Adaptive Primal Dual Optimization for Image Processing and Learning

Tom Goldstein
Rice University
tag7@rice.edu

Ernie Esser
University of British Columbia
eesser@eos.ubc.ca

Richard Baraniuk
Rice University
richb@rice.edu

Abstract

The Primal-Dual Hybrid Gradient method is a powerful splitting scheme for large-scale constrained and non-differentiable problems. We present practical adaptive variants of PDHG that converge more quicky and are easier to use than conventional splitting schemes. We also study the convergence of PDHG, and prove new results guaranteeing convergence of the method when adaptivity is used properly.

1 Introduction

This manuscript considers saddle-point optimization problems of form

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

where f and g are convex functions, $A \in \mathbb{R}^{M\times N}$ is a matrix, and $X \subset \mathbb{R}^N$ and $Y \subset \mathbb{R}^M$ are convex sets.

The formulation (1) is extremely useful for solving inverse problems involving the ℓ_1 norm and Total-Variation (TV). Many common minimization problems have a simple saddle-point form including image segmentation, TVL1 minimization, and general linear programing.

The Primal-Dual Hybrid Gradient (PDHG) [1, 2] solves (1) efficiently by addressing the terms f and g separately. One of the primary difficulties with PDHG is that it relies on step-size parameters that must be carefully chosen by the user. The speed of the method depends heavily on the choice of these parameters, and there is often no intuitive way to choose them.

We present practical adaptive schemes that optimize the convergence of PDHG automatically as the problem is solved. The new methods are not only easier to use in practice, but also result in faster convergence than conventional schemes. After introducing the adaptive methods, we prove new theoretical results that guarantee convergence of PDHG under very general circumstances, including adaptive stepsizes.

2 The Primal-Dual Hybrid Gradient Method

The PDHG method [3, 1, 4, 2] is listed in Algorithm 1. In steps (2-3), the method decreases the energy (1) in x by first taking a gradient descent step with respect to the inner product term in (1), and then taking a "backward" or proximal step for f. In steps (5-6), the energy (1) is increased by changing y. A gradient ascent step is taken with respect to the inner product term, and then a backward step is taken with respect to g.

Algorithm 1 Basic PDHG

```
Require: $x_0 \in \mathbb{R}^N$, $y_0 \in \mathbb{R}^M$, $\sigma_k, \tau_k > 0$
1: while Not Converged do
2:     $\hat{x}_{k+1} = x_k - \tau_k A^T y_k$                                   ▷ Forward descent
3:     $x_{k+1} = \arg\min_{x \in X} f(x) + \frac{1}{2\tau_k}\|x - \hat{x}_{k+1}\|^2$   ▷ Backward descent
4:     $\bar{x}_{k+1} = x_{k+1} + (x_{k+1} - x_k)$                              ▷ Prediction step
5:     $\hat{y}_{k+1} = y_k + \sigma_k A\bar{x}_{k+1}$                            ▷ Forward ascent
6:     $y_{k+1} = \arg\min_{y \in Y} g(y) + \frac{1}{2\sigma_k}\|y - \hat{y}_{k+1}\|^2$  ▷ Backward ascent
7: end while
```

Steps 3 and 6 of Algorithm 1 can be written compactly using *proximal* operators of f/g :

$$J_{\tau F}(\hat{x}) = \underset{x \in X}{\arg\min}\, f(x) + \frac{1}{2\tau}\|x - \hat{x}\|^2 \tag{2}$$

where $F = \partial f$. Algorithm 1 is convergent with constant stepsizes satisfying $\sigma\tau < \frac{1}{\rho(A^T A)}$ [1, 2, 4]. However, PDHG does not converge when non-constant stepsizes are used, even in the case that $\sigma_k \tau_k < \frac{1}{\rho(A^T A)}$. In this article, we identify the specific stepsize conditions that guarantee convergence and propose practical adaptive methods satisfying these conditions.

