A first-order primal-dual algorithm with linesearch

Yura Malitsky*, Thomas Pock*

August 31, 2016

Abstract

The paper proposes a linesearch for the primal-dual method. Each iteration of the linesearch requires to update only the dual (or primal) variable. For many problems, in particular for regularized least squares, the linesearch does not require any additional matrix-vector multiplications. We prove convergence of the proposed method under the standard assumptions. We also show an ergodic $O(1/N)$ rate of convergence for our method. In case when one of the prox-functions is strongly convex, we modify our basic method to get a better convergence rate. Finally, we propose the linesearch for a saddle point problem with an additional smooth term. Numerical experiments approve the efficiency of proposed methods.

2010 Mathematics Subject Classification: 49M29 65K10 65Y20 90C25
Keywords: Saddle-point problems, first order algorithms, primal-dual algorithms, linesearch, convergence rates, backtracking

1 Introduction

In this work we propose a linesearch procedure for the primal-dual algorithm (PDA) that was introduced in [4]. It is a simple first-order method that is widely used for solving nonsmooth composite minimization problems. Recently, it was shown the connection of PDA with proximal point algorithm [11] and ADMM [6]. Some generalizations of the method were considered in [5, 7, 9]. A survey of possible applications of the algorithm can be found in [6, 12].

The basic form of PDA uses fixed step sizes during all iterations. This requires knowing the operator norm of K, which has several drawbacks. First, we need to compute it, i.e., to compute $\lambda_{\max}(K^*K)$, which may be quite expensive for large scale dense matrices. Second, even if we know this norm, one can often use larger steps which usually yields a faster convergence. As a remedy for the first issue one can use a diagonal precondition [15], but still there is no strong evidence that such precondition improves or at least does not worsen the speed of convergence of PDA. Regarding the second issue, as we will see in our experiments, the speed improvement gained by using the linesearch sometimes can be significant.

*Institute for Computer Graphics and Vision, Graz University of Technology, 8010 Graz, Austria. e-mail: yura.malitsky@icg.tugraz.at, pock@icg.tugraz.at

Our proposed analysis of PDA exploits the idea of recent works [13, 14] where are proposed several algorithms for solving a monotone variational inequality. Those algorithms are different from the PDA, however, they use a similar extrapolation step $x^k + \theta(x^k - x^{k-1})$. Although our analysis of the primal-dual method is not so elegant as for example in [11], it gives a simple and a cheap way to incorporate the linesearch for defining the step sizes. Each inner iteration of the linesearch requires to update only the dual variable. Moreover, the step sizes may increase from iteration to iteration. We prove the convergence of the algorithm under quite general assumptions. Also we show that in many important cases the PDAL (primal-dual algorithm with linesearch) preserves the same complexity of iteration as PDA does. In particular, our method, applied for any regularized least-squares problem, uses the same number of matrix-vector multiplication per iteration as forward-backward method or FISTA (both with fixed step size) do, but does not require to know a matrix norm and, in addition, uses adaptive steps.

For the case when a primal or dual objective are strongly convex, we modify our linesearch procedure in order to construct accelerated versions of PDA. This is done in a similar way as in [4]. The obtained algorithms share the same complexity per iteration as PDAL does, but in many cases substantially outperform PDA and PDAL.

We also consider a more general primal-dual problem which involves an additional smooth function with Lipschitz-continuous gradient (see [7]). For this case we generalize our linesearch to avoid knowing that Lipschitz constant.

The authors in [9] also propose a linesearch for the primal-dual method, with the goal to vary the ratio between primal and dual steps such that primal and dual residuals remain roughly of the same size. The same idea was used in [10] for ADMM method. However, we should highlight that this principle is just a heuristic, as it is not clear that it in fact improves the speed of convergence. The linesearch proposed in [9] requires an update of both primal and dual variables, which may make the algorithm much more expensive than basic PDA. Also the authors proved the convergence of the iterates only when one of the sequences (x^k) or (y^k) is bounded. Although this is often the case, there are many problems which can not be encompassed by this assumption. At last, it is not clear how to derive accelerated versions of that algorithm.

As a byproduct of our analysis, we show how one can use a variable ratio between primal and dual steps and under which circumstances we can guarantee convergence. However, it was not the purpose of this paper to develop new strategies how to vary such ratio during iterations.

The paper is organized as follows. In the next section we introduce the notations and recall some useful facts. Section 3 presents our basic primal-dual algorithm with a linesearch. We prove its convergence, establish its ergodic convergence rate and consider some particular examples of how the linesearch works. In section 4 we propose the accelerated versions of PDAL under the assumption that the primal or dual problem is strongly convex. Section 5 concerns with more general saddle point problem which involves an additional smooth function. In section 6 we illustrate the efficiency of our methods for several typical problems.

2 Preliminaries

Let X, Y be two finite-dimensional real vector spaces equipped with an inner product $\langle \cdot, \cdot \rangle$ and norm $\|\cdot\| = \sqrt{\langle \cdot, \cdot \rangle}$. We are focusing on the following problem

$$\min_{x \in X} \max_{y \in Y} \langle Kx, y \rangle + g(x) - f^*(y), \tag{1}$$

where

- $K \colon X \to Y$ is a bounded linear operator, with operator norm $L = \|K\|$;
- $g \colon X \to (-\infty, +\infty]$ and $f^* \colon Y \to (-\infty, +\infty]$ are proper lower semicontinuous convex functions.

Note that f^* denotes the Legendre-Fenchel conjugate of a convex l.s.c. function f. We hope there will be no ambiguity for the reader when we write K^* to denote the adjoint of the operator K.

Throughout the paper we assume that problem (1) has a saddle point. (1) is equivalent to the following primal

$$\min_{x \in X} f(Kx) + g(x)$$

and respectively dual problem:

$$\min_{y \in Y} f^*(y) + g^*(-K^*x).$$

Recall that for a proper lower semi-continuous convex function $h \colon X \to (-\infty, +\infty]$ the proximal operator prox_h is defined as

$$\operatorname{prox}_h \colon X \to X \colon x \mapsto \operatorname{argmin}_z \{h(z) + \frac{1}{2} \|z - x\|^2\}.$$

The following important characteristic property of the proximal operator is well-known:

$$\bar{x} = \operatorname{prox}_h x \quad \Leftrightarrow \quad \langle \bar{x} - x, y - \bar{x} \rangle \geq h(\bar{x}) - h(y) \quad \forall y \in X. \tag{2}$$

We will often use the following identity (cosine rule):

$$2 \langle a - b, c - a \rangle = \|b - c\|^2 - \|a - b\|^2 - \|a - c\|^2 \quad \forall a, b, c \in X.$$

Let $(\hat{x}, \hat{y})$ be a saddle point of problem (1). Then by the definition of the saddle point we have

$$\begin{aligned} P_{\hat{x},\hat{y}}(x) &:= g(x) - g(\hat{x}) + \langle K^*\hat{y}, x - \hat{x} \rangle \geq 0 \quad \forall x \in X, \\ D_{\hat{x},\hat{y}}(y) &:= f^*(y) - f^*(\hat{y}) - \langle K\hat{x}, y - \hat{y} \rangle \geq 0 \quad \forall y \in Y. \end{aligned} \tag{3}$$

The expression $\mathcal{G}_{\hat{x},\hat{y}}(x, y) = P_{\hat{x},\hat{y}}(x) + D_{\hat{x},\hat{y}}(y)$ is known as a primal-dual gap. In certain cases when it is clear which saddle point is considered, we will omit the subscript in P, D, and $\mathcal{G}$. It is also important to highlight that for fixed $(\hat{x}, \hat{y})$ functions $P(\cdot)$, $D(\cdot)$ are convex.

Consider the original primal-dual method

$$\begin{aligned} y^{k+1} &= \operatorname{prox}_{\sigma f^*}(y^k + \sigma K \bar{x}^k) \\ x^{k+1} &= \operatorname{prox}_{\tau g}(x^k - \tau K^* y^{k+1}) \\ \bar{x}^{k+1} &= x^{k+1} + \theta(x^{k+1} - x^k). \end{aligned}$$

In [4] its convergence was proved under assumptions $\theta = 1$, $\tau, \sigma > 0$, and $\tau\sigma L^2 < 1$. In the next section we will show how to incorporate the linesearch into this method.

3 Linesearch

Formally the method is the following

Algorithm 1 *Primal-dual algorithm with linesearch*

Initialization: Choose $x^0 \in X$, $y^1 \in Y$, $\tau_0 > 0$, $\mu \in (0,1), \delta \in (0,1)$, and $\beta > 0$. Set $\theta_0 = 1$.

Main iteration:

1. Compute
$$x^k = \operatorname{prox}_{\tau_{k-1} g}(x^{k-1} - \tau_{k-1} K^* y^k).$$

2. Choose any $\tau_k \in [\tau_{k-1}, \tau_{k-1}\sqrt{1+\theta_{k-1}}]$ and run

   Linesearch:

   2.a. Compute
   $$\begin{aligned} \theta_k &= \frac{\tau_k}{\tau_{k-1}} \\ \bar{x}^k &= x^k + \theta_k(x^k - x^{k-1}) \\ y^{k+1} &= \operatorname{prox}_{\beta\tau_k f^*}(y^k + \beta\tau_k K \bar{x}^k) \end{aligned}$$

   2.b. Break linesearch if
   $$\sqrt{\beta}\tau_k \left\| K^* y^{k+1} - K^* y^k \right\| \le \delta \left\| y^{k+1} - y^k \right\| \tag{4}$$

   Otherwise, set $\tau_k := \tau_k \mu$ and go to 2.a.

End of linesearch

Given all information from the current iterate: x^k, y^k, τ_{k-1}, and θ_{k-1}, we first choose some trial step $\tau_k \in [\tau_{k-1}, \tau_{k-1}\sqrt{1+\theta_{k-1}}]$, then during every iteration of the linesearch it is decreased by μ. At the end of the linesearch we obtain a new iterate y^{k+1} and a step size τ_k which we will use to compute the next iterate x^{k+1}. In the Algorithm 1 there are two opposite options: always start the linesearch from the largest possible step $\tau_k = \tau_{k-1}\sqrt{1+\theta_{k-1}}$, or, on the contrary, never increase τ_k. Step 2 allows to find a compromise between them.

