On the Convergence of Primal-Dual Hybrid Gradient Algorithm

He, Bingsheng; You, Yanfei; Yuan, Xiaoming

Published in:
SIAM Journal on Imaging Sciences

DOI:
10.1137/140963467

Published: 03/12/2014

Document Version:
Publisher's PDF, also known as Version of record

SIAM J. IMAGING SCIENCES
Vol. 7, No. 4, pp. 2526–2537

On the Convergence of Primal-Dual Hybrid Gradient Algorithm*

Bingsheng He†, Yanfei You‡, and Xiaoming Yuan§

Abstract. The primal-dual hybrid gradient algorithm (PDHG) has been widely used, especially for some basic image processing models. In the literature, PDHG's convergence was established only under some restrictive conditions on its step sizes. In this paper, we revisit PDHG's convergence in the context of a saddle-point problem and try to better understand how to choose its step sizes. More specifically, we show by an extremely simple example that PDHG is not necessarily convergent even when the step sizes are fixed as tiny constants. We then show that PDHG with constant step sizes is indeed convergent if one of the functions of the saddle-point problem is strongly convex, a condition that does hold for some variational models in imaging. With this additional condition, we also establish a worst-case convergence rate measured by the iteration complexity for PDHG with constant step sizes.

Key words. primal-dual hybrid gradient algorithm, image restoration, total variation, saddle-point problem, convex optimization, convergence rate

AMS subject classifications. 90C25, 94A08

DOI. 10.1137/140963467

1. Introduction.

We consider a saddle-point problem

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}}\ \Phi(x,y) := \theta_1(x) - y^TAx - \theta_2(y), \tag{1.1}$$

where $A \in \Re^{m\times n}$, $\mathcal{X} \subseteq \Re^n$, $\mathcal{Y} \subseteq \Re^m$ are closed convex sets, and $\theta_1 : \Re^n \to \Re$ and $\theta_2 : \Re^m \to \Re$ are convex but not necessarily smooth functions. The solution set of (1.1) is assumed to be nonempty throughout our discussion. The model (1.1) captures a wide range of applications in different areas. For example, finding a saddle point for the Lagrange function of the canonical convex minimization model with linear equality or inequality constraints is a special case of (1.1), and a number of variational models with the total variation (TV) regularization in [21] arising in image restoration can also be reformulated as special cases of (1.1); see details in, e.g., [7, 22, 24, 25].

For many applications, the functions θ_1 and θ_2 usually have some special properties, and it is worthwhile to take advantage of them in algorithmic design. This idea has inspired some efficient splitting algorithms whose common advantage is that the functions θ_i are treated

*Received by the editors April 3, 2014; accepted for publication (in revised form) August 25, 2014; published electronically December 3, 2014.
http://www.siam.org/journals/siims/7-4/96346.html
†International Centre of Management Science and Engineering, and Department of Mathematics, Nanjing University, Nanjing, 210093, China (hebma@nju.edu.cn). This author was supported by NSFC grant 91130007 and MOEC fund 20110091110004.
‡Department of Mathematics, Nanjing University, Nanjing, 210093, China (yanfeiyou@gmail.com).
§Department of Mathematics, Hong Kong Baptist University, Hong Kong (xmyuan@hkbu.edu.hk). This author was supported by the General Research Fund from Hong Kong Research Grants Council 203613.

individually and thus the resulting subproblems are often easier, or sometimes easy enough to have closed-form solutions. In particular, for solving some TV image restoration models which are all special cases of the model (1.1), the primal-dual hybrid gradient algorithm (PDHG) was proposed in [25]. Slightly extending the original PDHG scheme in [25] to the model (1.1), we obtain the scheme

$$\left\{\begin{array}{l} x^{k+1} = \arg\min\{\Phi(x, y^k) + \frac{r}{2}\|x - x^k\|^2 \,|\, x \in \mathcal{X}\}, \\ y^{k+1} = \arg\max\{\Phi(x^{k+1}, y) - \frac{s}{2}\|y - y^k\|^2 \,|\, y \in \mathcal{Y}\}, \end{array}\right. \tag{1.2}$$

where r and s are positive scalars. As mentioned in [7], the PDHG scheme (1.2) corresponds to the classical Arrow–Hurwicz method in [1]. In the imaging literature, the efficiency of PDHG has been well demonstrated; see, e.g., [2, 9, 24, 25]. For some interesting variants of PDHG, we refer the reader to [7, 8, 11, 13, 19]. In this short paper, we focus our discussion on only the PDHG scheme (1.2).