3 Common Saddle-Point Problems

While the applications of saddle-point problems are vast, we focus here on several problems from statistics, image processing and signal processing.

3.1 Total-Variation Denoising

A common application of Total-Variation (TV) is the Rudin-Osher-Fatemi (ROF) denoising model [5]:

$$\min_x |\nabla x| + \frac{\mu}{2}\|x - f\|^2. \tag{3}$$

A noise contaminated image f is denoised by recovering an image x that is similar to y in the ℓ_2-sense, while having small TV.

The TV term can be written as a maximization over the "dual" variable $y \in \mathbb{R}^{2\times N}$, where the image $x \in \mathbb{R}^N$ has N pixels. The equation (3) then becomes

$$\max_{y \in C_\infty} \min_x \frac{\mu}{2}\|x - f\|^2 + y \cdot \nabla x \tag{4}$$

which is clearly of the form (1). To apply Algorithm 1, we need efficient solutions to the sub-problems in steps 3 and 6, which can be written

$$J_{\tau F}(\hat{x}) = \underset{x}{\arg\min}\, \frac{\mu}{2}\|x - f\|^2 + \frac{1}{2\tau}\|x - \hat{x}\|^2 = \frac{\tau}{\tau\mu + 1}(\mu f + \frac{1}{\tau}\hat{x}) \tag{5}$$

$$J_{\sigma G}(\hat{y}) = \underset{y \in C_\infty}{\arg\min}\, \frac{1}{2\sigma}\|y - \hat{y}\|^2 = \left(\frac{y_i}{\max\{y_i, 1\}}\right)_{i=1}^{M}. \tag{6}$$

3.2 Scaled Lasso

The square-root lasso [6] (or equivalently the scaled lasso [7]) is a variable selection regression that obtains sparse solutions to systems of linear equations. Given a data matrix D and a vector b, a sparse solution to the system $Dx = b$ is obtained by solving

$$\min_x |x| + \lambda\|Dx - b\| \tag{7}$$

Note that the ℓ_2 term in (7) is not squared as in the conventional lasso model. Using techniques from Section 3.1 we can write this energy as

$$\max_{\|y_1\|_\infty \le 1, \|y_2\| \le \lambda} \min \langle y_1, x\rangle + \langle y_2, Dx - b\rangle \tag{8}$$

which can be solved using PDHG.

3.3 Other Applications

We refer the reader to [8] for a discussion of various saddle-point formulations including compressed sensing for single-pixel cameras, TVL1 image restoration, image segmentation, ℓ_∞ minimization, and linear programming.

4 Convergence Theory

We begin by defining the following constants which quantify the relative change between stepsizes:

$$\delta_k = \min\left\{\frac{\tau_{k+1}}{\tau_k}, \frac{\sigma_{k+1}}{\sigma_k}, 1\right\}, \quad \text{and} \quad \phi_k = 1 - \delta_k \geq 0. \tag{9}$$

We can now state our main result. For a proof, see [8].

Theorem: Algorithm 1 converges if the following three requirements hold:

A The sequences $\{\tau_k\}$ and $\{\sigma_k\}$ are bounded.

B The sequence $\{\phi_k\}$ is summable, i.e. $\sum_{k\geq 0} \phi_k < \infty$.

C One of the following two conditions is met:

C1 There is a constant L such that for all $k > 0$

$$\tau_k \sigma_k < L < \rho(A^T A)^{-1}.$$

C2 Either X or Y is bounded, and there is a constant $c \in (0, 1)$ such that for all $k > 0$

$$c\sigma_k \|x_{k+1} - x_k\|^2 + c\tau_k \|y_{k+1} - y_k\|^2 \geq 2\tau_k\sigma_k \langle A(x_{k+1} - x_k), y_{k+1} - y_k \rangle.$$

Two conditions must always hold to guarantee convergence: **A**) The stepsizes must remain bounded, and **B**) the sequence $\{\phi_i\}$ must be summable. Together, these conditions ensure that the steps sizes do not oscillate too wildly as k gets large. In addition, either **C1** or **C2** must hold. When we know the spectral radius of $A^T A$, we can use condition **C1** to enforce that the method is stable. Otherwise, the backtracking condition **C2** can be used to enforce stability. It can be easily seen that Algorithm 2 satisfies the above convergence conditions. This is elaborated in [8].

