The primal-dual hybrid gradient method reduces to a primal method for linearly constrained optimization problems

Yura Malitsky*

Abstract

In this work, we show that for linearly constrained optimization problems the primal-dual hybrid gradient algorithm, analyzed by Chambolle and Pock [3], can be written as an entirely primal algorithm. This allows us to prove convergence of the iterates even in the degenerate cases when the linear system is inconsistent or when the strong duality does not hold. We also obtain new convergence rates which seem to improve existing ones in the literature. For a decentralized distributed optimization we show that the new scheme is much more efficient than the original one.

2010 Mathematics Subject Classification: 49M29 65K10 65Y20 90C25
Keywords: First-order algorithms, primal-dual algorithms, convergence rates, linearly constrained optimization problem, penalty methods, distributed optimization

1 Introduction

In this paper, we study nonsmooth optimization problems with linear constraints of the form

$$\min_{x\in\mathbb{R}^n} \quad g(x) \quad \text{s.t.} \quad Ax = b, \tag{1}$$

where $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, and $g\colon \mathbb{R}^n \to (-\infty, +\infty]$ is a proper lower semicontinuous convex function. Problem (1) is one of the most importance in convex optimization and, in particular, includes conic optimization, which in turn includes linear and semidefinite optimization. (1) often arises in machine learning, inverse problems, and distributed optimization. Notice that every composite optimization problem $\min_u g_1(u)+g_2(Ku)$ can be also written in the form (1) by setting $x = (u, v)^T$, $g(x) = g_1(u) + g_2(v)$, $A = [\ K^T \mid -I\ ]^T$, and $b = 0$. In the sequel, we restrict our attention to large-scale problems where computing the projection onto the subspace $\{x\colon Ax = b\}$ is expensive or even practically impossible.

In this work, we focus on the primal-dual hybrid algorithm (PDHG), analyzed by Chambolle and Pock in [3]. It is a popular method for solving convex-concave saddle problems with a bilinear term, owing to the fact that its iteration requires computing only two proximal operators and two matrix-vector multiplications. This is different from the alternating direction

*Institute for Numerical and Applied Mathematics, University of Göttingen, 37083 Göttingen, Germany, e-mail: y.malitskyi@math.uni-goettingen.de. The research was supported by the German Research Foundation grant SFB755-A4

method of multipliers (ADMM)—another popular approach to solve (1), where each iteration requires solving a nontrivial problem. However, on the other hand, the PDHG algorithm is a particular case of a more general proximal ADMM [1, 21]. For possible extensions and applications of the PDHG, we refer the reader to [4, 5, 10, 14, 26].

By introducing the Lagrange multiplier y, one can rewrite (1) as a saddle point problem

$$\min_{x\in\mathbb{R}^n}\max_{y\in\mathbb{R}^m} \quad g(x)+\langle Ax,y\rangle-\langle b,y\rangle. \tag{2}$$

Then the PDHG applied to (2) generates sequences (x^k), (y^k) according to

$$\begin{aligned} y^{k+1} &= y^k+\sigma(A\bar{x}^k-b)\\ x^{k+1} &= \operatorname{prox}_{\tau g}(x^k-\tau A^T y^{k+1}), \end{aligned} \tag{3}$$

where $x^0\in\mathbb{R}^n, y^0\in\mathbb{R}^m$ are arbitrary and $\bar{x}^k=2x^k-x^{k-1}$ for all $k\in\mathbb{N}$. The convergence result stated in [3] says that if there exists a saddle point for problem (2) and $\tau\sigma\|A\|^2<1$, then (x^k,y^k) converges to a saddle point of (2). Moreover, in this case, we have $O(1/k)$ ergodic rate for the primal-dual gap.

In this work, we are interested in cases which are not covered by the above statement. More precisely, we provide answers to the following two questions.

- *What will happen to the iterates of* (3) *if problem* (1) *is infeasible?*
- *What will happen to the iterates of* (3) *if there is no saddle point in* (2)*?*

Note that the standard analysis of the PDHG in [3], or alternative ones in [10,14], cannot resolve aforementioned issues. To answer these questions, we show that the primal-dual algorithm (3) can be reformulated as an entirely primal algorithm, *i.e.,* without resorting to dual variables. This is in fact our main result, as it views the primal-dual algorithm (3) from a new perspective. We reveal a connection of (3) to the accelerated proximal gradient method [24] for the composite minimization. A novel analysis yields new convergence rates which are more suitable in practice and seem to be better than existing ones. We show a connection of the PDHG to the diagonal penalty methods and inverse problems. The new scheme is simpler than the original one to implement and has a smaller memory footprint. In fact, at least from an algorithmic point of view, it might be the simplest existing scheme for solving such a generic problem (1). In contrast to the standard PDHG method, it can be applied to a decentralized distributed optimization problem with only one communication per iteration. Moreover, it also achieves a better complexity rate. The new scheme is favorable to new extensions. For example, in the subsequent paper [13], inspired by the coordinate extension of Tseng's method [24] proposed by Fercoq and Richtárik [7], we derive a coordinate extension of the PDHG for (1).

Paper outline. In section 2 we briefly recall the standard notation from convex analysis and establish several preparatory lemmas. Section 3 is dedicated to the new analysis of the PDHG algorithm (3) and its consequences. In section 4 we consider several generalizations of the PDHG method: when g is strongly convex and when the objective in (1) has an additional smooth term.

2 Preliminaries

Throughout the paper we will work in a finite-dimensional vector space $\mathbb{R}^n$ equipped with an inner product $\langle\cdot,\cdot\rangle$ and a norm $\|\cdot\|=\sqrt{\langle\cdot,\cdot\rangle}$. A function g is called γ–strongly convex function, if $g - \frac{\gamma}{2}\|\cdot\|^2$ is convex. For a convex lower semi-continuous (lsc) function $g\colon \mathbb{R}^n \to (-\infty, +\infty]$ we denote by $\operatorname{dom} g$ the domain of g, *i.e.*, the set $\{x\colon g(x) < +\infty\}$, and by prox_g the proximal operator of g that is $\operatorname{prox}_g(z) = \operatorname{argmin}_x\{g(x) + \frac{1}{2}\|x - z\|^2\}$. The following characteristic property (prox-inequality) will often be used:

$$\bar{x} = \operatorname{prox}_g z \quad \Leftrightarrow \quad \langle \bar{x} - z, x - \bar{x}\rangle \geq g(\bar{x}) - g(x) \quad \forall x \in \mathbb{R}^n. \tag{4}$$

When g is γ–strongly convex, the above inequality can be strengthened:

$$\bar{x} = \operatorname{prox}_g z \quad \Leftrightarrow \quad \langle \bar{x} - z, x - \bar{x}\rangle \geq g(\bar{x}) - g(x) + \frac{\gamma}{2}\|\bar{x} - x\|^2 \quad \forall x \in \mathbb{R}^n. \tag{5}$$

For a linear operator $A\colon \mathbb{R}^n \to \mathbb{R}^m$ with a slight abuse of notation we denote its operator norm as $\|A\|$. Throughout the paper $f(x) = \frac{1}{2}\|Ax - b\|^2$, where $x \in \mathbb{R}^n, b \in \mathbb{R}^m$, and $f_* := \min_x f(x)$. Since f is a quadratic, we have

$$\alpha f(x) + (1-\alpha) f(y) = f(\alpha x + (1-\alpha)y) + \frac{\alpha(1-\alpha)}{2}\|A(x-y)\|^2 \quad \forall x, y \in \mathbb{R}^n \ \forall \alpha \in \mathbb{R}. \tag{6}$$

$$f(x) - f_* = f(x) - f(\bar{x}) = \frac{1}{2}\|A(x - \bar{x})\|^2 \quad \forall x \in \mathbb{R}^n \ \forall \bar{x} \in \operatorname{argmin} f. \tag{7}$$

Another useful identity (the cosine law) obviously holds:

$$2\langle x - y, z - x\rangle = \|y - z\|^2 - \|x - y\|^2 - \|x - z\|^2 \qquad \forall x, y, z \in \mathbb{R}^n. \tag{8}$$

We conclude our preliminary section by two important lemmas.

