A General Framework for a Class of First Order Primal-Dual Algorithms for TV Minimization

Ernie Esser Xiaoqun Zhang Tony Chan

August 13, 2009

Abstract

We generalize the primal-dual hybrid gradient (PDHG) algorithm proposed by Zhu and Chan in [51], draw connections to similar methods and discuss convergence of several special cases and modifications. In particular, we point out a convergence result for a modified version of PDHG that has a similarly good empirical convergence rate for total variation (TV) minimization problems. Its convergence follows from interpreting it as an inexact Uzawa method discussed in [49]. We also prove a convergence result for PDHG applied to TV denoising with some restrictions on the PDHG step size parameters. It is shown how to interpret this special case as a projected averaged gradient method applied to the dual functional. We discuss the range of parameters for which the inexact Uzawa method and the projected averaged gradient method can be shown to converge. We also present some numerical comparisons of these algorithms applied to TV denoising, TV deblurring and constrained l_1 minimization problems.

1 Introduction

Total variation minimization problems arise in many image processing and compressive sensing applications for regularizing inverse problems where one expects the recovered image or signal to be piecewise constant or have sparse gradient. However, a lack of differentiability makes minimizing TV regularized functionals computationally challenging, and so there is considerable interest in efficient algorithms, especially for large scale problems.

The PDHG method [51] starts with a saddle point formulation of the problem and proceeds by alternating proximal steps that alternately maximize and minimize a penalized form of the saddle function. PDHG can generally be applied to saddle point formulations of inverse problems that can be formulated as minimizing a convex fidelity term plus a convex regularizing term. However, its performance for problems like TV denoising is of special interest since it compares favorably with other popular methods like Chambolle's method [9] and split Bregman [24].

PDHG is an example of a first order method, meaning it only requires functional and gradient evaluations. Other examples of first order methods popular for TV minimization include gradient descent, Chambolle's method and split Bregman. Second order methods like the method of Chan, Golub and Mulet (CGM) [10] work by essentially applying Newton's method to an appropriate formulation of the Euler Lagrange equations and therefore also require information about the Hessian. These can be quadratically convergent and are useful for computing benchmark solutions of high accuracy. However, the cost per iteration is much higher, so for large scale problems or when high accuracy is not required, these are often less practical than the first order methods that have much lower cost per iteration.

PDHG is also an example of a primal-dual method. Each iteration updates both a primal and a dual variable. It is thus able to avoid some of the difficulties that arise when working only on the primal or dual side. For example, for TV minimization, gradient descent applied to the primal functional has trouble where the gradient of the solution is zero because the functional is not differentiable there. Chambolle's method is a method on the dual that is very effective for TV denoising, but doesn't easily extend to applications where the dual problem is more complicated, such as TV deblurring. Primal-dual algorithms can avoid to some extent these difficulties. Other examples include CGM [10], split Bregman [24], and more generally other Bregman iterative algorithms [48] and Lagrangian-based methods.

An adaptive time stepping scheme for PDHG was proposed in [51] and shown to outperform other popular TV denoising algorithms like Chambolle's method, CGM and split Bregman in many numerical experiments with a wide variety of stopping conditions. Aside from some special cases of the PDHG algorithm like gradient projection and subgradient descent, the theoretical convergence properties were not known.

In this paper we show that we can make a small modification to the PDHG algorithm, which has little effect on its performance, but that allows the modified algorithm to be interpreted as an inexact Uzawa method of the type analyzed in [49]. After initially preparing this paper it was brought to our attention that the specific modified PDHG algorithm applied here has been previously proposed by Pock, Cremers, Bischof and Chambolle [35] for minimizing the Mumford-Shah functional. They also prove convergence for a special class of saddle point problems. Here, in a more general setting, we apply the convergence analysis for the inexact Uzawa method in [49] to show the modified PDHG algorithm converges for a range of fixed parameters. While this is nearly as effective as fixed parameter versions of PDHG, well chosen adaptive step sizes are an improvement. With more restrictions on the step size parameters, we prove a convergence result for PDHG applied to TV denoising by interpreting it as a projected averaged gradient method on the dual.

We additionally show that the modified PDHG method can be extended in the same ways as PDHG was extended in [51] to apply to additional problems like TV deblurring, l_1 minimization and constrained minimization problems. For these extensions we point out the range of parameters for which the convergence theory from [49] is applicable. We gain some insight into why the method works by putting it in a general framework and comparing it to related algorithms.

The organization of this paper is as follows. In Sections 2 and 3 we review the main idea of the PDHG algorithm and details about its application to TV deblurring type problems. Then in Section 4, we discuss primal-dual formulations for a more general problem. We define a general version of PDHG and discuss in detail the framework in which it can be related to other similar algorithms. These connections are diagrammed in Figure 1. In Section 5 we show how to interpret PDHG applied to TV denoising as a projected averaged gradient method on the dual and present a convergence result for a special case. Then in Sections 6 and 7, we discuss the application of the modified PDHG algorithm to TV deblurring type problems as well as constrained TV and l_1 minimization problems. Section 8 presents numerical experiments for TV denoising, TV deblurring and constrained l_1 minimization, comparing the performance of the modified PDHG algorithm with other methods.

2 Background and Notation

The PDHG algorithm in a general setting is a method for solving problems of the form

$$\min_{u\in\mathbb{R}^m} J(Au) + H(u),$$

where J and H are closed proper convex functions and $A \in \mathbb{R}^{n\times m}$. Usually, $J(Au)$ will correspond to a regularizing term of the form $\|Au\|$, in which case PDHG works by using duality to rewrite it as the saddle point problem

$$\min_{u\in\mathbb{R}^m} \max_{\|p\|_*\leq 1} \langle p, Au\rangle + H(u)$$

and then alternating dual and primal steps of the form

$$p^{k+1} = \max_{\|p\|_*\leq 1} \langle p, Au^k\rangle - \frac{1}{2\delta_k}\|p - p^k\|_2^2$$
$$u^{k+1} = \min_{u\in\mathbb{R}^m} \langle p^{k+1}, Au\rangle + H(u) + \frac{1}{2\alpha_k}\|u - u^k\|_2^2$$

for appropriate parameters α_k and δ_k. Here, $\|\cdot\|$ denotes an arbitrary norm on $\mathbb{R}^m$ and $\|\cdot\|_*$ denotes its dual norm defined by

$$\|x\|_* = \max_{\|y\|\leq 1} \langle x, y\rangle,$$

where $\langle\cdot,\cdot\rangle$ is the standard Euclidean inner product. Formulating the saddle point problem also uses the fact that $\|\cdot\|_{**} = \|\cdot\|$ [25], from which it follows that $\|Au\| = \max_{\|p\|_*\leq 1}\langle p, Au\rangle$.

The applications considered here are to solve constrained and unconstrained TV and l_1 minimization problems. In particular, we mainly focus on unconstrained TV minimization problems of the form

$$\min_{u\in\mathbb{R}^m} \|u\|_{TV} + \frac{\lambda}{2}\|Ku - f\|_2^2, \tag{1}$$

where $\|\cdot\|_{TV}$ denotes the discrete TV seminorm to be defined. If K is a linear blurring operator, this corresponds to a TV regularized deblurring model. It also includes the TV denoising case when $K = I$. These applications are analyzed in [51], which also mentions possible extensions such as to TV denoising with a constraint on the variance of u and also l_1 minimization.

The remainder of this section defines a discretization of the total variation seminorm and in particular defines a norm, $\|\cdot\|_E$, and a matrix, D, such that $\|u\|_{TV} = \|Du\|_E$. Therefore $\|u\|_{TV}$ is of the form $J(Au)$ with $J = \|\cdot\|_E$ and $A = D$. The details are included for completeness.

Define the discretized version of the total variation seminorm by

$$\|u\|_{TV} = \sum_{p=1}^{M_r}\sum_{q=1}^{M_c} \sqrt{(D_1^+ u_{p,q})^2 + (D_2^+ u_{p,q})^2} \tag{2}$$

for $u \in \mathbb{R}^{M_r\times M_c}$. Here, D_k^+ represents a forward difference in the k^{th} index and we assume Neumann boundary conditions. It will be useful to instead work with vectorized $u \in \mathbb{R}^{M_r M_c}$ and to rewrite $\|u\|_{TV}$. The convention for vectorizing an M_r by M_c matrix will be to associate the (p, q) element of the matrix with the $(q-1)M_r + p$ element of the vector. Consider a graph $G(\mathcal{E}, \mathcal{V})$ defined by an M_r by M_c grid with $\mathcal{V} = \{1, ..., M_r M_c\}$ the set of $m = M_r M_c$ nodes and $\mathcal{E}$ the set of

$e = 2M_rM_c - M_r - M_c$ edges. Assume the nodes are indexed so that the node corresponding to element (p, q) is indexed by $(q-1)M_r + p$. The edges, which will correspond to forward differences, can be indexed arbitrarily. Define $D \in \mathbb{R}^{e\times m}$ to be the edge-node adjacency matrix for this graph. So for a particular edge $\eta \in \mathcal{E}$ with endpoint indices $i, j \in \mathcal{V}$ and $i < j$, we have

$$D_{\eta,\nu} = \begin{cases} -1 & \text{for } \nu = i, \\ 1 & \text{for } \nu = j, \\ 0 & \text{for } \nu \neq i, j. \end{cases} \tag{3}$$

The matrix D is a discretization of the gradient and $-D^T$ is the corresponding discretization of the divergence.

Also define $E \in \mathbb{R}^{e\times m}$ such that

$$E_{\eta,\nu} = \begin{cases} 1 & \text{if } D_{\eta,\nu} = -1, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

The matrix E will be used to identify the edges used in each forward difference. Now define a norm on $\mathbb{R}^e$ by

$$\|w\|_E = \sum_{\nu=1}^{m} \left(\sqrt{E^T(w^2)} \right)_\nu . \tag{5}$$

Note that in this context, the square root and w^2 denote componentwise operations. Another way to interpret $\|w\|_E$ is as the sum of the l_2 norms of vectors w^ν, where $w^\nu = \begin{bmatrix} \vdots \\ w_e \\ \vdots \end{bmatrix}$ for e such that $E_{e,\nu} = 1$. Typically, away from the boundary, w^ν is of the form $w^\nu = \begin{bmatrix} w_{e_1^\nu} \\ w_{e_2^\nu} \end{bmatrix}$, where e_1^ν and e_2^ν are the edges used in the forward difference at node ν. So in terms of w^ν, $\|w\|_E = \sum_{\nu=1}^m \|w^\nu\|_2$. The discrete TV seminorm defined above (2) can be written in terms of $\|\cdot\|_E$ as

$$\|u\|_{TV} = \|Du\|_E.$$

Use of the matrix E is nonstandard, but also more general. For example, by redefining D, the same notation can apply to other discretizations of $\|u\|_{TV}$. Also, by adding edge weights the same notation easily extends to nonlocal TV.

By definition, the dual norm $\|\cdot\|_{E^*}$ to $\|\cdot\|_E$ is

$$\|x\|_{E^*} = \max_{\|y\|_E \leq 1} \langle x, y \rangle. \tag{6}$$

This dual norm arises in the saddle point formulation of (1) that the PDHG algorithm for TV deblurring is based on. If x^ν is defined analogously to w^ν, then

$$\|x\|_{E^*} = \max_\nu \|x^\nu\|_2.$$

To see this, note that by the Cauchy Schwarz inequality,

$$\max_{\|y\|_E \leq 1} \langle x, y \rangle = \max_{\sum_{\nu=1}^m \|y^\nu\|_2 \leq 1} \sum_{\nu=1}^{m} \langle x^\nu, y^\nu \rangle \leq \max_\nu \|x^\nu\|_2 = \|x^{\tilde{\nu}}\|_2 \text{ for some } \tilde{\nu}.$$

The the maximum is trivially attained if $\|x^{\tilde{\nu}}\|_2 = 0$ and otherwise the maximum is attained for y such that $y^\nu = \begin{cases} \frac{x^{\tilde{\nu}}}{\|x^{\tilde{\nu}}\|_2} & \text{if } \nu = \tilde{\nu} \\ 0 & \text{otherwise} \end{cases}$. In terms of the matrix E, the dual norm can be written as

$$\|x\|_{E^*} = \|\sqrt{E^T(x^2)}\|_\infty.$$

3 PDHG for TV Deblurring

In this section we review from [51] the application of PDHG to the TV deblurring and denoising problems, but using the present notation.

3.1 Saddle Point Formulations

For TV minimization problems, the saddle point formulation that the algorithm of Zhu and Chan in [51] is based on starts with the observation that

$$\|u\|_{TV} = \max_{p \in X} \langle p, Du \rangle, \tag{7}$$

where

$$X = \{p \in \mathbb{R}^e : \|p\|_{E^*} \leq 1\}. \tag{8}$$

The set X, which is the unit ball in the dual norm of $\|\cdot\|_E$, can also be interpreted as a Cartesian product of unit balls in the l_2 norm. For example, in order for Du to be in X, the discretized gradient $\begin{bmatrix} u_{p+1,q} - u_{p,q} \\ u_{p,q+1} - u_{p,q} \end{bmatrix}$ of u at each node (p, q) would have to have Euclidean norm less than or equal to 1. The dual norm interpretation is one way to explain (7) since

$$\max_{\|p\|_{E^*} \leq 1} \langle p, Du \rangle = \|Du\|_E,$$

which equals $\|u\|_{TV}$ by definition. Using duality to rewrite $\|u\|_{TV}$ is also the starting point for the primal-dual approach used by CGM [10] and a second order cone programming (SOCP) formulation used in [23]. Here it can be used to reformulate problem (1) as the min-max problem

$$\min_{u \in \mathbb{R}^m} \max_{p \in X} \Phi(u, p), \tag{9}$$

where

$$\Phi(u, p) = \langle p, Du \rangle + \frac{\lambda}{2}\|Ku - f\|_2^2.$$

Other saddle point formulations of (1) are possible. For example, another approach is to replace Du with a new variable w under the constraint that $w = Du$. One can then handle the constraint by forming the Lagrangian, incorporating a Lagrange multiplier p. This yields the following saddle point formulation:

$$\max_{p \in \mathbb{R}^e} \inf_{u \in \mathbb{R}^m, w \in \mathbb{R}^e} \|w\|_E + \frac{\lambda}{2}\|Ku - f\|_2^2 + \langle p, Du - w \rangle. \tag{10}$$

Methods such as the alternating direction method of multipliers (ADMM) [20, 22, 5, 14] and Split Bregman [24] are based on this saddle point formulation.