4.1 Adaptive PDHG

The first adaptive method is listed in Algorithm 2. The loop in Algorithm 2 begins with a standard PDHG step. Steps 4 and 5 compute the primal and dual residuals. If the primal residual is large compared to the dual, τ is increased by a factor of $(1-\alpha_k)^{-1}$, and the σ decreased by $(1-\alpha_k)$. If the primal residual is small compared to the dual, then τ is decreased and σ is increased. The stepsizes are only updated if the residuals differ by a factor greater then Δ. The sequence $\{\alpha_k\}$ controls the adaptivity level. When we update τ/σ, we multiply α by $\eta < 1$. In this way the adaptivity decreases over time and thus fulfills condition (B) of our convergence theorem.

We have found that $a_0 = 0.5$, $\Delta = 1.5$, and $\eta = 0.95$ is a fairly robust choice for the constants in Algorithm 2.

4.2 Backtracking PDHG

When $\|A\|$ is unknown, the backtracking condition **C2** can be used to enforce stability. See [8].

5 Numerical Results

We test the adaptive PDHG method using the ROF denoising problem (3). Numerical results for additional applications from imaging and machine learning are presented in [8]. We consider four

Algorithm 2 Adaptive PDHG

```
Require: $x_0 \in \mathbb{R}^N, y_0 \in \mathbb{R}^M, \sigma_0\tau_0 < \rho(A^TA)^{-1}, (\alpha_0, \eta) \in (0,1)^2, \Delta > 1, s > 0$
while $p_k, d_k >$ tolerance do
    $x_{k+1} = J_{\tau_k F}(x_k - \tau_k A^T y_k)$                          ▷ Begin with normal PDHG
    $y_{k+1} = J_{\sigma_k G}(y_k + \sigma_k A(2x_{k+1} - x_k))$
    $p_{k+1} = |(x_k - x_{k+1})/\tau_k - A^T(y_k - y_{k+1})|$              ▷ Compute primal residual
    $d_{k+1} = |(y_k - y_{k+1})/\sigma_k - A(x_k - x_{k+1})|$              ▷ Compute dual residual
    if $p_{k+1} > sd_{k+1}\Delta$ then                                   ▷ If primal residual is large...
        $\tau_{k+1} = \tau_k/(1 - \alpha_k)$                              ▷ Increase primal stepsize
        $\sigma_{k+1} = \sigma_k(1 - \alpha_k)$                           ▷ Decrease dual stepsize
        $\alpha_{k+1} = \alpha_k\eta$                                     ▷ Decrease adaptivity level
    end if
    if $p_{k+1} < sd_{k+1}/\Delta$ then                                  ▷ If dual residual is large...
        $\tau_{k+1} = \tau_k(1 - \alpha_k)$                               ▷ Decrease primal stepsize
        $\sigma_{k+1} = \sigma_k/(1 - \alpha_k)$                          ▷ Increase dual stepsize
        $\alpha_{k+1} = \alpha_k\eta$                                     ▷ Decrease adaptivity level
    end if
    if $sd_{k+1}/\Delta \le p_{k+1} \le sd_{k+1}\Delta$ then             ▷ If residuals are similar...
        $\tau_{k+1} = \tau_k$,$\sigma_{k+1} = \sigma_k$, $\alpha_{k+1} = \alpha_k$    ▷ Leave stepsizes the same
    end if
end while
```

