) = z, \quad (M^{-1} \text{ is the inverse of } M \in \mathbb{S}^p_{++}). \tag{2.2}$$

For any given real numbers $\sigma_1, \sigma_2, \ldots, \sigma_m > 0$, we denote $S_i := \sigma_i^{-1} I_{d_i}$, $i = 1, 2, \ldots, m$, where I_{d_i} stands for the $d_i \times d_i$ identity matrix, and define the block diagonal matrix $S := \operatorname{Diag}[S_1, S_2, \ldots, S_m] \in \mathbb{S}^d_{++}$ with $d = \sum_{i=1}^m d_i$.

The algorithm we propose consists of a predictor-corrector gradient step for handling the smooth part of K and a proximal step for handling the nonsmooth part.

PAPC: Proximal Alternating Predictor Corrector
Initialization. $\left(u^0, v^0\right) \in \mathbb{R}^n \times \mathbb{R}^d$ and let $\tau > 0, S \succ 0$.
General Step. For $k = 1, 2, \ldots$, compute

$$p^k = u^{k-1} - \tau\left(\mathcal{A}v^{k-1} + \nabla f\left(u^{k-1}\right)\right), \tag{2.3}$$

$$v^k = \operatorname{prox}_S^g\left(v^{k-1} + S^{-1}\mathcal{A}^T p^k\right), \tag{2.4}$$

$$u^k = u^{k-1} - \tau\left(\mathcal{A}v^k + \nabla f\left(u^{k-1}\right)\right). \tag{2.5}$$

The choice of the parameters τ and S will be made precise in Section 3.
A few remarks regarding the computational steps involved in the PAPC method are now in order.

- A major computational effort of the method is given in the second step (2.4). Since here $g\left(v\right) = \sum_{i=1}^m g_i\left(v_i\right)$, using the definition of the matrix S we immediately obtain that at any given point $x_i \in \mathbb{R}^{d_i}$, $i = 1, 2, \ldots, m$,

  $$\operatorname{prox}_S^g\left(x\right) = \left(\operatorname{prox}_{\sigma_1}^{g_1}\left(x_1\right), \operatorname{prox}_{\sigma_2}^{g_2}\left(x_2\right), \ldots, \operatorname{prox}_{\sigma_m}^{g_m}\left(x_m\right)\right),$$

  and hence the second step of the algorithm (2.4) decomposes accordingly and for all $i = 1, 2, \ldots, m$ we have

  $$v_i^k = \operatorname{prox}_{\sigma_i}^{g_i}\left(v_i^{k-1} + \sigma_i A_i^T p^k\right) = \operatorname{argmin}_{v_i \in \mathbb{R}^{d_i}}\left\{g_i\left(v_i\right) + \frac{1}{2\sigma_i}\left\|v_i - \left(v_i^{k-1} + \sigma_i A_i^T p^k\right)\right\|^2\right\}.$$

  Thus, the algorithm PAPC achieves full decomposition for the given structure of K in the sense that for each i, it avoids the much more difficult task of computing the proximal map of the composite function $g_i \circ A_i$, and only requires computing the proximal map of $g_i\left(\cdot\right)$.

- The algorithm uses only *one* evaluation of the gradient of the smooth function f, and a careful implementation requires only one application of the operator $\mathcal{A}$ and one application of the operator $\mathcal{A}^T$ per iteration. Thus, for large scale problems this potentially amounts to a considerable reduction of computation time compared to the straightforward implementation.

3 Convergence Results for PAPC

In this section, we establish the main convergence properties of the PAPC algorithm. In particular, we prove its global rate of convergence, showing that it shares the claimed $O(1/\varepsilon)$ efficiency estimate. As an easy by-product we also derive a global convergence of the sequence generated by PAPC to a saddle-point of $K\left(\cdot, \cdot\right)$. We start by recalling three elementary facts.