Lemma 1. Suppose that sequences $(x^k) \subset \mathbb{R}^n$, $(b_k) \subset \mathbb{R}$ and a set $D \subseteq \mathbb{R}^n$ satisfy:

(i) all cluster points of (x^k) belong to D;

(ii) for all $x \in D$ the sequence $(\|x^k - x\|^2 + b_k)$ is nonincreasing and bounded below.

Then the sequence (x^k) converges to some point in D.

Proof. Suppose, on the contrary, that there exist two different subsequences (x^{k_i}) and (x^{k_j}) such that $x^{k_i} \to \tilde{x}_1$, $x^{k_j} \to \tilde{x}_2$ and $\tilde{x}_1 \neq \tilde{x}_2$. Let $a_k(x) := \|x^k - x\|^2 + b_k$. By (ii), the sequence $(a_k(x))$ is convergent for any $x \in D$. Setting $x := \tilde{x}_1$, we obtain

$$\begin{aligned} \lim_{k\to\infty} a_k(\tilde{x}_1) &= \lim_{i\to\infty} a_{k_i}(\tilde{x}_1) = \lim_{i\to\infty}(\|x^{k_i} - \tilde{x}_1\|^2 + b_{k_i}) = \lim_{i\to\infty} b_{k_i} \\ &= \lim_{j\to\infty} a_{k_j}(\tilde{x}_1) = \lim_{j\to\infty}(\|x^{k_j} - \tilde{x}_1\|^2 + b_{k_j}) = \|\tilde{x}_2 - \tilde{x}_1\|^2 + \lim_{j\to\infty} b_{k_j}, \end{aligned}$$

from which $\lim_{i\to\infty} b_{k_i} = \|\tilde{x}_2 - \tilde{x}_1\|^2 + \lim_{j\to\infty} b_{k_j}$ follows. Setting $x = \tilde{x}_2$, we analogously derive

$$\lim_{j\to\infty} b_{k_j} = \|\tilde{x}_1 - \tilde{x}_2\|^2 + \lim_{i\to\infty} b_{k_i},$$

from which we conclude that $\tilde{x}_1 = \tilde{x}_2$. Therefore, the whole sequence (x^k) converges to some point in D. □

Lemma 2. Let $f(x) = \frac{1}{2}\|Ax - b\|^2$, $f_* = \min_x f(x)$ and $\bar{x} \in \operatorname{argmin} f$. Then for any $u, v \in \mathbb{R}^n$ it holds

$$\langle \nabla f(u), \bar{x} - v\rangle = 2f_* - f(u) - f(v) + \frac{1}{2}\|A(u - v)\|^2.$$

Proof. As $\bar{x} \in \operatorname{argmin} f$, $A^T A\bar{x} = A^T b$. Using this, we have

$$\begin{aligned} \langle \nabla f(u), \bar{x} - v\rangle &= \langle A^T(Au - b), \bar{x} - v\rangle = \langle A^T A(u - \bar{x}), \bar{x} - v\rangle = \langle A(u - \bar{x}), A(\bar{x} - v)\rangle \\ &= \frac{1}{2}\|A(u - v)^2\| - \frac{1}{2}\|A(u - \bar{x})\|^2 - \frac{1}{2}\|A(v - \bar{x})\|^2. \end{aligned}$$

Then the statement follows directly from (7). □

3 Main part

In this section, we describe several cases where known analyses of the PDHG are non applicable.

Infeasible problem. The first issue arises when the constraints in (1) are inconsistent: $\mathcal{P} := \{x\colon Ax = b\} = \varnothing$. Clearly, in this case problem (1), as a minimization problem over an empty set, does not have a lot of sense. If we know in advance that $\mathcal{P} = \varnothing$, probably the most natural thing is to consider the following generalization of (1):

$$\min_{x\in\mathbb{R}^n} \quad g(x) \quad \text{s.t.} \quad x \in \operatorname*{argmin}_{u\in\mathbb{R}^n} f(u), \tag{9}$$

where $f(x) = \frac{1}{2}\|Ax - b\|^2$. Now the constraints are always nonempty. Moreover, one can equivalently recast (9) as

$$\min_{x\in\mathbb{R}^n} \quad g(x) \quad \text{s.t.} \quad A^T Ax = A^T b. \tag{10}$$

Hence, we again have a problem of the same class as (1) but now it is always feasible. Thus, we still can apply the PDHG to (10), which yields the following recursion:

$$\begin{aligned} y^{k+1} &= y^k + \sigma(A^T A\bar{x}^k - A^T b) \\ x^{k+1} &= \operatorname{prox}_{\tau g}(x^k - \tau A^T A y^{k+1}), \end{aligned} \tag{11}$$

where $\tau\sigma\|A\|^4 < 1$ and, as before, $\bar{x}^k = x^k + (x^k - x^{k-1})$. This scheme has several drawbacks, compared to (3). First, it uses four matrix-vector multiplications. Of course, one can precompute $A^T A$, but for large-scaled problems it might be expensive and often not desirable. For instance, if A is sparse, $A^T A$ will probably have a worse sparsity, which will lead to a more expensive iteration. Third, it is known that the conditional number of $A^T A$ is a square of the conditional number of A [20, Chapter 8]. Hence, when A is ill-conditioned, working with $A^T A$ will be much harder than with A. Finally, the stepsizes in (11) have to satisfy a more restrictive inequality $\tau\sigma\|A\|^4 < 1$. Because of that, (11) will have worse estimates for the convergence rate.

Another reason not to apply algorithm (11) to solve (9) is the absence of a priori knowledge that the set $\mathcal{P}$ is empty. Thus, the best situation would be to have something meaningful from the iterates of (3). We show that this is indeed the case: the PDHG algorithm given by (3) solves a general problem (9), so there is no need to apply more expensive scheme (11).

No strong duality. The convergence of almost all widely-used methods (PDHG, ADMM, variational inequality methods) for solving (1) heavily relies on the duality arguments. The assumption that the strong duality holds is usually taken for granted, although sometimes it is not so easy to check whether it is satisfied. The standard condition that ensures that strong duality holds for problem (1) is $b \in \operatorname{ri}(A \operatorname{dom} g)$, where $\operatorname{ri}(C)$ stands for the relative interior of C. Whenever g is not full-domain, *i.e.,* not finite-valued, it becomes non-trivial to verify it.

3.1 Primal form of PDHG

Here we show that the primal-dual method applied to (1) can be seen as a modified Tseng's method [24], whose stepsizes tend to infinity.

Reducing the primal-dual algorithm (3) to the only primal form is in fact trivial. Iterating the first equation in (3), one can derive

$$\begin{aligned} y^{k+1} &= y^k + \sigma(A\bar{x}^k - b) = y^{k-1} + \sigma A(\bar{x}^k + \bar{x}^{k-1}) - 2\sigma b = \dots \\ &= y^0 + \sigma A(\bar{x}^k + \dots + \bar{x}^0) - (k+1)\sigma b. \end{aligned}$$