Note that in PDAL parameter β plays the role of the ratio $\dfrac{\sigma}{\tau}$ between fixed steps in PDA. Thus, we can rewrite the stopping criteria (4) as

$$\tau_k\sigma_k \left\| K^* y^{k+1} - K^* y^k \right\|^2 \le \delta^2 \left\| y^{k+1} - y^k \right\|^2,$$

where $\sigma_k = \beta\tau_k$. Of course in PDAL we can always choose fixed steps τ_k, σ_k with $\tau_k\sigma_k \leq \dfrac{\delta^2}{L^2}$ and set $\theta_k = 1$. In this case PDAL will coincide with PDA, though our proposed analysis seems to be new.

Remark 1. It is clear that each iteration of the linesearch requires computation of $\operatorname{prox}_{\beta\tau_k f^*}(\cdot)$ and K^*y^{k+1}. As in problem (1) we can always exchange primal and dual variables, it makes sense to choose for the dual variable in PDAL that one for which the respective prox-operator is simpler to compute. Note that during the linesearch we need to compute only once Kx^k and then use that $K\bar{x}^k = (1+\theta_k)Kx^k - \theta_k Kx^{k-1}$.

Remark 2. Note that when $\operatorname{prox}_{\sigma f^*}$ is a linear (or affine) operator, the linesearch becomes extremely simple: it does not require any additional matrix-vector multiplications. We itemize some examples below:

1. $f^*(y) = \langle c, y\rangle$. Then it is easy to verify that $\operatorname{prox}_{\sigma f^*} u = u - \sigma c$. Thus, we have

$$\begin{aligned} y^{k+1} &= \operatorname{prox}_{\sigma_k f^*}(y^k + \sigma_k K\bar{x}^k) = y^k + \sigma_k K\bar{x}^k - \sigma_k c \\ K^*y^{k+1} &= K^*y^k + \sigma_k(K^*K\bar{x}^k - K^*c) \end{aligned}$$

2. $f^*(y) = \frac{1}{2}\|y-b\|^2$. Then $\operatorname{prox}_{\sigma f^*} u = \frac{u+\sigma b}{1+\sigma}$ and we obtain

$$\begin{aligned} y^{k+1} &= \operatorname{prox}_{\sigma_k f^*}(y^k + \sigma_k K\bar{x}^k) = \frac{y^k + \sigma_k(K\bar{x}^k + b)}{1+\sigma_k} \\ K^*y^{k+1} &= \frac{1}{1+\sigma_k}(K^*y^k + \sigma_k(K^*K\bar{x}^k + K^*b)) \end{aligned}$$

3. $f^*(y) = \delta_H(y)$, the characteristic function of the hyperplane $H = \{u\colon \langle u, a\rangle = b\}$. Then $\operatorname{prox}_{\sigma f^*} u = P_H u = u + \frac{b-\langle u,a\rangle}{\|a\|^2}a$. And hence,

$$\begin{aligned} y^{k+1} &= \operatorname{prox}_{\sigma_k f^*}(y^k + \sigma_k K\bar{x}^k) = y^k + \sigma_k K\bar{x}^k + \frac{b - \left\langle a, y^k + \sigma_k K\bar{x}^k\right\rangle}{\|a\|^2}a \\ K^*y^{k+1} &= K^*y^k + \sigma_k K^*K\bar{x}^k + \frac{b - \left\langle a, y^k + \sigma_k K\bar{x}^k\right\rangle}{\|a\|^2}K^*a. \end{aligned}$$

Evidently, each iteration of PDAL for all the cases above requires only two matrix-vector multiplications. Namely, before the linesearch starts we have to compute Kx^k and K^*Kx^k and then during the linesearch should use the following relations

$$\begin{aligned} K\bar{x}^k &= (1+\theta_k)Kx^k - \theta_k Kx^{k-1} \\ K^*K\bar{x}^k &= (1+\theta_k)K^*Kx^k - \theta_k K^*Kx^{k-1}, \end{aligned}$$

where Kx^{k-1} and K^*Kx^{k-1} should be reused from the previous iteration. All other operations are comparably cheap and, hence, the cost per iteration of PDA and PDAL are almost the same. Of course, for problems with very sparse matrices such as finite differences operators, this is not true, since in that case the cost of a matrix-vector multiplication is comparable with the cost of a vector-vector addition.

One simple implication of these facts is that for a regularized least-squares problem

$$\min_x \frac{1}{2}\|Ax-b\|^2 + g(x) \tag{5}$$

our method does not require to know $\|A\|$ but at the same time does not require any additional matrix-vector multiplication as it is the case for standard first order methods with backtracking (e.g. proximal gradient method, FISTA). In fact, we can rewrite (5) as

$$\min_x \max_y g(x) + \langle Ax, y\rangle - \frac{1}{2}\|y+b\|^2. \tag{6}$$

Here $f^*(y) = \frac{1}{2}\|y+b\|^2$ and hence, we are in the situation where operator prox_{f^*} is affine.

By the construction of the algorithm we simply have the following:

Lemma 1. (i) The linesearch in the PDAL always terminates.

(ii) There exists $\tau > 0$ such that $\tau_k > \tau$ for all $k \geq 0$.

(iii) There exists $\theta > 0$ such that $\theta_k \leq \theta$ for all $k \geq 0$.

Proof. (i) In each iteration of the linesearch τ_k is multiplied by factor $\mu < 1$. Since, (4) is satisfied for any $\tau_k \leq \frac{\delta}{\sqrt{\beta}L}$, the inner loop can not run infinitely long.

(ii) Without loss of generality, assume that $\tau_0 > \frac{\delta\mu}{\sqrt{\beta}L}$. Our goal is to show that from $\tau_{k-1} > \frac{\delta\mu}{\sqrt{\beta}L}$ follows $\tau_k > \frac{\delta\mu}{\sqrt{\beta}L}$. Suppose that $\tau_k = \tau_{k-1}\sqrt{1+\theta_{k-1}}\mu^i$ for some $i \in \mathbb{Z}^+$. If $i = 0$ then $\tau_k > \tau_{k-1} > \frac{\delta\mu}{\sqrt{\beta}L}$. If $i > 0$ then $\tau_k' = \tau_{k-1}\sqrt{1+\theta_{k-1}}\mu^{i-1}$ does not satisfy (4). Thus, $\tau_k' > \frac{\delta}{\sqrt{\beta}L}$ and hence, $\tau_k > \frac{\delta\mu}{\sqrt{\beta}L}$.

(iii) From $\tau_k \leq \tau_{k-1}\sqrt{1+\theta_{k-1}}$ it follows that $\theta_k \leq \sqrt{1+\theta_{k-1}}$. From this it can be easily conclude that $\theta_k \leq \frac{\sqrt{5}+1}{2}$ for all $k \in \mathbb{Z}^+$. In fact, assume the contrary, and let r be the smallest number such that $\theta_r > \frac{\sqrt{5}+1}{2}$. Since $\theta_0 = 1$, we have $r \geq 1$, and hence $\theta_{r-1} \geq \theta_r^2 - 1 > \frac{\sqrt{5}+1}{2}$. This yields a contradiction. □

Theorem 1. *Let (x^k, y^k) be a sequence generated by PDAL. Then it is a bounded sequence in $X \times Y$ and all its cluster points are solutions of* (1)*. Moreover, if $g|_{\mathrm{dom}\, g}$ is continuous and (τ_k) is bounded from above then the whole sequence (x^k, y^k) converges to a solution of* (1)*.*

It seems that convergence will be still in place without these assumptions, however it is not clear how to prove this. The condition of $g|_{\mathrm{dom}\, g}$ to be continuous is not restrictive: it holds for any g with open $\mathrm{dom}\, g$ (this includes all finite-valued functions) or for an indicator δ_C of any closed convex set C. Also it holds for any separable convex l.s.c. function (Corollary 9.15, [2]). The boundedness of (τ_k) from the above is rather theoretical: clearly we can easily bound it in the PDAL.

Proof. Let $(\hat{x}, \hat{y})$ be any saddle point of (1). From (2) it follows that

$$\left\langle x^{k+1} - x^k + \tau_k K^* y^{k+1}, \hat{x} - x^{k+1}\right\rangle \geq \tau_k(g(x^{k+1}) - g(\hat{x})) \tag{7}$$

$$\left\langle \frac{1}{\beta}(y^{k+1} - y^k) - \tau_k K\bar{x}^k, \hat{y} - y^{k+1}\right\rangle \geq \tau_k(f^*(y^{k+1}) - f^*(\hat{y})). \tag{8}$$

Since $x^k = \operatorname{prox}_{\tau_{k-1}g}(x^{k-1} - \tau_{k-1}K^*y^k)$, we have again by (2) that for all $x \in X$

$$\left\langle x^k - x^{k-1} + \tau_{k-1}K^*y^k, x - x^k \right\rangle \geq \tau_{k-1}(g(x^k) - g(x)).$$

After substitution in the last inequality $x = x^{k+1}$ and $x = x^{k-1}$ we get

$$\left\langle x^k - x^{k-1} + \tau_{k-1}K^*y^k, x^{k+1} - x^k \right\rangle \geq \tau_{k-1}(g(x^k) - g(x^{k+1})), \tag{9}$$

$$\left\langle x^k - x^{k-1} + \tau_{k-1}K^*y^k, x^{k-1} - x^k \right\rangle \geq \tau_{k-1}(g(x^k) - g(x^{k-1})). \tag{10}$$

Adding (9), multiplied by $\theta_k = \frac{\tau_k}{\tau_{k-1}}$, and (10), multiplied by θ_k^2, we obtain

$$\left\langle \bar{x}^k - x^k + \tau_k K^*y^k, x^{k+1} - \bar{x}^k \right\rangle \geq \tau_k((1+\theta_k)g(x^k) - g(x^{k+1}) - \theta_k g(x^{k-1})), \tag{11}$$

where we have used that $\bar{x}^k = x^k + \theta_k(x^k - x^{k-1})$.