Fact 1 - Three points inequality: For any convex function h on $\mathbb{R}^p$ which is continuously differentiable with gradient ∇h which assumed to be L_h-Lipschitz continuous we have:

$$h\left(x\right) \leq h\left(y\right) + \left\langle \nabla h\left(z\right), x - y\right\rangle + \frac{L_h}{2}\left\|x - z\right\|^2, \ \forall\, x, y, z \in \mathbb{R}^p.$$

Fact 2 - Proximal Inequality: Let $h : \mathbb{R}^p \rightarrow \left(-\infty, \infty\right]$ be a proper, lsc and convex function. Given $M \in \mathbb{S}_+^p$ and $x \in \mathbb{R}^p$, let $z \in \operatorname{argmin}_{\xi \in \mathbb{R}^p}\left\{h\left(\xi\right) + \frac{1}{2}\left\|\xi - x\right\|_M^2\right\}$. Then, for all $\xi \in \mathbb{R}^p$, we have

$$h\left(z\right) - h\left(\xi\right) \leq \left\langle \xi - z, M\left(z - x\right)\right\rangle.$$

Fact 3 - Pythagoras identity: For any matrix $M \in \mathbb{S}_+^p$, we have

$$2\left\langle w - v, M\left(u - v\right)\right\rangle = \left\|w - v\right\|_M^2 - \left\|w - u\right\|_M^2 + \left\|u - v\right\|_M^2, \qquad \forall\, u, v, w \in \mathbb{R}^p. \tag{3.1}$$

The proofs of these facts are immediate. Fact 1 follows from the convexity of h and the well known Descent Lemma for smooth functions, see e.g., [4]. Fact 2 follows from the optimality condition which characterizes z and the convex subgradient inequality, while Fact 3 is simple algebra.

To establish the iteration complexity of the PAPC algorithm and the convergence of the generated sequence $\left\{\left(u^k,v^k\right)\right\}_{k\in\mathbb{N}}$, we consider the following quantity

$$\Gamma_k\left(u,v\right)=K\left(u^k,v\right)-K\left(u,v^k\right),\quad \forall\, u\in\mathbb{R}^n,\ v\in\mathbb{R}^d.$$

Our main task is to find an upper-bound for $\Gamma_k(u,v)$, $k\in\mathbb{N}$. Indeed, $\Gamma_k\left(u,v\right)\le 0$ for all $u\in\mathbb{R}^n$ and all $v\in\mathbb{R}^d$, implies that $K\left(u^k,v^k\right)\le K\left(u,v^k\right)$ for all $u\in\mathbb{R}^n$, and that $K\left(u^k,v\right)\le K\left(u^k,v^k\right)$ for all $v\in\mathbb{R}^d$, namely, that $\left(u^k,v^k\right)$ is a saddle-point of K with saddle-point value $K\left(u^k,v^k\right)$. We now proceed to prove two key inequalities which will be the basis for proving our main convergence results.

Lemma 3.1. *Let $\left\{\left(p^k,v^k,u^k\right)\right\}_{k\in\mathbb{N}}$ be the sequence generated by the PAPC algorithm, then for every $k\in\mathbb{N}$ and every $u\in\mathbb{R}^n, v\in\mathbb{R}^d$, the following hold:*

(i) $K\left(u^k,v^k\right)-K\left(u,v^k\right)\le\frac{1}{2\tau}\left(\left\|u-u^{k-1}\right\|^2-\left\|u-u^k\right\|^2\right)-\frac{1}{2}\left(\frac{1}{\tau}-L_f\right)\left\|u^k-u^{k-1}\right\|^2.$

(ii) $K\left(u^k,v\right)-K\left(u^k,v^k\right)\le\frac{1}{2}\left(\left\|v-v^{k-1}\right\|_G^2-\left\|v-v^k\right\|_G^2-\left\|v^k-v^{k-1}\right\|_G^2\right),$
where $G:=S-\tau\mathcal{A}^T\mathcal{A}$, is assumed positive semi-definite.

Proof. (i) Applying Fact 1 on the convex and differentiable function $h\left(u\right):=K\left(u,v^k\right)$ with $x:=u^k$, $y:=u$ and $z:=u^{k-1}$, yields

$$K\left(u^k,v^k\right)-K\left(u,v^k\right)\le\left\langle\nabla_u K\left(u^{k-1},v^k\right),u^k-u\right\rangle+\frac{L_f}{2}\left\|u^k-u^{k-1}\right\|^2.$$

Using the fact that $\nabla_u K\left(u^{k-1},v^k\right)=\mathcal{A}v^k+\nabla f\left(u^{k-1}\right)=\tau^{-1}\left(u^{k-1}-u^k\right)$, where the last equation follows from the definition of step (2.5), we get

$$K\left(u^k,v^k\right)-K\left(u,v^k\right)\le\frac{1}{\tau}\left\langle u^{k-1}-u^k,u^k-u\right\rangle+\frac{L_f}{2}\left\|u^k-u^{k-1}\right\|^2.$$

The desired result follows by using the identity (3.1) with $M\equiv I_n$, for the first term in the right hand side of the above inequality.

