A first-order primal-dual method with adaptivity to local Lipschitz smoothness

Anonymous Author(s)
Affiliation
Address
email

Abstract

We consider the problem of finding a saddle point for the convex-concave objective $\min_x \max_y f(x) + \langle Ax, y\rangle - g^*(y)$, where f is a convex differentiable function with locally-Lipschitz gradient and g is convex and possibly non-smooth. We propose an adaptive version of the celebrated Chambolle-Pock method, which alternates between primal gradient steps and dual proximal steps. The method achieves stepsize adaptivity through a simple rule involving $\|A\|$ and the local gradients of f and, under standard assumptions, we prove an $\mathcal{O}(k^{-1})$ ergodic convergence rate. Furthermore, when f is also locally strongly convex and A has full row rank, a minor change in the basic stepsize rule allows the method to achieve faster convergence. Preliminary numerical experiments are provided for illustrating the practical performance of the method.

1 Introduction

In this paper we study a particular instance of the composite minimization problem

$$\min_{x\in\mathcal{X}} f(x) + g(Ax), \tag{1}$$

where f and g are convex, proper and lower-semicontinuous (l.s.c.), and A is a linear operator.

Problems of the form (1) have been studied in the literature under various assumptions on f and g. For specific instances of A the function $h(x) := g(Ax)$ may be prox-friendly and if, in addition, f is L-smooth, the objective is suitable for applying forward-backward splitting algorithms. The most popular in this category are the Proximal Gradient algorithm and its accelerated counterpart [Nesterov, 2013, Beck and Teboulle, 2009]. When $h(x)$ does not have an efficient proximal operator and g is nonsmooth, however, A is usually decoupled from g and problem (1) is reformulated as the convex-concave saddle-point problem (2):

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \langle Ax, y\rangle + f(x) - g^*(y), \tag{2}$$

where g^* denotes the Fenchel conjugate of g. Objective (2) is typically addressed by primal-dual splitting algorithms which, under strong duality, can recover the solution to the original problem (1). In the particular case when f and g are prox-friendly and possibly non-smooth, the most popular optimization algorithm is arguably the Primal-Dual Hybrid Gradient (PDHG) proposed in [Chambolle and Pock, 2011], which was further extended to handle an additional L-smooth component in [Condat, 2013, Vũ, 2013] with convergence rates provided in [Chambolle and Pock, 2016a].

Together, these classes of algorithms cover a broad range of problems in diverse fields such as signal processing, machine learning, inverse problems, telecommunications and many others. As a result, a great amount of research effort has gone into addressing practical concerns such as robustness

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

to inexact oracles, acceleration and automation of stepsize selection, leading to numerous variants of these algorithms. For a comprehensive list of examples and theoretical details we refer the reader to review papers [Goldstein et al., 2014, Combettes and Pesquet, 2011, Parikh and Boyd, 2014, Komodakis and Pesquet, 2015, Chambolle and Pock, 2016b]. In this work, we focus on the latter line of investigation – that of stepsize regime automation, for primal-dual algorithms targeting problem (2).

In their basic form, primal-dual methods require as input a stepsize parameter belonging to a designated interval of stability, which depends on problem specific constants like the global smoothness parameter and $\| A \|$. Dependence on such constants is undesirable because they may be costly to compute and often can only access upper-bound estimates. This leaves the practitioner with the task of painstakingly searching for the optimal stepsize, which even so may be overly-conservative and thus impact convergence speed. Finally, when one of the terms is L-smooth, these methods rely on the knowledge of L for stepsize selection which prevents them from handling functions that do not satisfy this assumption globally.

Consequently, recent efforts have been directed towards devising methods with stepsizes that adapt to the local geometry on-the-go [Goldstein et al., 2013, 2015, Malitsky and Pock, 2018, Pedregosa and Gidel, 2018, Chang et al., 2021]. These approaches resort to linesearch procedures for finding larger stepsizes, and exhibit improved practical performance. It thus becomes apparent that better convergence comes at the cost of additional computation and an indeterminate number of extra steps spent in subprocedures at every iteration.