3.2 Existence of Saddle Point

One way to ensure that there exists a saddle point (u^*, p^*) of the convex-concave function Φ is to restrict u and p to be in bounded sets. Existence then follows from ([38] 37.6). The dual variable p is already required to lie in the set X. Assume that

$$\ker(D) \bigcap \ker(K) = \{0\}.$$

This is equivalent to assuming that $\ker(K)$ does not contain the vector of all ones, which is very reasonable for deblurring problems where K is an averaging operator. With this assumption, it follows that there exists $c \in \mathbb{R}$ such that the set

$$\left\{ u : \|Du\|_E + \frac{\lambda}{2}\|Ku - f\|_2^2 \leq c \right\}$$

is nonempty and bounded. Thus we can restrict u to a bounded set.

3.3 Optimality Conditions

If (u^*, p^*) is a saddle point of Φ, it follows that

$$\max_{p \in X} \langle p, Du^* \rangle + \frac{\lambda}{2}\|Ku^* - f\|_2^2 = \Phi(u^*, p^*) = \min_{u \in \mathbb{R}^m} \langle p^*, Du \rangle + \frac{\lambda}{2}\|Ku - f\|_2^2,$$

from which we can deduce the optimality conditions

$$D^T p^* + \lambda K^T (Ku^* - f) = 0 \tag{11}$$

$$p^* E \sqrt{E^T (Du^*)^2} = Du^* \tag{12}$$

$$p^* \in X. \tag{13}$$

The second optimality condition (12) with E defined by (4) can be understood as a discretization of

$$p^* |\nabla u^*| = \nabla u^*.$$

3.4 PDHG Algorithm

In [51] it is shown how to interpret the PDHG algorithm applied to (1) as a primal-dual proximal point method for solving (9) by iterating

$$p^{k+1} = \arg\max_{p \in X} \langle p, Du^k \rangle - \frac{1}{2\lambda\tau_k}\|p - p^k\|_2^2 \tag{14a}$$

$$u^{k+1} = \arg\min_{u \in \mathbb{R}^m} \langle p^{k+1}, Du \rangle + \frac{\lambda}{2}\|Ku - f\|_2^2 + \frac{\lambda(1-\theta_k)}{2\theta_k}\|u - u^k\|_2^2. \tag{14b}$$

The index k denotes the current iteration. Also, τ_k and θ_k are the dual and primal step sizes respectively. The above max and min problems can be explicitly solved, yielding

Algorithm: PDHG for TV Deblurring

$$p^{k+1} = \Pi_X \left(p^k + \tau_k \lambda D u^k \right) \tag{15a}$$

$$u^{k+1} = \left((1-\theta_k) + \theta_k K^T K \right)^{-1} \left((1-\theta_k) u^k + \theta_k (K^T f - \frac{1}{\lambda} D^T p^{k+1}) \right). \tag{15b}$$

Here, Π_X is the orthogonal projection onto X defined by

$$\Pi_X(q) = \arg\min_{p \in X} \|p - q\|_2^2 \tag{16}$$

$$= \frac{q}{E \max\left(\sqrt{E^T(q^2)}, 1 \right)}, \tag{17}$$

where the division and max are understood in a componentwise sense. For example, $\Pi_X(Du)$ can be thought of as a discretization of

$$\begin{cases} \frac{\nabla u}{|\nabla u|} & \text{if } |\nabla u| > 1 \\ \nabla u & \text{otherwise} \end{cases}.$$

In the denoising case where $K = I$, the p^{k+1} update remains the same and the u^{k+1} simplifies to

$$u^{k+1} = (1-\theta_k) u^k + \theta_k (f - \frac{1}{\lambda} D^T p^{k+1}).$$

For the initialization, we take $u^0 \in \mathbb{R}^m$ and $p^0 \in X$.

4 General Algorithm Framework

In this section we consider a more general class of problems that PDHG can be applied to. We define equivalent primal, dual and several primal-dual formulations. We also place PDHG in a general framework that connects it to other related alternating direction methods applied to saddle point problems.

4.1 Primal-Dual Formulations

PDHG can more generally be applied to what we will refer to as the primal problem

$$\min_{u \in \mathbb{R}^m} F_P(u), \tag{P}$$

where

$$F_P(u) = J(Au) + H(u), \tag{18}$$

$A \in \mathbb{R}^{n \times m}$, $J : \mathbb{R}^n \to (-\infty, \infty]$ and $H : \mathbb{R}^m \to (-\infty, \infty]$ are closed convex functions. Assume there exists a solution u^* to (P). We will pay special attention to the case where $J(Au) = \|Au\|$ for some norm $\|\cdot\|$, but this assumption is not required. $J(Au)$ reduces to $\|u\|_{TV}$ when $J = \|\cdot\|_E$ and

$A = D$ as defined in Section 2. Also in Section 2 when J was a norm, it was shown how to use the dual norm to define a saddle point formulation of (P) as

$$\min_{u\in\mathbb{R}^m} \max_{\|p\|_*\leq 1} \langle Au, p\rangle + H(u).$$

This can equivalently be written in terms of the Legendre-Fenchel transform, or convex conjugate, of J denoted by J^* and defined by

$$J^*(p) = \sup_{w\in\mathbb{R}^n} \langle p, w\rangle - J(w).$$

When J is a closed proper convex function, we have that $J^{**} = J$ [16]. Therefore,

$$J(Au) = \sup_{p\in\mathbb{R}^n} \langle p, Au\rangle - J^*(p).$$

So an equivalent saddle point formulation of (P) is

$$\min_{u\in\mathbb{R}^m} \sup_{p\in\mathbb{R}^n} L_{PD}(u, p), \tag{PD}$$

where

$$L_{PD} = \langle p, Au\rangle - J^*(p) + H(u).$$

This holds even when J is not a norm, but in the case when $J(w) = \|w\|$, we can then use the dual norm representation of $\|w\|$ to write

$$\begin{aligned} J^*(p) &= \sup_w \langle p, w\rangle - \max_{\|y\|_*\leq 1} \langle w, y\rangle \\ &= \begin{cases} 0 & \text{if } \|p\|_* \leq 1 \\ \infty & \text{otherwise} \end{cases}, \end{aligned}$$

in which case we can interpret J^* as the indicator function for the unit ball in the dual norm.

Let (u^*, p^*) be a saddle point of L_{PD}. In particular, this means

$$\max_{p\in\mathbb{R}^n} \langle p, Au^*\rangle - J^*(p) + H(u^*) = L_{PD}(u^*, p^*) = \min_{u\in\mathbb{R}^m} \langle p^*, Au\rangle + H(u),$$

from which we can deduce the equivalent optimality conditions

$$-A^T p^* \in \partial H(u^*) \tag{19}$$

$$Au^* \in \partial J^*(p^*), \tag{20}$$

where ∂ denotes the subdifferential. The subdifferential $\partial F(x)$ of a convex function $F : \mathbb{R}^m \to (-\infty, \infty]$ at the point x is defined by the set

$$\partial F(x) = \{q \in \mathbb{R}^m : F(y) \geq F(x) + \langle q, y - x\rangle \ \forall y \in \mathbb{R}^m\}.$$

Using the definitions of the Legendre transform and subdifferential, the optimality conditions (19) and (20) could also be be rewritten as

$$u^* \in \partial H^*(-A^T p^*) \tag{19a}$$

$$p^* \in \partial J(Au^*) \tag{20a}$$

Another useful saddle point formulation that we will refer to as the split primal problem is obtained by introducing the constraint $w = Au$ in (P) and forming the Lagrangian

$$L_P(u, w, p) = J(w) + H(u) + \langle p, Au - w \rangle. \tag{21}$$

The corresponding saddle point problem is

$$\max_{p \in \mathbb{R}^n} \inf_{u \in \mathbb{R}^m, w \in \mathbb{R}^n} L_P(u, w, p). \tag{$\mathrm{SP_P}$}$$

Although p was introduced in (21) as a Lagrange multiplier for the constraint $Au = w$, it has the same interpretation as the dual variable p in (PD). It follows immediately from the optimality conditions that if (u^*, w^*, p^*) is a saddle point for ($\mathrm{SP_P}$), then (u^*, p^*) is a saddle point for (PD).

The dual problem is

$$\max_{p \in \mathbb{R}^n} F_D(p), \tag{D}$$

where the dual functional $F_D(p)$ is a concave function defined by

$$F_D(p) = \inf_{u \in \mathbb{R}^m} L_{PD}(u, p) = \inf_{u \in \mathbb{R}^m} \langle p, Au \rangle - J^*(p) + H(u) = -J^*(p) - H^*(-A^T p). \tag{22}$$

Note that this is equivalent to defining the dual by

$$F_D(p) = \inf_{u \in \mathbb{R}^m, w \in \mathbb{R}^n} L_P(u, w, p). \tag{23}$$

Since we assumed there exists an optimal solution u^* to the convex problem (P), it follows from Fenchel Duality ([38] 31.2.1) that there exists an optimal solution p^* to (D) and $F_P(u^*) = F_D(p^*)$. Moreover, u^* solves (P) and p^* solves (D) if and only if (u^*, p^*) is a saddle point of $L_{PD}(u, p)$ ([38] 31.3).

By introducing the constraint $y = -A^T p$ in (D) and forming the corresponding Lagrangian

$$L_D(p, y, u) = J^*(p) + H^*(y) + \langle u, -A^T p - y \rangle, \tag{24}$$

we obtain yet another saddle point problem,

$$\max_{u \in \mathbb{R}^m} \inf_{p \in \mathbb{R}^n, y \in \mathbb{R}^m} L_D(p, y, u), \tag{$\mathrm{SP_D}$}$$

which we will refer to as the split dual problem. Although u was introduced in ($\mathrm{SP_D}$) as a Lagrange multiplier for the constraint $y = -A^T p$, it actually has the same interpretation as the primal variable u in (P). The optimality conditions for ($\mathrm{SP_D}$) are $Au^* \in \partial J^*(p^*)$, which agrees with (20), and $u^* \in \partial H^*(y^*)$, which is exactly (19a) since $y^* = -A^T p^*$. So if (p^*, y^*, u^*) is a saddle point for ($\mathrm{SP_D}$), then (u^*, p^*) is a saddle point for (PD). Note also that

$$F_P(u) = -\inf_{p \in \mathbb{R}^n, y \in \mathbb{R}^m} L_D(p, y, u).$$

4.2 Algorithm Framework and Connections to PDHG

In this section we define a general version of PDHG applied to (PD) and discuss connections to related algorithms that can be interpreted as alternating direction methods applied to ($\mathrm{SP_P}$) and ($\mathrm{SP_D}$). These connections are summarized in Figure 1.

A useful tool for drawing connections between the algorithms in this section is the Moreau decomposition [30, 12].

Theorem 4.1. *[12] If J is a closed proper convex function on $\mathbb{R}^m$ and $f \in \mathbb{R}^m$, then*

$$f = \arg\min_u J(u) + \frac{1}{2\alpha}\|u - f\|_2^2 + \alpha \arg\min_p J^*(p) + \frac{\alpha}{2}\|p - \frac{f}{\alpha}\|_2^2. \tag{25}$$

It was shown in [51] that PDHG applied to TV denoising can be interpreted as a primal-dual proximal point method applied to a saddle point formulation of the problem. More generally, applied to (PD) it yields

---

Algorithm: PDHG on (PD)

$$p^{k+1} = \arg\max_{p\in\mathbb{R}^n} -J^*(p) + \langle p, Au^k\rangle - \frac{1}{2\delta_k}\|p - p^k\|_2^2 \tag{26a}$$

$$u^{k+1} = \arg\min_{u\in\mathbb{R}^m} H(u) + \langle A^T p^{k+1}, u\rangle + \frac{1}{2\alpha_k}\|u - u^k\|_2^2, \tag{26b}$$

---

where $p^0 = 0$, u^0 is arbitrary, and $\alpha_k, \delta_k > 0$. The parameters τ_k and θ_k from (15) in terms of δ_k and α_k are

$$\theta_k = \frac{\lambda\alpha_k}{1 + \alpha_k\lambda} \qquad \tau_k = \frac{\delta_k}{\lambda}.$$

4.2.1 Proximal Forward Backward Splitting Special Cases of PDHG

Two notable special cases of PDHG are $\alpha_k = \infty$ and $\delta_k = \infty$. These special cases correspond to the proximal forward backward splitting method (PFBS) [28, 34, 12] applied to (D) and (P) respectively.

PFBS is an iterative splitting method that can be used to find a minimum of a sum of two convex functionals by alternating a (sub)gradient descent step with a proximal step. Applied to (D) it yields

$$p^{k+1} = \arg\min_{p\in\mathbb{R}^n} J^*(p) + \frac{1}{2\delta_k}\|p - (p^k + \delta_k A u^{k+1})\|_2^2, \tag{27}$$

where $u^{k+1} \in \partial H^*(-A^T p^k)$. Since $u^{k+1} \in \partial H^*(-A^T p^k) \Leftrightarrow -A^T p^k \in \partial H(u^{k+1})$, which is equivalent to

$$u^{k+1} = \arg\min_{u\in\mathbb{R}^m} H(u) + \langle A^T p^k, u\rangle,$$

(27) can be written as

Algorithm: PFBS on (D)

$$u^{k+1} = \arg\min_{u\in\mathbb{R}^m} H(u) + \langle A^T p^k, u\rangle \tag{28a}$$

$$p^{k+1} = \arg\min_{p\in\mathbb{R}^n} J^*(p) + \langle p, -Au^{k+1}\rangle + \frac{1}{2\delta_k}\|p - p^k\|_2^2. \tag{28b}$$

Even though the order of the updates is reversed relative to PDHG, since the initialization is arbitrary it is still a special case of (26) where $\alpha_k = \infty$.

If we assume that $J(\cdot) = \|\cdot\|$, we can interpret the p^{k+1} step as an orthogonal projection onto a convex set,

$$p^{k+1} = \Pi_{\{p:\|p\|_*\leq 1\}}\left(p^k + \delta_k A u^{k+1}\right).$$

Then PFBS applied to (D) can be interpreted as a (sub)gradient projection algorithm.

As a special case of ([12] Theorem 3.4), the following convergence result applies to (28).