For simplicity, assume that the PDHG starts from (x^0, y^0) with $\bar{x}^0 = x^0$ and $y^0 = 0$. Then the above equation is equivalent to $y^1 = \sigma(Ax^0 - b)$ if $k = 0$, to $y^2 = 2\sigma(Ax^1 - b)$ if $k = 1$, and to

$$y^{k+1} = \sigma A(\bar{x}^k + \dots + \bar{x}^0) - \sigma(k+1)b = \sigma A(2x^k + x^{k-1} + \dots + x^1) - \sigma(k+1)b,$$

if $k \geq 2$. Define a new sequence (z^k) with $z^0 = x^0$, $z^1 = x^1$ and $z^k = \frac{1}{k+1}(2x^k + x^{k-1} + \dots + x^1)$ for all $k \geq 2$. Then $y^{k+1} = (k+1)\sigma(Az^k - b)$ for all $k \geq 0$ and hence, the primal-dual scheme (3) can be written as

$$\begin{aligned} z^k &= \frac{k}{k+1} z^{k-1} + \frac{1}{k+1}\bar{x}^k \\ x^{k+1} &= \operatorname{prox}_{\tau g}(x^k - (k+1)\tau\sigma A^T(Az^k - b)), \end{aligned} \tag{12}$$

where $k \geq 0$. This is easy to see: the sequence (z^k), defined as above, obviously satisfies the first recurrent equation in (12) and the second equation is a direct consequence of one in (3). We continue transforming the scheme (12) by introducing another sequence (s^k) with $s^0 = x^0$ and $s^k = \frac{x^1 + \dots + x^k}{k}$ for $k \geq 1$. Then (12) can be cast as

$$\begin{aligned} z^k &= \frac{x^k + ks^k}{k+1} \\ x^{k+1} &= \operatorname{prox}_{\tau g}(x^k - (k+1)\lambda\nabla f(z^k)) \\ s^{k+1} &= \frac{x^{k+1} + ks^k}{k+1}, \end{aligned} \tag{13}$$

where $k \geq 0$ and as usually $f(x) = \frac{1}{2}\|Ax - b\|^2$ and $\lambda = \tau\sigma$. Remember that the iterates (x^k) in the scheme (13) are exactly the same as in the PDHG method (with $y^0 = 0$). In the theorem below we will show that these iterates converge in fact to a solution of

$$\min_{x \in \mathbb{R}^n} \quad g(x) \quad \text{s.t.} \quad x \in \operatorname{argmin} f, \tag{14}$$

which is a more general problem than the original (1).

Based on [2,12,17,18], Tseng in [24] proposed a simple and elegant way to analyze accelerated gradient methods of Nesterov for a problem of composite minimization $\min_x g(x) + h(x)$, where $h\colon \mathbb{R}^n \to \mathbb{R}$ is a convex smooth function. Among several schemes that Tseng proposed one was the following

$$\begin{aligned} z^k &= \theta_k x^k + (1 - \theta_k)s^k \\ x^{k+1} &= \operatorname{prox}_{\frac{\lambda}{\theta_k} g}(x^k - \frac{\lambda}{\theta_k}\nabla h(z^k)) \\ s^{k+1} &= \theta_k x^{k+1} + (1 - \theta_k)s^k. \end{aligned} \tag{15}$$

Convergence of (15) was proved under the assumption that ∇h is λ^{-1}–Lipschitz continuous and $\theta_k \in (0, 1]$ satisfies $\frac{1-\theta_k}{\theta_k^2} \leq \frac{1}{\theta_{k-1}^2}$. The simplest choice for such (θ_k) is $\theta_k = \frac{2}{k+2}$.

It is easy to see how similar (15) and (13) are. What are the differences? First, the PDHG algorithm uses $\theta_k = \frac{1}{k+1}$, which does not satisfy condition for (θ_k) in (15) and goes slightly faster to zero than $\theta_k = \frac{2}{k+2}$. This is only due to the fact that f is a quadratic function, for which we can use tighter estimates. In fact, the same can be done for the Tseng algorithm in the case $h = f$. Second, in every iteration Tseng's scheme (15) uses the same stepsizes for both g and h and this is natural, as both these functions are equal-right in the composite minimization problem. In contrast, in problem (9) f and g are not equivalent: f impose hard constraints, thus the stepsize for f goes to infinity. Third, in (15) h can be an arbitrary function (up to the restrictions above), while in (13) f is a quadratic function. In fact, in our subsequent work we extend the algorithm (13) to a more general problem where f need not be a quadratic. It is also interesting to remark that in the case $h \equiv f$ and g is the indicator function of some closed convex set, both schemes (13) and (15) coincide: all equations are the same, only (θ_k) will be slightly different, see the discussion above.

In fact, since ∇f is linear, we do not need variable z^k in (13) at all. Evidently, the scheme (13) can be cast in a simpler way as

$$\begin{aligned} x^{k+1} &= \operatorname{prox}_{\tau g}(x^k - \lambda \nabla f(x^k + k\lambda s^k)) \\ s^{k+1} &= (x^{k+1} + ks^k)/(k+1), \end{aligned} \tag{16}$$

where $k \geq 0$ and $x^0 = s^0$.

Recall that $f_* = \min_x f(x)$, $g_* = \min_{x \in \operatorname{argmin} f} g(x)$. For our analysis we will also need the following *penalty function*

$$F_k(x) := g(x) + \sigma k(f(x) - f_*).$$

Theorem 1. *Assume that the solution set S of (14) is nonempty and $\lambda\|A\|^2 < 1$. Then for sequences (x^k), (s^k), generated by (13) (or (16)), it holds*

(i) $F_k(s^k) - g_* = O(1/k)$.

(ii) If there exists a Lagrange multiplier for problem (14), then (x^k) and (s^k) converge to a solution of (14) and $f(x^k) - f_* = o(1/k)$, $f(s^k) - f_* = O(1/k^2)$, $|F_k(s^k) - g_*| = O(1/k)$, $|g(s^k) - g_*| = O(1/k)$.

(iii) If S is bounded and g is bounded below, then all cluster points of (s^k) belong to S and $f(s^k) - f_* = o(1/k)$.

Therefore, in the most general case one can consider s^k as an approximated minimizer of the problem $\min_x F_k(x)$. When the strong duality holds, it is possible to prove convergence of the iterates and derive some important rates. Finally, even when there is no strong duality, but the solution set S is bounded and the function g is bounded below, we still can say something meaningful about convergence of the iterates (s^k) and the rate of the feasibility gap $f(s^k) - f_*$. The latter conditions are usually easy to check in advance, in contrast to the strong duality.

The strong duality plays such an important role here because it allows us to obtain a key estimate to prove global convergence. Specifically, assume that the strong duality holds for problem (14) and $(x^*, u^*) \in \mathbb{R}^n \times \mathbb{R}^n$ is a saddle point of

$$\min_x \max_u g(x) + \langle u, A^T A - A^T b\rangle.$$

This means that $0 \in A^T A u^* + \partial g(x^*)$ and hence,

$$g(x) - g_* \geq \langle -A^T A u^*, x - x^*\rangle \geq -\|Au^*\| \cdot \|A(x - x^*)\| = -D_y \cdot \sqrt{2(f(x) - f_*)}, \tag{17}$$

where for simplicity $D_y := \|Au^*\|$. Notice that in the consistent case a saddle point for (1) is (x^*, y^*) with $y^* = Au^*$ Thus, the above estimate recovers a more common one $g(x) - g_* \geq -\|y^*\|\|Ax - b\|$ and therefore, in this case $D_y = \|y^*\|$.

Proof. Let $\bar{x} \in S$. By the prox-inequality (4) and linearity of ∇f,

$$\frac{1}{\lambda}\langle x^{k+1} - x^k, \bar{x} - x^{k+1}\rangle + \langle \nabla f(x^k), \bar{x} - x^{k+1}\rangle + k\langle \nabla f(s^k), \bar{x} - x^{k+1}\rangle \geq \frac{1}{\sigma}(g(x^{k+1}) - g_*). \tag{18}$$

From Lemma 2 it follows

$$\begin{aligned} \langle \nabla f(x^k), \bar{x} - x^{k+1}\rangle &= 2f_* - f(x^{k+1}) - f(x^k) + \frac{1}{2}\|A(x^{k+1} - x^k)\|^2, \\ \langle \nabla f(s^k), \bar{x} - x^{k+1}\rangle &= 2f_* - f(x^{k+1}) - f(s^k) + \frac{1}{2}\|A(x^{k+1} - s^k)\|^2. \end{aligned} \tag{19}$$

Using the above identities in (18), we deduce

$$\begin{aligned} (k+1)(f(x^{k+1}) - f_*) + (f(x^k) - f_*) + k(f(s^k) - f_*) + \frac{1}{\sigma}(g(x^{k+1}) - g_*) - \frac{1}{2}\|A(x^{k+1} - x^k)\|^2 \\ \leq \frac{1}{\lambda}\langle x^{k+1} - x^k, \bar{x} - x^{k+1}\rangle + \frac{k}{2}\|A(x^{k+1} - s^k)\|^2. \end{aligned} \tag{20}$$