(ii) Using the definition of $K\left(\cdot,\cdot\right)$, step (2.4) of PAPC can be written (after omitting constant terms) as

$$v^k=\operatorname{prox}_S^{-K(p^k,\cdot)}\left(v^{k-1}\right)=\operatorname{argmin}_{v\in\mathbb{R}^d}\left\{-K\left(p^k,v\right)+\frac{1}{2}\left\|v-v^{k-1}\right\|_S^2\right\}.$$

Applying Fact 2 to the convex function $h(\nu):=-K(p^k,\nu)$ with $\xi:=v$, $z:=v^k$ and $x:=v^{k-1}$, yields

$$K\left(p^k,v\right)-K\left(p^k,v^k\right)\le\left\langle v^k-v^{k-1},S\left(v-v^k\right)\right\rangle. \tag{3.2}$$

Now, from the definition of $K(\cdot,\cdot)$, simple algebra shows that the following identity holds

$$K\left(u^k,v\right)-K\left(u^k,v^k\right)+K\left(p^k,v^k\right)-K\left(p^k,v\right)=\left\langle u^k-p^k,\mathcal{A}\left(v-v^k\right)\right\rangle. \tag{3.3}$$

Using the definitions of p^k in (2.3) and u^k in (2.5) we have $u^k-p^k=\tau\mathcal{A}\left(v^{k-1}-v^k\right)$, hence, together with (3.2) and (3.3), we obtain

$$\begin{aligned}K\left(u^k,v\right)-K\left(u^k,v^k\right)&=\left\langle u^k-p^k,\mathcal{A}\left(v-v^k\right)\right\rangle+K\left(p^k,v\right)-K\left(p^k,v^k\right)\\&\le\tau\left\langle v^{k-1}-v^k,\mathcal{A}^T\mathcal{A}\left(v-v^k\right)\right\rangle+\left\langle v^k-v^{k-1},S\left(v-v^k\right)\right\rangle\\&=\left\langle v^k-v^{k-1},\left(S-\tau\mathcal{A}^T\mathcal{A}\right)\left(v-v^k\right)\right\rangle.\end{aligned}$$

Thus, with $G := S - \tau \mathcal{A}^T \mathcal{A}$, which assumed to be positive semidefinite, the desired result follows by using the identity (3.1) with $M \equiv G$. □

Before we proceed with the convergence results, we need some additional notations. For any sequence $\{x^k\}_{k\in\mathbb{N}}$ and any integer $N \geq 1$, we denote by $\bar{x}^N := \frac{1}{N}\sum_{k=1}^{N} x^k$ the average (ergodic) sequence associated with $\{x^k\}_{k\in\mathbb{N}}$. Let $\sigma := \max_{1\leq i\leq m} \sigma_i$.

Theorem 3.1. *Let $\left\{\left(p^k, u^k, v^k\right)\right\}_{k\in\mathbb{N}}$ be the sequence generated by the PAPC algorithm with $\tau L_f \leq 1$ and $\sigma\tau\sum_{i=1}^{m}\|A_i\|^2 \leq 1$. Then, $G = S - \tau\mathcal{A}^T\mathcal{A}$ is positive semidefinite and for every $u \in \mathbb{R}^n$ and $v \in \mathbb{R}^d$, we have*

$$K\left(\bar{u}^N, v\right) - K\left(u, \bar{v}^N\right) \leq \frac{\tau^{-1}\left\|u - u^0\right\|^2 + \left\|v - v^0\right\|_G^2}{2N}.$$

Proof. We begin by showing that the condition $\sigma\tau\sum_{i=1}^{m}\|A_i\|^2 \leq 1$ implies that the symmetric $d \times d$ matrix G is positive semi-definite. First, note that $G = S - \tau\mathcal{A}^T\mathcal{A} \succeq 0$ if $\lambda_{\min}(S) \geq \tau\lambda_{\max}\left(\mathcal{A}^T\mathcal{A}\right)$, where $\lambda_{\min}(\cdot)$ ($\lambda_{\max}(\cdot)$) stands for the minimal (maximal) eigenvalue of the given symmetric matrix. Recalling the definition of S given in Section 2.3, we obtain that $\lambda_{\min}(S) = \min_{1\leq i\leq m}\sigma_i^{-1} = \left(\max_{1\leq i\leq m}\sigma_i\right)^{-1} = \sigma^{-1}$, and hence the last condition reduces to $\sigma\tau\lambda_{\max}\left(\mathcal{A}^T\mathcal{A}\right) \leq 1$. On the other hand, using the definition of $\mathcal{A}$ we have