In this work, we address a particular instance of problem (2), in which f is locally smooth (or ∇f is locally Lipschitz) and g is prox-friendly. To illustrate the motivation of our framework, we take the following example often encountered in image processing settings:

$$\min_x \frac{1}{2}\| Kx - b \|^2 + \lambda \| Dx \|_{2,1}, \quad K \in \mathbb{R}^{m\times d}, D \in \mathbb{R}^{2d\times d},$$

where x is 2D image, K is a problem-specific measurement operator, b is the vector of (possibly noisy) observations and D is the discrete gradient operator and the regularization term represents the isotropic TV norm. In order to apply any of the aforementioned primal-dual algorithms, one needs to make a choice about how to decouple the linear operators. There are three options: decoupling with respect to K leaves us with having to compute the proximal operator of the TV norm for the primal step, which is an iterative procedure [Chambolle, 2004]. Decoupling D implies performing gradient steps on f, since in general its proximal operator is not efficient. Finally, decoupling with respect to both implies increasing the dimensionality of the dual variable to $m + 2d$, which is problematic for large d and m. The sensible choice is the second one (i.e., decoupling D), and the question we seek to answer with this work is

> *Under the aforementioned regularity assumptions, does there exist a method for addressing* (2) *that adapts to the local problem geometry without resorting to an iterative sub-procedure?*

Our contribution is to propose a first-order primal-dual scheme that answers this question in the affirmative and is accompanied by theoretical convergence guarantees. Using standard analysis techniques we show an ergodic convergence of $\mathcal{O}(k^{-1})$ when ∇f is *locally* Lipschitz and g is prox-friendly, and a linear convergence rate for the case of f being in addition *locally* strongly convex and A being full row rank. We provide proof-of-concept experiments for the image inpainting problem, as well as use our method as a heuristic for TV-regularized nonconvex phase retrieval.

The rest of the paper is structured as follows: Section 2 provides details about related work; Section 3 introduces notation, along with technical preliminaries and assumptions to be used in our analysis; Section 4 reports the main theoretical results alongside partial proofs; finally, numerical results are provided in Section 5.

2 Related Work

The result presented in this paper lie at the intersection of two lines of work, which we now describe.

Adaptive gradient methods Arguably the simplest and most widespread of optimization methods, Gradient Descent (GD), has attracted a lot of interest for addressing the issue of better stepsize

selection. For smooth functions, the stepsize τ of vanilla GD inversely depends on the global smoothness constant L which may yield overly conservative updates. Various solutions have been proposed to better adapt to the local geometry, usually involving either non-negligible additional computation (linesearch), or the knowledge of quantities such as the optimal value $f(x^*)$ ([Polyak, 1969]). In a recent work, Malitsky and Mishchenko [2020] propose an extremely simple and effective alternative. Their stepsize selection regime is:

$$\tau_k = \min\left\{\tau_{k-1}\sqrt{1+\frac{\tau_k}{\tau_{k-1}}}, \frac{\| x_k - x_{k-1} \|}{2\| \nabla f(x_k) - \nabla f(x_{k-1}) \|}\right\}. \tag{3}$$

Adaptivity essentially comes 'for free' in (3), as it involves solely quantities which have already been computed. Moreover, the analysis in [Malitsky and Mishchenko, 2020] does not require a global smoothness constant, thus extending the reach of provably-convergent GD to a wider class of differentiable functions. The method enjoys the well known $\mathcal{O}(k^{-1})$ convergence rate.

In the present work, we show that this technique can be extended to the analysis of primal-dual methods to achieve an algorithm whose stepsizes adapt to the local geometry of the objective's (locally) smooth component.

Adaptive first-order primal-dual methods As metioned in Section 1, there is a strong incentive to develop primal-dual methods with automated stepsize selection. Goldstein et al. [2013] and Goldstein et al. [2015] propose a heuristic criterion for varying the stepsizes, which states that maintaining a balance between the magnitude of the primal and dual residuals throughout iterations may yield improved convergence behavior. A potentially very costly linesearch procedure is proposed to this end, which requires updates in both the primal and dual variables.

Malitsky and Pock [2018], in turn, propose a linesearch procedure involving only dual variable updates and which, for certain problems such as regularized least squares does not require any additional matrix-vector multiplications. Their method, PDAL, can handle objectives composed of up to three terms, of which one is L-smooth. A characteristic of this algorithm is that it maintains a constant ratio between primal and dual stepsizes through a hyperparameter β: $\sigma_k = \beta\tau_k \ \forall k$. Our approach uses a similar technique, as we shall describe in Section 4.