Theorem 4.2. *Fix $p^0 \in \mathbb{R}^n$, $u^0 \in \mathbb{R}^m$ and let (u^k, p^k) be defined by (28). If H^* is differentiable, $\nabla(H^*(-A^T p))$ is Lipschitz continuous with Lipschitz constant equal to $\frac{1}{\beta}$, and $0 < \inf \delta_k \leq \sup \delta_k < 2\beta$, then $\{p^k\}$ converges to a solution of (D) and $\{u^k\}$ converges to the unique solution of (P).*

Proof. Convergence of $\{p^k\}$ to a solution of (D) follows from ([12] 3.4). From (28a), u^{k+1} satisfies $-A^T p^k \in \partial H(u^{k+1})$, which, from the definitions of the subdifferential and Legendre transform, implies that $u^{k+1} = \nabla H^*(-A^T p^k)$. So by continuity of ∇H^*, $u^k \to u^* = \nabla H^*(-A^T p^*)$. From (28b) and the convergence of $\{p^k\}$, $Au^* \in \partial J^*(p^*)$. Therefore (u^*, p^*) satisfies the optimality conditions (19,20) for (PD), which means u^* solves (P) ([38] 31.3). Uniqueness follows from the assumption that H^* is differentiable, which by ([38] 26.3) means that $H(u)$ in the primal functional is strictly convex. □

It will be shown later in Section 4.2.3 how to equate modified versions of the PDHG algorithm with convergent alternating direction methods, namely split inexact Uzawa methods from [49] applied to the split primal ($\mathrm{SP_P}$) and split dual ($\mathrm{SP_D}$) problems. The connection there is very similar to the equivalence from [44] between PFBS applied to (D) and what Tseng in [44] called the alternating minimization algorithm (AMA) applied to ($\mathrm{SP_P}$). AMA applied to ($\mathrm{SP_P}$) is an alternating direction method that alternately minimizes first the Lagrangian $L_P(u, w, p)$ with respect to u and then the augmented Lagrangian $L_P + \frac{\delta_k}{2}\|Au - w\|_2^2$ with respect to w before updating the Lagrange multiplier p.

Algorithm: AMA on ($\mathrm{SP_P}$)

$$u^{k+1} = \arg\min_{u\in\mathbb{R}^m} H(u) + \langle A^T p^k, u\rangle \tag{29a}$$

$$w^{k+1} = \arg\min_{w\in\mathbb{R}^n} J(w) - \langle p^k, w\rangle + \frac{\delta_k}{2}\|Au^{k+1} - w\|_2^2 \tag{29b}$$

$$p^{k+1} = p^k + \delta_k(Au^{k+1} - w^{k+1}) \tag{29c}$$

To see the equivalence between (28) and (29), first note that (29a) is identical to (28a), so it suffices to show that (29b) and (29c) are together equivalent to (28b). Combining (29b) and (29c) yields

$$p^{k+1} = (p^k + \delta_k A u^{k+1}) - \delta_k \arg\min_w J(w) + \frac{\delta_k}{2}\|w - \frac{(p^k + \delta_k A u^{k+1})}{\delta_k}\|_2^2.$$

By the Moreau decomposition (25), this is equivalent to

$$p^{k+1} = \arg\min_p J^*(p) + \frac{1}{2\delta_k}\|p - (p^k + \delta_k A u^{k+1})\|_2^2,$$

which is exactly (28b).

In [44], convergence of (u^k, w^k, p^k) satisfying (29) to a saddle point of $L_P(u, w, p)$ is directly proved under the assumption that H is strongly convex. H is strongly convex with modulus α if for $0 \leq \lambda \leq 1$,

$$\lambda H(u) + (1-\lambda)H(v) - H(\lambda u + (1-\lambda)v) \geq \alpha\lambda(1-\lambda)\|u - v\|_2^2 \qquad \forall u, v \in \mathbb{R}^m.$$

In fact, this assumption directly implies the condition on H^* in Theorem 4.2.

Theorem 4.3. *Suppose H is strongly convex with modulus $\frac{\beta\|A\|^2}{2} > 0$, where $\|A\|$ denotes the operator norm of A. Then H^* is differentiable and $\nabla(H^*(-A^T p))$ is Lipschitz continuous with Lipschitz constant equal to $\frac{1}{\beta}$.*

Proof. H^* is differentiable since H is strictly convex ([38] 26.3). Let p_1 and p_2 be arbitrary vectors in $\mathbb{R}^n$. Let $u_1 = \nabla H^*(-A^T p_1)$ and $u_2 = \nabla H^*(-A^T p_2)$, which means $-A^T p_1 \in \partial H(u_1)$ and $-A^T p_2 \in \partial H(u_2)$. From strong convexity, it follows that ∂H is strongly monotone with modulus $\beta\|A\|^2$, meaning

$$\langle u_2 - u_1, -A^T(p_2 - p_1)\rangle \geq \beta\|A\|^2\|u_2 - u_1\|_2^2,$$

it follows that

$$\begin{aligned} \|A(u_2 - u_1)\|_2^2 \leq \|A\|^2\|u_2 - u_1\|_2^2 &\leq \frac{1}{\beta}\langle u_2 - u_1, -A^T(p_2 - p_1)\rangle \\ &\leq \frac{1}{\beta}\|A(u_2 - u_1)\|_2\|p_2 - p_1\|_2 \\ \Rightarrow \|A\left(\nabla H^*(-A^T p_2) - \nabla H^*(-A^T p_1)\right)\|_2 &\leq \frac{1}{\beta}\|p_2 - p_1\|_2. \end{aligned}$$

□

The other special case of PDHG where $\delta_k = \infty$ can be analyzed in a similar manner. The corresponding algorithm is PFBS applied to (P),

Algorithm: PFBS on (P)

$$p^{k+1} = \arg\min_{p\in\mathbb{R}^n} J^*(p) + \langle -Au^k, p\rangle \tag{30a}$$

$$u^{k+1} = \arg\min_{u\in\mathbb{R}^m} H(u) + \langle u, A^T p^{k+1}\rangle + \frac{1}{2\alpha_k}\|u - u^k\|_2^2, \tag{30b}$$

which is analogously equivalent to AMA applied to ($\mathrm{SP_D}$).

Algorithm: AMA on ($\mathrm{SP_D}$)

$$p^{k+1} = \arg\min_{p\in\mathbb{R}^m} J^*(p) + \langle -Au^k, p\rangle \tag{31a}$$

$$y^{k+1} = \arg\min_{y\in\mathbb{R}^m} H^*(y) - \langle u^k, y\rangle + \frac{\alpha_k}{2}\|y + A^T p^{k+1}\|_2^2 \tag{31b}$$

$$u^{k+1} = u^k + \alpha_k(-A^T p^{k+1} - y^{k+1}) \tag{31c}$$

The equivalence follows in much the same way as before. The p^{k+1} update is already the same for both. By applying the Moreau decomposition again (25), it is possible to show that (31b) and (31c) together are equivalent to (30b). Also the analogous version of Theorem 4.2 applies to (30).

It's important to note that there are other ways to apply the algorithms described above. For example, when applying PFBS to (P), we could have applied the gradient step to $H(u)$ and the proximal step to $J(Au)$. This would have corresponded to swapping the roles of p and y in AMA applied to ($\mathrm{SP_D}$). There is a corresponding alternate version of AMA on ($\mathrm{SP_P}$). But these alternate versions aren't considered here because they aren't as closely connected to PDHG. In addition, those alternate versions involve more complicated minimization steps in the sense that variables are coupled by either the matrix A or A^T.

4.2.2 Reinterpretation of PDHG as Relaxed AMA

The general form of PDHG (26) can also be interpreted as alternating direction methods applied to ($\mathrm{SP_P}$) or ($\mathrm{SP_D}$). It differs from AMA only in that an additional proximal penalty is added to the step which minimizes the Lagrangian. This method will be referred to as relaxed AMA.

Algorithm: Relaxed AMA on ($\mathrm{SP_P}$)

$$u^{k+1} = \arg\min_{u\in\mathbb{R}^m} H(u) + \langle A^T p^k, u\rangle + \frac{1}{2\alpha_k}\|u - u^k\|_2^2 \tag{32a}$$

$$w^{k+1} = \arg\min_{w\in\mathbb{R}^n} J(w) - \langle p^k, w\rangle + \frac{\delta_k}{2}\|Au^{k+1} - w\|_2^2 \tag{32b}$$

$$p^{k+1} = p^k + \delta_k(Au^{k+1} - w^{k+1}) \tag{32c}$$

Algorithm: Relaxed AMA on ($\mathrm{SP_D}$)

$$p^{k+1} = \arg\min_{p\in\mathbb{R}^m} J^*(p) + \langle -Au^k, p\rangle + \frac{1}{2\delta_k}\|p - p^k\|_2^2 \tag{33a}$$

$$y^{k+1} = \arg\min_{y\in\mathbb{R}^m} H^*(y) - \langle u^k, y\rangle + \frac{\alpha_k}{2}\|y + A^T p^{k+1}\|_2^2 \tag{33b}$$

$$u^{k+1} = u^k + \alpha_k(-A^T p^{k+1} - y^{k+1}) \tag{33c}$$

The equivalence of these relaxed AMA algorithms to the general form of PDHG (26) follows by a similar argument as in Section 4.2.1.

Although equating PDHG to this relaxed AMA algorithm doesn't yield any direct convergence results for PDHG, it does show a close connection to the alternating direction method of multipliers (ADMM) [20, 22], which does have a well established convergence theory [14]. If, instead of adding proximal terms of the form $\frac{1}{2\alpha_k}\|u - u^k\|_2^2$ and $\frac{1}{2\delta_k}\|p - p^k\|_2^2$ to the first step of AMA applied to ($\mathrm{SP_P}$) and ($\mathrm{SP_D}$), we add the augmented Lagrangian penalties $\frac{\delta_k}{2}\|Au - w^k\|_2^2$ and $\frac{\alpha_k}{2}\|A^T p - y^k\|_2^2$, then we get exactly ADMM applied to ($\mathrm{SP_P}$) and ($\mathrm{SP_D}$) respectively.

Algorithm: ADMM on ($\mathrm{SP_P}$)

$$u^{k+1} = \arg\min_{u\in\mathbb{R}^m} H(u) + \langle A^T p^k, u\rangle + \frac{\delta_k}{2}\|Au - w^k\|_2^2 \tag{34a}$$

$$w^{k+1} = \arg\min_{w\in\mathbb{R}^n} J(w) - \langle p^k, w\rangle + \frac{\delta_k}{2}\|Au^{k+1} - w\|_2^2 \tag{34b}$$

$$p^{k+1} = p^k + \delta_k(Au^{k+1} - w^{k+1}) \tag{34c}$$

Algorithm: ADMM on ($\mathrm{SP_D}$)

$$p^{k+1} = \arg\min_{p\in\mathbb{R}^m} J^*(p) + \langle -Au^k, p\rangle + \frac{\alpha_k}{2}\|y^k + A^T p\|_2^2 \tag{35a}$$

$$y^{k+1} = \arg\min_{y\in\mathbb{R}^m} H^*(y) - \langle u^k, y\rangle + \frac{\alpha_k}{2}\|y + A^T p^{k+1}\|_2^2 \tag{35b}$$

$$u^{k+1} = u^k + \alpha_k(-A^T p^{k+1} - y^{k+1}) \tag{35c}$$

ADMM applied to ($\mathrm{SP_P}$) can be interpreted as Douglas Rachford splitting [13] applied to (D) and ADMM applied to ($\mathrm{SP_D}$) can be interpreted as Douglas Rachford splitting applied to (P) [19, 21, 15, 14]. It is also shown in [18, 41] how to interpret these as the split Bregman algorithm of [24]. A general convergence result for ADMM can be found in [14]. Assuming that we are most interested in finding a solution u^* to (P), when we apply ADMM to ($\mathrm{SP_P}$), we want to ensure

that (u^k, w^k, p^k) converges to a saddle point. One result from [18] that follows directly from the convergence analysis of Eckstein and Bertsekas in [14] is given by the next theorem. Recall we are assuming throughout that J and H are closed proper convex functions, and there exists a solution to (P).

Theorem 4.4. *[14, 18] Suppose $H(u) + \|Au\|_2^2$ is strictly convex and $\delta_k = \delta > 0$. Let p^0 and w^0 be arbitrary. Then (u^k, w^k, p^k) satisfying (34) converges to a saddle point of $L_P(u, w, p)$.*

On the other hand, when applying ADMM to ($\mathrm{SP_D}$), it isn't necessary to insist that (p^k, y^k, u^k) converge to a saddle point if we are only interested in the convergence of $\{u^k\}$ to a solution of (P). Instead we can make use of the convergence theory for the equivalent Douglas Rachford splitting method on (P). This method is used to find u such that

$$0 \in A^T \partial J(Au) + \partial H(u),$$

which is one way to solve (P). Formal application of the classical Douglas Rachford splitting method yields the iterations,

$$\begin{aligned} &\frac{\hat{u}^{k+1} - u^k}{\alpha_k} + A^T \partial J(A\hat{u}^{k+1}) + \partial H(u^k) \ni 0 \\ &\frac{u^{k+1} - u^k}{\alpha_k} + A^T \partial J(A\hat{u}^{k+1}) + \partial H(u^{k+1}) \ni 0, \end{aligned}$$

where δ is thought of as a time step. Although algorithm (35) can be interpreted as Douglas Rachford splitting applied to (P), there may be other ways to formally satisfy the Douglas Rachford iterations. An interesting way to arrive at the version that corresponds exactly to ADMM applied to ($\mathrm{SP_D}$) is to apply ADMM to yet another Lagrangian formulation of (P), namely

$$\max_y \inf_{v,u} L_{P_{DR}}(v, u, y) := J(Av) + H(u) + \langle y, v - u \rangle.$$

This also yields a more implementable way of writing the Douglas Rachford splitting algorithm [15].

Algorithm: Douglas Rachford on (P)

$$v^{k+1} = \arg\min_{v \in \mathbb{R}^m} J(Av) + \frac{1}{2\alpha_k}\|v - u^k + \alpha_k y^k\|_2^2 \tag{36a}$$

$$u^{k+1} = \arg\min_{u \in \mathbb{R}^m} H(u) + \frac{1}{2\alpha_k}\|u - v^{k+1} - \alpha_k y^k\|_2^2 \tag{36b}$$

$$y^{k+1} = y^k + \frac{1}{\alpha_k}(v^{k+1} - u^{k+1}) \tag{36c}$$

The following theorem from [15] shows that convergence of u^k can be ensured with very few assumptions.

Theorem 4.5. *[15] Let $\alpha_k = \alpha > 0$ and let (u^0, y^0) be arbitrary. Suppose (v^k, u^k, y^k) satisfies (36). Then $\{u^k\}$ converges to a solution of (P).*

4.2.3 Modifications of PDHG

In this section we show that two slightly modified versions of the PDHG algorithm, denoted PDHGMp and PDHGMu, can be interpreted as a split inexact Uzawa method from [49] applied to ($\mathrm{SP_P}$) and ($\mathrm{SP_D}$) respectively. In the constant step size case, PDHGMp replaces p^{k+1} in the u^{k+1} step (26b) with $2p^{k+1} - p^k$ whereas PDHGMu replaces u^k in the p^{k+1} step (26a) with $2u^k - u^{k-1}$. The variable step size case will also be discussed. The advantage of these modified algorithms is that for appropriate parameter choices they are nearly as efficient as PDHG numerically, and some known convergence results [49] can be applied. Convergence of PDHGMu for a special class of saddle point problems is also proved in [35] based on an argument in [36].