As delineated in the literature (see, e.g., [2, 7, 24, 25]), $1/r$ and $1/s$ are the step sizes associated with gradient-type methods (or subgradient-type methods if θ_1 and/or θ_2 are/is not smooth) for solving the decomposed subproblems in (1.2). It is important to choose appropriate values for r and s to ensure PDHG's efficiency. In [9], it was shown that PDHG is related to the inexact Uzawa method in [1], and the convergence of PDHG was proved for some variational imaging models under some asymptotical conditions on the step size sequences. In [7], the worst-case $O(1/k)$ convergence rate measured by the iteration complexity was established for PDHG where k is the iteration counter,[1] and sharper convergence rates such as the $O(1/k^2)$ and $O(1/e^k)$ rates were also established under stronger assumptions. In [2], the convergence of a more general scheme of PDHG was established under some asymptotical conditions on the step size sequences. In particular, the efficiency of the step size rules in [25] was well explained because they satisfy the conditions posed for the analysis in [2]. In general, specifying a step size sequence is more difficult than choosing a constant step size for implementing PDHG practically.

Our goal in this short paper is to better understand how to choose step sizes for PDHG. First, we show by an extremely simple example that PDHG is not necessarily convergent even when its step sizes are fixed as tiny constants. Then, we show that PDHG with constant step sizes is convergent if one function in the model (1.1) is strongly convex. We also establish a worst-case $O(1/k)$ convergence rate measured by the iteration complexity for PDHG with this additional condition. This is further theoretical support for the empirical efficiency of PDHG.

2. Preliminaries.

In this section, we show that the saddle-point problem (1.1) can be reformulated as a variational inequality (VI) problem, on which our theoretical discussion on PDHG's convergence will be based.

Let $(x^*, y^*) \in \mathcal{X} \times \mathcal{Y}$ be a solution point of the saddle-point problem (1.1). Then, we have

$$\Phi_{y\in\mathcal{Y}}(x^*, y) \le \Phi(x^*, y^*) \le \Phi_{x\in\mathcal{X}}(x, y^*). \tag{2.1}$$

[1] As in [16, 17] and many other works, a worst-case $O(1/k)$ convergence rate means that the accuracy to a solution under certain criteria is of the order $O(1/k)$ after k iterations of an iterative scheme, or, equivalently, it requires at most $O(1/\epsilon)$ iterations to achieve an approximate solution with an accuracy of ϵ.

Note that the second inequality in (2.1) implies that

$$x^* \in \mathcal{X}, \ \ \theta_1(x) - \theta_1(x^*) + (x - x^*)^T(-A^T y^*) \geq 0 \quad \forall\, x \in \mathcal{X}.$$

Analogously, the first inequality in (2.1) implies that

$$y^* \in \mathcal{Y}, \ \ \theta_2(y) - \theta_2(y^*) + (y - y^*)^T(Ax^*) \geq 0 \quad \forall\, y \in \mathcal{Y}.$$

Therefore, finding a solution point (x^*, y^*) of (1.1) is equivalent to solving the VI: Find $u^* = (x^*, y^*)$ such that

$$\text{VI}(\Omega, \theta, M): \quad u^* \in \Omega, \quad \theta(u) - \theta(u^*) + (u - u^*)^T M u^* \geq 0 \quad \forall\, u \in \Omega, \tag{2.2a}$$

where

$$u := \begin{pmatrix} x \\ y \end{pmatrix}, \quad \theta(u) := \theta_1(x) + \theta_2(y), \quad M := \begin{pmatrix} 0 & -A^T \\ A & 0 \end{pmatrix}, \quad \text{and} \quad \Omega := \mathcal{X} \times \mathcal{Y}. \tag{2.2b}$$

Note that the matrix M given in (2.2b) is skew-symmetric. We denote by Ω^* the solution set of $\text{VI}(\Omega, \theta, M)$. Then it is nonempty under our nonempty assumption on the solution set of (1.1).

Let $\partial\theta_1(x)$ and $\partial\theta_2(y)$ be the subdifferentials of $\theta_1(x)$ and $\theta_2(y)$, respectively. Then, $\text{VI}(\Omega, \theta, M)$ can be alternatively expressed as finding $u^* \in \Omega$, $\xi^* \in \partial\theta_1(x^*)$, and $\eta^* \in \partial\theta_2(y^*)$ such that

$$\text{VI}(\Omega, F): \quad (u - u^*)^T F(u^*) \geq 0 \quad \forall\, u \in \Omega, \tag{2.3a}$$

where

$$F(u) := \begin{pmatrix} \xi - A^T y \\ \eta + Ax \end{pmatrix} \quad \text{with} \ \ \xi \in \partial\theta_1(x) \ \ \text{and} \ \ \eta \in \partial\theta_2(y). \tag{2.3b}$$

Because of the convexity of $\theta_1(x)$ and $\theta_2(y)$, the mapping F given in (2.3b) is monotone.

3. The nonconvergence of PDHG—An illustrative example.

In this section, we show by a special example of the abstract model (1.1) that PDHG (1.2) is not necessarily convergent even when r and s are fixed as very large values (equally, the step sizes $1/r$ and $1/s$ are extremely small).