$$\lambda_{\max}\left(\mathcal{A}^T\mathcal{A}\right) = \left\|\mathcal{A}^T\mathcal{A}\right\| = \left\|\sum_{i=1}^{m} A_i^T A_i\right\| \leq \sum_{i=1}^{m}\left\|A_i^T A_i\right\| = \sum_{i=1}^{m}\|A_i\|^2,$$

and the first part of the claim follows. Now, let $u \in \mathbb{R}^n$ and $v \in \mathbb{R}^d$. Since $G \succeq 0$ and we assume that $\tau L_f \leq 1$, adding the two inequalities (i)-(ii) of Lemma 3.1 it immediately follows that for all $k \in \mathbb{N}$, and for all $u \in \mathbb{R}^n$, $v \in \mathbb{R}^d$

$$K\left(u^k, v\right) - K\left(u, v^k\right) \leq \frac{1}{2\tau}\left(\left\|u - u^{k-1}\right\|^2 - \left\|u - u^k\right\|^2\right) + \frac{1}{2}\left(\left\|v - v^{k-1}\right\|_G^2 - \left\|v - v^k\right\|_G^2\right).$$

Now, since $K(u, v)$ is convex-concave, using the definition of $\left(\bar{u}^N, \bar{v}^N\right)$, by the Jensen inequality and summing the last inequality over $k = 1, 2, \ldots, N$ it follows that

$$K\left(\bar{u}^N, v\right) - K\left(u, \bar{v}^N\right) \leq \frac{1}{N}\sum_{k=1}^{N}\left(K\left(u^k, v\right) - K\left(u, v^k\right)\right) \leq \frac{1}{2\tau N}\left\|u - u^0\right\|^2 + \frac{1}{2N}\left\|v - v^0\right\|_G^2.$$

This proves the desired result. □

Two important consequences of this global upper-bound established in Theorem 3.1 can be deduced. Let $\varepsilon > 0$, then following [14], a point $(u_\varepsilon, v_\varepsilon)$ is called an ε-saddle-point for K if

$$\sup\left\{K\left(u_\varepsilon, v\right) - K\left(u, v_\varepsilon\right) \;:\; u \in S_P, v \in S_D\right\} \leq \varepsilon,$$

where S_P is the optimal solutions set of the primal problem and S_D is the optimal solutions set of the dual problem associated to the saddle-point function K (see Section 2). We thus immediately obtain from Theorem 3.1 the following efficiency estimate result.

Corollary 3.1. *Let $\left\{\left(p^k, u^k, v^k\right)\right\}_{k\in\mathbb{N}}$ be the sequence generated by the PAPC algorithm with $\tau L_f \leq 1$ and $\sigma\tau\sum_{i=1}^{m}\|A_i\|^2 \leq 1$. Assume that both optimal solutions sets S_P and S_D associated to the saddle-point problem (M) are compact. Then, given a desired accuracy $\varepsilon > 0$, the PAPC method produces an ε-saddle-point $\left(\bar{u}^N, \bar{v}^N\right)$ of K in $N = O(1/\varepsilon)$ iterations.*

Remark 3.1. Note that under standard qualification conditions, which is our standing assumption (cf. Section 2.2), the optimal set S_D of the dual problem associated to (M) is always compact.

Another easy consequence of Theorem 3.1 is a convergence result of the sequence generated by PAPC to a saddle-point of problem (M).

Corollary 3.2. *Let $\left\{\left(p^k, u^k, v^k\right)\right\}_{k\in\mathbb{N}}$ be the sequence generated by the PAPC algorithm with $\tau L_f < 1$ and $\sigma\tau\sum_{i=1}^{m}\|A_i\|^2 < 1$. Then, the sequence $\left\{\left(u^k, v^k\right)\right\}_{k\in\mathbb{N}}$ converges to a saddle-point $(\widetilde{u}, \widetilde{v})$ of K.*