Consider the following identity

$$\tau_k \left\langle K^*y^{k+1} - K^*\hat{y}, \bar{x}^k - \hat{x} \right\rangle - \tau_k \left\langle K\bar{x}^k - K\hat{x}, y^{k+1} - \hat{y} \right\rangle = 0. \tag{12}$$

Summing (7), (8), (11), and (12), we get

$$\begin{aligned}\left\langle x^{k+1} - x^k, \hat{x} - x^{k+1} \right\rangle + \frac{1}{\beta}\left\langle y^{k+1} - y^k, \hat{y} - y^{k+1} \right\rangle + \left\langle \bar{x}^k - x^k, x^{k+1} - \bar{x}^k \right\rangle \\ + \tau_k \left\langle K^*y^{k+1} - K^*y^k, \bar{x}^k - x^{k+1} \right\rangle - \tau_k \left\langle K^*y, \bar{x}^k - \hat{x} \right\rangle + \tau_k \left\langle K\hat{x}, y^{k+1} - y \right\rangle \\ \geq \tau_k\big(f^*(y^{k+1}) - f^*(\hat{y}) + (1+\theta_k)g(x^k) - \theta_k g(x^{k-1}) - g(\hat{x})\big).\end{aligned} \tag{13}$$

Using that

$$f^*(y^{k+1}) - f^*(\hat{y}) - \left\langle K\hat{x}, y^{k+1} - \hat{y} \right\rangle = D_{\hat{x},\hat{y}}(y^{k+1})$$

and

$$\begin{aligned}&(1+\theta_k)g(x^k) - \theta_k g(x^{k-1}) - g(\hat{x}) + \left\langle K^*y, \bar{x}^k - \hat{x} \right\rangle \\ &= (1+\theta_k)\left(g(x^k) - g(\hat{x}) + \left\langle K^*\hat{y}, x^k - \hat{x} \right\rangle\right) - \theta_k\left(g(x^{k-1}) - g(\hat{x}) + \left\langle K^*\hat{y}, x^{k-1} - \hat{x} \right\rangle\right) \\ &= (1+\theta_k)P_{\hat{x},\hat{y}}(x^k) - \theta_k P_{\hat{x},\hat{y}}(x^{k-1}),\end{aligned}$$

we can rewrite (13) as (we will not further write the subscripts for P and D unless it is unclear)

$$\begin{aligned}\left\langle x^{k+1} - x^k, \hat{x} - x^{k+1} \right\rangle + \frac{1}{\beta}\left\langle y^{k+1} - y^k, \hat{y} - y^{k+1} \right\rangle + \left\langle \bar{x}^k - x^k, x^{k+1} - \bar{x}^k \right\rangle \\ + \tau_k \left\langle K^*y^{k+1} - K^*y^k, \bar{x}^k - x^{k+1} \right\rangle \\ \geq \tau_k((1+\theta_k)P(x^k) - \theta_k P(x^{k-1}) + D(y^{k+1})).\end{aligned} \tag{14}$$

Let ε_k denotes the right-hand side of (14). Using the cosine rule for every item in the first line of (14), we obtain

$$\begin{aligned}\frac{1}{2}(\|x^k - \hat{x}\|^2 - \|x^{k+1} - \hat{x}\|^2 - \|x^{k+1} - x^k\|^2) + \frac{1}{2\beta}(\|y^k - \hat{y}\|^2 - \|y^{k+1} - \hat{y}\|^2 - \|y^{k+1} - y^k\|^2) \\ + \frac{1}{2}(\|x^{k+1} - x^k\|^2 - \|\bar{x}^k - x^k\|^2 - \|x^{k+1} - \bar{x}^k\|^2) \\ + \tau_k \left\langle K^*y^{k+1} - K^*y^k, \bar{x}^k - x^{k+1} \right\rangle \geq \varepsilon_k.\end{aligned}$$

By (4), Cauchy-Schwarz, and Cauchy's inequalities we get

$$\begin{aligned}\tau_k \left\langle K^* y^{k+1} - K^* y^k, \bar{x}^k - x^{k+1} \right\rangle &\leq \frac{\delta}{\sqrt{\beta}} \left\| x^{k+1} - \bar{x}^k \right\| \left\| y^{k+1} - y^k \right\| \\ &\leq \frac{1}{2} \left\| x^{k+1} - \bar{x}^k \right\|^2 + \frac{\delta^2}{2\beta} \left\| y^{k+1} - y^k \right\|^2,\end{aligned}$$

from which we derive that

$$\begin{aligned}\frac{1}{2}(\left\| x^k - \hat{x} \right\|^2 - \left\| x^{k+1} - \hat{x} \right\|^2) + \frac{1}{2\beta}(\left\| y^k - \hat{y} \right\|^2 - \left\| y^{k+1} - \hat{y} \right\|^2) \\ - \frac{1}{2} \left\| \bar{x}^k - x^k \right\|^2 - \frac{1-\delta^2}{2\beta} \left\| y^{k+1} - y^k \right\|^2 \geq \varepsilon_k. \quad (15)\end{aligned}$$

Since $(\hat{x}, \hat{y})$ is a saddle point, $D(y^k) \geq 0$ and $P(x^k) \geq 0$ and hence (15) yields

$$\begin{aligned}\frac{1}{2}(\left\| x^k - \hat{x} \right\|^2 - \left\| x^{k+1} - \hat{x} \right\|^2) + \frac{1}{2\beta}(\left\| y^k - \hat{y} \right\|^2 - \left\| y^{k+1} - \hat{y} \right\|^2) \\ - \frac{1}{2} \left\| \bar{x}^k - x^k \right\|^2 - \frac{1-\delta^2}{2\beta} \left\| y^{k+1} - y^k \right\|^2 \geq \tau_k((1+\theta_k)P(x^k) - \theta_k P(x^{k-1})) \quad (16)\end{aligned}$$

or, taking into account $\theta_k \tau_k \leq (1+\theta_{k-1})\tau_{k-1}$,

$$\begin{aligned}\frac{1}{2} \left\| x^{k+1} - \hat{x} \right\|^2 + \frac{1}{2\beta} \left\| y^{k+1} - \hat{y} \right\|^2 + \tau_k(1+\theta_k)P(x^k) \leq \\ \frac{1}{2} \left\| x^k - \hat{x} \right\|^2 + \frac{1}{2\beta} \left\| y^k - \hat{y} \right\|^2 + \tau_{k-1}(1+\theta_{k-1})P(x^{k-1}) \\ - \frac{1}{2} \left\| \bar{x}^k - x^k \right\|^2 - \frac{1-\delta^2}{2\beta} \left\| y^{k+1} - y^k \right\|^2 \quad (17)\end{aligned}$$

From this we deduce that (x^k), (y^k) are bounded sequences and $\lim_{k\to\infty} \left\| \bar{x}^k - x^k \right\| = 0$, $\lim_{k\to\infty} \left\| y^k - y^{k-1} \right\| = 0$. Also notice that

$$\frac{x^{k+1} - x^k}{\tau_k} = \frac{\bar{x}^{k+1} - x^{k+1}}{\tau_{k+1}} \to 0 \quad \text{as } k \to \infty,$$

where the latter holds because τ_k is separated from 0 by Lemma 1. Let (x^{k_i}, y^{k_i}) be a subsequence that converges to some cluster point (x^*, y^*). Passing to the limit in

$$\begin{aligned}\left\langle \frac{1}{\tau_{k_i}}(x^{k_i+1} - x^{k_i}) + K^* y^{k_i}, x - x^{k_i+1} \right\rangle \geq g(x^{k_i+1}) - g(x) \quad \forall x \in X, \\ \left\langle \frac{1}{\beta\tau_{k_i}}(y^{k_i+1} - y^{k_i}) - K\bar{x}^{k_i}, y - y^{k_i+1} \right\rangle \geq f^*(y^{k_i+1}) - f^*(y) \quad \forall y \in Y,\end{aligned}$$

we obtain that (x^*, y^*) is a saddle point of (1).

When $g|_{\operatorname{dom} g}$ is continuous, $g(x^{k_i}) \to g(x^*)$, and hence, $P_{x^*,y^*}(x^{k_i}) \to 0$. From (17) it follows that the sequence $a_k = \frac{1}{2} \left\| x^{k+1} - x^* \right\|^2 + \frac{1}{2\beta} \left\| y^{k+1} - y^* \right\|^2 + \tau_k(1+\theta_k)P_{x^*,y^*}(x^k)$ is monotone. Taking into account boundedness of (τ_k) and (θ_k), we obtain

$$\lim_{k\to\infty} a_k = \lim_{i\to\infty} a_{k_i} = 0,$$

which means that $x^k \to x^*$, $y^k \to y^*$. □

Theorem 2 (Ergodic convergence)**.** *Let (x^k, y^k) be a sequence generated by PDAL and $(\hat{x}, \hat{y})$ be any saddle point of* (1)*. Then for the ergodic sequence (X^N, Y^N) holds*

$$\mathcal{G}_{\hat{x},\hat{y}}(X^N, Y^N) \leq \frac{1}{s_N}\left(\frac{1}{2}\left\|x^1 - \hat{x}\right\|^2 + \frac{1}{2\beta}\left\|y^1 - \hat{y}\right\|^2 + \tau_1\theta_1 P_{\hat{x},\hat{y}}(x^0)\right),$$

where $s_N = \sum_{k=1}^N \tau_k$, $X^N = \dfrac{\tau_1\theta_1 x^0 + \sum_{k=1}^N \tau_k \bar{x}^k}{\tau_1\theta_1 + s_N}$, $Y^N = \dfrac{\sum_{k=1}^N \tau_k y^k}{s_N}$.

Proof. Summing (15) from $k = 1$ to N, we get

$$\frac{1}{2}(\left\|x^1 - \hat{x}\right\|^2 - \left\|x^{N+1} - \hat{x}\right\|^2) + \frac{1}{2\beta}(\left\|y^1 - \hat{y}\right\|^2 - \left\|y^{N+1} - \hat{y}\right\|^2) \geq \sum_{k=1}^N \varepsilon_k \tag{18}$$

The right side in (18) can be expressed as

$$\sum_{k=1}^N \varepsilon_k = \tau_N(1+\theta_N)P(x^N) + \sum_{k=2}^N [(1+\theta_{k-1})\tau_{k-1} - \theta_k\tau_k)]P(x^{k-1}) - \theta_1\tau_1 P(x^0) + \sum_{k=1}^N \tau_k D(y^{k+1})$$

By convexity of P,

$$\begin{aligned}\tau_N(1+\theta_N)P(x^N) + \sum_{k=2}^N [(1+\theta_{k-1})\tau_{k-1} - \theta_k\tau_k)]P(x^{k-1}) \\ \geq (\tau_1\theta_1 + s_N)P\left(\frac{\tau_1(1+\theta_1)x^1 + \sum_{k=2}^N \tau_k\bar{x}^k}{\tau_1\theta_1 + s_N}\right) \\ = (\tau_1\theta_1 + s_N)P\left(\frac{\tau_1\theta_1 x^0 + \sum_{k=1}^N \tau_k\bar{x}^k}{\tau_1\theta_1 + s_N}\right) \geq s_N P(X^N),\end{aligned} \tag{19}$$

where $s_N = \sum_{k=1}^N \tau_k$. Similarly,

$$\sum_{k=1}^N \tau_k D(y^k) \geq s_N D\left(\frac{\sum_{k=1}^N \tau_k y^k}{s_N}\right) = s_N D(Y^N). \tag{20}$$

Hence

$$\sum_{k=1}^N \varepsilon_k \geq s_N(P(X^N) + D(Y^N)) - \tau_1\theta_1 P(x^0)$$

and we conclude

$$\mathcal{G}(X^N, Y^N) = P(X^N) + D(Y^N) \leq \frac{1}{s_N}\left(\frac{1}{2}\left\|x^1 - \hat{x}\right\|^2 + \frac{1}{2\beta}\left\|y^1 - \hat{y}\right\|^2 + \tau_1\theta_1 P(x^0)\right).$$

□

Clearly, we have the same $O(1/N)$ rate of convergence as in [4,5], though with the ergodic sequence (X^N, Y^N) defined in a different way.