The split inexact Uzawa method from [49] applied to ($\mathrm{SP_D}$) can be thought of as a modification of ADMM (35) that adds $\frac{1}{2}\|p-p^k\|_{D_0}$ to (35a), where D_0 is a positive definite matrix and $\|x\|_{D_0}^2$ is defined to be $\langle D_0 x, x\rangle$. Applying the main idea of the Bregman operator splitting algorithm from [50], a useful choice of D_0 is one that simplifies the minimization step by decoupling variables. To apply it to ($\mathrm{SP_D}$) for example, we choose $D_0 = (\frac{1}{\delta_k} - \alpha_k AA^T)$, where $0 < \delta_k < \frac{1}{\alpha_k \|A\|^2}$ so that D_0 is positive definite. Altogether, the objective functional for the first step of ADMM on ($\mathrm{SP_D}$) is modified by adding $\frac{1}{2}\langle p-p^k, (\frac{1}{\delta_k} - \alpha_k AA^T)(p-p^k)\rangle$. By combining terms, the new update for p^{k+1} can be written as

$$p^{k+1} = \arg\min_{p\in\mathbb{R}^n} J^*(p) + \frac{1}{2\delta_k}\|p - p^k - \delta_k A u^k + \alpha_k \delta_k A(A^T p^k + y^k)\|_2^2.$$

The updates for y^{k+1} and u^{k+1} remain the same. Altogether, the resulting algorithm is given by

---

Algorithm: Split Inexact Uzawa applied to ($\mathrm{SP_D}$)

$$p^{k+1} = \arg\min_{p\in\mathbb{R}^n} J^*(p) + \frac{1}{2\delta_k}\|p - p^k - \delta_k A u^k + \alpha_k \delta_k A(A^T p^k + y^k)\|_2^2 \tag{37a}$$

$$y^{k+1} = \arg\min_{y\in R^m} H^*(y) - \langle u^k, y\rangle + \frac{\alpha_k}{2}\|y + A^T p^{k+1}\|_2^2 \tag{37b}$$

$$u^{k+1} = u^k + \alpha_k(-A^T p^{k+1} - y^{k+1}). \tag{37c}$$

The above algorithm can be shown to converge at least for fixed step sizes α and δ satisfying $0 < \delta < \frac{1}{\alpha\|A\|^2}$.

Theorem 4.6. *[49] Let $\alpha_k = \alpha > 0$, $\delta_k = \delta > 0$ and $0 < \delta < \frac{1}{\alpha\|A\|^2}$. Let (p^k, y^k, u^k) satisfy (37). Also let p^* be optimal for (D) and $y^* = -A^T p^*$. Then*

- $\|A^T p^k + y^k\|_2 \to 0$
- $J^*(p^k) \to J^*(p^*)$
- $H^*(y^k) \to H^*(y^*)$

and all convergent subsequences of (p^k, y^k, u^k) converge to a saddle point of L_D (24).

Moreover, the split inexact Uzawa algorithm can be rewritten in a form that is very similar to PDHG. Since the y^{k+1} (37b) and u^{k+1} (37c) steps are the same as those for AMA on ($\mathrm{SP_D}$) (31), then by the same argument they are equivalent to the u^{k+1} update in PDHG (26b). From (37c), we have that

$$y^k = \frac{u^{k-1}}{\alpha_{k-1}} - \frac{u^k}{\alpha_{k-1}} - A^T p^k. \tag{38}$$

Substituting this into (37a), we see that (37) is equivalent to a modified form of PDHG where u^k is replaced by $\left((1+\frac{\alpha_k}{\alpha_{k-1}})u^k - \frac{\alpha_k}{\alpha_{k-1}}u^{k-1}\right)$ in (26a). The resulting form of the algorithm will be denoted PDHGMu.

Algorithm: PDHGMu

$$p^{k+1} = \arg\min_{p\in\mathbb{R}^n} J^*(p) + \langle p, -A\left((1+\frac{\alpha_k}{\alpha_{k-1}})u^k - \frac{\alpha_k}{\alpha_{k-1}}u^{k-1}\right)\rangle + \frac{1}{2\delta_k}\|p-p^k\|_2^2 \tag{39a}$$

$$u^{k+1} = \arg\min_{u\in\mathbb{R}^m} H(u) + \langle A^T p^{k+1}, u\rangle + \frac{1}{2\alpha_k}\|u-u^k\|_2^2, \tag{39b}$$

Similarly, the corresponding split inexact Uzawa method applied to ($\mathrm{SP_P}$) is obtained by adding $\frac{1}{2}\langle u-u^k, (\frac{1}{\alpha_k} - \delta_k A^T A)(u-u^k)\rangle$ to the u^{k+1} step of ADMM applied to ($\mathrm{SP_P}$) (34a).

Algorithm: Split Inexact Uzawa applied to ($\mathrm{SP_P}$)

$$u^{k+1} = \arg\min_{u\in\mathbb{R}^m} H(u) + \frac{1}{2\alpha_k}\|u - u^k - \alpha_k A^T p^k + \delta_k\alpha_k A^T(Au^k - w^k)\|_2^2 \tag{40a}$$

$$w^{k+1} = \arg\min_{w\in\mathbb{R}^n} J(w) - \langle p^k, w\rangle + \frac{\delta_k}{2}\|Au^{k+1} - w\|_2^2 \tag{40b}$$

$$p^{k+1} = p^k + \delta_k(Au^{k+1} - w^{k+1}) \tag{40c}$$

Again by Theorem 4.6, the above algorithm converges for fixed stepsizes α and δ with $0 < \alpha < \frac{1}{\delta\|A\|^2}$. Note this requirement is equivalent to requiring $0 < \delta < \frac{1}{\alpha\|A\|^2}$.

Since from (40c), we have that

$$w^k = \frac{p^{k-1}}{\delta_{k-1}} - \frac{p^k}{\delta_{k-1}} + Au^k, \tag{41}$$

a similar argument shows that (40) is equivalent to a modified form of PDHG where p^k is replaced by $\left((1+\frac{\delta_k}{\delta_{k-1}})p^k - \frac{\delta_k}{\alpha_{k-1}}p^{k-1}\right)$. The resulting form of the algorithm will be denoted PDHGMp.

Algorithm: PDHGMp

$$u^{k+1} = \arg\min_{u\in\mathbb{R}^m} H(u) + \langle A^T\left((1+\frac{\delta_k}{\delta_{k-1}})p^k - \frac{\delta_k}{\alpha_{k-1}}p^{k-1}\right), u\rangle + \frac{1}{2\alpha_k}\|u-u^k\|_2^2, \tag{42a}$$

$$p^{k+1} = \arg\min_{p\in\mathbb{R}^n} J^*(p) - \langle p, Au^{k+1}\rangle + \frac{1}{2\delta_k}\|p-p^k\|_2^2 \tag{42b}$$

The modifications to u^k and p^k in the split inexact Uzawa methods are reminiscent of the predictor-corrector step in Chen and Teboulle's predictor corrector proximal method (PCPM) [11]. Despite some close similarities, however, the algorithms are not equivalent. The modified PDHG algorithms are more implicit than PCPM.

The connections between the algorithms discussed so far are diagrammed in Figure 1. For simplicity, constant step sizes are assumed in the diagram.

5 Interpretation of PDHG as Projected Averaged Gradient Method for TV Denoising

Even though we know of a convergence result (4.6) for the modified PDHG algorithms PDHGMu (39) and PDHGMp (42), it would be nice to show convergence of the original PDHG method (26) because PDHG still has some numerical advantages. Empirically, the stability requirements for the step size parameters are less restrictive for PDHG, so there is more freedom to tune the parameters to improve the rate of convergence. In this section, we restrict attention to PDHG applied to TV denoising and prove a convergence result assuming certain conditions on the parameters.

5.1 Projected Gradient Special Case

In the case of TV denoising, problem (P) becomes

$$\min_{u\in\mathbb{R}^m} \|u\|_{TV} + \frac{\lambda}{2}\|u-f\|_2^2, \tag{43}$$

with $J = \|\cdot\|_E$, $A = D$ and $H(u) = \frac{\lambda}{2}\|u-f\|_2^2$, in which case PFBS on (D) simplifies to

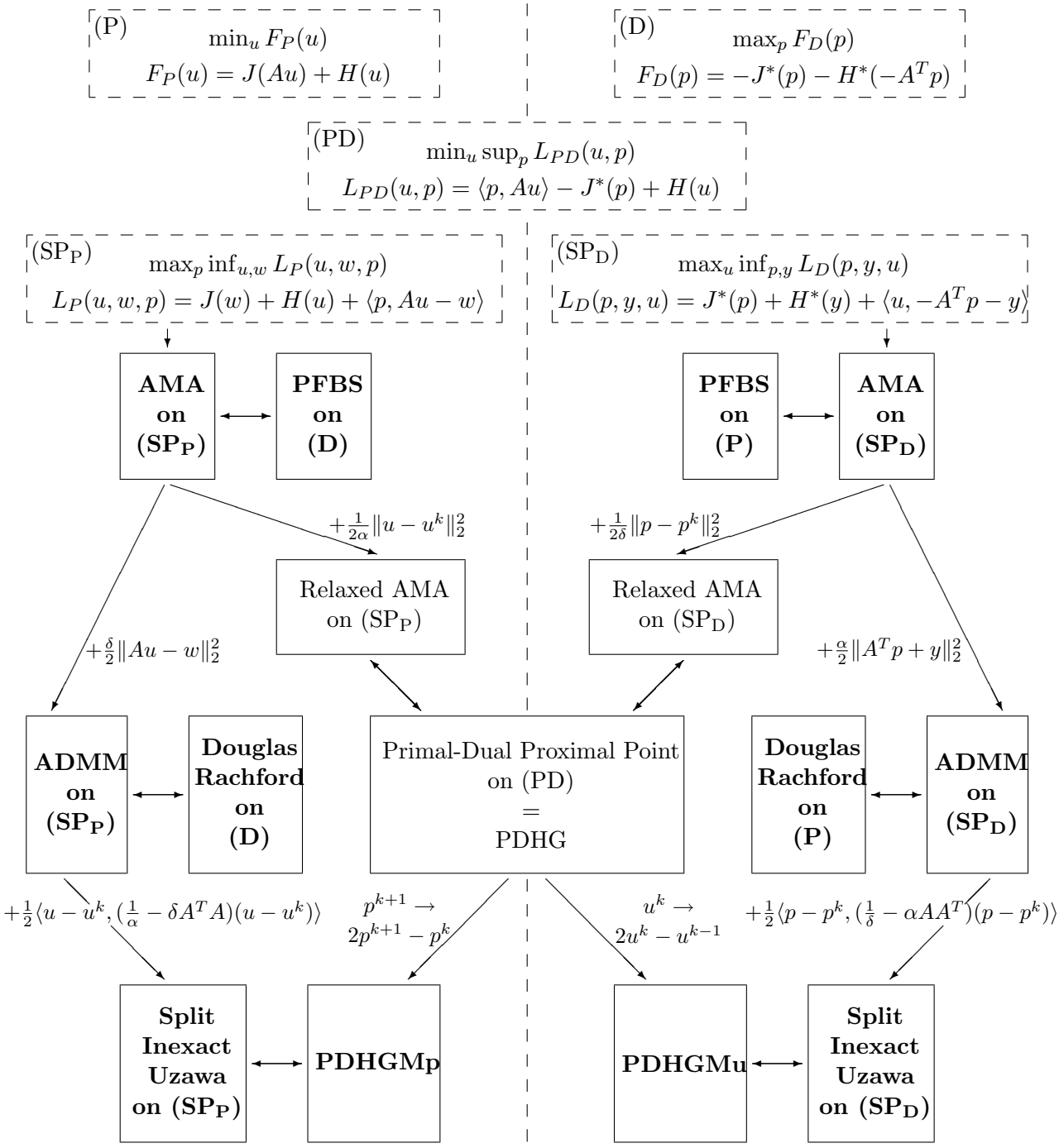
$$p^{k+1} = \arg\min_{p\in\mathbb{R}^n} J^*(p) + \frac{1}{2\delta_k}\|p - (p^k + \delta_k D\nabla H^*(-D^T p^k))\|_2^2.$$

Since J^* is the indicator function for the unit ball, denoted X (8), in the dual norm $\|\cdot\|_{E^*}$, this is exactly an orthogonal projection onto the convex set X (17). Letting $\tau_k = \frac{\delta_k}{\lambda}$ and using also that

$$H^*(-D^T p) = \frac{1}{2\lambda}\|\lambda f - D^T p\|_2^2 - \frac{\lambda}{2}\|f\|_2^2,$$

the algorithm simplifies to

Figure 1: PDHG-Related Algorithm Framework



Legend:
- (P): Primal
- (D): Dual
- (PD): Primal-Dual
- (SP_P): Split Primal
- (SP_D): Split Dual

AMA: Alternating Minimization Algorithm (4.2.1)
PFBS: Proximal Forward Backward Splitting (4.2.1)
ADMM: Alternating Direction Method of Multipliers (4.2.2)
PDHG: Primal Dual Hybrid Gradient (4.2)
PDHGM: Modified PDHG (4.2.3)
Bold: Well Understood Convergence Properties

Algorithm: Gradient Projection for TV Denoising

$$p^{k+1} = \Pi_X \left(p^k - \tau_k D(D^T p^k - \lambda f) \right). \tag{44}$$

Many variations of gradient projection applied to TV denoising are discussed in [52]. As already noted in [51], algorithm PDGH applied to TV denoising reduces to projected gradient descent when $\theta_k = 1$. Equivalence to (15) in the $\theta_k = 1$ case can be seen by plugging $u^k = (f - \frac{1}{\lambda} D^T p^k)$ into the update for p^{k+1}. This can also be interpreted as projected gradient descent applied to

$$\min_{p \in X} G(p) := \frac{1}{2} \|D^T p - \lambda f\|_2^2, \tag{45}$$

an equivalent form of the dual problem.