Let us consider the linear programming in $\Re$:

$$\begin{array}{rl} \min & x \\ \text{subject to (s.t.)} & x \geq 1, \\ & x \geq 0. \end{array} \tag{3.1}$$

The dual problem of (3.1) is

$$\begin{array}{rl} \max & y \\ \text{s.t.} & y \leq 1, \\ & y \geq 0. \end{array} \tag{3.2}$$

Obviously, the unique optimal solutions of (3.1) and (3.2) are $x^* = 1$ and $y^* = 1$, respectively. The Lagrange function of (3.1) is

$$L(x, y) = x - y(x - 1), \tag{3.3}$$

which is defined on $\Re_+ \times \Re_+$, and $(x^*, y^*) = (1, 1)$ is the unique saddle point of the Lagrange function.

Finding the saddle point of $L(x, y)$ is obviously a special case of the model (1.1) with $\theta_1(x) = x$, $\theta_2(y) = -y$, $A = 1$, $\mathcal{X} = \mathcal{Y} = \Re_+$. If we apply PDHG (1.2) with $r = s = 1$, the scheme (1.2) reduces to

$$\begin{cases} x^{k+1} = \arg\min\{L(x, y^k) + \frac{r}{2}\|x - x^k\|^2 \,|\, x \geq 0\} = \max\{(x^k + \frac{1}{r}(y^k - 1)), 0\}, \\ y^{k+1} = \arg\max\{L(x^{k+1}, y) - \frac{s}{2}\|y - y^k\|^2 \,|\, y \geq 0\} = \max\{(y^k - \frac{1}{s}(x^{k+1} - 1)), 0\}. \end{cases} \tag{3.4}$$

Then, starting with $(x^0, y^0) = (0, 0)$, the sequence $\{(x^k, y^k)\}$ generated by (3.4) is

$$\begin{pmatrix} x^0 \\ y^0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} x^1 \\ y^1 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \begin{pmatrix} x^2 \\ y^2 \end{pmatrix} = \begin{pmatrix} 0 \\ 2 \end{pmatrix}, \quad \cdots, \quad \begin{pmatrix} x^7 \\ y^7 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} x^1 \\ y^1 \end{pmatrix}, \quad \cdots.$$

This is actually a cyclic sequence with six distinct iterates:

$$\begin{pmatrix} x^{k+6} \\ y^{k+6} \end{pmatrix} = \begin{pmatrix} x^k \\ y^k \end{pmatrix} \quad \forall\, k \geq 1.$$

Hence, the sequence $\{(x^k, y^k)\}$ generated by PDHG (1.2) with $r = s = 1$ does not converge to the solution point $(1, 1)$ as depicted in Figure 1.

In fact, even when r and s are as large as 5000, the distance between any iterate generated by PDHG (1.2) and the saddle point $(1, 1)$ is still at least 0.5 after 1000 iterations. We have additionally tested the cases where $r = s = 10^i$ with $i = 1, 2, 3, 4$, and all cases fail to converge to the solution point after 1000 iterations. This is sufficient to illustrate the nonconvergence of PDHG (1.2) even with extremely tiny constant step sizes. This example also shows the difficulty of proving the convergence of PDHG with constant step sizes without further assumptions on the model (1.1).

For comparison purposes, we also show the result when the modified PDHG scheme proposed in [7] is applied to finding the saddle point of $L(x, y)$. For solving (1.1), the modified PDHG scheme in [7] is

$$\begin{cases} x^{k+1} = \arg\min\{\Phi(x, y^k) + \frac{r}{2}\|x - x^k\|^2 \,|\, x \in \mathcal{X}\}, \\ y^{k+1} = \arg\max\{\Phi(x^{k+1} + \alpha(x^{k+1} - x^k), y) - \frac{s}{2}\|y - y^k\|^2 \,|\, y \in \mathcal{Y}\}, \end{cases} \tag{3.5}$$

in which $\alpha \in [0, 1]$ is a combination parameter. In [13], the range of α was extended to $[-1, 1]$ under the condition that the output of (3.5) should be further corrected. Let us just focus on (3.5) with $\alpha = 1$. For this case, if we also take $r = s = 1$, the scheme (3.5) for finding the saddle point of $L(x, y)$ reads as

$$\begin{cases} x^{k+1} = \max\{(x^k + y^k - 1), 0\}, \\ y^{k+1} = \max\{[y^k - (2x^{k+1} - x^k) + 1], 0\}. \end{cases}$$

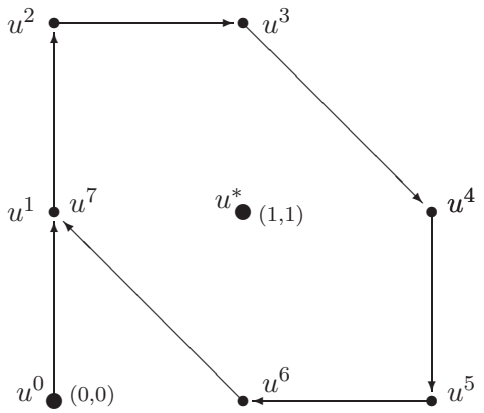

Figure 1. *The divergence of PDHG* (1.2) *with* $r = s = 1$.