It is well-known that in many cases the speed of convergence of PDA crucially depends on the ratio between primal and dual steps $\beta = \dfrac{\sigma}{\tau}$. Motivated by this, paper [9] proposed an adaptive strategy how to choose β in every iteration. Although it is not the goal of this work to study the strategies for defining β, we show that analysis of PDAL allows to incorporate such strategies in a very natural way.

Theorem 3. *Let $(\beta_k) \subset (0,\infty)$ be a bounded sequence and (x^k, y^k) be a sequence generated by PDAL with variable (β_k). Suppose that one of the following holds:*

(i) (β_k) is monotone;

(ii) (y^k) is bounded.

Then the statement of Theorem 1 holds.

Proof. Let (β_k) be nondecreasing. Then using $\frac{1}{\beta_k} \le \frac{1}{\beta_{k-1}}$, we get from (17)

$$\frac{1}{2}\left\|x^{k+1}-\hat{x}\right\|^2 + \frac{1}{2\beta_k}\left\|y^{k+1}-\hat{y}\right\|^2 + \tau_k(1+\theta_k)P(x^k) \le \frac{1}{2}\left\|x^k-\hat{x}\right\|^2 + \frac{1}{2\beta_{k-1}}\left\|y^k-\hat{y}\right\|^2 + \tau_{k-1}(1+\theta_{k-1})P(x^{k-1}) - \frac{1}{2}\left\|\bar{x}^k-x^k\right\|^2 - \frac{1-\delta^2}{2\beta_k}\left\|y^{k+1}-y^k\right\|^2.$$

Since (β_k) is separated from zero and bounded from above, the conclusion in Theorem 1 simply follows.

If (β_k) is decreasing than the above arguments should be modified. Consider (16), multiplied by β_k,

$$\frac{\beta_k}{2}(\left\|x^k-\hat{x}\right\|^2 - \left\|x^{k+1}-\hat{x}\right\|^2) + \frac{1}{2}(\left\|y^k-\hat{y}\right\|^2 - \left\|y^{k+1}-\hat{y}\right\|^2) - \frac{\beta_k}{2}\left\|\bar{x}^k-x^k\right\|^2 - \frac{1-\delta^2}{2}\left\|y^{k+1}-y^k\right\|^2 \ge \beta_k\tau_k((1+\theta_k)P(x^k) - \theta_k P(x^{k-1})). \quad (21)$$

As $\beta_k < \beta_{k-1}$, we have $\theta_k\beta_k\tau_k \le (1+\theta_{k-1})\beta_{k-1}\tau_{k-1}$ that in turn implies

$$\frac{\beta_k}{2}\left\|x^{k+1}-\hat{x}\right\|^2 + \frac{1}{2}\left\|y^{k+1}-\hat{y}\right\|^2 + \tau_k\beta_k(1+\theta_k)P(x^k) \le \frac{\beta_{k-1}}{2}\left\|x^k-\hat{x}\right\|^2 + \frac{1}{2}\left\|y^k-\hat{y}\right\|^2 + \tau_{k-1}\beta_{k-1}(1+\theta_{k-1})P(x^{k-1}) - \frac{\beta_k}{2}\left\|\bar{x}^k-x^k\right\|^2 - \frac{1-\delta^2}{2}\left\|y^{k+1}-y^k\right\|^2.$$

Due to the given properties of (β_k), the rest is trivial.

The case when (y^k) is a bounded sequence is simpler. We only need to observe that it still follows from (17) (with β_k instead of β) that (x^k) is bounded and $\lim_{k\to\infty}\left\|\bar{x}^k-x^k\right\| = 0$, $\lim_{k\to\infty}\left\|y^k-y^{k-1}\right\| = 0$. These are the only ingredients which we need to prove convergence. □

Furthermore, we can obtain a similar result when (x^k) is bounded. However, this requires some modification of Step 2 in PDAL. Namely, we have to ensure that $\tau_k = \tau_{k-1}\sqrt{\frac{\beta_{k-1}}{\beta_k}(1+\theta_{k-1})}$. This yields $\theta_k\beta_k\tau_k \le (1+\theta_{k-1})\beta_{k-1}\tau_{k-1}$, similar to what we have used in (21).

Most often we know that (y^k) is bounded a priori when f^* is a characteristic function of some bounded convex set. It is natural to ask if it is possible to use nonmonotone (β_k)

without the requirement for (y^k) therefor to be bounded? To answer this question, we can easily use the strategies from [10] where ADMM with variable steps was proposed. One way is to use any β_k during a finite number of iterations and then switch to monotone (β_k). Another way is to relax monotonicity of (β_k) to the following:

$$\text{there exists } (\rho_k) \subset \mathbb{R}_+ \quad \text{such that:} \sum_k \rho_k < \infty \quad \text{and}$$
$$\beta_k \le \beta_{k-1}(1+\rho_k) \quad \forall k \in \mathbb{N} \quad \text{or}$$
$$\beta_{k-1} \le \beta_k(1+\rho_k) \quad \forall k \in \mathbb{N}.$$

We suppose it should be quite straightforward to prove convergence of PDAL with the latter strategy.

4 Acceleration

It is shown [4] that in case g or f^* are strongly convex, one can modify the primal-dual algorithm and derive a better convergence rate. We show that the same holds for PDAL. The main difference of the accelerated variant APDAL from the basic PDAL is that now we have to vary β in every iteration.

Of course due to the symmetry of the primal and dual variables in (1), we can always suppose that the primal objective is strongly convex. However, from the computational point of view it might not be desirable to exchange g and f^* in the PDAL because of Remark 2. Therefore, we discuss two cases separately. Also notice that both accelerated algorithms below coincide with PDAL when a parameter of strong convexity $\gamma = 0$.

4.1 g is strongly convex

Assume that g is γ–strongly convex, i.e.,

$$g(x_2) - g(x_1) \ge \langle u, x_2 - x_1\rangle + \frac{\gamma}{2}\|x_2 - x_1\|^2 \quad \forall x_1, x_2 \in X, u \in \partial g(x_1).$$

Below we assume that the parameter γ is known. The following algorithm (APDAL) exploits the strong convexity of g:

Note that in contrast to PDAL, we set $\delta = 1$, as in any case we will not be able to prove convergence of (y^k).

Instead of (7), now one can use a stronger inequality:

$$\left\langle x^{k+1} - x^k + \tau_k K^* y^{k+1}, \hat{x} - x^{k+1}\right\rangle \ge \tau_k(g(x^{k+1}) - g(\hat{x}) + \frac{\gamma}{2}\left\|x^{k+1} - \hat{x}\right\|^2). \tag{23}$$

In turn, (23) yields a stronger version of (15) (also with β_k instead of β):

$$\frac{1}{2}(\left\|x^k - \hat{x}\right\|^2 - \left\|x^{k+1} - \hat{x}\right\|^2) + \frac{1}{2\beta_k}(\left\|y^k - \hat{y}\right\|^2 - \left\|y^{k+1} - \hat{y}\right\|^2)$$
$$- \frac{1}{2}\left\|\bar{x}^k - x^k\right\|^2 \ge \varepsilon_k + \frac{\gamma\tau_k}{2}\left\|x^{k+1} - \hat{x}\right\|^2$$

Algorithm 2 *Accelerated primal-dual algorithm with linesearch: g is strongly convex*

Initialization: Choose $x^0 \in X$, $y^1 \in Y$, $\mu \in (0,1)$, $\tau_0 > 0$, $\beta > 0$. Set $\theta_0 = 1$.
Main iteration:
1. Compute

$$x^k = \operatorname{prox}_{\tau_{k-1} g}(x^{k-1} - \tau_{k-1} K^* y^k)$$
$$\beta_k = \beta_{k-1}(1 + \gamma \tau_{k-1})$$

2. Choose any $\tau_k \in [\tau_{k-1}\sqrt{\frac{\beta_{k-1}}{\beta_k}}, \tau_{k-1}\sqrt{\frac{\beta_{k-1}}{\beta_k}(1+\theta_{k-1})}]$ and run
Linesearch:
2.a. Compute

$$\theta_k = \frac{\tau_k}{\tau_{k-1}}$$
$$\bar{x}^k = x^k + \theta_k(x^k - x^{k-1})$$
$$y^{k+1} = \operatorname{prox}_{\beta_k \tau_k f^*}(y^k + \beta_k \tau_k K \bar{x}^k)$$

2.b. Break linesearch if

$$\sqrt{\beta_k}\tau_k \left\| K^* y^{k+1} - K^* y^k \right\| \le \left\| y^{k+1} - y^k \right\| \tag{22}$$

Otherwise, set $\tau_k := \tau_k \mu$ and go to 2.a.
End of linesearch

or, alternatively,

$$\frac{1}{2}\left\|x^k - \hat{x}\right\|^2 + \frac{1}{2\beta_k}\left\|y^k - \hat{y}\right\|^2 - \frac{1}{2}\left\|\bar{x}^k - x^k\right\|^2$$
$$\ge \varepsilon_k + \frac{1+\gamma\tau_k}{2}\left\|x^{k+1} - \hat{x}\right\|^2 + \frac{\beta_{k+1}}{\beta_k}\frac{1}{2\beta_{k+1}}\left\|y^{k+1} - \hat{y}\right\|^2. \tag{24}$$

Note that the algorithm provides that $\frac{\beta_{k+1}}{\beta_k} = 1 + \gamma\tau_k$. For brevity let

$$A_k = \frac{1}{2}\left\|x^k - \hat{x}\right\|^2 + \frac{1}{2\beta_k}\left\|y^k - \hat{y}\right\|^2.$$