Theorem 5.1. *Fix $p^0 \in \mathbb{R}^n$. Let p^k be defined by (44) with $0 < \inf \tau_k \leq \sup \tau_k < \frac{1}{4}$, and define $u^{k+1} = f - \frac{D^T p^k}{\lambda}$. Then $\{p^k\}$ converges to a solution of (45), and $\{u^k\}$ converges to a solution of (43).*

Proof. Since G is Lipschitz continuous with Lipschitz constant $\|DD^T\|$ and $u^{k+1} = \nabla H^*(-D^T p^k) = f - \frac{D^T p^k}{\lambda}$, then by Theorem 4.2 the result follows if $0 < \inf \tau_k \leq \sup \tau_k < \frac{2}{\|DD^T\|}$. We can bound $\|DD^T\|$ by the largest eigenvalue of $D^T D$, which is minus the discrete Laplacian corresponding to Neumann boundary conditions. The matrix $D^T D$ from its definition has only the numbers 2, 3 and 4 on its main diagonal. All the off diagonal entries are 0 or -1, and the rows sum to zero. Therefore, by the Gersgorin Circle Theorem, all eigenvalues are in the interval $[0, 8]$. Thus $\|DD^T\| \leq 8$, so $\frac{1}{4} \leq \frac{2}{\|DD^T\|}$. □

5.1.1 AMA Equivalence and Soft Thresholding Interpretation

By the general equivalence between PFBS and AMA, (44) is equivalent to

Algorithm: AMA for TV Denoising

$$u^{k+1} = f - \frac{D^T p^k}{\lambda} \tag{46a}$$

$$w^{k+1} = \tilde{S}_{\frac{1}{\delta_k}} (Du^{k+1} + \frac{1}{\delta_k} p^k) \tag{46b}$$

$$p^{k+1} = p^k + \delta_k (Du^{k+1} - w^k), \tag{46c}$$

where $\tilde{S}$ denotes the soft thresholding operator for $\|\cdot\|_E$ defined by

$$\tilde{S}_\alpha(f) = \arg\min_z \|z\|_E + \frac{1}{2\alpha} \|z - f\|_2^2.$$

The general equivalence of these algorithms, which has already been demonstrated, also provides a way to define the soft thresholding operator in terms of a projection.

A direct application of Moreau's decomposition (25) shows that $\tilde{S}_\alpha(f)$ can be defined by

$$\tilde{S}_\alpha(f) = f - \alpha \Pi_X(\frac{f}{\alpha}), \tag{47}$$

with Π_X defined by (17). Similar derivations work for other norms. For example, this can be used to define the well known soft thresholding operator corresponding to l_1-l_2 minimization. Let

$$S_\alpha(f) = \arg\min_u \|u\|_1 + \frac{1}{2\alpha}\|u - f\|_2^2.$$

Then

$$S_\alpha(f) = f - \alpha \Pi_{\{p:\|p\|_\infty \leq 1\}}(\frac{f}{\alpha}),$$

where

$$\Pi_{\{p:\|p\|_\infty \leq 1\}}(p) = \frac{p}{\max(|p|, 1)}. \tag{48}$$

In fact, it's not necessary to assume that J is a norm to obtain similar projection interpretations. It's enough that J be a convex 1-homogeneous function, as Chambolle points out in [9] when deriving a projection formula for the solution of the TV denoising problem. By letting $z = D^T p$, the dual problem (45) is solved by the projection

$$z = \Pi_{\{z:z=D^T p, \|p\|_{E^*} \leq 1\}}(\lambda f),$$

and the solution to the TV denoising problem is given by

$$u^* = f - \frac{1}{\lambda}\Pi_{\{z:z=D^T p, \|p\|_{E^*} \leq 1\}}(\lambda f).$$

However, the projection is nontrivial to compute.

5.2 Projected Averaged Gradient

In the $\theta \neq 1$ case, still for TV denoising, the projected gradient descent interpretation of PDHG extends to an interpretation as a projected averaged gradient descent algorithm. Consider for simplicity parameters τ and θ that are independent of k. Then plugging u^{k+1} into the update for p yields

$$p^{k+1} = \Pi_X\left(p^k - \tau d_\theta^k\right) \tag{49}$$

where

$$d_\theta^k = \theta \sum_{i=1}^{k} (1-\theta)^{k-i} \nabla G(p^i) + (1-\theta)^k \nabla G(p^0)$$

is a convex combination of gradients of G at the previous iterates p^i. Note that d_θ^k is not necessarily a descent direction.

In the following section, the connection to a projected average gradient method on the dual is made for the more general case when the parameters are allowed to depend on k. Convergence results are presented for some special cases.

This kind of averaging of previous iterates suggests a connection to Nesterov's method [31]. Several recent papers study variants of his method and their applications. Weiss, Aubert and Blanc-Féraud in [46] apply a variant of Nesterov's method [32] to smoothed TV functionals. Beck and Teboulle in [1] and Becker, Bobin and Candes in [2] also study variants of Nesterov's method that apply to l_1 and TV minimization problems. Tseng gives a unified treatment of accelerated proximal gradient methods like Nesterov's in [45]. However, despite some tantalizing similarities to PDGH, it appears that none is equivalent.

5.2.1 Convergence

For a minimizer $\overline{p}$, the optimality condition for the dual problem (45) is

$$\overline{p} = \Pi_X(\overline{p} - \tau\nabla G(\overline{p})), \quad \forall\tau \geq 0, \tag{50}$$

or equivalently

$$\langle \nabla G(\overline{p}), p - \overline{p}\rangle \geq 0, \quad \forall p \in X.$$

In the following, we denote $\overline{G} = \min_{p\in X} G(p)$ and let X^* denote the set of minimizers. As mentioned above, the PDHG algorithm (15) for TV denoising is related to a projected gradient method on the dual variable p. When τ and θ are allowed to depend on k, the algorithm can be written as

$$p^{k+1} = \Pi_X\left(p^k - \tau_k d^k\right) \tag{51}$$

where

$$d^k = \sum_{i=0}^{k} s_k^i \nabla G(p^i), \quad s_k^i = \theta_{i-1}\prod_{j=i}^{k-1}(1-\theta_j).$$

Note that

$$\sum_{i=0}^{k} s_k^i = 1, \quad s_k^i = (1-\theta_{k-1})s_{k-1}^i \qquad \forall k \geq 0, i \leq k, \qquad \text{and} \tag{52}$$

$$d^k = (1-\theta_{k-1})d^{k-1} + \theta_{k-1}\nabla G(p^k). \tag{53}$$

As above, the direction d^k is a linear (convex) combination of gradients of all previous iterates. We will show d^k is an ϵ-gradient at p^k. This means d^k is an element of the ϵ-differential (ϵ-subdifferential for nonsmooth functionals), $\partial_\epsilon G(p)$, of G at p^k defined by

$$G(q) \geq G(p^k) + \langle d^k, q - p^k\rangle - \epsilon, \forall q \in X$$

When $\epsilon = 0$ this is the definition of d^k being a sub-gradient (in this case, the gradient) of G at p^k.

For p and q, the Bregman distance based on G between p and q is defined as

$$D(p,q) = G(p) - G(q) - \langle \nabla G(q), p - q\rangle \quad \forall p, q \in X \tag{54}$$

From (45), the Bregman distance (54) reduces to

$$D(p,q) = \frac{1}{2}\|D^T(p-q)\|_2^2 \leq \frac{L}{2}\|p-q\|^2,$$

where L is the Lipschitz constant of ∇G.

Lemma 5.2. *For any $q \in X$, we have*

$$G(q) - G(p^k) - \langle d^k, q - p^k \rangle = \sum_{i=0}^{k} s_k^i (D(q, p^i) - D(p^k, p^i)).$$

Proof. For any $q \in X$,

$$\begin{aligned} G(q) - G(p^k) - \langle d^k, q - p^k \rangle &= G(q) - G(p^k) - \langle \sum_{i=0}^{k} s_k^i \nabla G(p^i), q - p^k \rangle \\ &= \sum_{i=0}^{k} s_k^i G(q) - \sum_{i=0}^{k} s_k^i G(p^i) - \sum_{i=0}^{k} s_k^i \langle \nabla G(p^i), q - p^i \rangle \\ &\quad + \sum_{i=0}^{k} s_k^i (G(p^i) - G(p^k) - \langle \nabla G(p^i), p^i - p^k \rangle \\ &= \sum_{i=0}^{k} s_k^i (D(q, p^i) - D(p^k, p^i)) \end{aligned}$$

□

Lemma 5.3. *The direction d^k is a ϵ_k-gradient of p^k where $\epsilon_k = \sum_{i=0}^{k} s_k^i D(p^k, p^i)$.*

Proof. By Lemma 5.2,

$$G(q) - G(p^k) - \langle d^k, q - p^k \rangle \geq -\sum_{i=0}^{k} s_k^i D(p^k, p^i) \quad \forall q \in X.$$

By the definition of ϵ-gradient, we obtain that d^k is a ϵ_k-gradient of G at p^k, where

$$\epsilon_k = \sum_{i=0}^{k} s_k^i D(p^k, p^i).$$

□

Lemma 5.4. *If $\theta_k \to 1$, then $\epsilon_k \to 0$.*

Proof. Let $h_k = G(p^k) - G(p^{k-1}) - \langle d^{k-1}, p^k - p^{k-1} \rangle$, then using the Lipschitz continuity of ∇G and the boundedness of d^k, we obtain

$$|h_k| = |D(p^k, p^{k-1}) + \langle (\nabla G(p^{k-1}) - d^{k-1}, p^k - p^{k-1} | \rangle| \leq \frac{L}{2} \|p^k - p^{k-1}\|_2^2 + C_1 \|p^k - p^{k-1}\|_2,$$

where L is the Lipschitz constant of ∇G, and C_1 is some positive constant. Since $\epsilon_k = \sum_{i=0}^{k} s_k^i D(p^k, p^i)$, and $\sum_{i=0} s_k^i = 1$, then ϵ_k is bounded for any k.

Meanwhile, by replacing q with p^k and p^k by p^{k-1} in Lemma 5.2, we obtain $h_k = \sum_{i=0}^{k-1} s_{k-1}^i (D(p^k, p^i) - D(p^{k-1}, p^i))$. From

$$s_k^i = (1 - \theta_{k-1}) s_{k-1}^i, \ \forall \ 1 \leq i \leq k-1,$$

we get

$$\begin{aligned}\epsilon_k &= (1-\theta_{k-1})\sum_{i=0}^{k-1} s_{k-1}^i D(p^k,p^i)\\ &= (1-\theta_{k-1})\epsilon_{k-1} + (1-\theta_{k-1})\sum_{i=0}^{k-1} s_{k-1}^i (D(p^k,p^i)-D(p^{k-1},p^i))\\ &= (1-\theta_{k-1})(\epsilon_{k-1}+h_k).\end{aligned}$$

By the boundness of h_k and ϵ_k, we get immediately that if $\theta_{k-1}\to 1$, then $\epsilon_k \to 0$. □

Since $\epsilon_k \to 0$, the convergence of p^k follows directly from classical [42, 27] ϵ-gradient methods. Possible choices of the step size τ_k are given in the following theorem:

Theorem 5.5. *[42, 27][Convergence to the optimal set using divergent series τ_k] Let $\theta_k \to 1$ and let τ_k be chosen according to one of the following cases:*

1. $\tau_k = \lambda_k \frac{G(p^k)-\overline{G}}{|d^k|^2}, 0 < r_1 \le \lambda_k \le 2 - r_2 < 2.$
2. $\tau_k > 0$, $\lim_{k\to\infty} \tau_k |d_k| = 0$ *and* $\sum_{k=1}^{\infty} \tau_k |d_k| = \infty$.
3. $\tau_k > 0$, $\lim_{k\to\infty} \tau_k = 0$ *and* $\sum_{k=1}^{\infty} \tau_k = \infty$.

Then the sequence p^k generated by the method (51) satisfies $G(p^k) \to \overline{G}$ and $dist\{p^k, X^*\} \to 0$.

Since we require $\theta_k \to 1$, the algorithm is equivalent to projected gradient descent in the limit. The conditions on τ_k also generally require $\tau_k \to 0$. It is well known that a divergent step size is slow and we can expect a better convergence rate without letting τ_k go to 0. In the following, we prove a different convergence result that doesn't require $\tau_k \to 0$ but still requires $\theta_k \to 1$.

Lemma 5.6. *For p^k defined by (51), we have $\langle d^k, p^{k+1}-p^k\rangle \le -\frac{1}{\tau_k}\|p^{k+1}-p^k\|_2^2$.*

Proof. Since p^{k+1} is the projection of $p^k - \tau_k d^k$ onto X, it follows that

$$\langle p^k - \tau_k d^k - p^{k+1}, p - p^{k+1}\rangle \le 0, \quad \forall p \in X.$$

Replacing p with p^k, we thus get

$$\langle d^k, p^{k+1}-p^k\rangle \le -\frac{1}{\tau_k}\|p^{k+1}-p^k\|_2^2.$$

□

Lemma 5.7. *Let p^k be generated by the method (51), then*

$$G(p^{k+1}) - G(p^k) - \frac{\beta_k^2}{\alpha_k}\|p^k - p^{k-1}\|_2^2 \le -\frac{(\alpha_k+\beta_k)^2}{\alpha_k}\|p^k - (\frac{\alpha_k}{\alpha_k+\beta_k}p^{k+1} + \frac{\beta_k}{\alpha_k+\beta_k}p^{k-1})\|_2^2$$

where

$$\alpha_k = \frac{1}{\tau_k\theta_{k-1}} - \frac{L}{2}, \quad \beta_k = \frac{1-\theta_{k-1}}{2\theta_{k-1}\tau_{k-1}} \tag{55}$$

Proof. By using the Taylor expansion and the Lipschiz continuity of ∇G (or directly from the fact that G is quadratic function), we have

$$G(p^{k+1}) - G(p^k) \leq \langle \nabla G(p^k), p^{k+1} - p^k \rangle + \frac{L}{2}\|p^{k+1} - p^k\|_2^2,$$

Since $\nabla G(p^k) = \frac{1}{\theta_{k-1}}(d^k - (1-\theta_{k-1})d^{k-1})$, we have

$$\begin{aligned} G(p^{k+1}) - G(p^k) &\leq \frac{1}{\theta_{k-1}}\langle d^k, p^{k+1} - p^k \rangle - \frac{1-\theta_{k-1}}{\theta_{k-1}}\langle d^{k-1}, p^{k+1} - p^k \rangle + \frac{L}{2}\|p^{k+1} - p^k\|_2^2, \\ &= (\frac{L}{2} - \frac{1}{\tau_k \theta_{k-1}})\|p^{k+1} - p^k\|_2^2 - \frac{1-\theta_{k-1}}{\theta_{k-1}}\langle d^{k-1}, p^{k+1} - p^k \rangle. \end{aligned}$$

On the other hand, since p^k is the projection of $p^{k-1} - \tau_{k-1}d^{k-1}$, we get

$$\langle p^{k-1} - \tau_{k-1}d^{k-1} - p^k, p - p^k \rangle \leq 0, \quad \forall p \in X.$$

Replacing p with p^{k+1}, we thus get

$$\langle d^{k-1}, p^{k+1} - p^k \rangle \geq \frac{1}{\tau_{k-1}}\langle p^{k-1} - p^k, p^{k+1} - p^k \rangle.$$

This yields

$$\begin{aligned} G(p^{k+1}) - G(p^k) &\leq -\alpha_k \|p^{k+1} - p^k\|^2 - 2\beta_k \langle p^{k-1} - p^k, p^{k+1} - p^k \rangle \\ &= -\frac{(\alpha_k + \beta_k)^2}{\alpha_k}\|p^k - (\frac{\alpha_k}{\alpha_k + \beta_k}p^{k+1} + \frac{\beta_k}{\alpha_k + \beta_k}p^{k-1})\|^2 + \frac{\beta_k^2}{\alpha_k}\|p^k - p^{k-1}\|^2. \end{aligned}$$

where α_k and β_k are defined as (55).