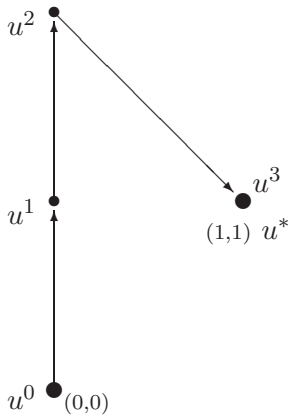

Figure 2. *The convergence of PDHG* (3.5) *with* $\alpha = 1$, $r = s = 1$.

If we take the initial iterate as $(x^0, y^0) = (0, 0)$, then we have

$$\begin{pmatrix} x^0 \\ y^0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} x^1 \\ y^1 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \begin{pmatrix} x^2 \\ y^2 \end{pmatrix} = \begin{pmatrix} 0 \\ 2 \end{pmatrix}, \quad \begin{pmatrix} x^3 \\ y^3 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

which implies

$$\begin{pmatrix} x^k \\ y^k \end{pmatrix} = \begin{pmatrix} x^* \\ y^* \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad \forall\, k \geq 3.$$

Therefore, after three iterations, we obtain the solution point $(1, 1)$ as depicted in Figure 2.

4. The convergence of PDHG. In this section, we show that PDHG with constant step sizes is convergent if one function in (1.1) is strongly convex. We also illustrate that the strong convexity assumption is satisfied by some popular applications of the model (1.1).

4.1. Additional assumptions. Let us first present the additional assumptions needed to ensure the convergence of PDHG with constant step sizes.

- $\theta_1(x)$ is strongly convex with the modulus $\tau > 0$. That is, there is a positive constant τ such that for any $\xi \in \partial\theta_1(x)$, we have

$$\theta_1(\tilde{x}) - \theta_1(x) \geq (\tilde{x} - x)^T \xi + \frac{\tau}{2}\|\tilde{x} - x\|^2 \quad \forall\, x, \tilde{x} \in \mathcal{X}. \tag{4.1a}$$

- For the given matrix A and the scalar $\tau > 0$, the parameter s in PDHG (1.2) satisfies

$$s > \frac{\rho(A^T A)}{\tau}, \tag{4.1b}$$

  where $\rho(\cdot)$ is the spectral radius of a matrix.

We will show that with the strong convexity assumption of θ_1, PDHG (1.2) is convergent with any positive constants r and s as long as s is greater than a fixed number. Neither of them needs to be adjusted iteratively and asymptotically tend to 0 or infinity. Clearly, similar conditions can be assumed on θ_2 and the parameter r because of the symmetry in the scheme (1.2). We omit the details for the sake of succinctness.

In fact, many concrete cases of (1.1) satisfy the assumption (4.1a). For example, consider the application of the linearized Bregman scheme in [12] to some sparse or low-rank optimization models (e.g., [3, 4, 5]), and various splitting versions of the augmented Lagrangian method in [14, 20] to some convex minimization models with linear constraints (e.g., [3, 15, 23]).

To show an application satisfying assumption (4.1a), let us take the standard TV-denoising model

$$\min \|\nabla x\|_1 + \frac{\mu}{2}\|x - f\|^2, \tag{4.2}$$

where $x \in \Re^n$ is the vector representation of a digital image, f is an observed image corrupted by Gaussian noise, ∇ is a discrete gradient operator (see, e.g., [21]) and thus $\nabla u \in \Re^{n\times 2}$, $\|\cdot\|_1$ and $\|\cdot\|$ are the standard l_1- and l_2-norm, respectively, and $\mu > 0$ is a trade-off parameter. Note that the l_2-term in the objective function of (4.2) reflects the data fidelity, and the l_1-term therein reflects the TV of pixel values and thus represents the sparsity of ∇x. Let

$$C_\infty := \left\{ y \in \Re^{n\times 2} \,|\, y_{1,i}^2 + y_{2,i}^2 \leq 1 \right\}.$$

As shown in many works such as [6, 9, 24, 25], we have

$$\|\nabla x\| = \max_{\|y\|_\infty \leq 1} \{y^T \nabla x\} = \max_{y \in C_\infty} \{y^T \nabla x\}.$$

Thus, the TV-denoising model (4.2) can be reformulated as

$$\min_{x\in\Re^n} \max_{y\in C_\infty} \left\{ \frac{\mu}{2}\|x - f\|^2 + y^T \nabla x \right\},$$

which is a special case of (1.1) with $\theta_1(x) = \frac{\mu}{2}\|x - f\|^2$, $A = -\nabla$, $\theta_2(y) = 0$, $X = \Re^n$, and $\mathcal{Y} = C_\infty$. It is clear that the assumption (4.1a) is satisfied with $\tau = \mu$. Moreover, it is known that $\|A^T A\| \leq 8$ (see [6]). Therefore, to apply PDHG (1.2) to the TV-denoising model (4.2), any positive r and $s > 8/\mu$ can ensure the convergence.

4.2. The contraction property. The goal of this section is to show that the sequence generated by PDHG (1.2) with constant step sizes is contractive with respect to the solution set Ω^* under the assumptions in (4.1). Based on this conclusion, we can prove its convergence and worst-case convergence rate measured by the iteration complexity.