Then from (24) follows

$$\frac{\beta_{k+1}}{\beta_k} A_{k+1} + \varepsilon_k \le A_k$$

or

$$\beta_{k+1} A_{k+1} + \beta_k \varepsilon_k \le \beta_k A_k.$$

Thus, summing the above from $k = 1$ to N, we get

$$\beta_{N+1} A_{N+1} + \sum_{k=1}^{N} \beta_k \varepsilon_k \le \beta_1 A_1.$$

Using the convexity in the same way as in (19), (20), we obtain

$$\sum_{k=1}^{N}\beta_k\varepsilon_k = \beta_N\tau_N(1+\theta_N)P(x^N)+\sum_{k=2}^{N}[(1+\theta_{k-1})\beta_{k-1}\tau_{k-1}-\theta_k\beta_k\tau_k)]P(x^{k-1})-\theta_1\beta_1\tau_1P(x^0)$$
$$+\sum_{k=1}^{N}\beta_k\tau_kD(y^{k+1}) \geq s_N(P(X^N)+D(Y^N))-\theta_1\sigma_1P(x^0),$$

where

$$\sigma_k = \beta_k\tau_k \qquad s_N = \sum_{k=1}^{N}\sigma_k$$
$$X^N = \frac{\sigma_1\theta_1x^0+\sum_{k=1}^{N}\sigma_k\bar{x}^k}{\sigma_1\theta_1+s_N} \qquad Y^N = \frac{\sum_{k=1}^{N}\sigma_ky^{k+1}}{s_N}$$

Hence,

$$\beta_{N+1}A_{N+1}+s_N\mathcal{G}(X^N,Y^N) \leq \beta_1A_1+\theta_1\sigma_1P(x^0).$$

From this we deduce that the sequence $(\|y^k-\hat{y}\|)$ is bounded and

$$\mathcal{G}(X^N,Y^N) \leq \frac{1}{s_N}(\beta_1A_1+\theta_1\sigma_1P(x^0))$$
$$\|x^{N+1}-\hat{x}\|^2 \leq \frac{1}{\beta_{N+1}}(\beta_1A_1+\theta_1\sigma_1P(x^0)).$$

Our next goal is to derive asymptotics for β_N and s_N. Obviously, (22) holds for any τ_k, β_k such that $\tau_k \leq \frac{1}{\sqrt{\beta_kL}}$. Since in each iteration of the linesearch we decrease τ_k by factor of μ, τ_k can not be less than $\frac{\mu}{\sqrt{\beta_kL}}$. Hence, we have

$$\beta_{k+1} = \beta_k(1+\gamma\tau_k) \geq \beta_k(1+\gamma\frac{\mu}{L\sqrt{\beta_k}}) = \beta_k+\frac{\gamma\mu}{L}\sqrt{\beta_k}. \tag{25}$$

By induction, one can show that there exists $C>0$ such that $\beta_k \geq Ck^2$ for all $k>0$. Then for some constant $C_1>0$ we have

$$\|x^{N+1}-\hat{x}\|^2 \leq \frac{C_1}{(N+1)^2}.$$

From (25) it follows $\beta_{k+1}-\beta_k \geq \frac{\gamma\mu}{L}\sqrt{C}k$. As $\sigma_k = \frac{\beta_{k+1}-\beta_k}{\gamma}$, we obtain $\sigma_k \geq \frac{\mu}{L}\sqrt{C}k$ and thus $s_N = \sum_{k=1}^{N}\sigma_k = O(N^2)$. This means that for some constant $C_2>0$

$$\mathcal{G}(X^N,Y^N) \leq \frac{C_2}{N^2}.$$

We have shown the following result:

Theorem 4. *Let (x^k,y^k) be a sequence generated by Algorithm 2. Then $\|x^N-\hat{x}\| = O(1/N)$ and $\mathcal{G}(X^N,Y^N) = O(1/N^2)$.*

Algorithm 3 *Accelerated primal-dual algorithm with linesearch: f^* is strongly convex*

Initialization: Choose $x^0 \in X$, $y^1 \in Y$, $\mu \in (0,1)$, $\tau_0 > 0$, $\beta_0 > 0$. Set $\theta_0 = 1$.

Main iteration:

1. Compute

$$x^k = \operatorname{prox}_{\tau_{k-1} g}(x^{k-1} - \tau_{k-1} K^* y^k)$$
$$\beta_k = \frac{\beta_{k-1}}{1 + \gamma \beta_{k-1} \tau_{k-1}}$$

2. Choose any $\tau_k \in [\tau_{k-1}, \tau_{k-1}\sqrt{1+\theta_{k-1}}]$ and run

Linesearch:

2.a. Compute

$$\theta_k = \frac{\tau_k}{\tau_{k-1}}, \quad \sigma_k = \beta_k \tau_k$$
$$\bar{x}^k = x^k + \theta_k (x^k - x^{k-1})$$
$$y^{k+1} = \operatorname{prox}_{\sigma_k f^*}(y^k + \sigma_k K \bar{x}^k)$$

2.b. Break linesearch if

$$\sqrt{\beta_k}\tau_k \left\| K^* y^{k+1} - K^* y^k \right\| \le \left\| y^{k+1} - y^k \right\|$$

Otherwise, set $\tau_k := \tau_k \mu$ and go to 2.a.

End of linesearch

4.2 f^* is strongly convex

The case when f^* is γ–strongly convex can be treated in a similar way.

Note that again we set $\delta = 1$.

Dividing (15) over τ_k and taking into account strong convexity of f^*, that has to be used in (8), we deduce

$$\frac{1}{2\tau_k}(\left\|x^k - \hat{x}\right\|^2 - \left\|x^{k+1} - \hat{x}\right\|^2) + \frac{1}{2\sigma_k}(\left\|y^k - \hat{y}\right\|^2 - \left\|y^{k+1} - \hat{y}\right\|^2) - \frac{1}{2\tau_k}\left\|\bar{x}^k - x^k\right\|^2 \ge \frac{\varepsilon_k}{\tau_k} + \frac{\gamma}{2}\left\|y^{k+1} - \hat{y}\right\|^2,$$

which can be rewritten as

$$\frac{1}{2\tau_k}\left\|x^k - \hat{x}\right\|^2 + \frac{1}{2\sigma_k}\left\|y^k - \hat{y}\right\|^2 - \frac{1}{2\tau_k}\left\|\bar{x}^k - x^k\right\|^2 \ge \frac{\varepsilon_k}{\tau_k} + \frac{\tau_{k+1}}{\tau_k}\frac{1}{2\tau_{k+1}}\left\|x^{k+1} - \hat{x}\right\|^2 + \frac{\sigma_{k+1}}{\sigma_k}(1+\gamma\sigma_k)\frac{1}{2\sigma_{k+1}}\left\|y^{k+1} - \hat{y}\right\|^2, \quad (26)$$

Note that by construction of (β_k) in Algorithm 3, we have

$$\frac{\tau_{k+1}}{\tau_k} = \frac{\sigma_{k+1}}{\sigma_k}(1 + \gamma\sigma_k).$$

Let $A_k = \frac{1}{2\tau_k}\left\|x^k - \hat{x}\right\|^2 + \frac{1}{2\sigma_k}\left\|y^k - \hat{y}\right\|^2$. Then (26) is equivalent to

$$\frac{\tau_{k+1}}{\tau_k} A_{k+1} + \frac{\varepsilon_k}{\tau_k} \le A_k - \frac{1}{2\tau_k}\left\|\bar{x}^k - x^k\right\|^2$$

or

$$\tau_{k+1}A_{k+1} + \varepsilon_k \leq \tau_k A_k - \frac{1}{2}\left\|\bar{x}^k - x^k\right\|^2.$$

Finally, summing the above from $k = 1$ to N, we get

$$\tau_{N+1}A_{N+1} + \sum_{k=1}^{N}\varepsilon_k \leq \tau_1 A_1 - \frac{1}{2}\sum_{k=1}^{N}\left\|\bar{x}^k - x^k\right\|^2$$

From this we conclude that the sequence (x^k) is bounded, $\lim_{k\to\infty}\left\|\bar{x}^k - x^k\right\| = 0$, and

$$\begin{aligned}
\mathcal{G}(X^N, Y^N) &\leq \frac{1}{s_N}(\tau_1 A_1 + \theta_1\tau_1 P(x^0)),\\
\left\|y^{N+1} - \hat{y}\right\|^2 &\leq \frac{\sigma_{N+1}}{\tau_{N+1}}(\tau_1 A_1 + \theta_1\tau_1 P(x^0)) = \beta_{N+1}(\tau_1 A_1 + \theta_1\tau_1 P(x^0)),
\end{aligned}$$

where X^N, Y^N, s_N are the same as in Theorem 2.

Let us turn to the derivation of the asymptotics of (τ_N) and (s_N). Analogously, we have that $\tau_k \geq \frac{\mu}{\sqrt{\beta_k}L}$ and thus

$$\beta_{k+1} = \frac{\beta_k}{1 + \gamma\beta_k\tau_k} \leq \frac{\beta_k}{1 + \gamma\frac{\mu}{L}\sqrt{\beta_k}}.$$

Again it is not difficult to show by induction that $\beta_k \leq \frac{C}{k^2}$ for some constant $C > 0$. In fact, as $\phi(\beta) = \frac{\beta}{1+\gamma\frac{\mu}{L}\sqrt{\beta}}$ is increasing, it is sufficient to show that

$$\beta_{k+1} \leq \frac{\beta_k}{1 + \gamma\frac{\mu}{L}\sqrt{\beta_k}} \leq \frac{\frac{C}{k^2}}{1 + \gamma\frac{\mu}{L}\sqrt{\frac{C}{k^2}}} \leq \frac{C}{(k+1)^2}.$$

The latter inequality is equivalent to $\sqrt{C} \geq (2 + \frac{1}{k})\frac{L}{\gamma\mu}$, which obviously holds for C large enough (of course C must also satisfy the induction basis).

The obtained asymptotics for (β_k) yields

$$\tau_k \geq \frac{\mu}{\sqrt{\beta_k}L} \geq \frac{\mu k}{\sqrt{C}L},$$

from which we deduce $\sigma_N = \sum_{k=1}^{N}\tau_k \geq \frac{\mu}{\sqrt{C}L}\sum_{k=1}^{N}k$. Finally, we obtain the following result:

Theorem 5. *Let (x^k, y^k) be a sequence generated by Algorithm 3. Then $\left\|y^N - \hat{y}\right\| = O(1/N)$ and $\mathcal{G}(X^N, Y^N) = O(1/N^2)$.*

Remark 3. In case both g and f^* are strongly convex, one can derive a new algorithm, combining ideas of Algorithm 2 and Algorithm 3 (see [4] for more details).