□

Theorem 5.8. *If α_k and β_k defined as (55) such that $\alpha_k > 0, \beta_k \geq 0$ and*

$$\sum_{k=0}^{\infty} \frac{(\alpha_k + \beta_k)^2}{\alpha_k} = \infty, \quad \sum_{k=0}^{\infty} \frac{\beta_k^2}{\alpha_k} < \infty, \quad \lim_{k\to\infty} \frac{\beta_k}{\alpha_k} = 0. \tag{56}$$

then every limit point pair (p^∞, d^∞) of a subsequence of (p^k, d^k) is such that p^∞ is a minimizer of (45) and $d^\infty = \nabla G(p^\infty)$.

Proof. The proof is adapted from [4](Proposition 2.3.1,2.3.2) and Lemma 5.7. Since p^k and d^k are bounded, the subsequence (p^k, d^k) has a convergent subsequence. Let (p^∞, d^∞) be a limit point of the pair (p^k, d^k), and let (p^{k_m}, d^{k_m}) be a subsequence that converges to (p^∞, d^∞). For $k_m > n_0$, lemma 5.7 implies that

$$G(p^{k_m}) - G(p^{n_0}) \leq -\sum_{k=n_0}^{k_m} \frac{(\alpha_k + \beta_k)^2}{\alpha_k}\|p^k - (\frac{\alpha_k}{\alpha_k + \beta_k}p^{k+1} + \frac{\beta_k}{\alpha_k + \beta_k}p^{k-1})\|_2^2 + \sum_{k=n_0}^{k_m} \frac{\beta_k^2}{\alpha_k}\|p^{k-1} - p^k\|_2^2.$$

By the boundness of the constraint set X, the conditions (56) for α_k and β_k and the fact that $G(p)$ is bounded from below, we conclude that

$$\|p^k - (\frac{\alpha_k}{\alpha_k+\beta_k}p^{k+1} + \frac{\beta_k}{\alpha_k+\beta_k}p^{k-1})\|_2 \to 0.$$

Given $\epsilon > 0$, we can choose m large enough such that $\|p^{k_m} - p^\infty\|_2 \leq \frac{\epsilon}{3}$, $\|p^k - (\frac{\alpha_k}{\alpha_k+\beta_k}p^{k+1} + \frac{\beta_k}{\alpha_k+\beta_k}p^{k-1})\|_2 \leq \frac{\epsilon}{3}$ for all $k \geq k_m$, and $\frac{\beta_{k_m}}{\alpha_{k_m}+\beta_{k_m}}\|(p^{k_m-1} - p^\infty)\|_2 \leq \frac{\epsilon}{3}$. This third requirement is possible because $\lim_{k\to\infty}\frac{\beta_k}{\alpha_k} = 0$. Then

$$\|(p^{k_m} - p^\infty) - \frac{\alpha_{k_m}}{\alpha_{k_m}+\beta_{k_m}}(p^{k_m+1} - p^\infty) - \frac{\beta_{k_m}}{\alpha_{k_m}+\beta_{k_m}}(p^{k_m-1} - p^\infty)\|_2 \leq \frac{\epsilon}{3}$$

implies

$$\|\frac{\alpha_{k_m}}{\alpha_{k_m}+\beta_{k_m}}(p^{k_m+1} - p^\infty) + \frac{\beta_{k_m}}{\alpha_{k_m}+\beta_{k_m}}(p^{k_m-1} - p^\infty)\|_2 \leq \frac{2}{3}\epsilon.$$

Since $\frac{\beta_{k_m}}{\alpha_{k_m}+\beta_{k_m}}\|(p^{k_m-1} - p^\infty)\|_2 \leq \frac{\epsilon}{3}$, we have

$$\|p^{k_m+1} - p^\infty\|_2 \leq \frac{\alpha_{k_m}+\beta_{k_m}}{\alpha_{k_m}}\epsilon.$$

Note that $k_m + 1$ is not necessarily an index for the subsequence $\{p^{k_m}\}$. Since $\lim_k \frac{\alpha_k+\beta_k}{\alpha_k} = 1$, then we have $\|p^{k_m+1} - p^\infty\|_2 \to 0$ when $m \to \infty$. According (51), the limit point p^∞, d^∞ is therefore such that

$$p^\infty = \Pi_X(p^\infty - \tau d^\infty) \tag{57}$$

for $\tau > 0$.

It remains to show that the corresponding subsequence $d^{k_m} = (1-\theta_{k_m-1})d^{k_m-1} + \theta_{k_m-1}\nabla G(p^{k_m})$ converges to $\nabla G(p^\infty)$. By the same technique, and the fact that $\theta_k \to 1$, we can get $\|\nabla G(p^{k_m}) - d^\infty\| \leq \epsilon$. Thus $\nabla G(p^{k_m}) \to d^\infty$. On the other hand, $\nabla G(p^{k_m}) \to \nabla G(p^\infty)$. Thus $d^\infty = \nabla G(p^\infty)$. Combining with (57) and the optimal condition (50), we conclude that p^∞ is a minimizer. □

In summary, the overall conditions on θ_k and τ_k are:

- $\theta_k \to 1, \tau_k > 0$,
- $0 < \tau_k\theta_k < \frac{2}{L}$,
- $\sum_{k=0}^\infty \frac{(\alpha_k+\beta_k)^2}{\alpha_k} = \infty$,
- $\lim_{k\to\infty}\frac{\beta_k}{\alpha_k} = 0$,
- $\sum_{k=0}^\infty \frac{\beta_k^2}{\alpha_k} < \infty$,

where

$$\alpha_k = \frac{1}{\tau_k\theta_{k-1}} - \frac{L}{2}, \quad \beta_k = \frac{1-\theta_{k-1}}{2\theta_{k-1}\tau_{k-1}}. \tag{58}$$

Finally, we have $\theta_k \to 1$, and for τ_k, the classical condition for the projected gradient descent algorithm, $0 < \tau_k < \frac{2}{L}$ and divergent stepsize $\lim_k \tau_k \to 0, \sum_k \tau_k \to \infty$, are special cases of the above conditions. Note that even though the convergence with $0 < \theta_k \leq c < 1$ and even $\theta_k \to 0$ is numerically demonstrated in [51], a theoretical proof is still an open problem.

6 Modified PDHG for TV Deblurring

The TV deblurring problem (1), which includes denoising as the $K = I$ special case, was one of the main applications of PDHG discussed in [51]. In the notation of problem (P), it corresponds to $J = \|\cdot\|_E$, $A = D$ and $H(u) = \frac{\lambda}{2}\|Ku - f\|_2^2$. In this section we consider the applications of PDHGMu and PDHGMp to this problem and point out when the convergence theory for the split inexact Uzawa method can be applied. In Section (8), some numerical experiments will be presented to compare the empirical performance of the modified PDHG algorithms to the original one.

In the variable step size case, the algorithms are given by

Algorithm: PDHGMp for TV Deblurring

$$u^{k+1} = (\frac{1}{\alpha_k} + \lambda K^T K)^{-1}\left(\lambda K^T f - D^T\left((1 + \frac{\delta_k}{\delta_{k-1}})p^k - \frac{\delta_k}{\delta_{k-1}}p^{k-1}\right) + \frac{u^k}{\alpha_k}\right) \tag{59a}$$

$$p^{k+1} = \Pi_X\left(p^k + \delta_k D u^{k+1}\right), \tag{59b}$$

---

where u^0, p^0 and p^{-1} are arbitrary, and

Algorithm: PDHGMu for TV Deblurring

$$p^{k+1} = \Pi_X\left(p^k + \delta_k D\left((1 + \frac{\alpha_k}{\alpha_{k-1}})u^k - \frac{\alpha_k}{\alpha_{k-1}}u^{k-1}\right)\right) \tag{60a}$$

$$u^{k+1} = (\frac{1}{\alpha_k} + \lambda K^T K)^{-1}(\lambda K^T f - D^T p^{k+1} + \frac{u^k}{\alpha_k}), \tag{60b}$$

---

where p^0, u^0 and u^{-1} are arbitrary. Theorem 4.6 applies when $\alpha_k = \alpha > 0$, $\delta_k = \delta > 0$ and $\delta < \frac{1}{\alpha\|D\|^2}$. Choosing δ is straightforward. Since $\|D\|^2 \leq 8$ (5.1), a safe choice for δ is to let $0 < \delta \leq \frac{1}{r\alpha}$ where $r > 8$. It would be interesting to extend the convergence analysis to the case where the parameters depend on k. Letting α_k be proportional to $\frac{1}{k}$ and fixing the product $\alpha_k\delta_k < \frac{1}{\|D\|^2}$ would result in parameters similar to those empirically optimized for PDHG in [51].

Note that from (38) and (39b), $y^{k+1} = \nabla H(u^{k+1})$, which we can substitute instead of (38) into (37a) to get an equivalent version of PDHGMu, whose updates only depend on the previous iteration instead of the previous two.

7 Extension to Constrained Minimization

The extension of PDHG to constrained minimization problems is discussed in [51] and applied for example to TV denoising with a constraint of the form $\|u - f\|^2 \leq \sigma^2$ with σ^2 an estimate of the variance of the Gaussian noise. In the context of our general primal problem (P), if u is constrained to be in a convex set S, then this still fits in the framework of (P) since the indicator function for S can be incorporated into the definition of $H(u)$.

7.1 General Convex Constraint

Consider the case when $H(u)$ is exactly the indicator function for a convex set $S \subset \mathbb{R}^m$, which would mean

$$H(u) = \begin{cases} 0 & \text{if } u \in S \\ \infty & \text{otherwise} \end{cases}.$$

Applying PDHG results in a primal step that can be interpreted as an orthogonal projection onto S. We could also apply the modified PDHG algorithms (42, 39). For example, PDHGMu would yield

---

Algorithm: PDHGMu for constrained minimization ($u \in S$)

$$p^{k+1} = \arg\min_{p \in \mathbb{R}^n} J^*(p) + \langle p, -A\left((1 + \frac{\alpha_k}{\alpha_{k-1}})u^k - \frac{\alpha_k}{\alpha_{k-1}}u^{k-1}\right)\rangle + \frac{1}{2\delta_k}\|p - p^k\|_2^2 \tag{61a}$$

$$u^{k+1} = \Pi_S\left(u^k - \alpha_k A^T p^{k+1}\right). \tag{61b}$$

For this algorithm to be practical, the projection Π_S must be straightforward to compute. Suppose the constraint on u is of the form $\|Ku - f\|_2 \leq \epsilon$ for some matrix K and $\epsilon > 0$. Then

$$\Pi_S(z) = (I - K^\dagger K)z + K^\dagger \begin{cases} Kz & \text{if } \|Kz - f\|_2 \leq \epsilon \\ f + r\left(\frac{Kz - KK^\dagger f}{\|Kz - KK^\dagger f\|_2}\right) & \text{otherwise} \end{cases},$$

where

$$r = \sqrt{\epsilon^2 - \|(I - KK^\dagger)f\|_2^2}$$

and $K^\dagger$ denotes the pseudoinverse of K. Also note that $(I - K^\dagger K)$ represents the orthogonal projection onto $\ker(K)$. A special case where this projection is easily computed is when $K = R\Phi$ where R is a row selector and Φ is orthogonal. Then $KK^T = I$ and $K^\dagger = K^T$. In this case, the projection onto S simplifies to

$$\Pi_S(z) = (I - K^T K)z + K^T \begin{cases} Kz & \text{if } \|Kz - f\|_2 \leq \epsilon \\ f + \epsilon\left(\frac{Kz - f}{\|Kz - f\|_2}\right) & \text{otherwise} \end{cases}.$$

7.2 Constrained l_1-Minimization

Compressive sensing problems [8] that seek to find a sparse solution satisfying some data constraints sometimes use the type of constraint described in the previous section. A simple example of such a problem is

$$\min_{z \in \mathbb{R}^m} \|\Psi z\|_1 \qquad \text{such that} \qquad \|R\Gamma z - f\|_2 \leq \epsilon, \tag{62}$$

where Ψ is an orthogonal matrix representing the basis in which we expect z to be sparse, R is a row selector and Γ is orthogonal. $R\Gamma$ can be thought of as a measurement matrix that represents a selection of some coefficients in an orthonormal basis. Since Ψ is orthogonal, problem (62) is equivalent to

$$\min_{u \in \mathbb{R}^m} \|u\|_1 \qquad \text{such that} \qquad \|Ku - f\|_2 \leq \epsilon, \tag{63}$$

where $K = R\Gamma\Psi^T$.

7.2.1 Applying PDHGMu

Letting $J = \|\cdot\|_1$, $A = I$, $S = \{u : \|Ku - f\|_2 \leq \epsilon\}$ and $H(u)$ equal the indicator function for S, application of PDHGMu yields

---

Algorithm: PDHGMu for Constrained l_1-Minimization

$$p^{k+1} = \Pi_{\{p:\|p\|_\infty \leq 1\}} \left(p^k + \delta_k \left((1 + \frac{\alpha_k}{\alpha_{k-1}}) u^k - \frac{\alpha_k}{\alpha_{k-1}} u^{k-1} \right) \right) \tag{64a}$$

$$u^{k+1} = \Pi_S \left(u^k - \alpha_k p^{k+1} \right), \tag{64b}$$

---

where

$$\Pi_{\{p:\|p\|_\infty \leq 1\}}(p) = \frac{p}{\max(|p|, 1)}$$

and

$$\Pi_S(u) = (I - K^T K)u + K^T \left(f + \frac{Ku - f}{\max\left(\frac{\|Ku-f\|_2}{\epsilon}, 1\right)} \right).$$

As before, Theorem 4.6 applies when $\alpha_k = \alpha > 0$, $\delta_k = \delta > 0$ and $\delta \leq \frac{1}{\alpha}$. Also, since $A = I$, the case when $\delta = \frac{1}{\alpha}$ is exactly ADMM applied to ($\mathrm{SP_D}$), which is equivalent to Douglas Rachford splitting on (P).