3 Preliminaries

Consider problem (2) and let $\mathcal{X}, \mathcal{Y}$ be finite dimensional real vector spaces equipped with the standard inner product $\langle\cdot,\cdot\rangle$ and the associated Euclidean norm $\|\cdot\| = \sqrt{\langle\cdot,\cdot\rangle}$. We denote by g^* the Fenchel conjugate of g in (1) defined as $g^*(y) = \sup_{x\in\mathrm{dom}(g)}\{\langle x, y\rangle - g(x)\}$. We use A^* to denote the adjoint operator of A.

One can easily see that (2) is a primal-dual formulation of the following primal and dual optimization problems, of which the former is the same as (1):

$$\min_{x\in\mathcal{X}} f(x) + g(Ax), \qquad \max_{y\in\mathcal{Y}} -(f^*(-A^*y) + g^*(y)).$$

A solution $(x^*, y^*) \in \mathcal{X}\times\mathcal{Y}$ of problem (2) satisfies the following optimality conditions:

$$-A^*y^* = \nabla f(x^*), \qquad Ax^* \in \partial g^*(y^*). \tag{4}$$

Let $(x^*, y^*) \in \mathcal{X}\times\mathcal{Y}$ be a saddle-point of (2). We define the following quantities for which the saddle-point property ensures non-negativity:

$$P_{x^*,y^*}(x) := f(x) - f(x^*) + \langle x - x^*, A^*y^*\rangle \geq 0,$$
$$D_{x^*,y^*}(y) := g^*(y) - g^*(y^*) - \langle Ax^*, y - y^*\rangle \geq 0.$$

Moreover, we define $\mathcal{G}_{x^*,y^*}(x,y) := P_{x^*,y^*}(x) + D_{x^*,y^*}(y)$ as the primal-dual gap relative to (x^*, y^*), which is itself non-negative. Note that for fixed (x^*, y^*), $P_{x^*,y^*}(\cdot)$, $D_{x^*,y^*}(\cdot)$ and $\mathcal{G}_{x^*,y^*}(\cdot,\cdot)$ are convex.

Given a function f and $L > 0$, we say that f is L-smooth is its gradient ∇f is Lipschitz continuous over $\mathrm{dom}(f)$: $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \forall x, y$. Furthermore, f is locally smooth if ∇f is Lipschitz continuous on any compact subset of $\mathrm{dom}(f)$. In particular, in the context of

consecutive iterates x_{k-1}, x_k of the algorithm, the associated local Lipschitz constant is given by $L_k := \frac{\| \nabla f(x_k) - \nabla f(x_{k-1}) \|}{\| x_k - x_{k-1} \|}$.

We also say that f is μ-strongly convex if $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \| x - y \|^2$. Similarly, f is locally strongly convex if it is strongly convex on any compact subset of $\mathrm{dom}(f)$.

We define the proximal operator of a convex function g as $\mathrm{prox}_g(x) = \arg\min_{z \in \mathrm{dom}(g)} \left\{ g(z) + \frac{1}{2} \| x - z \|^2 \right\}$, and say that g is 'prox-friendly' if $\mathrm{prox}_g(x)$ has a closed form solution or can be efficiently computed to high accuracy.

Finally, the two following blanket assumptions will hold throughout the paper:

Assumption 3.1. *Function f is differentiable and ∇f is locally Lipschitz, while g possibly non-smooth and prox-friendly.*

Assumption 3.2. *We assume that a saddle-point exists for problem* (2) *and thus strong duality holds.*

Note that the first part of Assumption 3.1 is much less restrictive than the regular global L-smoothness premise, and thus enlarges the category of admissible functions with examples such as: $x \mapsto \log(x)$, $\exp(x)$, $\tan(x)$ and arbitrary convex polynomials. Assumption 3.2, on the other hand, is standard in the literature (see e.g., [Chambolle and Pock, 2011]).