Convexity of $F_{k+1}(x) = g(x) + \sigma(k+1)(f(x) - f_*)$ and the property (6) for f yield

$$F_{k+1}(x^{k+1}) + kF_{k+1}(s^k) \geq (k+1)F_{k+1}(s^{k+1}) + \frac{k}{2(k+1)}\|A(x^{k+1} - s^k)\|^2. \tag{21}$$

Applying (21) to (20) and using that $kF_{k+1}(s^k) = kF_k(s^k) + k(f(s^k) - f_*)$, we obtain

$$\begin{aligned} \frac{k+1}{\sigma}F_{k+1}(s^{k+1}) + (f(x^k) - f_*) - \frac{1}{2}\|A(x^{k+1} - x^k)\|^2 - \frac{1}{\sigma}g_* \\ \leq \frac{1}{\lambda}\langle x^{k+1} - x^k, \bar{x} - x^{k+1}\rangle + \frac{k}{\sigma}F_k(s^k). \end{aligned}$$

Finally, using the cosine law (8) and $\|A(x^{k+1} - x^k)\| \leq \|A\|\|x^{k+1} - x^k\|$, we arrive at

$$\begin{aligned} \frac{1}{2\lambda}\|x^{k+1} - \bar{x}\|^2 + \frac{k+1}{\sigma}(F_{k+1}(s^{k+1}) - g_*) + \frac{1 - \lambda\|A\|^2}{2\lambda}\|x^{k+1} - x^k\|^2 + (f(x^k) - f_*) \\ \leq \frac{1}{2\lambda}\|x^k - \bar{x}\|^2 + \frac{k}{\sigma}(F_k(s^k) - g_*), \end{aligned} \tag{22}$$

which after multiplying by σ and setting $\beta = \frac{1-\lambda\|A\|^2}{2\tau}$ we can rewrite as

$$\frac{1}{2\tau}\|x^{k+1} - \bar{x}\|^2 + (k+1)(F_{k+1}(s^{k+1}) - g_*) + \beta\|x^{k+1} - x^k\|^2 + \sigma(f(x^k) - f_*) \\ \leq \frac{1}{2\tau}\|x^k - \bar{x}\|^2 + k(F_k(s^k) - g_*).$$

Iterating the above, we obtain

$$\frac{1}{2\tau}\|x^{k+1} - \bar{x}\|^2 + (k+1)\Big(F_{k+1}(s^{k+1}) - g_*\Big) + \beta\sum_{i=0}^{k}\|x^{i+1} - x^i\|^2 + \sigma\sum_{i=0}^{k}(f(x^i) - f_*) \\ \leq \frac{1}{2\tau}\|x^0 - \bar{x}\|^2 = \frac{D_x^2}{2\tau}, \quad (23)$$

where $D_x = \|x^0 - \bar{x}\|$. It follows that

$$F_k(s^k) - g_* \leq \frac{D_x^2}{2\tau k}. \quad (24)$$

(ii) There exists a Lagrange multiplier $u^* \in \mathbb{R}^n$ for problem (14). Applying the estimate obtained in (17), we derive

$$\frac{1}{2\tau}\|x^k - \bar{x}\|^2 + \sigma k^2(f(s^k) - f_*) - D_y k\sqrt{2(f(s^k) - f_*)} + \beta\sum_{i=0}^{k-1}\|x^{i+1} - x^i\|^2 \\ + \sigma\sum_{i=0}^{k-1}(f(x^i) - f_*) \leq \frac{D_x^2}{2\tau}. \quad (25)$$

Let $t = k\sqrt{f(s^k) - f_*}$. Then from (25) it follows that $\sigma t^2 - \sqrt{2}D_y t \leq D_x^2/2\tau$ and therefore, we have

$$t = k\sqrt{f(s^k) - f_*} \leq \frac{D_y + \sqrt{D_y^2 + \sigma D_x^2/\tau}}{\sqrt{2}\sigma}. \quad (26)$$

By this, we show that $f(s^k) - f_* = O(1/k^2)$. Since $t \mapsto \sigma t^2 - \sqrt{2}D_y t$ is bounded below by the constant $-\frac{D_y^2}{2\sigma}$, from (25) we conclude that (x^k) is bounded, $\|x^k - x^{k-1}\| \to 0$, $f(x^k) - f_* = o(1/k)$, and

$$-\frac{D_y^2}{2\sigma} \leq k(F_k(s^k) - g_*) \leq \frac{D_x^2}{2\tau}. \quad (27)$$

From the last inequality we have that $|F_k(s^k) - g_*| = O(1/k)$ and since $f(s^k) - f_* = O(1/k^2)$, we may deduce that $|g(s^k) - g_*| = O(1/k)$. By the definition of z^k, we also obtain $f(z^k) - f_* = O(1/k^2)$. It only remains to prove that (x^k) is convergent. First we show that all cluster points of (x^k) belong to S. Let (x^{k_i}) be any subsequence that converges to $\tilde{x}$. By the above, we know that $\tilde{x}$ is feasible, that is $f(\tilde{x}) = f_*$. By the prox-inequality, we have

$$\langle x^{k_i} - x^{k_i-1}, \bar{x} - x^{k_i}\rangle + k_i\sigma\langle A^T(Az^{k_i} - b), \bar{x} - x^{k_i}\rangle \geq \tau(g(x^{k_i}) - g_*). \quad (28)$$

If we want to tend $k_i \to \infty$, we need to know how to estimate the second term in the left-hand side of (28). Using that $A^T b = A^T A\bar{x}$, we derive

$$k_i\langle A^T(Az^{k_i} - b), \bar{x} - x^{k_i}\rangle = k_i\langle A(z^{k_i} - \bar{x}), A(\bar{x} - x^{k_i})\rangle \le k_i\|A(z^{k_i} - \bar{x})\| \cdot \|A(x^{k_i} - \bar{x})\|$$
$$= k_i\sqrt{f(z^{k_i}) - f_*}\sqrt{f(x^{k_i}) - f_*} \to 0,$$

due to the obtained asymptotics for $f(z^k)$ and $f(x^k)$. Hence, passing to the limit in (28) and using that $x^k - x^{k-1} \to 0$, we deduce $0 \ge \tau(g(\tilde{x}) - g_*)$. This means that $\tilde{x} \in S$ and therefore, all cluster points of (x^k) belong to S. From (22) it follows that

$$\frac{1}{2\tau}\|x^{k+1} - \bar{x}\|^2 + (k+1)(F_{k+1}(s^{k+1}) - g_*) \le \frac{1}{2\tau}\|x^k - \bar{x}\|^2 + k(F_k(s^k) - g_*).$$

As $(F_k(s^k) - g_*)_k$ is bounded below, we can apply Lemma 1 and conclude that the sequence (x^k) converges to some element in S. The convergence of (s^k) follows immediately.

(iii) First, we observe that from (24) we have

$$\sigma(f(s^k) - f_*) \le \frac{D_x^2}{2\tau k} + \frac{g_* - g(s^k)}{k}. \tag{29}$$

Since, g is bounded below, we obtain that $f(s^k) - f_* = O(1/k)$. Now we show that the sequence $(s^i)_{i\in\mathcal{I}}$ with $\mathcal{I} = \{i\colon g(s^i) < g_*\}$ is bounded. To this end, we use arguments from [22]. By our assumption the set $S = \{x\colon g(x) \le g_*, f(x) \le f_*\}$ is nonempty and bounded. Consider the convex function $\varphi(x) = \max\{g(x) - g_*, f(x) - f_*\}$. Notice that S coincides with the level set $\mathcal{L}(0)$ of φ:

$$S = \mathcal{L}(0) = \{x\colon \varphi(x) \le 0\}.$$

Since $\mathcal{L}(0)$ is bounded, $\mathcal{L}(c) = \{x\colon g(x) \le c\}$ is bounded for any $c \in \mathbb{R}$ as well. Fix any $c \ge 0$ such that $f(s^k) - f_* \le c$ for all k. As $g(s^i) - g_* < 0 \le c$ for $i \in \mathcal{I}$, we have that $s^i \in \mathcal{L}(c)$, which is a bounded set. Hence, $(s^i)_{i\in\mathcal{I}}$ is bounded.