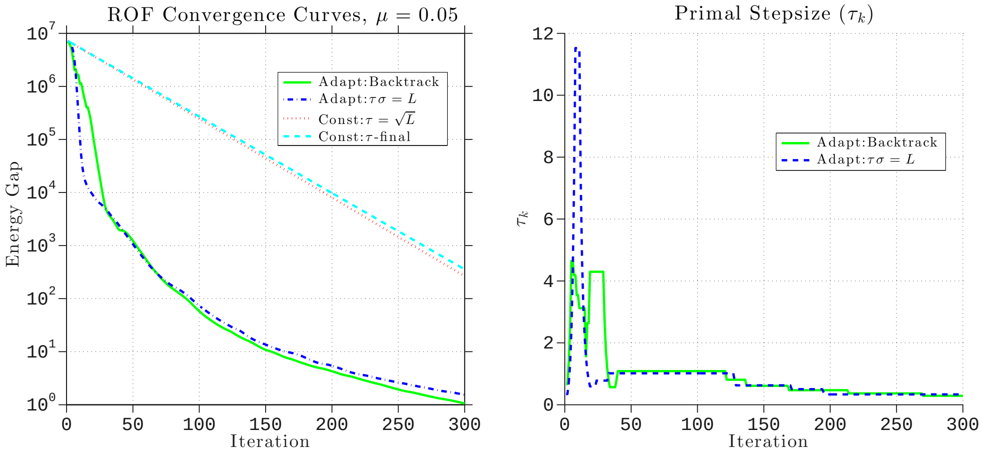

Figure 1: (left) Convergence curves for the Rudin-Osher-Fatemi denoising experiment with $\mu = 0.05$. The y-axis displays the difference between the value of the ROF objective function (3) at the kth iterate and the optimal objective value. (right) Stepsize sequences, $\{\tau_k\}$, for both adaptive schemes.

variants of PDHG. "Adapt:Backtrack" is Algorithm 2 with backtracking. The method "Adapt: $\tau\sigma = L$" is adaptive PDHG without backtracking. "Const: $\tau, \sigma = \sqrt{L}$" is the non-adaptive method with $\tau = \sigma = \rho(A^TA)^{-\frac{1}{2}}$. "Const: τ-final" refers to the constant-stepsize method, where the stepsizes are chosen to be the final values of the stepsizes used by "Adapt: $\tau\sigma = L$". This final method is meant to demonstrate the performance on PDHG with a stepsize that is customized to the problem instance at hand, but still non-adaptive. Note the superior performance of the adaptive methods. Because the optimal stepsize choice depends on the active set (which evolves over time) the adaptive scheme is even able to out-perform the "tuned" stepsize choice used by "Const: τ-final".

References

[1] E. Esser, X. Zhang, and T. Chan, "A general framework for a class of first order primal-dual algorithms for TV minimization," *UCLA CAM Report 09-67*, 2009.

[2] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *Convergence*, vol. 40, no. 1, pp. 1–49, 2010.

[3] M. Zhu and T. Chan, "An efficient primal-dual hybrid gradient algorithm for total variation image restoration," *UCLA CAM technical report, 08-34*, 2008.

[4] B. He and X. Yuan, "Convergence analysis of primal-dual algorithms for a saddle-point problem: From contraction perspective," *SIAM J. Img. Sci.*, vol. 5, pp. 119–149, Jan. 2012.

[5] L. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica. D.*, vol. 60, pp. 259–268, 1992.

[6] A. Belloni, V. Chernozhukov, and L. Wang, "Square-root lasso: pivotal recovery of sparse signals via conic programming," *Biometrika*, vol. 98, no. 4, pp. 791–806, 2011.

[7] T. Sun and C.-H. Zhang, "Scaled sparse linear regression," *Biometrika*, vol. 99, no. 4, pp. 879–898, 2012.

[8] T. Goldstein, E. Esser, and R. Baraniuk, "Adaptive Primal-Dual Hybrid Gradient Methods for Saddle-Point Problems," *Available at Arxiv.org (arXiv:1305.0546)*, 2013.