Proof. Adding the two inequalities of Lemma 3.1 it immediately follows that for any $u \in \mathbb{R}^n$, $v \in \mathbb{R}^d$ and for all $k \in \mathbb{N}$

$$K\left(u^k, v\right) - K\left(u, v^k\right) \leq \frac{1}{2\tau}\left(\left\|u^{k-1} - u\right\|^2 - \left\|u^k - u\right\|^2\right) - \frac{1}{2}\left(\frac{1}{\tau} - L_f\right)\left\|u^k - u^{k-1}\right\|^2 + \frac{1}{2}\left(\left\|v^{k-1} - v\right\|_G^2 - \left\|v^k - v\right\|_G^2 - \left\|v^k - v^{k-1}\right\|_G^2\right). \tag{3.4}$$

In particular, let $(u^*, v^*) \in \mathbb{R}^n \times \mathbb{R}^d$ be an arbitrary saddle-point of the function K, then by the saddle-point property, $K\left(u^k, v^*\right) - K\left(u^*, v^k\right) \geq 0$, it follows from the last inequality that

$$\left(\frac{1}{\tau} - L_f\right)\left\|u^k - u^{k-1}\right\|^2 + \left\|v^k - v^{k-1}\right\|_G^2 \leq D\left(w^{k-1}, w^*\right) - D\left(w^k, w^*\right), \tag{3.5}$$

where $w := (u, v) \in \mathbb{R}^n \times \mathbb{R}^d$ and we define $D\left(w_1, w_2\right) := \frac{1}{\tau}\left\|u_1 - u_2\right\|^2 + \left\|v_1 - v_2\right\|_G^2$. As a consequence of (3.5), the sequence $\left\{D\left(w^k, w^*\right)\right\}_{k\in\mathbb{N}}$ is non-increasing and therefore the sequence $\left\{w^k\right\}_{k\in\mathbb{N}}$ is bounded. On the other hand, summing (3.5) for any $k = 1, 2, \ldots, N$, since $G \succ 0$ and $L_f\tau < 1$, it follows that

$$\lim_{k\to\infty}\left\|u^k - u^{k-1}\right\| = 0 \quad \text{and} \quad \lim_{k\to\infty}\left\|v^k - v^{k-1}\right\|_G = 0. \tag{3.6}$$

Since the sequence $\left\{w^k\right\}_{k\in\mathbb{N}}$ is bounded, it has at least one limit point. Suppose that $\widetilde{w} = (\widetilde{u}, \widetilde{v})$ is a limit point of the sequence $\left\{w^k\right\}_{k\in\mathbb{N}}$, then taking the limit in (3.4) over the appropriate subsequences and using (3.6) yields $K(\widetilde{u}, v) - K(u, \widetilde{v}) \leq 0$ for all $u \in \mathbb{R}^n$ and $v \in \mathbb{R}^d$, which proves that $(\widetilde{u}, \widetilde{v})$ is a saddle-point of K. To complete the proof it only remains to show that $\left\{w^k\right\}_{k\in\mathbb{N}}$ has a unique limit point. This follows by a standard argument, see e.g., [19, page 885]. □

4 Composite Minimization via Saddle-Point

Our main purpose in this section is, on one hand, to illustrate the flexibility of the model (M), and on the other hand, the simplicity of the resulting PAPC algorithm when applied to these problems. To do so, we start with a simple but key observation, which for ease of reference we call the *Dual Transportation Trick.*

4.1 The Dual Transportation Trick

A fundamental convex analysis result [18] states that the *bi-conjugate* of a proper, lsc and convex function $h : \mathbb{R}^p \to (-\infty, \infty]$ coincides with itself, i.e., $h^{**} = h$. Thus, any proper, lsc and convex function h admits the following variational max-representation

$$h(x) = \max_{u\in\mathbb{R}^p}\left\{\langle u, x\rangle - h^*(u)\right\}.$$

Our main observation is that this well known and fundamental relation is in fact the key player not only for handling constraints, but also for deriving "full splitting" of most optimization problems involving composition with linear maps through their saddle-point representation in the form (M). This is described next, and then we illustrate it on various optimization models.

Let $U \subset \mathbb{R}^n$ be a closed and convex set, consider the following constrained convex problem

$$(C) \qquad \min_{u \in \mathbb{R}^p} \left\{ F\left(u\right) : \; u \in U \right\}.$$