Sometimes, we make the additional assumption that

Assumption 3.3. *Function f is locally strongly convex and operator A has full row-rank.*

4 Algorithm and convergence

Algorithm 1 A(daptive)P(primal)D(ual)A(lgorithm)

Input: $x_0 \in \mathcal{X}, y_0 \in \mathcal{Y}, \tau_{\text{init}} > 0, \tau_0 = \infty, \theta_0 = 1, \beta > 0, c \in (0, 1)$
$x_1 = x_0 - \tau_{\text{init}}(\nabla f(x_0) + A^T y_0)$
for $k = 1, 2, \ldots$ **do**
  Set $\tau_k = \min \left\{ \frac{1}{2\sqrt{L_k^2 + (\beta/(1-c))\| A \|^2}}, \tau_{k-1}\sqrt{1 + \theta_{k-1}} \right\}$, $\sigma_k = \beta \tau_k$, $\theta_k = \frac{\tau_k}{\tau_{k-1}}$
  $\tilde{x}_k = x_k + \theta_k(x_k - x_{k-1})$
  $y_{k+1} = \mathrm{prox}_{\sigma_k g^*}(y_k + \sigma_k A \tilde{x}_k)$
  $x_{k+1} = x_k - \tau_k(\nabla f(x_k) + A^T y_{k+1})$
end for

4.1 High level ideas

The primal-dual method proposed for solving problem (2) under assumptions 3.1 and 3.2, is provided in Algorithm 1 under the abbreviation APDA, which we use from here onwards. APDA follows the same structure as the basic PDHG [Chambolle and Pock, 2016a] for the given assumptions. Notice that if we restrict Assumption 3.1 to L-smooth functions f, we can in fact recover PDHG by setting $\theta_k = \theta = 1$ and $\tau_k = \tau$, $\sigma_k = \sigma$ fixed such that $\left(\frac{1}{\tau} - L\right) \frac{1}{\sigma} \geq \| A \|^2$. In this context, Chambolle and Pock [2016b] suggest setting the stepsizes by finding a constant $p > 0$ such that $\tau = \frac{1}{L/p + \| A \|}$ and $\sigma = \frac{1}{p \| A \|}$ satisfy the aforementioned condition. We can observe how p enforces a trade-off between balancing out a potentially large Lipschitz constant and taking big-enough steps in the dual update.

Our algorithm, on the other hand, disposes of the global condition of [Chambolle and Pock, 2016a] and relies instead on a criterion dictated by $\| A \|$ and the local geometry of f:

$$\tau_k \leq \frac{1}{2\sqrt{L_k^2 + (\beta/(1-c))\| A \|^2}}, \forall k, \tag{5}$$

where L_k is the local Lipschitz constant defined in the previous section. Here, $\beta = \frac{\sigma_k}{\tau_k}$ corresponds to the fixed ratio of PDHG stepsizes, playing a similar role to p above. Intuitively, this rule demands

that τ_k does not overstep a constant related to the local curvature, thus allowing for adaptive and potentially bigger stepsizes, while also removing the requirement of L-smoothness. Moreover, we have a free choice with respect to β, as our stepsizes need not satisfy a global stability condition as in the case of PDHG.

By itself, the local condition (5) does not ensure convergence, since overly-aggressive and possibly destabilizing stepsizes might occur in near-linear regions. This issue is addressed by the second term of the minimization defining τ_k, which enforces, informally, that the stepsize does not increase 'too fast' in consecutive iterations. Specifically, the increase factor is at most $\sqrt{1+\theta_k}$, where $\theta_k = \frac{\tau_k}{\tau_{k-1}}$.

Remark 4.1. *While τ_k does not adapt to $\| A \|$, for many practical problems this fact is not a big hindrance. Function f typically represents the data fidelity term, whose smoothness constant L (should it exist) can far exceed $\| A \|$ – the linear operator enforcing structured regularization on x. A specific example are TV-regularized imaging problems, where A is the discrete gradient operator whose norm is bounded by $\sqrt{8}$ [Chambolle, 2004], while the data fidelity term may involve a very large number of measurements and a larger norm, consequently.*

Remark 4.2. *APDA takes an additional primal step prior to the for-loop, which is controlled by τ_{init} given as input. This is needed for estimating L_1 in the first iteration. In practice we set $\tau_{init} = 10^{-9}$, a sufficiently small value to ensure that x_1 does not depart too far from x_0 and yield a good estimate of L_1. Furthermore, the setting of $\tau_0 = \infty$ simply ensures that in the first step, $\tau_1 = \frac{1}{2\sqrt{L_1^2 + (\beta/(1-c))\| A \|^2}}$ and has no impact on further steps. Finally, in our experiments we set $c =$ `1e-15` – this is a parameter introduced for theoretical purposes as explained in the following section.*