Now we prove the boundedness of the whole sequence $(s^k)_{k\in\mathbb{N}}$. Let $M > 0$ be any constant that bounds from above $(\|s^i\|)_{i\in\mathcal{I}}$ and $D_x + \|\bar{x}\|$. For every index k we have two alternatives: either $g(s^k) < g_*$ or $g(s^k) \ge g_*$. If the latter holds, then

$$\|x^k\| \le \|x^k - \bar{x}\| + \|\bar{x}\| \le \|x^0 - \bar{x}\| + \|\bar{x}\| = D_x + \|\bar{x}\| \le M,$$

where the second inequality holds because of (23). If the former holds, then by the above arguments we know that $\|s^k\| \le M$. Assume that for the index k, $\|s^k\| \le M$. If for the index $k+1$, $g(s^{k+1}) < g_*$, then we are done: $k+1 \in \mathcal{I}$ and hence, $\|s^{k+1}\| \le M$. If $g(s^{k+1}) \ge g_*$, then $\|x^{k+1}\| \le M$. Observe that

$$\|s^{k+1}\| = \frac{\|ks^k + x^{k+1}\|}{k+1} \le \frac{k}{k+1}M + \frac{1}{k+1}M = M,$$

which completes the proof that (s^k) is bounded. As $f(s^k) - f_* = O(1/k)$, all cluster points of (s^k) are feasible. Taking the limit in (24) and using that g is lsc, we can also conclude that all cluster points of (s^k) belong to S. This guarantees that the whole sequence $(g(s^k))$ converges to g_*. By this, one can improve the obtained estimate for $f(s^k) - f_*$. In particular, now from (29) we have $f(s^k) - f_* = o(1/k)$ and the proof is complete. □

Remark 1.

(a) Note that all known proofs of the PDHG algorithm cover only the case (ii), but even in that case they can show convergence only in the consistent case, *i.e.,* when $f_* = 0$.

(b) Let g be the indicator function δ_C of some closed convex set C. Then in this case the performance of the primal-dual method does not depend on the ratio σ/τ and the strong duality assumption. This is natural, as now one can formulate problem (1) as a constrained least squares problem.

(c) Notice that one can easily make a solution set of problem (14) bounded by adding $\rho\|x\|^2$ to the objective g for some small number $\rho > 0$. In this case, the solution will be unique $S = \{\bar{x}\}$, and hence $s^k \to \bar{x}$. Quite often it is possible to show that $\bar{x}$ will be not far from the actual solution of (14).

3.2 Consequences

Complexity estimates. Consider the case when problem (1) is feasible, *i.e.*, $f_* = 0$, and the strong duality holds. For this case we will derive explicit estimates for ε-optimality and compare them with existing ones.

A vector $x \in \mathbb{R}^n$ is called an ε–approximate solution of (1) if it satisfies

$$|g(x) - g_*| \le \varepsilon \quad \text{and} \quad \|Ax - b\| \le \varepsilon.$$

Let (x^*, y^*) be any saddle point of (1). As $\|As^k - b\| = \sqrt{2(f(s^k) - f_*)}$, from (26) we derive

$$\|As^k - b\| \le \frac{D_y + \sqrt{D_y^2 + \sigma D_x^2/\tau}}{\sigma k}, \tag{30}$$

where we recall $D_x = \|x^0 - x^*\|$ and $D_y = \|y^*\|$. Similarly, we can obtain

$$-\frac{D_y^2 + D_y\sqrt{D_y^2 + \sigma D_x^2/\tau}}{\sigma k} \le g(s^k) - g_* \le \frac{D_x^2}{2\tau k}. \tag{31}$$

From (30) and (31) it is clear that we need $O(1/\varepsilon)$ iterations to obtain ε-solution. Quite remarkably, there are several papers [6,23] for solving problem (1) that also use Tseng's method, where it is applied to the dual smoothed problem. Moreover, we observe that our estimate for the feasibility gap and the lower estimate for the objective $g(s^k) - g_*$ are exactly the same as the ones obtained in [23, Theorem 3]. However, our upper bound for the objective $g(s^k) - g_*$ is still tighter than the one in [23]. Since the authors in [23] claim that their method achieves the best-known rate for the non-smooth settings, we believe our estimates improve their findings.

Implementation details. It is interesting to remark that the schemes (13) or (16) require even less memory than the original PDHG method. In particular, at every moment we have to keep only two vectors $x^k, s^k \in \mathbb{R}^n$. In contrast, for the PDHG algorithm, as one can see from (3), we have to store $x^k, x^{k-1} \in \mathbb{R}^n$ and $y^k \in \mathbb{R}^m$.

Moreover, in the case $m \ll n$, it might be more efficient to switch to the dual variables by using $\tilde{x}^k = Ax^k$, $\tilde{s}^k = As^k$. In this notation, the scheme (16) can be rewritten as

$$
\begin{aligned}
\tilde{x}^k &= Ax^k \\
x^{k+1} &= \operatorname{prox}_{\tau g}(x^k - \lambda A^T(\tilde{x}^k + k\tilde{s}^k - b)) \\
\tilde{s}^{k+1} &= (\tilde{x}^{k+1} + k\tilde{s}^k)/(k+1).
\end{aligned}
$$

This scheme preserves the same amount of computation per iteration, but requires us to store only one primal variable $x^k \in \mathbb{R}^n$ and two dual variables $\tilde{x}^k, \tilde{s}^k \in \mathbb{R}^m$, which in the case $2m < n$ is cheaper than the schemes (3) or (13) do. Finally, in the case $m \gg n$ there is another possibility to precompute $A^T A \in \mathbb{R}^{n\times n}$ and use it in all iterations of (13) or (16).

Connection to penalty methods. Another approach to solve (1) or more general problem (14) is the penalty method. It consists in solving a sequence of unconstrained optimization problems

$$\min_x g(x) + \rho_k(f(x) - f_*), \quad \rho_k > 0 \tag{32}$$

for some increasing sequence $\rho_k \to \infty$ as $k \to \infty$. Intuitively it is clear that with larger ρ_k, solutions of (32) become closer to a solution of our constrained problem (14). For more rigorous treatment on this subject, see [8, 19]. In general, penalty methods do not use duality arguments, although one still needs them in order to obtain some convergence rates or even to prove global convergence [8]. As an exception, there is a recent paper [16] that studies the conic optimization problem without assuming that there exists a Lagrange multiplier. The authors applied the accelerated gradient method for the penalized objective (although different from (32)) and derived $O(1/k^{\frac{2}{3}})$ estimate for the feasibility gap.

Generally speaking, we cannot solve just one problem (32) for ρ_k large enough. First, because we do not know which ρ_k is large enough for approximation of the true solution. And second, because solving (32) in practice becomes difficult for large ρ_k. Thus, penalty methods require solving a sequence of optimization problems (32), which can be quite costly. Clearly, for a specific choice $\rho_k = k\sigma$ problem (32) becomes nothing more than just $\min_x F_k(x)$. What we have shown is that the primal-dual method provides a nice alternative to penalty methods. Instead of solving a sequence of problems (32), it runs one iteration of something similar to the proximal gradient method for each of the problems (32) with $\rho_k = k\sigma$. In the literature this is known as a *diagonal penalty* method, see a nice overview of such methods in [9]. In general, diagonal penalty methods are a modification of some known algorithms with an appropriate penalty function; proving their convergence can be tricky and usually it requires additional assumptions like strong convexity. Thus, it is quite remarkably that the vanilla PDHG method is unintentionally a diagonal penalty method.

Connection to inverse problems. A central problem in inverse problems is solving a linear system $Ax^\dagger + \varepsilon = b$, where the matrix $A \in \mathbb{R}^{m\times n}$ is given, the vector $b \in \mathbb{R}^m$ is observed, and $\varepsilon \in \mathbb{R}^m$ is some noise. Since in most cases these problems are ill-posed and the noise ε is unknown, in order to solve them we have to impose an appropriate regularization. The most common approach is to consider the Tikhonov regularization:

$$\min_x g(x) + \frac{\gamma}{2}\|Ax - b\|^2, \tag{33}$$

where g is the regularizer that promotes some desirable properties of a solution such as sparsity, smoothness, etc., and $\gamma > 0$ is the regularization parameter. The question of how to choose this parameter is the main concern of such approach. In a theory the best thing would be to solve a sequence of problems (33) with different $(\gamma_k)_{k\in\mathbb{N}}$ and choose among solutions the best; apparently it is not the most practical way. In our notation it means that we would like to choose the "best" $\hat{x}^k$ (whatever it means) among all $k \in \mathbb{N}$ such that

$$\hat{x}^k \in \operatorname*{argmin}_x F_k(x) := g(x) + \sigma k(f(x) - f_*). \tag{34}$$

It is clear that the problem (33) with $\gamma = \sigma k$ is equivalent to (34).