Let δ_U denotes the usual *convex indicator function* of the set U (i.e., 0 if $u \in U$ and ∞ otherwise), and recall that $\delta_U^* = \sigma_U$, the so-called *support function* of the set U, which is proper, lsc and convex when U is a closed and convex set, thus in this case $\sigma_U^* = \delta_U$. Equipped with these basic objects, problem (C) can be written

$$\begin{aligned} \min_{u \in \mathbb{R}^p} \left\{ F\left(u\right) + \delta_U\left(u\right) \right\} &= \min_{u \in \mathbb{R}^p} \max_{v \in \mathbb{R}^p} \left\{ F\left(u\right) + \langle u, v \rangle - \delta_U^*\left(v\right) \right\} && (4.1) \\ &= \min_{u \in \mathbb{R}^p} \max_{v_1, v_2 \in \mathbb{R}^p} \left\{ \langle u, v_1 \rangle - F^*\left(v_1\right) - \langle u, v_2 \rangle - \delta_U^*\left(v_2\right) \right\}. && (4.2) \end{aligned}$$

Clearly, both saddle-point representations given in (4.1) and (4.2) can be seen as particular cases of model (M), yet they illustrate two important different goals: (4.1) provides a way to reinterpret a constrained optimization problem as an unconstrained saddle-point problem. It allows to transport the primal constraint variable $u \in U$ into the objective function, but with an additional nonsmooth convex function $\sigma_U = \delta_U^*$ in the dual space of the saddle-point function. The representation (4.2) provides a way to continue and further decompose the problem to fit our model (M) in the case where F is also nonsmooth. As described below, this elementary dual transportation trick allows to easily apply PAPC on constrained SD as well as quite general composite optimization models.

4.2 Handling Constrained Saddle-Point Problems

Let $U \subseteq \mathbb{R}^n$ be a closed and convex set, consider the following constrained saddle-point problem (here for simplicity of exposition it is enough to look at $m = 1$, which means that, $d \equiv d_1$, $\mathcal{A} = A_1 \equiv A$, $g_1\left(v\right) \equiv g\left(v\right)$ and $\sigma_1 = \sigma$):

$$\text{(CM)} \qquad \min_{u \in U} \max_{v \in \mathbb{R}^d} \left\{ K\left(u, v\right) = f\left(u\right) + \langle u, Av \rangle - g\left(v\right) \right\}.$$

Using the dual transportation trick we obtain the following equivalent unconstrained saddle-point problem which is compatible with the requirements of the PAPC method, namely unconstrained in u:

$$\text{(CM}'\text{)} \qquad \min_{u \in \mathbb{R}^n} \max_{v \in \mathbb{R}^d, w \in \mathbb{R}^n} \left\{ K'\left(u; v, w\right) := f\left(u\right) + \langle u, Av \rangle - g\left(v\right) + \langle u, w \rangle - \sigma_U\left(w\right) \right\}.$$

Observing that the inner maximization problem in (CM′) is *separable* in the variables v and w, the PAPC method for solving the constrained saddle-point problem (CM) can be formulated as follows.

PAPC: constrained version
Initialization. $\left(u^0, v^0, w^0\right) \in \mathbb{R}^n \times \mathbb{R}^d \times \mathbb{R}^n$ and $\tau, \sigma > 0$.
General Step ($k = 1, 2, \ldots$)

$$p^k = u^{k-1} - \tau\left(Av^{k-1} + w^{k-1} + \nabla f\left(u^{k-1}\right)\right), \qquad (4.3)$$

$$v^k = \mathrm{prox}_\sigma^g\left(v^{k-1} + \sigma A^T p^k\right), \qquad (4.4)$$

$$w^k = \mathrm{prox}_\sigma^{\sigma_U}\left(w^{k-1} + \sigma p^k\right) = w^{k-1} + \sigma p^k - \sigma P_U\left(\frac{w^{k-1} + \sigma p^k}{\sigma}\right). \qquad (4.5)$$

$$u^k = u^{k-1} - \tau\left(Av^k + w^k + \nabla f\left(u^{k-1}\right)\right). \qquad (4.6)$$

Theorem 3.1 holds on problem (P′) with the parameters $\tau L_f \leq 1$ and $\sigma\tau\left(\|A\|^2 + 1\right) \leq 1$.