4.2 Analysis – the base case

In short, the main steps of our analysis are: First, we establish an inequality that characterizes the dynamics of APDA given in Lemma 4.1 below. Based on it, we are able to prove the boundedness of sequences $\{x_k\}$ and $\{y_k\}$ in Theorem 4.1. In turn, sequence boundedness alongside the local smoothness property of f allows us to conclude that there exists a constant $L > 0$ such that f is L-smooth on the compact set $\overline{\text{Conv}}(\{x^*, x_0, x_1, \ldots\})$ – the closed convex hull generated by $\{x^*, x_0, x_1, \ldots\}$. Finally, we leverage this information to show that (x_k, y_k) converges to a saddle point of (2) and derive the associated ergodic convergence rates presented in Theorem 4.1.

Lemma 4.1. *Consider APDA along with Assumptions 3.1 and 3.2, and let $(x^*, y^*) \in X \times \mathcal{Y}$ be a saddle point of problem 2. Then, for all k*

1. $\frac{\beta\tau_k\| A \|}{1-c} < \frac{1}{2\tau_k\| A \|} - \frac{L_k}{\| A \|}$,
2. $$\begin{aligned}\| x_{k+1} - x^* \|^2 + \frac{1}{\beta}\| y_{k+1} - y^* \|^2 + (1 - \epsilon_k\tau_k\| A \| - \tau_k L_k)\| x_k - x_{k+1} \|^2 \\ + \left(\frac{1}{\beta} - \frac{\tau_k\| A \|}{\epsilon_k}\right)\| y_{k+1} - y_k \|^2 + 2\tau_k(1+\theta_k)P_{x^*,y^*}(x_k) + 2\tau_k D_{x^*,y^*}(y_{k+1}) \\ \leq \| x_k - x^* \|^2 + \frac{1}{\beta}\| y_k - y^* \|^2 + \frac{1}{2}\| x_k - x_{k-1} \|^2 + 2\tau_k\theta_k P_{x^*,y^*}(x_{k-1}),\end{aligned}$$

where $\epsilon_k \in \left(\frac{\beta\tau_k\| A \|}{1-c}, \frac{1}{2\tau_k\| A \|} - \frac{L_k}{\| A \|}\right)$. Moreover, $1 - \epsilon_k\tau_k\| A \| - \tau_k L_k > \frac{1}{2}$ and $\frac{1}{\beta} - \frac{\tau_k\| A \|}{\epsilon_k} > \frac{c}{\beta} > 0$.

Proof sketch. The full proof is deferred to the appendix. We use basic algebraic manipulations, the method's update rules, the Cauchy-Schwarz and Young inequalities and basic properties of the prox operator to obtain the recurrence

$$\begin{aligned}\| x_{k+1} - x^* \|^2 + \frac{1}{\beta}\| y_{k+1} - y^* \|^2 + \left(1 - \tau_k\| A \|\epsilon - \frac{\tau_k L_k}{\delta}\right)\| x_k - x_{k+1} \|^2 \\ + \left(\frac{1}{\beta} - \frac{\tau_k\| A \|}{\epsilon}\right)\| y_{k+1} - y_k \|^2 + 2\tau_k(1+\theta_k)P_{x^*,y^*}(x_k) + 2\tau_k D_{x^*,y^*}(y_{k+1})\end{aligned}$$

$$\leq \| x_k - x^* \|^2 + \frac{1}{\beta} \| y_k - y^* \|^2 + \tau_k L_k \delta \| x_k - x_{k-1} \|^2 + 2\tau_k \theta_k P_{x^*,y^*}(x_{k-1}), \quad (6)$$

where $\epsilon, \delta > 0$ are constants involved in Young's inequality.

Notice that the recurrence resembles an energy, with all quantities on the LHS being non-negative apart from the second and third terms. In order to obtain anything worthwhile, we must choose $\epsilon,\ \delta > 0$ such that the corresponding terms are positive and the energy on the left hand-side decreases. More precisely, the choice needs to satisfy:

$$\begin{cases} \frac{1}{\beta} - \frac{\tau_k \| A \|}{\epsilon} > 0 \\ 1 - \tau_k \| A \| \epsilon - \frac{\tau_k L_k}{\delta} > \tau_k L_k \delta \end{cases} \quad (7)$$