7.2.2 Reversing Roles of J and H

A related approach for problem (63) is to define

$$H(z) = \begin{cases} 0 & \|z - f\|_2 \leq \epsilon \\ \infty & \text{otherwise} \end{cases} \tag{65}$$

and then apply PDHGMu to the problem of minimizing $H(Ku) + J(u)$ with the roles of J and H reversed. This will no longer satisfy the constraint at each iteration, but it does greatly simplify the projection step. The resulting algorithm is

---

Algorithm: PDHGRMu (reversed role version) for Constrained l_1-Minimization

$$p^{k+1} = \arg\min_p H^*(p) + \langle p, -K \left((1 + \frac{\alpha_k}{\alpha_{k-1}}) u^k - \frac{\alpha_k}{\alpha_{k-1}} u^{k-1} \right) \rangle + \frac{1}{2\delta_k} \|p - p^k\|_2^2 \tag{66a}$$

$$u^{k+1} = \arg\min_u J(u) + \langle K^T p^{k+1}, u \rangle + \frac{1}{2\alpha_k} \|u - u^k\|_2^2 \tag{66b}$$

This can be explicitly written in terms of projections as

$$p^{k+1} = p^k + \delta_k K \left((1 + \frac{\alpha_k}{\alpha_{k-1}}) u^k - \frac{\alpha_k}{\alpha_{k-1}} u^{k-1} \right) - \delta_k \Pi_{\{z:\|z-f\|_2 \leq \epsilon\}} \left(\frac{p^k}{\delta_k} + K \left((1 + \frac{\alpha_k}{\alpha_{k-1}}) u^k - \frac{\alpha_k}{\alpha_{k-1}} u^{k-1} \right) \right)$$

and

$$u^{k+1} = u^k - \alpha_k K^T p^{k+1} - \alpha_k \Pi_{\{p:\|p\|_\infty \leq 1\}} \left(\frac{u^k}{\alpha_k} - K^T p^{k+1} \right),$$

where

$$\Pi_{\{z:\|z-f\|_2 \leq \epsilon\}}(z) = f + \frac{z-f}{\max\left(\frac{\|z-f\|_2}{\epsilon}, 1\right)}.$$

This variant of PDHGMu is still an application of the split inexact Uzawa method (37). Also, since $\|K\| \leq 1$, the conditions for convergence are the same as for (64). Moreover, since $KK^T = I$, if $\delta = \frac{1}{\alpha}$, then this method can again be interpreted as ADMM applied to the split dual problem.

Note that $\Pi_{\{z:\|z-f\|_2 \leq \epsilon\}}$ is much simpler to compute than Π_S. The benefit of simplifying the projection step is more important for problems where $K^\dagger$ is not as practical to deal with numerically.

An example of such a problem would be a constrained TV deblurring problem of the form

$$\min_u \|Du\|_E + H(Ku)$$

with H from (65) and K a matrix representing the blurring operator. Letting

$$J_1(u) = 0, \qquad J_2(z) = J_2(w, v) = \|w\|_E + H(v)$$

and

$$z = \begin{bmatrix} w \\ v \end{bmatrix} = Bu = \begin{bmatrix} D \\ K \end{bmatrix} u,$$

we can directly apply the split inexact Uzawa method or PDHGMp to minimize $J_1(u) + J_2(z)$ subject to $Bu = z$.

8 Numerical Experiments

We perform three numerical experiments to show the modified and unmodified PDHG algorithms have similar performance and applications. The first is a comparison between PDHG (26), PDHGMu (39) and ADMM (34) applied to TV denoising. The second compares the application of PDHG and PDHGMu to an unconstrained TV deblurring problem. The performance of PDHGMp (42) for these examples is essentially identical to that of PDHGMu and therefore not included. The third experiment applies PDHGMu to a compressive sensing problem formulated as a constrained l_1 minimization problem.

8.1 PDHGM, PDHG and ADMM for TV denoising

Here, we closely follow the numerical example presented in Table 4 of [51], which compares PDHG to Chambolle's method [9] and CGM [10] for TV denoising. We use the same 256×256 cameraman image with intensities in $[0, 255]$. The image is corrupted with zero mean gaussian noise having standard deviation 20. We also use the same parameter $\lambda = .053$. Both adaptive and fixed stepsize strategies are compared. In all examples, we initialize $u^0 = f$ and $p^0 = 0$. Figure 2 shows the clean and noisy images along with a benchmark solution for the denoised image.

Figure 2: Original, noisy and denoised cameraman images

Recall the PDHG algorithm for the TV denoising problem (43) is given by (15) with $K = I$. The adaptive strategy used for PDHG is the same one proposed in [51] where

$$\tau_k = .2 + .008k \qquad \theta_k = \frac{.5 - \frac{5}{15+k}}{\tau_k}. \tag{67}$$

These can be related to the step sizes δ_k and α_k in (26) by

$$\delta_k = \lambda\tau_k \qquad \alpha_k = \frac{\theta_k}{\lambda(1-\theta_k)}.$$

These time steps don't satisfy the requirements of Theorem 5.8, which requires $\theta_k \to 1$. However, we find that the adaptive PDHG strategy (67), for which $\theta_k \to 0$, is better numerically for TV denoising.

The PDHGMu algorithm for TV denoising is given by (60) with $K = I$. For PDHGMu, δ_k is always taken to be

$$\delta_k = \frac{1}{8.01\alpha_k}.$$

Due to the stability requirement for PDHGMu, using the same adaptive time steps of (67) can be unstable. Instead the adaptive strategy we use for PDHGMu is

$$\alpha_k = \frac{1}{\lambda(1 + .5k)}. \tag{68}$$

Unfortunately, no adaptive strategy for PDHGMu can satisfy the requirements of Theorem 4.6, which assumes fixed time steps. However, the rate of convergence of the adaptive PDHGMu strategy for TV denoising is empirically better than the fixed parameter strategies.

We also perform some experiments with fixed α and δ. A comparison is made to gradient projection (44). An additional comparison is made to ADMM as applied to (10) with $K = I$. This algorithm alternates soft thresholding, solving a Poisson equation and updating the Lagrange

multiplier. The explicit iterations are given by

$$\begin{aligned} w^{k+1} &= \tilde{S}_{\frac{1}{\alpha}}(Du^k + \frac{p^k}{\alpha}) \\ u^{k+1} &= (\lambda - \alpha\Delta)^{-1}(\lambda f + \alpha D^T w^{k+1} - D^T p^k) \\ p^{k+1} &= p^k + \alpha(w^{k+1} - Du^{k+1}), \end{aligned} \tag{69}$$

where $\tilde{S}$ is defined as in (47). This is equivalent to the split Bregman algorithm [24], which was compared to PDHG elsewhere in [51]. However, by working with the ADMM form of the algorithm, it's easier to use the duality gap as a stopping condition since u and p have the same interpretations in both algorithms. As in [51] we use the relative duality gap R for the stopping condition defined by

$$\begin{aligned} R(u,p) &= \frac{F_P(u) - F_D(p)}{F_D(p)} \\ &= \frac{\left(\|u\|_{TV} + \frac{\lambda}{2}\|u-f\|_2^2\right) - \left(\frac{\lambda}{2}\|f\|_2^2 - \frac{1}{2\lambda}\|D^T p - \lambda f\|_2^2\right)}{\frac{\lambda}{2}\|f\|_2^2 - \frac{1}{2\lambda}\|D^T p - \lambda f\|_2^2}, \end{aligned}$$

which is the duality gap divided by the dual functional. The duality gap is defined to be the difference between the primal and dual functionals. This quantity is always nonnegative, and is zero if and only if (u,p) is a saddle point of (9) with $K = I$. Table 1 shows the number of iterations required for the relative duality gap to fall below tolerances of 10^{-2}, 10^{-4} and 10^{-6}. Note that the complexity of the PDHG and PDHGMu iterations scale like $O(m)$ whereas the ADMM iterations scale like $O(m \log m)$. Results for PDHGMp were identical to those for PDHGMu and are therefore not included in the table.

From Table 1, we see that PDHG and PDHGMu both benefit from adaptive stepsize schemes. The adaptive versions of these algorithms are compared in Figure 3, which plots the l_2 distance to the benchmark solution versus number of iterations. PDHG with the adaptive stepsizes outperforms all the other numerical experiments, but for identical fixed parameters, PDHGMu performed slightly better than PDHG. However, for fixed α the stability requirement, $\delta < \frac{1}{\alpha\|D\|^2}$ for PDHGMu places an upper bound on δ which is empirically about four times less than for PDHG. Table 1 shows that for fixed α, PDHG with larger δ outperforms PDHGMu. The stability restriction for PDHGMu is also why the same adaptive time stepping scheme used for PDHG could not be used for PDHGMu.

Table 1 also demonstrates that larger α is more effective when the relative duality gap is large, and smaller α is better when this duality gap is small. Since PDHG for large α is similar to projected gradient descent, roughly speaking this means the adaptive PDHG algorithm starts out closer to being gradient projection on the dual problem, but gradually becomes more like a form of subgradient descent on the primal problem.

8.2 TV Deblurring Example

PDHGMu and PDHG also perform similarly for unconstrained TV deblurring (1). For this example we use the same cameraman image from the previous section and let K be a convolution operator corresponding to a normalized Gaussian blur with a standard deviation of 3 in a 17 by 17 window. Letting h denote the clean image, the given data f is taken to be $f = Kh + \eta$, where η is mean zero Gaussian noise with standard deviation 1. We set the fidelity parameter $\lambda = 100$. For the

Algorithm	tol $= 10^{-2}$	tol $= 10^{-4}$	tol $= 10^{-6}$
PDHG (adaptive)	14	70	310
PDHGMu (adaptive)	19	92	365
PDHG $\alpha = 5, \delta = .025$	31	404	8209
PDHG $\alpha = 1, \delta = .125$	51	173	1732
PDHG $\alpha = .2, \delta = .624$	167	383	899
PDHGMu $\alpha = 5, \delta = .025$	21	394	8041
PDHGMu $\alpha = 1, \delta = .125$	38	123	1768
PDHGMu $\alpha = .2, \delta = .624$	162	355	627
PDHG $\alpha = 5, \delta = .1$	22	108	2121
PDHG $\alpha = 1, \delta = .5$	39	123	430
PDHG $\alpha = .2, \delta = 2.5$	164	363	742
PDHGMu $\alpha = 5, \delta = .1$	unstable		
PDHGMu $\alpha = 1, \delta = .5$	unstable		
PDHGMu $\alpha = .2, \delta = 2.5$	unstable		
Proj. Grad. $\delta = .0132$	48	750	15860
ADMM $\delta = .025$	17	388	7951
ADMM $\delta = .125$	22	100	1804
ADMM $\delta = .624$	97	270	569

Table 1: Iterations Required for TV Denoising

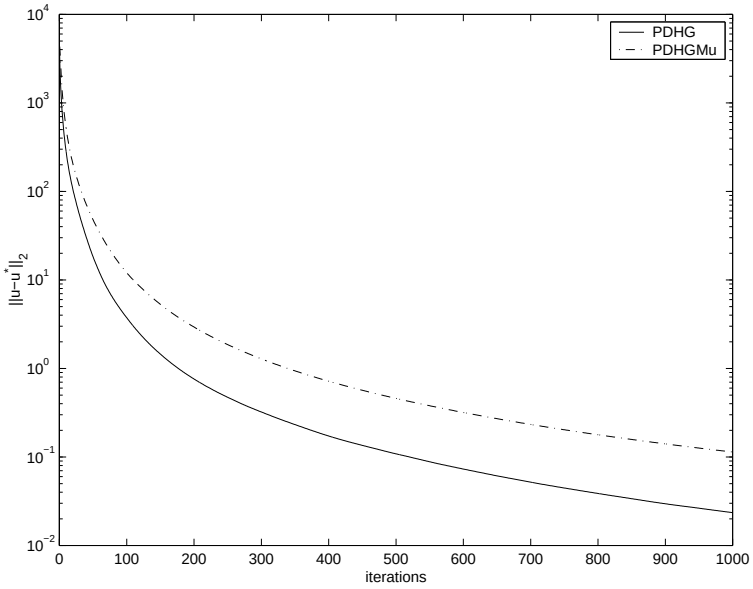


Figure 3: l_2 error versus iterations for denoising

numerical experiments we used the fixed parameter versions of PDHG and PDHGMu with $\alpha = .2$ and $\delta = .495$. The images h, f and the benchmark recovered image from 50000 iterations of PDHGMu are shown in Figure 4. Figure 5 compares the l_2 error to the benchmark solution as a

Figure 4: Original, blurry/noisy and recovered cameraman images

function of number of iterations for PDHG and PDHGMu. Empirically, with the same parameters, the performance of these two algorithms is nearly identical, and the curves are indistinguishable in Figure 5.

8.3 PDHGMu for Constrained l_1 Minimization

Here we compare PDHGMu (64) and the reversed role version, PDHGRMu (66), applied to the compressive sensing problem given by (63) with $\epsilon = .01$. Let $K = R\Gamma\Psi^T$, where R is a row selector, Γ is an orthogonal 2D discrete cosine transform and Ψ is orthogonal 2D Haar wavelet transform.

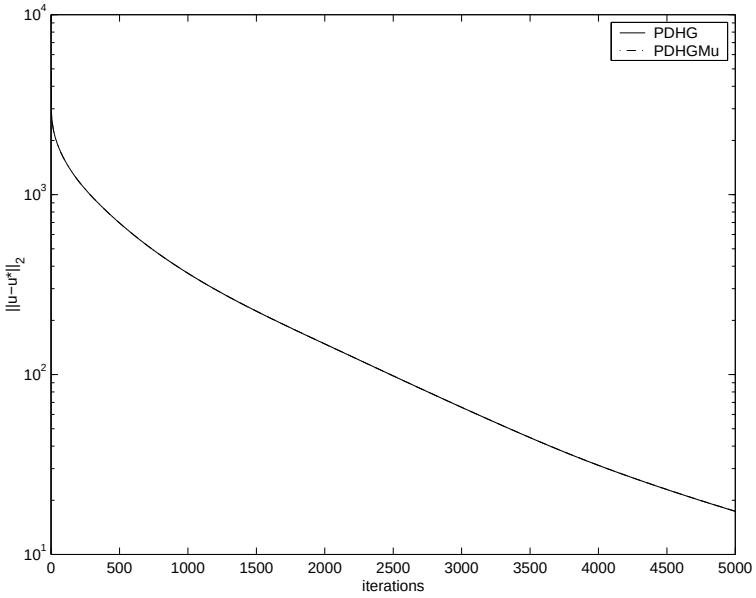


Figure 5: l_2 error versus iterations for deblurring

It follows that $KK^T = I$ and $K^\dagger = K^T$. R selects about ten percent of the DCT measurements, mostly low frequency ones. The constrained l_1 minimization model aims to recover a sparse signal in the wavelet domain that is consistent with these partial DCT measurements [7].