4.3 Minimization with Finite Sum of Composite Functions

Let $U \subseteq \mathbb{R}^n$ be a closed and convex set. The problem of interest can be described as follows

$$\text{(Gen)} \qquad \min_{u \in \mathbb{R}^p} \left\{ F(u) + \sum_{i=1}^{m} H_i(B_i u) : \ u \in U \right\},$$

where F is a smooth and convex function on $\mathbb{R}^p$ (see (ii) of the problem setting in Section 2), H_i, $i = 1, 2, \ldots, m$, is a proper, lsc and convex function over $\mathbb{R}^{d_i}$ (extended valued) and $B_i \in \mathbb{R}^{d_i} \times \mathbb{R}^p$. This model is quite general and covers many interesting problems (e.g., in imaging sciences and machine learning) and also includes convex problems with separable structure in the objective and coupling linear constraints of the form

$$\text{(SC)} \qquad \min_{x_i} \left\{ \sum_{i=1}^{m} \psi_i(x_i) : \ \sum_{i=1}^{m} B_i x_i = b \right\}.$$

Indeed, a direct computation shows that a dual formulation of problem (SC) also fits the model (Gen) with $F := \langle u, b \rangle$, $H_i := \psi_i^*$ and $B_i \leftarrow -B_i^T$ (for convenience, the dual problem is written as minimiztion problems after an appropriate change of sign).

Using the dual transportation trick for the constraint $u \in U$, and the fact that H_i is proper, lsc and convex, problem (Gen) can be written as

$$\min_{u \in \mathbb{R}^p} \max_{y_i \in \mathbb{R}^{d_i}, w \in \mathbb{R}^p} \left\{ F(u) + \sum_{i=1}^{m} \left\langle B_i^T y_i, u \right\rangle + \langle w, u \rangle - \sum_{i=1}^{m} H_i^*(y_i) - \sigma_U(w) \right\},$$

which clearly reduces to a saddle-point problem in the form (M) with saddle-point function $K(u, v) = F(u) + \langle u, \mathcal{A}v \rangle - g(v)$ through the identification $v := (y_1, y_2, \ldots, y_m, w)$, $\mathcal{A}v := \sum_{i=1}^{m} B_i^T y_i + w$ and $g(v) := \sum_{i=1}^{m} H_i^*(y_i) + \sigma_U(w)$. Observe that this yields a fully separable nonsmooth part in the variables y_i, $i = 1, 2, \ldots, m$ and w, which allows for adequate decomposition in the main computational step of PAPC. In particular, this eliminates the difficulty of computing the proximal map of the composition of a convex function with a linear map.

4.4 Constrained Composite Minimization

Another interesting model is the following constrained composite convex minimization problem

$$\text{(C-Gen)} \qquad \min_{u \in \mathbb{R}^p} \left\{ F(u) : \ \sum_{i=1}^{m} H_i(B_i u) \leq \alpha \right\},$$

where F, H_i $(i = 1, 2, \ldots, m)$ and B_i $(i = 1, 2, \ldots, m)$ as in the (Gen) model (note that thanks to the dual transportation trick (cf. (4.2)), we can also easily consider the case where nonsmooth terms would be present in the objective), and here $H_i(\cdot)$ is finite valued. To tackle this problem, we first reformulate the constraint set as the intersection of adequate closed and convex sets defined as follows: $C_i := \{(y, t) \in \mathbb{R}^p \times \mathbb{R} : \ H_i(y) \leq t\}$,

$$\Delta_m := \left\{ z \in \mathbb{R}^m : \ \sum_{i=1}^{m} z_i \leq \alpha \right\} \quad \text{and} \quad D_i := \{(u, y) \in \mathbb{R}^p \times \mathbb{R}^m : \ B_i u = y\}.$$

Then with these sets, problem (C-Gen) can be written as follows:

$$\min_{u \in \mathbb{R}^p, z_i \in \mathbb{R}, y_i \in \mathbb{R}^{d_i}, \, i=1,2\ldots,m} \left\{ F(u) + \delta_{\Delta_m}(z) + \sum_{i=1}^{m} \delta_{C_i}(y_i, z_i) + \sum_{i=1}^{m} \delta_{D_i}(u, y_i) \right\}.$$

As previously explained, using the dual transportation trick, it is then easy to see that the later can then be written as a saddle-point problem of the form (M) which will involve a separable sum of support functions. Applying the PAPC algorithm on the resulting minimax formulation of problem (C-Gen), requires in this case (thanks to Moreau proximal identity) the computation of the projection onto each set Δ_m, C_i and D_i, $i = 1, 2, \ldots, m$. The projection onto Δ_m and D_i admits a closed form solution. Furthermore, the projection onto a set of the form C_i, namely the epigraph of H_i, can also be computed via the following result whose simple proof is left to the reader.