For theoretical purposes related to controlling the sequences $\|y_{k+1} - y_k\|^2$ and $\|x_{k+1} - x_k\|^2$, we strengthen the first inequality to be $\frac{1}{\beta} - \frac{\tau_k \| A \|}{\epsilon} > \frac{c}{\beta},\ c \in (0,1)$. In practice, we choose c to be as close to 0 as possible. Using the same reasoning for the second inequality coupled with the observation that $\tau_k L_k$ is an increasing function of L_k whose limit at infinity is $\frac{1}{2}$, we consider the strengthened version $1 - \tau_k \| A \| \epsilon - \frac{\tau_k L_k}{\delta} > \frac{1}{2}\delta$. The strengthened inequalities, rearranged for ϵ are thus

$$\begin{cases} \epsilon > \frac{\beta \tau_k \| A \|}{1-c}, \\ \epsilon < \frac{(1-\delta/2)}{\tau_k \| A \|} - \frac{L_k}{\delta \| A \|}. \end{cases} \quad (8)$$

If we can show that $\forall k$, τ_k satisfies $\frac{\beta \tau_k \| A \|}{1-c} < \frac{(1-\delta/2)}{\tau_k \| A \|} - \frac{L_k}{\delta \| A \|}$, it means we can always choose an $\epsilon = \epsilon_k$ in this interval such that the energy on the LHS of (6) decreases. Setting Young's inequality parameter $\delta = 1$ and evaluating the roots of the implied quadratic inequality in τ_k, this condition amounts to $\tau_k \in \left(0, \dfrac{1}{L_k + \sqrt{L_k^2 + 2(\beta/(1-c))\| A \|^2}}\right)$. Finally, we show that indeed our choice of τ_k satisfies this constraint and the proof is concluded. □

We are now ready to state the main convergence result, given in Theorem 4.1 below.

Theorem 4.1. *Consider APDA along with Assumptions 3.1 and 3.2, and let $(x^*, y^*) \in X \times \mathcal{Y}$ be a saddle point of problem* (2)*. Then, for all k*

1. ***Boundedness.*** *The sequence $\{(x_k, y_k)\}$ is bounded. Specifically, for all k,*
$$\| x_k - x^* \|^2 + \| y_k - y^* \|^2 \leq D,$$
where $D := \| x_1 - x^* \|^2 + \frac{1}{\beta} \| y_1 - y^* \|^2 + \frac{1}{2} \| x_1 - x_0 \|^2 + 2\tau_1 \theta_1 P_{x^*,y^*}(x_0) < \infty$.

2. ***Convergence to a saddle point.*** *The sequence $\{(x_k, y_k)\}$ converges to a saddle point of* (2).

3. ***Ergodic convergence.*** *Let sequence $\{(x_k, y_k)\}$ be generated by APDA. Define* $S_k := \sum_{i=1}^{k} \tau_i$,
$$X_k := \frac{1}{S_k} \left(\tau_k (1+\theta_k) x_k + \sum_{i=1}^{k-1} \left(\tau_i (1+\theta_i) - \tau_{i+1} \theta_{i+1} \right) x_i \right) \text{ and } Y_k := \frac{1}{S_k} \sum_{i=1}^{k} \tau_i y_{i+1}.$$
Then, for all k
$$\mathcal{G}_{x^*,y^*}(X_k, Y_k) \leq \frac{D\sqrt{L^2 + (\beta/(1-c))\| A \|^2}}{k}$$
where L is the Lipschitz constant of ∇f over the compact set $\overline{Conv}(\{x_*, x_0, x_1, \ldots\})$.

We briefly discuss the results here and defer the proof to the appendix due to lack of space. The boundedness result of Theorem 4.1 point 1. implies that the closed set $\mathcal{C} = \overline{\text{Conv}}(\{x^*, x_0, x_1, \ldots\})$ is also bounded and hence compact. The local smoothness condition for f then ensures that there exists $L > 0$ such that f is L-smooth over $\mathcal{C}$. Using this fact, we can show inductively that for all k, $\tau_k \geq \dfrac{1}{2\sqrt{L^2 + (\beta/(1-c))\| A \|^2}} > 0$, a fact that is instrumental in deriving the subsequent convergence results, as well as Theorem 4.2. We emphasize that the appearance of constant L in the provided rates is a consequence of the boundedness, whose proof does not require its knowledge.