5 A more general problem

In this section we show how to apply the linesearch procedure for the more general problem

$$\min_{x\in X}\max_{y\in Y} \langle Kx, y\rangle + g(x) - f^*(y) - h(y), \tag{27}$$

where in addition to the previous assumptions, we suppose that $h\colon Y \to \mathbb{R}$ is a smooth convex function with L_h–continuous gradient ∇h. Using the idea of [11], Condat in [7] proposed an extension of primal-dual method to solve (27):

$$\begin{aligned} y^{k+1} &= \operatorname{prox}_{\sigma f^*}(y^k + \sigma(K\bar{x}^k - \nabla h(y^k))) \\ x^{k+1} &= \operatorname{prox}_{\tau g}(x^k - \tau K^* y^{k+1}) \\ \bar{x}^{k+1} &= 2x^{k+1} - x^k. \end{aligned}$$

This scheme was proved to converge under the condition $\tau\sigma \|K\|^2 \le 1-\sigma L_h$. Originally the smooth function was added to the primal part and not to the dual as in (27). For us it is more convenient to consider precisely that form due to the nonsymmetry of the proposed linesearch procedure. However, simply exchanging max and min in (27), we recover the form which was considered in [7].

In addition to the issues related to the operator norm of K which motivated us to derive the PDAL, here we also have to know the Lipschitz constant L_h of ∇h. This has several drawbacks. First, its computation might be expensive. Second, our estimation of L_h might be very conservative and will result in smaller steps. Third, using local information about h instead of global L_h often allows to use larger steps. Therefore, the introduction of a linesearch to the algorithm above is of great practical interest.

The algorithm below exploits the same idea as Algorithm 1 does. However, its stopping criteria is more involved. The interested reader may find out it as a combination of the stopping criteria (4) and a descent lemma for smooth function h.

Note that in case $h \equiv 0$, Algorithm 4 corresponds exactly to Algorithm 1.

We briefly sketch the proof of convergence. Let $(\hat{x}, \hat{y})$ be any saddle point of (1). Similarly to (7), (8), and (11), we get

$$\left\langle x^{k+1} - x^k + \tau_k K^* y^{k+1}, \hat{x} - x^{k+1}\right\rangle \ge \tau_k(g(x^{k+1}) - g(\hat{x})) \tag{29}$$

$$\left\langle \frac{1}{\beta}(y^{k+1} - y^k) - \tau_k K\bar{x}^k + \tau_k \nabla h(y^k), \hat{y} - y^{k+1}\right\rangle \ge \tau_k(f^*(y^{k+1}) - f^*(\hat{y})). \tag{30}$$

$$\left\langle \theta_k(x^k - x^{k-1}) + \tau_k K^* y^k, x^{k+1} - \bar{x}^k\right\rangle \ge \tau_k((1+\theta_k)g(x^k) - g(x^{k+1}) - \theta_k g(x^{k-1})) \tag{31}$$

Summation of (29), (30), (31), and identity (12) yields

$$\begin{aligned} &\left\langle x^{k+1} - x^k, \hat{x} - x^{k+1}\right\rangle + \frac{1}{\beta}\left\langle y^{k+1} - y^k, \hat{y} - y^{k+1}\right\rangle + \theta_k \left\langle x^k - x^{k-1}, x^{k+1} - \bar{x}^k\right\rangle \\ &\quad + \tau_k \left\langle K^* y^{k+1} - K^* y^k, \bar{x}^k - x^{k+1}\right\rangle - \tau_k \left\langle K^* y, \bar{x}^k - \hat{x}\right\rangle + \tau_k \left\langle K\hat{x}, y^{k+1} - y\right\rangle \\ &+ \tau_k \left\langle \nabla h(y^k), \hat{y} - y^{k+1}\right\rangle \ge \tau_k\big(f^*(y^{k+1}) - f^*(\hat{y}) + (1+\theta_k)g(x^k) - \theta_k g(x^{k-1}) - g(\hat{x})\big). \end{aligned} \tag{32}$$

Algorithm 4 *General primal-dual algorithm with a linesearch*

Initialization: Choose $x^0 \in X$, $y^1 \in Y$, $\tau_0 > 0$, $\mu \in (0,1), \delta \in (0,1)$ and $\beta > 0$. Set $\theta_0 = 1$.

Main iteration:

1. Compute
$$x^k = \operatorname{prox}_{\tau_{k-1} g}(x^{k-1} - \tau_{k-1} K^* y^k).$$

2. Choose any $\tau_k \in [\tau_{k-1}, \tau_{k-1}\sqrt{1+\theta_{k-1}}]$ and run

Linesearch:

2.a. Compute
$$\begin{aligned} \theta_k &= \frac{\tau_k}{\tau_{k-1}}, \quad \sigma_k = \beta\tau_k \\ \bar{x}^k &= x^k + \theta_k(x^k - x^{k-1}) \\ y^{k+1} &= \operatorname{prox}_{\sigma_k f^*}(y^k + \sigma_k(K\bar{x}^k - \nabla h(y^k))) \end{aligned}$$

2.b. Break linesearch if
$$\tau_k\sigma_k \left\|K^* y^{k+1} - K^* y^k\right\|^2 + 2\sigma_k[h(y^{k+1}) - h(y^k) - \left\langle \nabla h(y^k), y^{k+1} - y^k \right\rangle] \le \delta \left\|y^{k+1} - y^k\right\|^2 \tag{28}$$

Otherwise, set $\tau_k := \tau_k\mu$ and go to 2.a.

End of linesearch

By convexity of h, we have $\tau_k(h(\hat{y}) - h(y^k) - \left\langle \nabla h(y^k), \hat{y} - y^k \right\rangle) \ge 0$. Combining it with inequality (28), divided over 2β, we get

$$\frac{\delta}{2\beta}\left\|y^{k+1} - y^k\right\|^2 - \frac{\tau_k^2}{2}\left\|K^* y^{k+1} - K^* y^k\right\|^2 \ge \tau_k[h(y^{k+1}) - h(\hat{y}) - \left\langle \nabla h(y^k), y^{k+1} - \hat{y} \right\rangle].$$

Adding the above inequality to (32) gives us

$$\begin{aligned} &\left\langle x^{k+1} - x^k, \hat{x} - x^{k+1} \right\rangle + \frac{1}{\beta}\left\langle y^{k+1} - y^k, \hat{y} - y^{k+1} \right\rangle + \theta_k \left\langle x^k - x^{k-1}, x^{k+1} - \bar{x}^k \right\rangle \\ &\quad + \tau_k \left\langle K^* y^{k+1} - K^* y^k, \bar{x}^k - x^{k+1} \right\rangle - \tau_k \left\langle K^* y, \bar{x}^k - \hat{x} \right\rangle + \tau_k \left\langle K\hat{x}, y^{k+1} - y \right\rangle \\ &\qquad + \frac{\delta}{2\beta}\left\|y^{k+1} - y^k\right\|^2 - \frac{\tau_k^2}{2}\left\|K^* y^{k+1} - K^* y^k\right\|^2 \\ &\quad \ge \tau_k\big((f^* + h)(y^{k+1}) - (f^* + h)(\hat{y}) + (1+\theta_k)g(x^k) - \theta_k g(x^{k-1}) - g(\hat{x})\big). \end{aligned} \tag{33}$$

Note that for problem (27) instead of (3) we have to use

$$D_{\hat{x},\hat{y}}(y) := f^*(y) + h(y) - f^*(\hat{y}) - h(\hat{y}) - \langle K\hat{x}, y - \hat{y} \rangle \ge 0 \quad \forall y \in Y,$$

which is true by definition of $(\hat{x}, \hat{y})$. Now we can rewrite (33) as

$$\begin{aligned} &\left\langle x^{k+1} - x^k, \hat{x} - x^{k+1} \right\rangle + \frac{1}{\beta}\left\langle y^{k+1} - y^k, \hat{y} - y^{k+1} \right\rangle + \theta_k \left\langle x^k - x^{k-1}, x^{k+1} - \bar{x}^k \right\rangle \\ &\quad + \tau_k \left\langle K^* y^{k+1} - K^* y^k, \bar{x}^k - x^{k+1} \right\rangle + \frac{\delta}{2\beta}\left\|y^{k+1} - y^k\right\|^2 - \frac{\tau_k^2}{2}\left\|K^* y^{k+1} - K^* y^k\right\|^2 \\ &\qquad \ge \tau_k\big((1+\theta_k)P(x^k) - \theta_k P(x^{k-1}) + D(y^{k+1})\big). \end{aligned} \tag{34}$$

Applying cosine rules for all inner products in the first line in (34) and using that $\theta_k(x^k - x^{k-1}) = \bar{x}^k - x^k$, we obtain

$$\begin{aligned}\frac{1}{2}(\|x^k - \hat{x}\|^2 - \|x^{k+1} - \hat{x}\|^2 - \|x^{k+1} - x^k\|^2) + \frac{1}{2\beta}(\|y^k - \hat{y}\|^2 - \|y^{k+1} - \hat{y}\|^2 - \|y^{k+1} - y^k\|^2) \\ + \frac{1}{2}(\|x^{k+1} - x^k\|^2 - \|\bar{x}^k - x^k\|^2 - \|x^{k+1} - \bar{x}^k\|^2) \\ + \tau_k \|K^* y^{k+1} - K^* y^k\| \|x^{k+1} - \bar{x}^k\| + \frac{\delta}{2\beta}\|y^{k+1} - y^k\|^2 - \frac{\tau_k^2}{2}\|K^* y^{k+1} - K^* y^k\|^2 \\ \geq \tau_k\big((1+\theta_k)P(x^k) - \theta_k P(x^{k-1}) + D(y^{k+1})\big). \quad (35)\end{aligned}$$

Finally, applying Cauchy's inequality in (35) and using that $\tau_k\theta_k \leq \tau_{k-1}(1+\theta_{k-1})$, we get

$$\begin{aligned}\frac{1}{2}\|x^{k+1} - \hat{x}\|^2 + \frac{1}{2\beta}\|y^{k+1} - \hat{y}\|^2 + \tau_k(1+\theta_k)P(x^k) \leq \\ \frac{1}{2}\|x^k - \hat{x}\|^2 + \frac{1}{2\beta}\|y^k - \hat{y}\|^2 + \tau_{k-1}(1+\theta_{k-1})P(x^{k-1}) \\ - \frac{1}{2}\|\bar{x}^k - x^k\|^2 - \frac{1-\delta}{2\beta}\|y^{k+1} - y^k\|^2,\end{aligned}$$

from which the convergence of (x^k) and (y^k) to a saddle point of (27) can be derived in a similar way as in Theorem 1.

6 Numerical Experiments

This section collects several numerical tests which will illustrate the performance of the proposed methods. Computations[1] were performed using Python 2.7 on an Intel Core i5-5200U CPU 2.20GHz running 64-bit Linux Mint 17.3.