Proposition 4.1. *Let $H : \mathbb{R}^p \to \mathbb{R}$ be convex and let $C := \{(y, t) \in \mathbb{R}^p \times \mathbb{R} : \; H(y) \leq t\}$. For any $(x, s) \notin C$, let $(\bar{y}, \bar{t}) = P_C((x, s))$ be the projection of (x, s) onto C. Then,*

$$\bar{y} = \operatorname{argmin}_{y \in \mathbb{R}^n} \left\{ \|y - x\|^2 + (H(y) - s)^2 \right\} \quad \textit{and} \quad \bar{t} = H(\bar{y}).$$

For example, when H is a norm, i.e., $H(\cdot) := \|\cdot\|$, then C is the second-order cone for which it is well known that P_C admits an explicit closed form.

Acknowledgment The work of Shoham Sabach was partially supported by a Tel Aviv University postdoctoral fellowship and the work of Marc Teboulle was partially supported by the Israel Science Foundation, ISF Grant 998-12.

References

[1] A. Auslender and M. Teboulle. *Asymptotic Cones and Functions in Optimization and Variational Inequalities.* Springer Monographs in Mathematics. New York: Springer, 2003.

[2] A. Auslender and M. Teboulle. Interior projection-like methods for monotone variational inequalities. *Mathematical Programming*, 104(1):39–68, 2005.

[3] A. Beck and M. Teboulle. Smoothing and first order methods: a unified framework. *SIAM J. Optim.*, 22(2):557–580, 2012.

[4] D. P. Bertsekas. *Nonlinear Programming.* Belmont MA: Athena Scientific, second edition, 1999.

[5] C. Chen, B. He, Y. Ye, and X. Yuan. The direct extension of ADMM for multi-block convex minimization problems is not necessarily convergent. Optimization Online, September 2013.

[6] G. Chen and M. Teboulle. A proximal-based decomposition method for convex minimization problems. *Math. Programming*, 64(1, Ser. A):81–101, 1994.

[7] J. Eckstein and D. P. Bertsekas. On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55(1):293–318, 1992.

[8] F. Facchinei and J. S. Pang. *Finite-dimensional variational inequalities and complementarity problems, Vol. II.* Springer Series in Operations Research. Springer-Verlag, New York, 2003.

[9] D. Gabay. Applications of the method of multipliers to variational inequalities. In M. Fortin and R. Glowinski, editors, *Augmented Lagrangian Methods: Applications to the Solution of Boundary-Valued Problems*, pages 299–331. North-Holland, Amsterdam, 1983, 1983.

[10] G. M. Korpelevitch. The extra gradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976.

[11] P. L. Lions and B. Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM J. Numer. Anal.*, 16(6):964–979, 1979.

[12] J. J. Moreau. Proximité et dualité dans un espace hilbertien. *Bull. Soc. Math. France*, 93:273–299, 1965.

[13] A. Nemirovski. 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 J. Optim.*, 15(1):229–251, 2004.

[14] A. S. Nemirovsky and D. B. Yudin. *Problem complexity and method efficiency in optimization.* A Wiley-Interscience Publication. John Wiley & Sons Inc., New York, 1983. Translated from the Russian and with a preface by E. R. Dawson, Wiley-Interscience Series in Discrete Mathematics.

[15] Y. Nesterov. Smooth minimization of non-smooth functions. *Math. Programming*, 103(1, Ser. A):127–152, 2005.

[16] D. P. Palomar and Y. C. Eldar. *Convex optimization in signal processing and communications.* Cambridge University Press, Cambridge, 2010.

[17] G. B. Passty. Ergodic convergence to a zero of the sum of monotone operators in Hilbert space. *J. Math. Anal. Appl.*, 72(2):383–390, 1979.

[18] R. T. Rockafellar. *Convex Analysis.* Princeton NJ: Princeton Univ. Press, 1970.

[19] R. T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM J. Control Optim.*, 14(5):877–898, 1976.

[20] R. T. Rockafellar and J. B. R. Wets. *Variational analysis.* Springer, 2004.

[21] R. Shefi and M. Teboulle. Rate of convergence analysis of decomposition methods based on the proximal method of multipliemu for convex minimization. *SIAM J. Optimization*, 24:269 – 297, 2014.

[22] S. Sra, S. Nowozin, and S. J. Wright. *Optimization for Machine Learning.* The MIT Press, Cambridge, 2011.

[23] P. Tseng. Applications of a splitting algorithm to decomposition in convex programming and variational inequalities. *SIAM J. Control Optim.*, 29(1):119–138, 1991.

[24] P. Tseng. Alternating projection-proximal methods for convex programming and variational inequalities. *SIAM J. Optim.*, 7(4):951–965, 1997.