4.3 Analysis under the additional Assumption 3.3

In the proof sketch of Lemma 4.1 we have seen that any $\tau_k \in \left(0, \frac{1}{L_k + \sqrt{L_k^2 + 2(\beta/(1-c))\| A \|^2}}\right)$ ensures that the energy on the LHS of (6) decreases, which in turn ensures sequence boundedness and convergence to a saddle point in Theorem 4.1. In particular, the setting for τ_k used by APDA satisfies this requirement. We can, however, choose any other value within this interval and maintain the same guarantees.

In particular, taking $\tau_k^+ = \min\left\{\frac{1}{2\sqrt{(L_k + 1/L_k)^2 + (\beta/(1-c))\| A \|^2}}, \tau_{k-1}\sqrt{1+\theta_{k-1}}\right\}$ – a smaller stepsize than the original one, Lemma 4.1 will continue to hold, along with points 1. and 2. of Theorem 4.1.

If we additionally consider Assumption 3.3, we now ask if we can obtain a different bound for point 3. of Theorem 4.1. This is indeed the case, and we show that running APDA with τ_k^+ yields a linear convergence under this additional assumption. The key is to derive a strengthened version of the inequality in Lemma 4.1, using the implication of this additional assumption that f is μ-strongly convex over $\overline{\text{Conv}}(\{x^*, x_0, x_1, \ldots\})$. Along with A being full row rank, the strengthened version of Lemma 4.1's result is

$$\begin{aligned}\| x_{k+1} - x^* \|^2 + \left(\frac{1}{\beta} + q_1\right)\| y_{k+1} - y^* \|^2 + \left(\frac{1}{2} + q_2\right)\| x_k - x_{k+1} \|^2 + q_3\| y_{k+1} - y_k \|^2 \\ + 2\tau_k(1+\theta_k)P_{x^*,y^*}(x_k) + 2\tau_k D_{x^*,y^*}(y_{k+1}) \\ \leq (1-q_4)\| x_k - x^* \|^2 + \frac{1}{\beta}\| y_k - y^* \|^2 + \left(\frac{1}{2} - q_5\right)\| x_k - x_{k-1} \|^2 + 2\tau_k\theta_k P_{x^*,y^*}(x_{k-1}),\end{aligned}$$

where $q_1, q_2, q_3, q_4, q_5 > 0$ are some appropriate constants detailed in the proof given the appendix. This new inequality represents in fact a contraction, which ensures the convergence rate in Theorem 4.2 which follows next.

Theorem 4.2. *Consider APDA along with Assumptions 3.1, 3.2 and 3.3. Let $(x^*, y^*) \in \mathcal{X} \times \mathcal{Y}$ be a saddle point of problem (2). Furthermore, let $\tau_k = \min\left\{\frac{1}{2\sqrt{(L_k + 1/L_k)^2 + (\beta/(1-c))\| A \|^2}}, \tau_{k-1}\sqrt{1+\theta_{k-1}}\right\}$, $s = \left(L + \frac{1}{\mu}\right)^2 + \frac{\beta\| A \|^2}{1-c}$ and $t = \left(\mu + \frac{1}{L}\right)^2 + \frac{\beta\| A \|^2}{1-c}$, where μ, L are the strong convexity and smoothness constants of f over the compact set $\overline{\text{Conv}}(\{x^*, x_0, x_1, \ldots\})$.*

Then, the iteration complexity required to get $\| x_k - x^* \|^2 + \| y_k - y^* \|^2 \leq \epsilon$ is $\mathcal{O}\left(\frac{1 - \min(p,q,r)}{\min(p,q,r)}\log(\epsilon^{-1})\right)$, for constants:

$$p = \frac{\mu}{4\sqrt{s}}, \quad q = \frac{\beta\sigma_{\min}^2(A)}{\beta\sigma_{\min}^2(A) + 4L\sqrt{s}\,(1+2Ls)}, \quad r = \frac{\sqrt{t}}{(\sqrt{t}+\sqrt{s})}.$$

5 Experiments

We now present some proof-of-concept experiments conducted for Algorithm 1. Additional experimental results and details are also included the appendix. The experiments were implemented in Python 3.9 and executed on a MacBook Pro with 32 GB RAM and a 2,9 GHz 6-Core Intel Core i9 processor. Part of the experime

The baseline we compare against is the PDHG in [Chambolle and Pock, 2016a], specifically Algorithm 1 for $g \equiv 0$. For obtaining x^* we ran one of the algorithms for a large number of iterations.