The analysis is presented in the VI context. In the following three lemmas, we first prove some important inequalities for the sequence generated by PDHG (1.2).

Lemma 4.1. *For given $u^k = (x^k, y^k)$, let u^{k+1} be generated by PDHG (1.2). Let $\theta(u)$, M, and Ω be given in (2.2b). Then we have*

$$u^{k+1} \in \Omega, \quad (u^{k+1} - u)^T Q(u^k - u^{k+1}) \geq \theta(u^{k+1}) - \theta(u) + (u^{k+1} - u)^T M u^{k+1} \quad \forall\, u \in \Omega, \tag{4.3}$$

where

$$Q = \begin{pmatrix} rI_n & A^T \\ 0 & sI_m \end{pmatrix}. \tag{4.4}$$

Proof. It follows from the first-order optimality conditions of the subproblems in (1.2) that

$$x^{k+1} \in \mathcal{X}, \quad \theta_1(x) - \theta_1(x^{k+1}) + (x - x^{k+1})^T \{-A^T y^k + r(x^{k+1} - x^k)\} \geq 0 \quad \forall\, x \in \mathcal{X},$$

and

$$y^{k+1} \in \mathcal{Y}, \quad \theta_2(y) - \theta_2(y^{k+1}) + (y - y^{k+1})^T \{Ax^{k+1} + s(y^{k+1} - y^k)\} \geq 0 \quad \forall\, y \in \mathcal{Y}.$$

Combining the above two inequalities, we have $u^{k+1} = (x^{k+1}, y^{k+1}) \in \Omega$ and

$$\theta(u) - \theta(u^{k+1}) + \begin{pmatrix} x - x^{k+1} \\ y - y^{k+1} \end{pmatrix}^T \left\{ \begin{pmatrix} -A^T y^{k+1} \\ Ax^{k+1} \end{pmatrix} + \begin{pmatrix} r(x^{k+1} - x^k) + A^T(y^{k+1} - y^k) \\ s(y^{k+1} - y^k) \end{pmatrix} \right\} \geq 0 \ \forall\, u \in \Omega.$$

Using the notation of the matrices M and Q (see (2.2b) and (4.4)), the last inequality can be written as

$$u^{k+1} \in \Omega, \quad \theta(u) - \theta(u^{k+1}) + (u - u^{k+1})^T M u^{k+1} \geq (u - u^{k+1})^T Q(u^k - u^{k+1}) \quad \forall\, u \in \Omega. \tag{4.5}$$

The assertion (4.3) follows from the above inequality immediately. ∎

Lemma 4.2. *For given $u^k = (x^k, y^k)$, let u^{k+1} be generated by PDHG (1.2). If the assumptions in (4.1) are satisfied, then we have*

$$(u^{k+1} - u)^T D(u^k - u^{k+1}) \geq (u^{k+1} - u)^T F(u) - \frac{1}{2\tau}\|A^T(y^k - y^{k+1})\|^2 \quad \forall\, u \in \Omega, \tag{4.6}$$

where

$$D = \begin{pmatrix} rI_n & 0 \\ 0 & sI_m \end{pmatrix} \tag{4.7}$$

and $F(u)$ is defined as in (2.3b).

Proof. First, we treat the right-hand side of (4.3). Recall the assumption (4.1a) and the convexity of $\theta_2(y)$. Then, for any $\xi \in \partial\theta_1(x)$ and $\eta \in \partial\theta_2(y)$, we have

$$\theta_1(x^{k+1}) - \theta_1(x) \geq \frac{\tau}{2}\|x^{k+1} - x\|^2 + (x^{k+1} - x)^T\xi$$

and

$$\theta_2(y^{k+1}) - \theta_2(y) \geq (y^{k+1} - y)^T\eta.$$

That is, for any $\xi \in \partial\theta_1(x)$ and $\eta \in \partial\theta_2(y)$, we have

$$\theta(u^{k+1}) - \theta(u) \geq \begin{pmatrix} x^{k+1} - x \\ y^{k+1} - y \end{pmatrix}^T \begin{pmatrix} \xi \\ \eta \end{pmatrix} + \frac{\tau}{2}\|x^{k+1} - x\|^2. \tag{4.8}$$

Recall that the matrix M is skew-symmetric. We thus have

$$(u^{k+1} - u)^T M u^{k+1} = (u^{k+1} - u)^T M u. \tag{4.9}$$

Now, adding (4.8) and (4.9), and using the notation of $F(u)$ (see (2.3b)), we obtain

$$\theta(u^{k+1}) - \theta(u) + (u^{k+1} - u)^T M u^{k+1} \geq (u^{k+1} - u)^T F(u) + \frac{\tau}{2}\|x^{k+1} - x\|^2.$$

Substituting this into the right-hand side of (4.3) yields that

$$(u^{k+1} - u)^T Q(u^k - u^{k+1}) \geq (u^{k+1} - u)^T F(u) + \frac{\tau}{2}\|x^{k+1} - x\|^2 \quad \forall\, u \in \Omega.$$