Instead of choosing parameter γ or solving a sequence of problems (34), one can apply the PDHG directly to the problem (1). Of course due to the noise, the linear system $Ax = b$ might be inconsistent, however this is not our concern, as we have already shown that the iterates (x^k) of the PDHG method will still converge to a solution of a more general problem (14). Moreover, we can show that $|F_k(\hat{x}^k) - F_k(s^k)| = O(1/k)$.

Assume that the strong duality holds. The same estimation as in Theorem 1 (ii) provides us

$$-\frac{D_y^2}{2\sigma} \le k(F_k(\hat{x}^k) - g_*).$$

From this it follows that $g_* \le \frac{D_y^2}{2\sigma k} + F_k(\hat{x}^k)$. Since $F_k(\hat{x}^k) \le F_k(s^k)$, from (27) we have

$$-\frac{D_y^2}{2\sigma k} \le F_k(\hat{x}^k) - g_* \le F_k(s^k) - g_* \le \frac{D_x^2}{2\tau k}.$$

Combining the latter two inequalities, we can conclude

$$F_k(\hat{x}^k) \le F_k(s^k) \le \frac{D_x^2}{2\tau k} + \frac{D_y^2}{2\sigma k} + F_k(\hat{x}^k),$$

and hence $|F_k(s^k) - F_k(\hat{x}^k)| = O(1/k)$. This means that by applying the PDHG algorithm only one time for one problem (1), we approach to each of the solutions $\hat{x}^k$ of the regularized problem.

Distributed optimization. Here we will show that the new scheme (16) applied to the distributed optimization enjoys much better properties than the original PDHG method. Assume we have a connected simple graph $G = (V, E)$ of $n = |V|$ computing units v_i, each having access to a convex lsc function $g_i : \mathbb{R}^d \to (-\infty, +\infty]$. Our aim is to find a consensus on the minimum of the aggregate objective $g_1(x_1) + \cdots + g_n(x_n)$ in a decentralized way. This problem can be written as

$$\min_{x_1,\dots,x_n} g_1(x_1) + \cdots + g_n(x_n) \quad \text{s.t.} \quad x_1 = \cdots = x_n. \tag{35}$$

In order to rely on the decentralized computation, we only allow communication between adjacent nodes. The standard way to impose the network topology is to reformulate problem (35) exploiting the *Laplacian matrix* $L \in \mathbb{R}^{n\times n}$ of G, which is zero everywhere, except $L_{ii} = \deg v_i$ and $L_{ij} = -1$ if $(i,j) \in E$. Then one can formulate problem (35) as

$$\min_{x\in\mathbb{R}^{n\times d}} \quad g(x) \quad \text{s.t.} \quad Lx = 0, \tag{36}$$

where $x = (x_1, \dots, x_n)^T \in \mathbb{R}^{n\times d}$, $g(x) = g_1(x_1) + \dots + g_n(x_n)$. It is not difficult to see [15] that condition $Lx = 0$ is equivalent to $x_1 = \dots = x_n$. Notice that multiplication of L with the current iterate x^k amounts to the exchange of information between the respective nodes. Of course, problem (36) is a particular case of (1). The PDHG (3) applied to (36) requires each node v_i to store x_i^k, x_i^{k-1}, and y_i^k in each iteration. Most important is that we have two matrix-vector multiplications $L\bar{x}^k$ and Ly^{k+1}, which implies two communications per iteration.

Can scheme (16) provide us something new? In fact, yes. Note that instead of problem (36) one can consider an equivalent problem

$$\min_{x\in\mathbb{R}^{n}} \quad g(x) \quad \text{s.t.} \quad \sqrt{L}x = 0. \tag{37}$$

Since L is symmetric semidefinite, $\sqrt{L}$ is well-defined and symmetric semidefinite as well. As $\sqrt{L}\sqrt{L} = L$, the scheme (16) boils down to

$$\begin{aligned} x^{k+1} &= \operatorname{prox}_{\tau g}(x^k - \lambda L(x^k + ks^k)) \\ s^{k+1} &= (x^{k+1} + ks^k)/(k+1), \end{aligned}$$

which means that we need only one communication between the nodes! The reason why we could not apply the original PDHG method (3) to problem (37) is obvious: in general $\sqrt{L}x^k$ is not related to communication of nodes: it might require exchanging of information between nodes that are not directly connected.

It remains to notice that the complexity estimates will be also much better for problem (37) than for (36). Indeed, (36) requires $\tau\sigma\|L\|^2 < 1$ while for (37) we need only $\tau\sigma\|L\| < 1$. For Laplacian matrices of graphs we know [15] that $\|L\| \geq d_{\max} + 1 > 1$, where $d_{\max}$ denotes the largest degree of the nodes of G. If we assume that the strong duality holds for (36) and a (x^*, u^*) is a saddle point, then it also holds for (37) and $(x^*, \sqrt{L}u^*)$ would be a respective saddle point. Thus, the bound for D_y in (17) for (37) is also not bigger than the one for (36). Hence, when the strong duality holds, from (30) and (31) we can conclude that the complexity rates for (37) are better than the ones for (36).

The idea to use constraints $\sqrt{L}x = 0$ in (37) is quite standard, see for example [11, 25]. However such algorithms require either more restrictive assumptions (strong convexity, dual-friendliness, etc.) or they have more expensive iterations.

As a remark, we note that the above discussion will be also valid if instead of L one considers a more general weighted Laplacian matrix.

4 Generalization

4.1 Additional smooth term

Assume now that we are given more structure in problem (1):

$$\min_{x\in\mathbb{R}^n} \quad g(x)+h(x) \quad \text{s.t.} \quad Ax=b, \tag{38}$$

where in addition to the previous settings we assume that $h\colon \mathbb{R}^n \to \mathbb{R}$ is a convex differentiable function with β–Lipschitz gradient. In most cases computing the $\operatorname{prox}_{g+h}$ is not practical anymore, thus the vanilla PDHG method will be not efficient for this problem. Condat and Vũ in [5, 26] proposed an extension of the PDHG algorithm to deal with such cases. Applied to (38), this algorithm is given by

$$\begin{aligned} y^{k+1} &= y^k + \sigma(A\bar{x}^k - b)\\ x^{k+1} &= \operatorname{prox}_{\tau g}(x^k - \tau(A^T y^{k+1} + \nabla h(x^k))). \end{aligned} \tag{39}$$

Its convergence can be proved under the assumptions that the solution set is nonempty, the strong duality holds and $\tau\sigma\|A\|^2 < 1-\tau\beta$.

It is clear that in the same way as in section 3.1 one can transform (39) into entirely primal algorithm:

$$\begin{aligned} z^k &= \frac{x^k + ks^k}{k+1}\\ x^{k+1} &= \operatorname{prox}_{\tau g}(x^k - \tau\nabla h(x^k) - (k+1)\lambda\nabla f(z^k))\\ s^{k+1} &= \frac{x^{k+1}+ks^k}{k+1}. \end{aligned} \tag{40}$$

We will not repeat the proof, rather give a key ingredient. It is the following inequality

$$\begin{aligned} \langle \nabla h(x^k), x - x^{k+1}\rangle &\le h(x) - h(x^{k+1}) + \frac{\beta}{2}\|x^{k+1}-x^k\|^2\\ &\le k(h(x)-h(s^k)) - (k+1)(h(x)-h(s^{k+1})) + \frac{\beta}{2}\|x^{k+1}-x^k\|^2, \end{aligned}$$

where the first inequality follows from the descent lemma and convexity of h and is standard in proving convergence of the proximal gradient method, and the second follows from convexity of h and the definition of s^{k+1}. Combining this inequality with the similar ones as in the proof of Theorem 1 we can show convergence of (40). Evidently, in the same way we can show that algorithm (40), and hence, (39), in fact solves a more general problem

$$\min_{x\in\mathbb{R}^n} \quad g(x)+h(x) \quad \text{s.t.} \quad x\in \operatorname{argmin} f. \tag{41}$$

4.2 g is strongly convex

When g is γ–strongly convex, we can obtain even better convergence rates. Although our results presented below will be valid for the general case as in (41), for the clarity of presentation we consider the case when $h\equiv 0$. Hence, now our problem reads as:

$$\min_{x\in\mathbb{R}^n} \quad g(x) \quad \text{s.t.} \quad x\in \operatorname{argmin} f, \tag{42}$$

where g is 1–strongly convex function that we assume without loss of generality.