For the numerical experiments, we let $\alpha = 1$ and $\delta = 1$. That means that both versions of PDHGMu applied this problem can be interpreted as different applications of ADMM to ($\mathrm{SP_D}$), or equivalently Douglas Rachford splitting applied to (P). Let h denote the clean image, which is a 32 by 32 synthetic image shown in figure 6. The data f is taken to be $R\Gamma h$. For the initialization, let $p^0 = 0$ and let $u^0 = \Psi z^0$, where $z^0 = \Gamma^T R^T R\Gamma h$ is the backprojection obtained by taking the inverse DCT of f with the missing measurements replaced by 0. Let u^* denote the solution obtained by 25000 iterations of PDHGRMu. Figure 6 shows h, z^0 and z^*, where $z^* = \Psi^T u^*$.

Both versions of PDHGMu applied to this problem have simple iterations that scale like $O(m)$, but they behave somewhat differently. PDHGMu (64) by definition satisfies the constraint at each

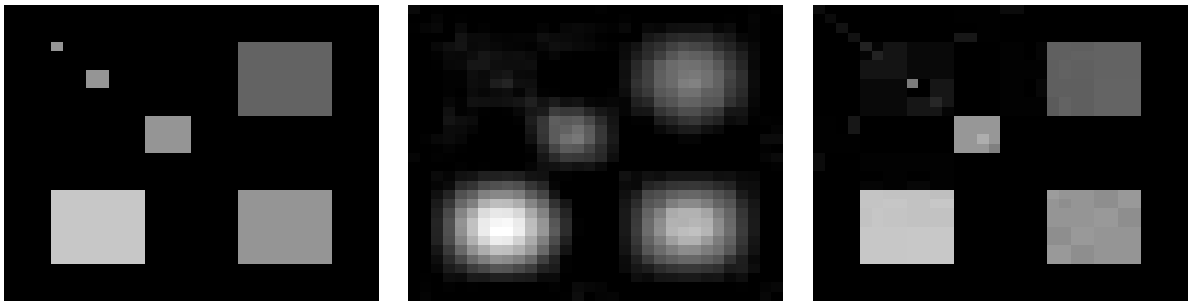
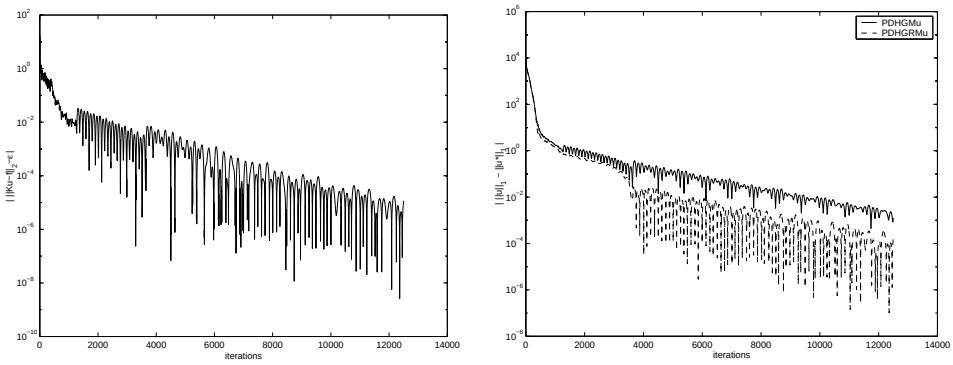


Figure 6: Original, damaged and recovered images



(a) Constraint versus iterations for PDHGRMu

(b) l_1 Norm Comparison

Figure 7: Comparison of PDHGRMu and PDHGMu

iteration. However, these projections onto the constraint set destroy the sparsity of the approximate solution so it can be a little slower to recover a sparse solution. PDHGRMu (66) on the other hand more quickly finds a sparse approximate solution but can take a long time to satisfy the constraint to a high precision.

To compare the two approaches, we compare plots of how the constraint and l_1 norm vary with iterations. Figure 7(a) plots $|\|Ku^k - f\|_2 - \epsilon|$ against the iterations k for PDHGRMu. Note this is always zero for PDHGMu, which stays on the constraint set. Figure 7(b) compares the differences $|\|u^k\|_1 - \|u^*\|_1|$ for both algorithms on a semilog plot, where $\|u^*\|_1$ is the l_1 norm of the benchmark solution. The empirical rate of convergence to $\|u^*\|_1$ was similar for both algorithms despite the many oscillations. PDHGRMu was a little faster to recover a sparse solution, but PDHGMu has the advantage of staying on the constraint set. For different applications with more complicated K, the simpler projection step for PDHGRMu would be an advantage of that approach.

Acknowledgements:
This work was supported by ONR N00014-03-1-0071, NSF DMS-0610079, NSF CCF-0528583 and NSF DMS-0312222. Thanks to Paul Tseng for pointing out some helpful references.

References

[1] Beck, A., and Teboulle, M., *A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems*, http://iew3.technion.ac.il/ becka/papers/revised-4.pdf, 2008.

[2] Becker, S., Bobin, J., and Candes, E. J., *NESTA: A Fast and Accurate First-Order Method for Sparse Recovery*, http://www.acm.caltech.edu/ emmanuel/papers/NESTA.pdf, 2009.

[3] Bertsekas, D., *Constrained Optimization and Lagrange Multiplier Methods*, Athena Scientific, 1996.

[4] Bertsekas, D., *Nonlinear Programming*, Athena Scientific, Second Edition. 1999.

[5] Bertsekas, D., and Tsitsiklis, J., *Parallel and Distributed Computation*, Prentice Hall, 1989.

[6] Boyd, S., and Vandenberghe, L., *Convex Analysis*, Cambridge University Press, 2006.

[7] Candes, E., Romberg, J., *Practical Signal Recovery from Random Projections*, IEEE Trans. Signal Processing, 2005.

[8] Candes, E., Romberg, J., and Tao, T., *Stable signal recovery from incomplete and inaccurate measurements*, Comm. Pure Appl. Math., 59, 1207-1223, 2005.

[9] Chambolle, A., *An Algorithm for Total Variation Minimization and Applications*, Journal of Mathematical Imaging and Vision, Vol. 20, pp. 89-97, 2004.

[10] Chan, T. F., Golub. G. H., and Mulet, P., *A nonlinear primal dual method for total variation based image restoration*, SIAM J. Sci. Comput., 20, 1999.

[11] Chen, G., and Teboulle, M., *A Proximal-Based Decomposition Method for Convex Minimization Problems*, Mathematical Programming, Vol., 64, pp. 81-101, 1994.

[12] Combettes, P., and Wajs, W., *Signal Recovery by Proximal Forward-Backward Splitting*, Multiscale Modelling and Simulation, 2006.

[13] Douglas, J., and Rachford, H. H., *On the Numerical Solution of Heat Conduction Problems in Two and Three Space Variables*, Transactions of the American mathematical Society 82, 1956, pp. 421-439.

[14] Eckstein, J., and Bertsekas, D., *On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators*, Mathematical Programming 55, North-Holland, 1992.

[15] Eckstein, J., *Splitting Methods for Monotone Operators with Applications to Parallel Optimization*, Ph. D. Thesis, Massachusetts Institute of Technology, Dept. of Civil Engineering, http://hdl.handle.net/1721.1/14356, 1989.

[16] Ekeland, I., and Temam, R. *Convex Analysis and Variational Problems*, SIAM, Classics in Applied Mathematics, 28, 1999.

[17] Elmoataz, A., Lezoray, O., and Bougleux, S., *Nonlocal Discrete Regularization on Weighted Graphs: A framework for Image and Manifold Processing*, IEEE, Vol. 17, No. 7, July 2008.

[18] Esser, E., *Applications of Lagrangian-Based Alternating Direction Methods and Connections to Split Bregman*, UCLA CAM Report [09-31], April 2009.

[19] Gabay, D., *Methodes numeriques pour l'optimisation non-lineaire*, These de Doctorat d'Etat et Sciences Mathematiques, Universite Pierre et Marie Curie, 1979.

[20] Gabay, D., and Mercier, B., *A dual algorithm for the solution of nonlinear variational problems via finite-element approximations*, Comp. Math. Appl., 2 (1976), pp. 17-40.

[21] Glowinski, R., and Le Tallec, P., *Augmented Lagrangian and Operator-splitting Methods in Nonlinear Mechanics*, SIAM 1989.

[22] Glowinski, R., and Marrocco, A., *Sur lapproximation par elements finis dordre un, et la resolution par penalisation-dualite dune classe de problemes de Dirichlet nonlineaires, Rev. Francaise dAut.* Inf. Rech. Oper., R-2 (1975), pp. 41-76.

[23] Goldfarb, D., and Yin, W., *Second-order Cone Programming Methods for Total Variation-Based Image Restoration*, SIAM J. Sci. Comput. Vol. 27, No. 2, (2005), pp. 622-645.

[24] Goldstein, T., and Osher, S., *The Split Bregman Algorithm for L1 Regularized Problems*, UCLA CAM Report [08-29], April 2008.

[25] Horn, R. A., and Johnson, C. R., *Matrix Analysis*, Cambridge University Press, 1985.

[26] Kim, S., Koh, K., Lustig, M., and Boyd, S., *An Interior-Point Method for Large-Scale l_1-Regularized Least Squares*, IEEE Journal of Selected Topics in Signal Processing, Vol. 1, No. 4, Dec. 2007.

[27] Larsson, F., Patriksson, M. and Stromberg,A.-B., *On the convergence of conditional ϵ-subgradient methods for convex programs and convex-concave saddle-point problems*, European Journal of operational Research, Vol. 151,No. 3, pp. 461 - 473, 2003.

[28] Lions, P., L., and Mercier, B., *Algorithms for the Sum of Two Nonlinear Operators*, SIAM Journal on Numerical Analysis, Vol. 16, No. 6, Dec., 1979, pp. 964-979.

[29] Malgouyres, F., and Zeng, T., *A Predual Proximal Point Algorithm Solving a Non Negative Basis Pursuit Denoising Model*, http://www.math.univ-paris13.fr/~malgouy/download/BP_Uzawa.pdf, 2008.

[30] Moreau, J. J., *Proximité et dualité dans un espace hilbertien*, Bull. Soc. Math. France, 93, 1965, pp. 273-299.

[31] Nesterov, Y., *Dual extrapolation and its applications to solving variational inequalities and related problems*, Math. Program., Ser. B, Vol. 119, pp. 319-344, 2007.

[32] Nesterov, Y., *Smooth Minimization of Non-Smooth Functions*, Mathematic Programming, Ser. A, No. 103, pp. 127-152, 2005.

[33] Nocedal, J., and Wright, S., *Numerical Optimization*, Springer, 1999.

[34] Passty, G. B., *Ergodic Convergence to a Zero of the Sum of Monotone Operators in Hilbert Space*, Journal of Mathematical Analysis and Applications, 72, 1979, pp. 383-390.

[35] Pock, T., Cremers, D., Bischof, H., and Chambolle, A., *An Algorithm for Minimizing the Mumford-Shah Functional*, ICCV, 2009.

[36] Popov, L., *A Modification of the Arrow-Hurwicz Method for Search of Saddle Points*, Mathematical Notes, 28(5), 1980, pp. 845-848.

[37] Rockafellar, R. T, *Augmented Lagrangians and Applications of the Proximal Point Algorithm in Convex Programming*, Mathematics of Operations Research, Vol. 1, No. 2, 1976, pp. 97-116.

[38] Rockafellar, R., T., *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[39] Rockafellar, R. T., *Monotone Operators and the Proximal Point Algorithm*, SIAM J. Control and Optimization, Vol. 14, No. 5, 1976.

[40] Rudin, L., Osher, S., and Fatemi, E., *Nonlinear Total Variation Based Noise Removal Algorithms*, Physica D, 60, 1992, pp. 259-268.

[41] Setzer, S., *Split Bregman Algorithm, Douglas-Rachford Splitting and Frame Shrinkage*, http://kiwi.math.uni-mannheim.de/~ssetzer/pub/setzer_fba_fbs_frames08.pdf, 2008.

[42] Shor, N. Z., Kiwiel, K. C., and Ruszcyǹski, *Minimization Methods for Non-Differentiable Functions*, Springer-Verlag New York, Inc. 1985.

[43] Tseng, P., *Alternating Projection-Proximal Methods for Convex Programming and Variational Inequalities*, SIAM J. Optim., Vol. 7, No. 4, pp. 951-965, Nov. 1997.

[44] Tseng, P., *Applications of a Splitting Algorithm to Decomposition in Convex Programming and Variational Inequalities*, SIAM J. Control and Optimization, Vol. 29, No. 1, pp. 119-138, Jan. 1991.

[45] Tseng, P., *On Accelerated Proximal Gradient Methods for Convex-Concave Optimization*, 2008.

[46] Weiss, P., Aubert, G., and Blanc-Féraud, L., *Efficient Schemes for Total Variation Minimization Under Constraints in Image Processing*, INRIA, No. 6260, July 2007.

[47] Yin, W., *Analysis and Generalizations of the Linearized Bregman Method*, UCLA CAM Report [09-42], May 2009.

[48] Yin, W., Osher, S., Goldfarb, D., Darbon, J., *Bregman Iterative Algorithms for l_1-Minimization with Applications to Compressed Sensing*, UCLA CAM Report [07-37], 2007.

[49] Zhang, X., *A Unified Primal-Dual Algorithm Based on l_1 and Bregman Iteration*, (Private Communication), April 2009.

[50] Zhang, X., Burger, M., Bresson, X., Osher, S., *Bregmanized Nonlocal Regularization for Deconvolution and Sparse Reconstruction*, UCLA CAM Report [09-03] 2009.

[51] Zhu, M., and Chan, T.F., *An Efficient Primal-Dual Hybrid Gradient Algorithm for Total Variation Image Restoration*, UCLA CAM Report [08-34], May 2008.

[52] Zhu, M., Wright, S.J., and Chan, T.F., *Duality-Based Algorithms for Total-Variation-Regularized Image Restoration*, Computational Optimization and Applications, Springer Netherlands, 2008.