Using the definitions of D and Q, it follows from the last inequality that

$$(u^{k+1}-u)^T D(u^k-u^{k+1}) \geq (u^{k+1}-u)^T F(u)+\frac{\tau}{2}\|x^{k+1}-x\|^2-(Ax^{k+1}-Ax)^T(y^k-y^{k+1}). \tag{4.10}$$

Furthermore, using the Cauchy–Schwarz inequality, we have

$$-(Ax^{k+1} - Ax)^T(y^k - y^{k+1}) \geq -\frac{\tau}{2}\|x^{k+1} - x\|^2 - \frac{1}{2\tau}\|A^T(y^k - y^{k+1})\|^2.$$

Substituting this into the right-hand side of (4.10), we obtain

$$(u^{k+1} - u)^T D(u^k - u^{k+1}) \geq (u^{k+1} - u)^T F(u) - \frac{1}{2\tau}\|A^T(y^k - y^{k+1})\|^2.$$

The assertion (4.6) is proved. ∎

Lemma 4.3. *For given $u^k = (x^k, y^k)$, let u^{k+1} be generated by PDHG (1.2). If the assumptions in (4.1) are satisfied, then we have*

$$(u - u^{k+1})^T F(u) \geq \frac{1}{2}\big(\|u - u^{k+1}\|_D^2 - \|u - u^k\|_D^2\big) + \frac{1}{2}\|u^k - u^{k+1}\|_G^2 \quad \forall\, u \in \Omega, \tag{4.11}$$

where the matrix D is defined by (4.7) and

$$G = \begin{pmatrix} rI_n & 0 \\ 0 & \left(s - \frac{\rho(A^TA)}{\tau}\right)I_m \end{pmatrix}. \tag{4.12}$$

Proof. First, it follows from (4.6) that

$$(u-u^{k+1})^T F(u) \geq (u-u^{k+1})^T D(u^k-u^{k+1}) - \frac{\rho(A^TA)}{2\tau}\|y^k-y^{k+1}\|^2 \quad \forall\, u\in\Omega. \tag{4.13}$$

Applying the identity

$$b^T D(b-a) = \frac{1}{2}\big(\|b\|_D^2 - \|a\|_D^2 + \|a-b\|_D^2\big)$$

to the term $(u-u^{k+1})^T D(u^k-u^{k+1})$ on the right-hand side (4.13) with

$$a = u-u^k \qquad \text{and} \qquad b = u-u^{k+1},$$

we thus obtain

$$(u-u^{k+1})^T F(u) \geq \frac{1}{2}\big(\|u-u^{k+1}\|_D^2 - \|u-u^k\|_D^2\big) + \frac{1}{2}\|u^k-u^{k+1}\|_D^2 - \frac{\rho(A^TA)}{2\tau}\|y^k-y^{k+1}\|^2. \tag{4.14}$$

Using the notation of the matrix G, from the last two terms on the right-hand side of (4.14), we have that

$$\begin{aligned}
&\|u^k-u^{k+1}\|_D^2 - \frac{\rho(A^TA)}{\tau}\|y^k-y^{k+1}\|^2 \\
&= r\|x^k-x^{k+1}\|^2 + \left(s-\frac{\rho(A^TA)}{\tau}\right)\|y^k-y^{k+1}\|^2 \\
&= \|u^k-u^{k+1}\|_G^2.
\end{aligned}$$

Substituting this in (4.14), we prove the assertion of this lemma. ∎

Based on the proved lemmas, it is easy to show that the sequence generated by PDHG (1.2) with constant step sizes is contractive with respect to the solution set Ω^*. We summarize the contraction property in the following theorem.

Theorem 4.4. *For given $u^k = (x^k, y^k)$, let u^{k+1} be generated by PDHG (1.2). If the assumptions in (4.1) are satisfied, then we have*

$$\|u^{k+1}-u^*\|_D^2 \leq \|u^k-u^*\|_D^2 - \|u^k-u^{k+1}\|_G^2 \quad \forall\, u^*\in\Omega^*, \tag{4.15}$$

where the matrices D and G are defined as in (4.7) and (4.12), respectively.

Proof. First, it follows from (4.11) that

$$\|u-u^k\|_D^2 - \|u-u^{k+1}\|_D^2 - \|u^k-u^{k+1}\|_G^2 \geq 2(u^{k+1}-u)^T F(u) \quad \forall\, u\in\Omega.$$

Setting $u = u^*$ in the last inequality and using the fact $(u^{k+1}-u^*)^T F(u^*) \geq 0$, we thus obtain

$$\|u^k-u^*\|_D^2 - \|u^{k+1}-u^*\|_D^2 - \|u^k-u^{k+1}\|_G^2 \geq 0 \quad \forall\, u^*\in\Omega^*.$$

The assertion (4.15) is proved. ∎

4.3. Convergence. The convergence of PDHG (1.2) with constant step sizes can be easily derived based on the conclusion in Theorem 4.4.