First, let us consider the case when the linear system $Ax = b$ is consistent and there exists a Lagrange multiplier. In this case, one can apply the accelerated PDHG method [3]:

$$\begin{aligned} y^{k+1} &= y^k + \sigma_k(A\bar{x}^k - b) \\ x^{k+1} &= \operatorname{prox}_{\tau_k g}(x^k - \tau_k A^T y^{k+1}), \end{aligned} \tag{43}$$

where $\bar{x}^k = x^k + \theta_k(x^k - x^{k-1})$, $\theta_k = \frac{\tau_k}{\tau_{k-1}}$ and

$$\tau_k = \frac{\tau_{k-1}}{\sqrt{1+\tau_{k-1}}}, \qquad \tau_k\sigma_k = \lambda, \quad \lambda\|A\|^2 \le 1 \qquad \forall k \ge 0. \tag{44}$$

Iterating the first equation in (43), one can derive

$$\begin{aligned} y^{k+1} &= y^k + \sigma_k(A\bar{x}^k - b) = y^{k-1} + A(\sigma_k\bar{x}^k + \sigma_{k-1}\bar{x}^{k-1}) - (\sigma_k + \sigma_{k-1})b = \dots \\ &= y^0 + A(\sigma_k\bar{x}^k + \dots + \sigma_0\bar{x}^0) - (\sigma_k + \dots + \sigma_0)b. \end{aligned}$$

Similarly as in section 3.1, one may introduce z^k, defined by $z^0 = x^0$, $z^1 = x^1$ and

$$z^k = \frac{\sigma_k\bar{x}^k + \dots + \sigma_0\bar{x}^0}{\sigma_k + \dots + \sigma_0} = \frac{(\sigma_k + \sigma_{k-1})x^k + \sigma_{k-2}x^{k-1} + \dots + \sigma_0 x^1}{\sigma_k + \dots + \sigma_0},$$

for $k \ge 2$. Here we have used that $\bar{x}^k = x^k + \theta_k(x^k - x^{k-1}) = x^k + \frac{\sigma_{k-1}}{\sigma_k}(x^k - x^{k-1})$. Let $\Sigma_k := \sigma_k + \dots + \sigma_0$ for $k \ge 0$ and $\Sigma_{-1} = 0$. Define sequence (s^k) as $s^0 = x^0$ and $s^k = \frac{\sigma_{k-1}x^k + \dots + \sigma_0 x^1}{\Sigma_{k-1}}$. For simplicity, we again assume that $y^0 = 0$, thus $y^{k+1} = \Sigma_k(Az^k - b)$ for all $k \ge 0$. Then the primal-dual scheme (43) might be written in the primal form:

$$\begin{aligned} z^k &= (\sigma_k x^k + \Sigma_{k-1}s^k)/\Sigma_k \\ x^{k+1} &= \operatorname{prox}_{\tau_k g}(x^k - \tau_k\Sigma_k\nabla f(z^k)) \\ s^{k+1} &= (\sigma_k x^{k+1} + \Sigma_{k-1}s^k)/\Sigma_k, \end{aligned} \tag{45}$$

where $k \ge 0$, τ_k, σ_k satisfy (44) and $\Sigma_k = \sigma_k + \dots + \sigma_0$. We show that this algorithm in fact solves (42).

Let g_* be the optimal value of (42), $F_k(x) = g(x) + \Sigma_k(f(x) - f_*)$ be the penalty function, and $\hat{x}^k$ be the unique minimizer of F_k.

Theorem 2. *Let (x^k), (s^k) be generated by (45), $\lambda\|A\|^2 \le 1$, and the solution set $S = \{\bar{x}\}$. Then it holds*

(i) (s^k) converges to $\bar{x}$, $F_k(\hat{x}^k) - g_* \le F_k(s^k) - g_* = O(1/k^2)$, $f(s^k) - f_* = o(1/k^2)$.

(ii) If there exists a Lagrange multiplier for problem (42), then (x^k) also converges to $\bar{x}$ at the rate $\|x^k - \bar{x}\| = O(1/k)$ and $f(x^k) - f_* = o(1/k^3)$, $f(s^k) - f_* = O(1/k^4)$, $|F_k(s^k) - g_*| = O(1/k^2)$.

Proof. By the prox-inequality (5),

$$\frac{1}{\tau_k}\langle x^{k+1} - x^k, \bar{x} - x^{k+1}\rangle + \sigma_k\langle\nabla f(x^k), \bar{x} - x^{k+1}\rangle + \Sigma_{k-1}\langle\nabla f(s^k), \bar{x} - x^{k+1}\rangle \\ \ge g(x^{k+1}) - g(\bar{x}) + \frac{1}{2}\|x^{k+1} - \bar{x}\|^2.$$

Using identities (19), we obtain

$$\Sigma_k(f(x^{k+1})-f_*)+\sigma_k(f(x^k)-f_*)+\Sigma_{k-1}(f(s^k)-f_*)+(g(x^{k+1})-g_*)+\frac{1}{2}\|x^{k+1}-\bar{x}\|^2 \\ -\frac{\sigma_k}{2}\|A(x^{k+1}-x^k)\|^2 \le \frac{1}{\tau_k}\langle x^{k+1}-x^k, \bar{x}-x^{k+1}\rangle + \frac{\Sigma_{k-1}}{2}\|A(x^{k+1}-s^k)\|^2. \quad (46)$$

Convexity of $F_{k+1}(x)=g(x)+\Sigma_k(f(x)-f_*)$ and the property (6) for f yield

$$--\sigma_k F_{k+1}(x^{k+1})+\Sigma_{k-1}F_{k+1}(s^k)\ge \Sigma_k F_{k+1}(s^{k+1})+\frac{\sigma_k\Sigma_{k-1}}{2\Sigma_k}\|A(x^{k+1}-s^k)\|^2. \quad (47)$$

Applying (47) to (46) and using that $\Sigma_{k-1}F_{k+1}(s^k)=\Sigma_{k-1}F_k(s^k)+\sigma_k\Sigma_{k-1}(f(s^k)-f_*)$, we obtain

$$\frac{\Sigma_k}{\sigma_k}F_{k+1}(s^{k+1})-g_*)+\sigma_k(f(x^k)-f_*)-\frac{\sigma_k}{2}\|A(x^{k+1}-x^k)\|^2+\frac{1}{2}\|x^{k+1}-\bar{x}\|^2 \\ \le \frac{1}{\tau_k}\langle x^{k+1}-x^k,\bar{x}-x^{k+1}\rangle+\frac{\Sigma_{k-1}}{\sigma_k}(F_k(s^k)-g_*).$$

From the cosine law (8) and $\|A(x^{k+1}-x^k)\|\le\|A\|\|x^{k+1}-x^k\|$ it follows that

$$\frac{1+\tau_k}{2\tau_k}\|x^{k+1}-x\|^2+\frac{\Sigma_k}{\sigma_k}(F_{k+1}(s^{k+1})-g_*)+\frac{1-\lambda\|A\|^2}{2\tau_k}\|x^{k+1}-x^k\|^2+\sigma_k(f(x^k)-f_*) \\ \le\frac{1}{2\tau_k}\|x^k-x\|^2+\frac{\Sigma_{k-1}}{\sigma_k}(F_k(s^k)-g_*). \quad (48)$$