For PDAL we initialize the input data as $\mu = 0.7$, $\delta = 0.99$, $\tau_0 = \frac{\sqrt{\min\{m,n\}}}{\|A\|_F}$. The latter is easy to compute and it is an upper bound of $\frac{1}{\|A\|}$. The parameter β for PDAL is always taken as $\frac{\sigma}{\tau}$ in PDA with fixed steps σ and τ. A trial step τ_k in Step 2 is always chosen as $\tau_k = \tau_{k-1}\sqrt{1+\theta_{k-1}}$.

6.1 Matrix game

We are interested in the following min-max matrix game

$$\min_{x\in\Delta_n}\max_{y\in\Delta_m}\langle Ax, y\rangle, \quad (36)$$

where $x \in \mathbb{R}^n$, $y \in \mathbb{R}^m$, $A \in \mathbb{R}^{m\times n}$, and Δ_m, Δ_n denote the standard unit simplices in $\mathbb{R}^m$ and $\mathbb{R}^n$ respectively.

[1]Codes can be found on `https://gitlab.icg.tugraz.at/malitsky/primal_dual_linesearch`

For this problem we study the performance of PDA, PDAL (Alg. 1), Tseng's FBF method [16], and PEGM [14]. For a comparison we use a primal-dual gap $\mathcal{G}(x, y)$ which can be easily computed for a feasible pair (x, y)

$$\mathcal{G}(x, y) = \max_i (Ax)_i - \min_j (A^* y)_j.$$

Since iterates obtained by Tseng's method may be infeasible, we used an auxiliary point (see [16]) to compute the primal-dual gap.

The initial point in all cases was chosen as $x^0 = \frac{1}{n}(1, \ldots, 1)$ and $y^0 = \frac{1}{m}(1, \ldots, 1)$. In order to compute projection onto the unit simplex we used the algorithm from [8]. For PDA we use $\tau = \sigma = 1/\|A\| = 1/\sqrt{\lambda_{\max}(A^* A)}$, which we compute in advance. The input data for FBF and PEGM are taken the same as in [14]. Note that these methods also use a linesearch.

We consider 4 differently generated samples of the matrix $A \in \mathbb{R}^{m \times n}$:

1. $m = n = 100$. All entries of A are generated independently from the uniform distribution in $[-1, 1]$.

2. $m = n = 100$. All entries of A are generated independently from the the normal distribution $\mathcal{N}(0, 1)$.

3. $m = 500$, $n = 100$. All entries of A are generated independently from the normal distribution $\mathcal{N}(0, 10)$.

4. $m = 100$, $n = 200$. All entries of A are generated independently from the uniform distribution in $[0, 1]$.

For every case we report the primal-dual gap vs the number of iterations. The results are presented on Figure 1. Each method was run for 50000 iterations. The execution time of all iterations for PDA and PEGM is almost the same, FBF is more than in 2 times expensive, and PDAL is about 1.5 times more expensive than PDA.

6.2 l_1–regularized least squares

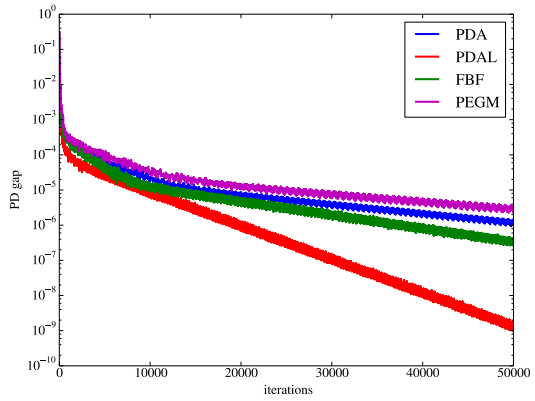
We study the following l_1–regularized problem

$$\min_x \phi(x) := \frac{1}{2} \|Ax - b\|^2 + \lambda \|x\|_1, \tag{37}$$

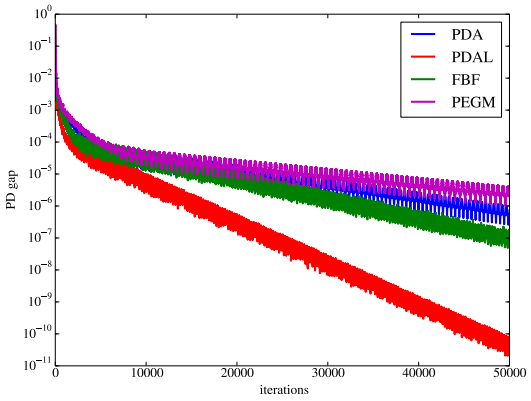
where $A \in \mathbb{R}^{m \times n}$, $x \in \mathbb{R}^n$, $b \in \mathbb{R}^m$. Let $g(x) = \lambda \|x\|_1$, $f(p) = \frac{1}{2} \|p - b\|^2$. Analogously to (5) and (6), we can rewrite (37) as

$$\min_x \max_y g(x) + \langle Ax, y \rangle - f^*(y),$$

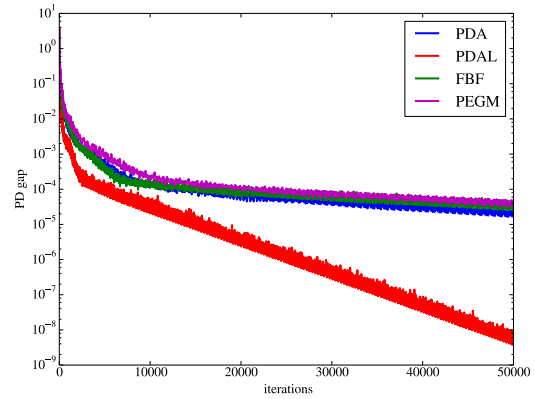
where $f^*(y) = \frac{1}{2} \|y\|^2 + (b, y) = \frac{1}{2} \|y + b\|^2 - \frac{1}{2} \|b\|^2$. Clearly, the last term does not have any impact on the prox-term and we can conclude that $\operatorname{prox}_{\lambda f^*}(y) = \frac{y - \lambda b}{1 + \lambda}$. This means that the linesearch in Alg. 1 does not require any additional matrix-vector multiplication (see Remark 2).



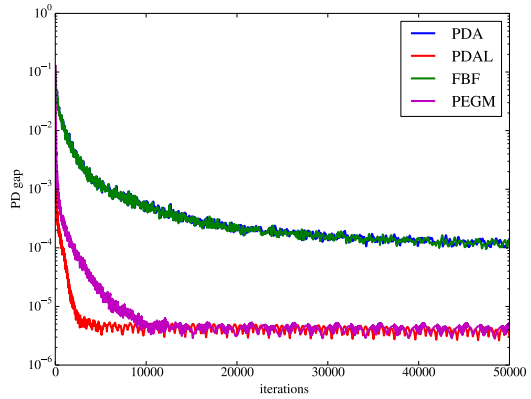
(a) Example 1



(b) Example 2



(c) Example 3



(d) Example 4

Figure 1: Convergence plots for problem (36)

We generate three instances of problem (37), for which we compare the performance of PDA, PDAL, PGM (proximal gradient method), and FISTA [3]. All methods except PDAL require predefined steps. For this we compute in advance $\|A\| = \sqrt{\lambda_{\max}(A^*A)}$. For PGM and FISTA we use a fixed step size $\alpha = \frac{1}{\|A\|^2}$, for PDA we use $\sigma = \frac{1}{20\|A\|}, \tau = \frac{20}{\|A\|}$. For PDAL we set $\beta = 400$.

For all cases below we generate some random $w \in \mathbb{R}^n$ in which s random coordinates are drawn from $\mathcal{N}(0,1)$ and the rest are zeros. Then we generate $\nu \in \mathbb{R}^m$ with entries drawn from $\mathcal{N}(0, 0.1)$ and set $b = Aw + \nu$. The initial points for all methods were $x^0 = (0, \ldots, 0)$, $y^0 = Ax^0 - b$.

The matrix $A \in \mathbb{R}^{m\times n}$ is constructed in one of the following ways:

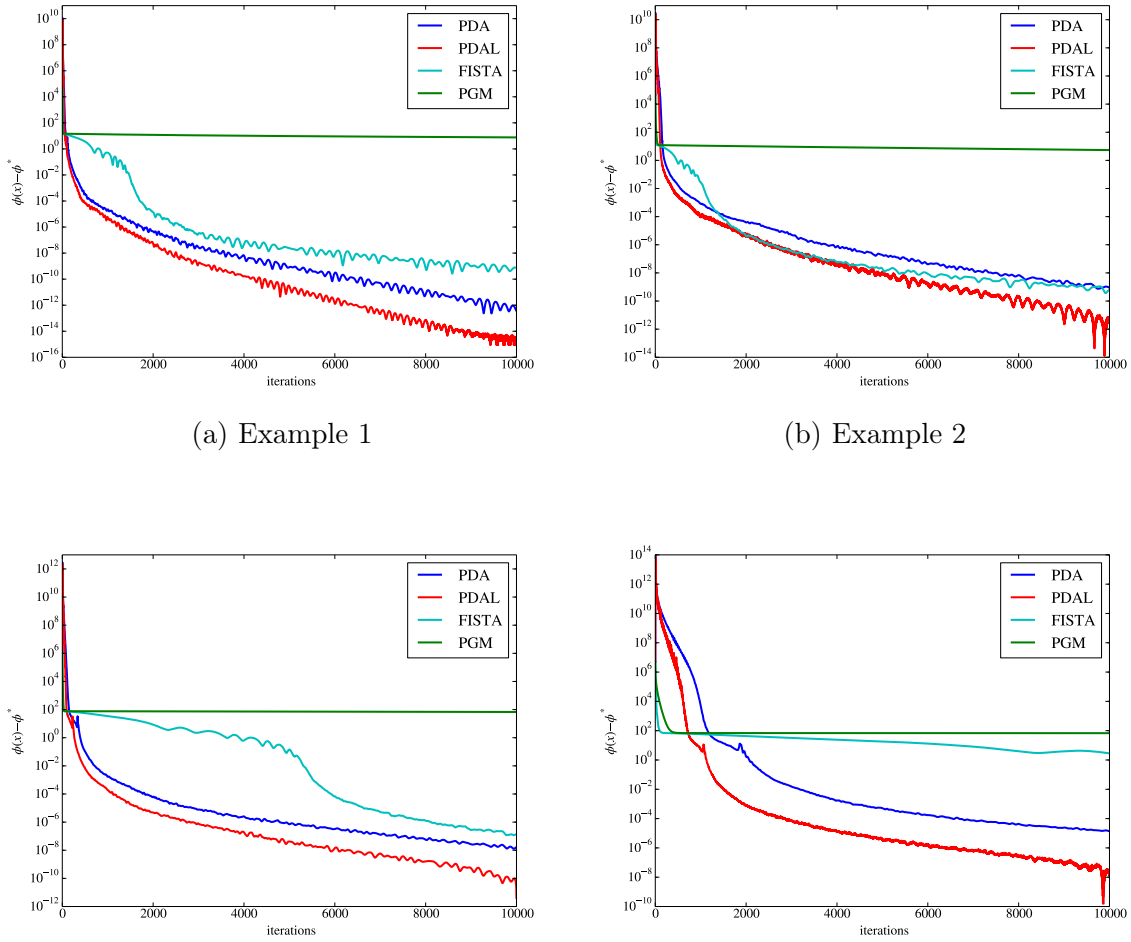
1. $n = 1000$, $m = 200$, $s = 10$, $\lambda = 0.1$. All entries of A are generated independently from $\mathcal{N}(0,1)$. The s entries of w are drawn from the uniform distribution in $[-10, 10]$.