Theorem 4.5. *For given $u^k = (x^k, y^k)$, let u^{k+1} be generated by PDHG (1.2). If the assumptions in (4.1) are satisfied, then the sequence $\{u^k\}$ converges to some u^∞ which belongs to Ω^*.*

Proof. According to (4.15), it holds that $\{u^k\}$ is bounded and

$$\lim_{k\to\infty} \|u^k - u^{k+1}\| = 0. \tag{4.16}$$

Thus, $\{u^{k+1}\}$ is also bounded. Let u^∞ be a cluster point of $\{u^{k+1}\}$, and let $\{\tilde{u}^{k_j}\}$ be the subsequence converging to u^∞. Since the matrix Q is not singular, it follows from (4.5) and the continuity of $\theta(u)$ that

$$u^\infty \in \Omega, \quad \theta(u) - \theta(u^\infty) + (u - u^\infty)^T M u^\infty \geq 0 \quad \forall\, u \in \Omega.$$

The above VI indicates that u^∞ is a solution point of $\mathrm{VI}(\Omega, \theta, M)$. By using (4.16) and $\lim_{j\to\infty} u^{k_j} = u^\infty$, the subsequence $\{u^{k_j}\}$ converges to u^∞. Due to (4.15), we have

$$\|u^{k+1} - u^\infty\|_D \leq \|u^k - u^\infty\|_D,$$

and thus $\{u^k\}$ converges to u^∞. The proof is complete. ∎

4.4. Convergence rate. In this subsection, we establish a worst-case $O(1/k)$ convergence rate measured by the iteration complexity in the ergodic sense for PDHG (1.2) under the assumptions in (4.1). The assertion in Lemma 4.3 will be used.

First, we need a characterization of a solution point of $\mathrm{VI}(\Omega, F)$.

Theorem 4.6. *The solution set of $\mathrm{VI}(\Omega, F)$ is convex and can be characterized as*

$$\Omega^* := \bigcap_{u\in\Omega} \{\tilde{u} \in \Omega : (u - \tilde{u})^T F(u) \geq 0\}. \tag{4.17}$$

Proof. See Theorem 2.3.5 in [10] or Theorem 2.1 in [13]. ∎

Theorem 4.6 thus implies that $\tilde{u} \in \Omega$ is an approximate solution of $\mathrm{VI}(\Omega, F)$ with an accuracy of ϵ if it satisfies

$$(u - \tilde{u})^T F(u) \geq -\epsilon \quad \forall\, u \in \mathcal{D}_\Omega(\tilde{u}), \tag{4.18}$$

where

$$\mathcal{D}_\Omega(\tilde{u}) = \{u \in \Omega \,|\, \|u - \tilde{u}\| \leq 1\}.$$

For a similar definition of the ϵ-approximate solution, we refer the reader to [18].

In the next theorem, we show an inequality from which a worst-case $O(1/k)$ convergence rate for PDHG (1.2) in the ergodic sense can be derived.

Theorem 4.7. *Let $\{u^k\}$ be the sequence generated by PDHG (1.2) under the assumptions in (4.1). For any integer $t > 0$, let*

$$u_t = \frac{1}{t+1} \sum_{k=0}^{t} u^{k+1}. \tag{4.19}$$

Then, we have

$$(u_t - u)^T F(u) \le \frac{1}{2(t+1)} \|u - u^0\|_D^2 \quad \forall\, u \in \Omega. \tag{4.20}$$

Proof. First, it follows from (4.11) that

$$(u - u^{k+1})^T F(u) \ge \frac{1}{2}\big(\|u - u^{k+1}\|_D^2 - \|u - u^k\|_D^2\big) \quad \forall\, u \in \Omega,$$

and thus

$$(u^{k+1} - u)^T F(u) + \frac{1}{2}\|u - u^{k+1}\|_D^2 \le \frac{1}{2}\|u - u^k\|_D^2 \quad \forall\, u \in \Omega.$$

Summing the above inequality over $k = 0, 1, \ldots, t$, we obtain

$$\left(\sum_{k=0}^{t} u^{k+1} - (t+1)u\right)^T F(u) + \frac{1}{2}\|u - u^{t+1}\|_D^2 \le \frac{1}{2}\|u - u^0|_D^2 \quad \forall\, u \in \Omega.$$

Because $u^k \in \Omega$, we have $u_t \in \Omega$, and the assertion (4.20) follows directly from the above inequality. ∎

It follows from the conclusion (4.20) and the definition (4.18) that after k iterations, the average of all k iterations, i.e., u_t given by (4.19), is an approximate solution of VI(Ω, F) with an accuracy of $O(1/k)$. Hence, a worst-case $O(1/k)$ convergence rate in the ergodic sense is established for PDHG (1.2) under the assumptions in (4.1).

5. Conclusions.

We revisited the primal-dual hybrid gradient (PDHG) algorithm in the context of a saddle-point problem and discussed its convergence. It was shown by an example that PDHG is not necessarily convergent even when its step sizes are fixed as extremely small constants. Then, we showed that PDHG with constant step sizes is convergent if one function in the saddle-point problem is strongly convex. Under this strong convexity assumption, we also derived a worst-case convergence rate measured by the iteration complexity in the ergodic sense for PDHG with constant step sizes. It would be interesting to know whether there is any worst-case convergence rate in a nonergodic sense for PDHG with constant step sizes under the strong convexity assumption on the model (1.1), which is in general stronger than the ergodic convergence rate derived in this paper. We leave this as a topic for future work.