By the definition of (τ_k), we have $\frac{\tau_k\sigma_{k+1}}{\sigma_k\tau_{k+1}}=\frac{\tau_k^2}{\tau_{k+1}^2}=(1+\tau_k)$ and $\tau_k\sigma_k=\lambda$. Multiplying the left and right hand-sides of (48) by σ_k and using the latter identities, we deduce

$$\frac{\sigma_{k+1}}{2\tau_{k+1}}\|x^{k+1}-\bar{x}\|^2+\Sigma_k(F_{k+1}(s^{k+1})-g_*)+\sigma_k^2(f(x^k)-f_*) \\ \le\frac{\sigma_k}{2\tau_k}\|x^k-\bar{x}\|^2+\Sigma_{k-1}(F_k(s^k)-g_*).$$

Iterating the above and recalling that $\Sigma_{-1}=0$, we obtain

$$\frac{\sigma_k}{2\tau_k}\|x^k-\bar{x}\|^2+\Sigma_{k-1}(F_k(s^k)-g_*)+\sum_{i=0}^{k-1}\sigma_i^2(f(x^i)-f_*)\le\frac{\sigma_0}{2\tau_0}\|x^0-\bar{x}\|^2=\frac{\sigma_0}{2\tau_0}D_x^2 \quad (49)$$

For simplicity, assume that $\tau_0=1$. Then it is not difficult to prove by induction that $\frac{2}{k+2}\le\tau_k\le\frac{3}{k+2}$. In the general case, all results will be the same up to some constants, as it is known from [3] that $\tau_k\sim 1/k$. In the case $\tau_0=1$, we have $\frac{k+2}{2}\lambda\ge\sigma_k\ge\frac{k+2}{3}\lambda$, hence

$$\frac{(k+4)(k+1)}{4}\lambda\ge\Sigma_k\ge\frac{(k+4)(k+1)}{6}\lambda.$$

First, let us estimate the term $F_k(s^k)-g_*$. From (49) it follows that $\Sigma_k(F_{k+1}(s^{k+1})-g_*)\le\frac{\lambda}{2}D_x^2$, where we took $\sigma_0=\lambda$, due to $\tau_0=1$. Thus,

$$F_{k+1}(s^{k+1})-g_*\le\frac{\lambda D_x^2}{2\Sigma_k}\le\frac{3D_x^2}{(k+4)(k+1)}=O(1/k^2). \quad (50)$$

As g is strongly convex, it is bounded below. The set $S = \{\bar{x}\}$ is of course bounded, thus we can use the same arguments as in part (iii) of Theorem 1 to conclude that (s^k) is bounded, $s^k \to \bar{x}$ and $g(s^k) \to g_*$. Equation (50) also yields $f(s^{k+1}) \leq \frac{\lambda D_x^2}{2\Sigma_k^2} + \frac{g_* - g(s^k)}{\Sigma_k} = o(1/k^2)$, since $g(s^k) \to g_*$ and $1/\Sigma_k = O(1/k^2)$.

Case (ii). There exists a Lagrange multiplier $u^* \in \mathbb{R}^n$. Applying inequality (17) to (49), we obtain

$$\frac{\sigma_k}{2\tau_k}\|x^k - \bar{x}\|^2 + \Sigma_{k-1}^2(f(s^k) - f_*) - D_y\Sigma_{k-1}\sqrt{2(f(s^k) - f_*)} + \sum_{i=0}^{k-1}\sigma_i^2(f(x^i) - f_*) \leq \frac{\lambda D_x^2}{2}.$$

Let $t = \Sigma_{k-1}\sqrt{f(s^k) - f_*}$. Then from the last equation it follows that $t^2 - \sqrt{2}D_y t \leq \frac{\lambda D_x^2}{2}$, from which one can derive that $t \leq \frac{D_y + \sqrt{D_y^2 + \lambda D_x^2}}{\sqrt{2}}$. By this, we show that $f(s^k) - f_* = O(1/k^4)$. Since $t^2 - \sqrt{2}D_y t$ is bounded below by the constant $-\frac{D_y^2}{2}$ and $\frac{\tau_k}{\sigma_k} \leq \frac{(k+2)^2}{9\lambda}$, we conclude that

$$\|x^k - \bar{x}\|^2 \leq \frac{\tau_k}{\sigma_k}(\lambda D_x^2 + D_y^2) \leq (D_x^2 + \frac{D_y^2}{\lambda})\frac{9}{(k+2)^2}$$

and $f(x^k) - f_* = o(1/k^3)$. From

$$-\frac{D_y^2}{2} \leq \Sigma_{k-1}(F_k(s^k) - g_*) \leq \frac{\lambda D_x^2}{2}$$

we observe that $|F_k(s^k) - g_*| = O(1/k^2)$ and due to the asymptotic of $f(s^k) - f_*$ we have $|g(s^k) - g_*| = O(1/k^2)$. □

Remark 2. We note that for both cases mentioned in this section it is straightforward to derive similar results as in section 3.2.

References

[1] S. Banert, R. I. Bot, and E. R. Csetnek. Fixing and extending some recent results on the admm algorithm. *arXiv:1612.05057*, 2016.

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

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

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

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

[6] P. Dvurechensky, A. Gasnikov, and A. Kroshnin. Computational optimal transport: Complexity by accelerated gradient descent is better than by Sinkhorn's algorithm. *arXiv:1802.04367*, 2018.

[7] O. Fercoq and P. Richtárik. Accelerated, parallel, and proximal coordinate descent. *SIAM Journal on Optimization*, 25(4):1997–2023, 2015.

[8] R. Fletcher. *Practical methods of optimization*. John Wiley & Sons, 2013.

[9] G. Garrigos, L. Rosasco, and S. Villa. Iterative regularization via dual diagonal descent. *Journal of Mathematical Imaging and Vision*, 60(2):189–215, 2018.

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

[11] D. Jakovetić, J. M. Moura, and J. Xavier. Linear convergence rate of a class of distributed augmented lagrangian algorithms. *IEEE Transactions on Automatic Control*, 60(4):922–936, 2015.

[12] G. Lan, Z. Lu, and R. D. Monteiro. Primal-dual first-order methods with iteration-complexity for cone programming. *Mathematical Programming*, 126(1):1–29, 2011.

[13] D. R. Luke and Y. Malitsky. Block-coordinate primal-dual method for the nonsmooth minimization over linear constraints. *arXiv:1801.04782*, 2018.

[14] Y. Malitsky and T. Pock. A first-order primal-dual algorithm with linesearch. *SIAM Journal on Optimization*, 28(1):411–432, 2018.

[15] R. Merris. Laplacian matrices of graphs: a survey. *Linear algebra and its applications*, 197:143–176, 1994.

[16] I. Necoara, A. Patrascu, and F. Glineur. Complexity of first-order inexact Lagrangian and penalty methods for conic convex programming. *Optimization Methods and Software*, pages 1–31, 2017.

[17] Y. Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$. *Doklady AN SSSR*, 269(3):543–547, 1983.

[18] Y. Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013.

[19] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

[20] Y. Saad. *Iterative Methods for Sparse Linear Systems*. Society for Industrial and Applied Mathematics, second edition, 2003.

[21] R. Shefi and M. Teboulle. Rate of convergence analysis of decomposition methods based on the proximal method of multipliers for convex minimization. *SIAM Journal on Optimization*, 24(1):269–297, 2014.

[22] M. Solodov. An explicit descent method for bilevel convex optimization. *Journal of Convex Analysis*, 14(2):227, 2007.

[23] Q. Tran-Dinh, O. Fercoq, and V. Cevher. A smooth primal-dual optimization framework for nonsmooth composite convex minimization. *SIAM Journal on Optimization*, 28(1):96–134, 2018.

[24] P. Tseng. On accelerated proximal gradient methods for convex-concave optimization, 2008.

[25] C. A. Uribe, S. Lee, A. Gasnikov, and A. Nedić. Optimal algorithms for distributed optimization. *arXiv:1712.00232*, 2017.

[26] B. C. Vũ. A splitting algorithm for dual monotone inclusions involving cocoercive operators. *Advances in Computational Mathematics*, 38(3):667–681, 2013.