2. $n = 10000$, $m = 1000$, $s = 100$, $\lambda = 0.1$. All entries of A are generated independently from $\mathcal{N}(0,1)$. The s entries of w are drawn from $\mathcal{N}(0,1)$.

3. $n = 5000$, $m = 1000$, $s = 50$, $\lambda = 0.1$. First, we generate the matrix B with entries from $\mathcal{N}(0, 1)$. Then for any $p \in (0, 1)$ we construct the matrix A by columns A_j, $j = 1, \ldots, n$ as follows: $A_1 = \frac{B_1}{\sqrt{1-p^2}}$, $A_j = p * A_{j-1} + B_j$. As p increases, A becomes more ill-conditioned (see [1] where this example was considered). In this experiment we take $p = 0.5$. Entries of w are chosen the same as in the first example.

4. The same as previous example, but with $p = 0.9$.

Figure 2 collects the convergence plots. Note that all methods require roughly the same amount of computation per iteration. Although, for the considered instances of (37) primal-dual methods show better performance, they require a tuning the parameter β.



(a) Example 1

(b) Example 2

(c) Example 3

(d) Example 4

Figure 2: Convergence plots for problem (37)

6.3 Nonnegative least squares

Consider another regularized least squares problem

$$\min_{x \geq 0} \phi(x) := \frac{1}{2} \|Ax - b\|^2, \tag{38}$$

where $A \in \mathbb{R}^{m \times n}$, $x \in \mathbb{R}^n$, $b \in \mathbb{R}^m$. Similarly as before, we can express it as

$$\min_x \max_y g(x) + \langle Ax, y \rangle - f^* y),$$

where $g(x) = \delta_{\mathbb{R}^n_+}(x)$, $f^*(y) = \frac{1}{2} \|y + b\|^2$. For this problem we consider 3 real data examples from the Matrix Market library[2]:

1. *WELL1033*: sparse matrix with $m = 1033$, $n = 320$;
2. *WELL1850*: sparse matrix with $m = 1850$, $n = 712$;
3. *ILLC1033*: sparse matrix with $m = 1033$, $n = 320$;
4. *ILLC1850*: sparse matrix with $m = 1850$, $n = 712$.

For all cases the entries of vector $b \in \mathbb{R}^m$ are generated independently from $\mathcal{N}(0, 1)$. The initial points are $x^0 = (0, \ldots, 0)$ and $y^0 = Ax^0 - b = -b$.

Since f^* is strongly convex, in addition to PDA, PDAL, and FISTA, we include in our comparison APDA and APDAL (Alg. 3). We take parameter of strong convexity as $\frac{1}{2}$. For PDA, APDA, and FISTA we compute $\|A\|$ and set $\tau = \sigma = \frac{1}{\|A\|}$, $\alpha = \frac{1}{\|A\|^2}$. For PDAL we set $\beta = 1$ (the same as σ/τ in PDA). As (38) is again just a regularized least squares problem, the linesearch does not require any additional matrix-vector multiplications. The results are presented on Figure 3. Here we can see how the usage of the linesearch improves the rate of convergence.

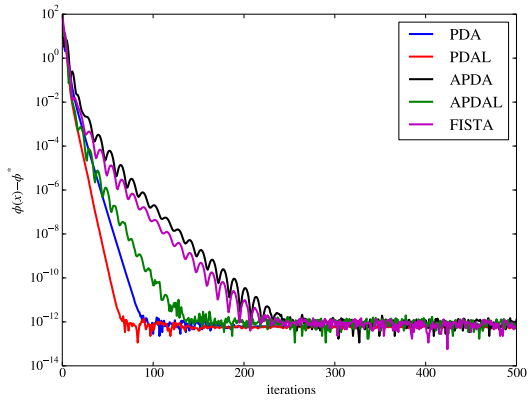
It is interesting to highlight that sometimes non-accelerated methods can be better than their accelerated variants. Also notice how similar the performance of APDA and FISTA for all examples.

7 Conclusion

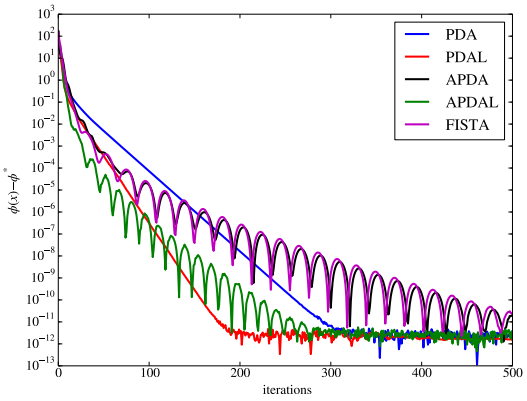
In this work, we have presented several primal-dual algorithms with linesearch. On the one hand, this allows us to avoid the evaluation of the operator norm, and on the other hand, it allows to make larger steps. The proposed linesearch is very simple and in many important cases does not require additional expensive operations (as matrix-vector multiplications or prox-operators). For all methods we have proved convergence. Our numerical experiments approve the efficiency of the proposed methods.

Acknowledgements: The work is supported by Austrian science fund (FWF) under the project "Efficient Algorithms for Nonsmooth Optimization in Imaging" (EANOI) No. I1148.

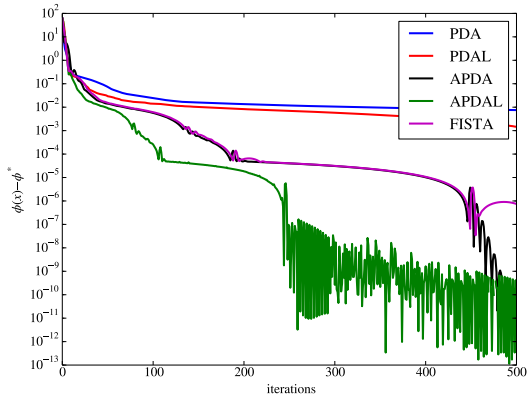
[2] `http://math.nist.gov/MatrixMarket/`



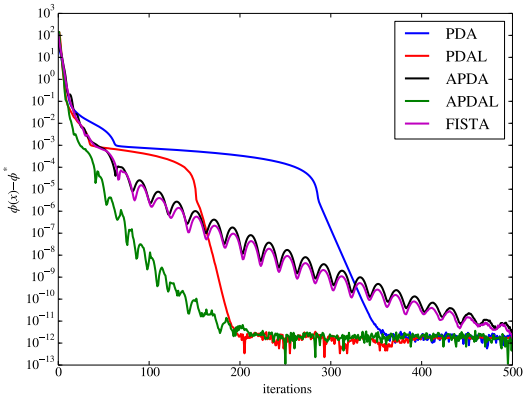
(a) Example 1



(b) Example 2



(c) Example 3



(d) Example 4

Figure 3: Convergence plots for problem (38)

References

[1] Agarwal, A., Negahban, S., and Wainwright, M. J. Fast global convergence rates of gradient methods for high-dimensional statistical recovery. In *Advances in Neural Information Processing Systems* (2010), pp. 37–45.

[2] Bauschke, H. H., and Combettes, P. L. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, New York, 2011.

[3] Beck, A., and Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problem. *SIAM Journal on Imaging Sciences 2*, 1 (2009), 183–202.

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

[5] Chambolle, A., and Pock, T. On the ergodic convergence rates of a first-order primal–dual algorithm. *Mathematical Programming* (2015), 1–35.

[6] Chambolle, A., and Pock, T. An introduction to continuous optimization for imaging. *Acta Numerica 25* (2016), 161–319.

[7] Condat, L. A primal–dual splitting method for convex optimization involving lipschitzian, proximable and linear composite terms. *Journal of Optimization Theory and Applications 158*, 2 (2013), 460–479.

[8] Duchi, J., Shalev-Shwartz, S., Singer, Y., and Chandra, T. Efficient projections onto the l 1-ball for learning in high dimensions. In *Proceedings of the 25th international conference on Machine learning* (2008), pp. 272–279.

[9] Goldstein, T., Li, M., Yuan, X., Esser, E., and Baraniuk, R. Adaptive primal-dual hybrid gradient methods for saddle-point problems. *arXiv preprint arXiv:1305.0546* (2013).

[10] He, B., Yang, H., and Wang, S. Alternating direction method with self-adaptive penalty parameters for monotone variational inequalities. *Journal of Optimization Theory and applications 106*, 2 (2000), 337–356.

[11] He, B., and Yuan, X. Convergence analysis of primal-dual algorithms for a saddle-point problem: From contraction perspective. *SIAM Journal on Imaging Sciences 5*, 1 (2012), 119–149.

[12] Komodakis, N., and Pesquet, J. C. Playing with duality: An overview of recent primal-dual approaches for solving large-scale optimization problems. *IEEE Signal Processing Magazine 32*, 6 (2015), 31–54.

[13] Malitsky, Y. Reflected projected gradient method for solving monotone variational inequalities. *SIAM Journal on Optimization 25*, 1 (2015), 502–520.

[14] Malitsky, Y. Proximal extrapolated gradient methods for variational inequalities. *arXiv preprint arXiv:1601.04001* (2016).

[15] Pock, T., and Chambolle, A. Diagonal preconditioning for first order primal-dual algorithms in convex optimization. In *Computer Vision (ICCV), 2011 IEEE International Conference on* (2011), IEEE, pp. 1762–1769.

[16] Tseng, P. A modified forward-backward splitting method for maximal monotone mappings. *SIAM Journal on Control and Optimization 38* (2000), 431–446.