REFERENCES

[1] K. J. Arrow, L. Hurwicz, and H. Uzawa, *Studies in Linear and Non-linear Programming*, Stanford Mathematical Studies in the Social Sciences II, Stanford University Press, Stanford, CA, 1958.

[2] S. Bonettini and V. Ruggiero, *On the convergence of primal-dual hybrid gradient algorithms for total variation image restoration*, J. Math. Imaging Vision, 44 (2012), pp. 236–253.

[3] J.-F. Cai, E. J. Candès, and Z. Shen, *A singular value thresholding algorithm for matrix completion*, SIAM J. Optim., 20 (2010), pp. 1956–1982.

[4] J.-F. Cai, S. Osher, and Z. Shen, *Linearized Bregman iterations for compressed sensing*, Math. Comp., 78 (2009), pp. 1515–1536.

[5] J.-F. Cai, S. Osher, and Z. Shen, *Linearized Bregman iterations for frame-based image deblurring*, SIAM J. Imaging Sci., 2 (2009), pp. 226–252.

[6] A. Chambolle, *An algorithm for total variation minimization and applications*, J. Math. Imaging Vision, 20 (2004), pp. 89–97.
[7] A. Chambolle and T. Pock, *A first-order primal-dual algorithm for convex problems with applications to imaging*, J. Math. Imaging Vision, 40 (2011), pp. 120–145.
[8] Y. Chen, G. Lan, and Y. Ouyang, *Optimal Primal-Dual Methods for a Class of Saddle Point Problems*, manuscript, 2013.
[9] E. Esser, X. Zhang, and T. F. Chan, *A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science*, SIAM J. Imaging Sci., 3 (2010), pp. 1015–1046.
[10] F. Facchinei and J.-S. Pang, *Finite-Dimensional Variational Inequalities and Complementarity Problems*, Vol. I, Springer Ser. Oper. Res., Springer-Verlag, New York, 2003.
[11] T. Goldstein, E. Esser, and R. Baraniuk, *Adaptive Primal-Dual Hybrid Gradient Methods for Saddle-Point Problems*, preprint, arXiv:1305.0546v1 [math.NA], 2013.
[12] T. Goldstein and S. Osher, *The split Bregman method for L1-regularized problems*, SIAM J. Imaging Sci., 2 (2009), pp. 323–343.
[13] B. He and X. Yuan, *Convergence analysis of primal-dual algorithms for a saddle-point problem: From contraction perspective*, SIAM J. Imaging Sci., 5 (2012), pp. 119–149.
[14] M. R. Hestenes, *Multiplier and gradient methods*, J. Optim. Theory Appl., 4 (1969), pp. 303–320.
[15] N. J. Higham, *Computing the nearest correlation matrix—A problem from finance*, IMA J. Numer. Anal., 22 (2002), pp. 329–343.
[16] A. S. Nemirovsky and D. B. Yudin, *Problem Complexity and Method Efficiency in Optimization*, Wiley-Intersci. Ser. Discrete Math., John Wiley & Sons, New York, 1983.
[17] Yu. E. Nesterov, *A method for solving the convex programming problem with convergence rate $O(1/k^2)$*, Dokl. Akad. Nauk SSSR, 269 (1983), pp. 543–547 (in Russian).
[18] Yu. E. Nesterov, *Gradient methods for minimizing composite functions*, Math. Program., 140 (2013), pp. 125–161.
[19] T. Pock and A. Chambolle, *Diagonal preconditioning for first order primal-dual algorithms in convex optimization*, in Proceedings of the IEEE International Conference on Computer Vision, 2011, pp. 1762–1769.
[20] M. J. D. Powell, *A method for nonlinear constraints in minimization problems*, in Optimization, R. Fletcher, ed., Academic Press, New York, 1969, pp. 283–298.
[21] L. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, Phys. D, 60 (1992), pp. 227–238.
[22] P. Weiss, L. Blanc-Féraud, and G. Aubert, *Efficient schemes for total variation minimization under constraints in image processing*, SIAM J. Sci. Comput., 31 (2009), pp. 2047–2080.
[23] H. Zhang, J.-F. Cai, L. Cheng, and J. Zhu, *Strongly convex programming for exact matrix completion and robust principal component analysis*, Inverse Probl. Imaging, 6 (2012), pp. 357–372.
[24] X. Zhang, M. Burger, and S. Osher, *A unified primal-dual algorithm framework based on Bregman iteration*, J. Sci. Comput., 46 (2010), pp. 20–46.
[25] M. Zhu and T. F. Chan, *An Efficient Primal-Dual Hybrid Gradient Algorithm for Total Variation Image Restoration*, CAM Report 08-34, UCLA, Los Angeles, CA, 2008.