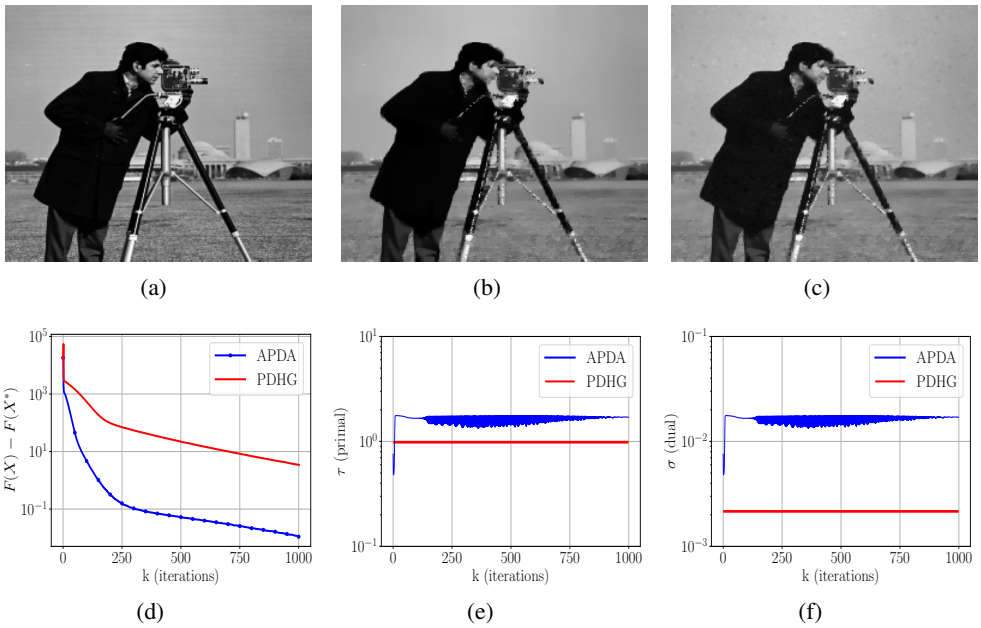


Figure 1: (a) Original image (can be downloaded from `http://www.cs.tut.fi/~foi/GCF-BM3D/`). (b) APDA reconstruction, PSNR = 25.53, SSIM = 0.9. (c) PDHG reconstruction, PSNR = 25.47, SSIM = 0.9. (d) Convergence rate. (e) Primal stepsize comparison. (f) Dual stepsize comparison.

5.1 Image inpainting

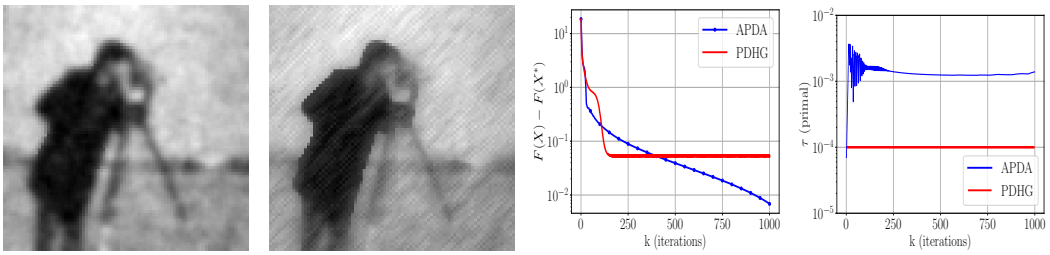
Image inpainting consists in reconstructing the missing parts of a subsampled image $b = P_\Omega(X^\natural)$, where $P_\Omega : \mathbb{R}^{m\times n} \in \mathbb{R}^{m\times n}$ is an operator that selects a subset $p \ll mn$ of pixels from the original image $X^\natural \in \mathbb{R}^{m\times n}$. This problem can be formulated as a regularized optimization objective:

$$\min_{X\in\mathbb{R}^{m\times n}} F(x) := \underbrace{\frac{1}{2}\|b - P_\Omega(X)\|_2^2}_{f(X)} + \underbrace{\lambda\|DX\|_{2,1}}_{g(X)\equiv\|\cdot\|_{\text{TV}}}, \tag{9}$$

where D represents the discrete gradient operator, and $\|Dx\|_{2,1} = \sum_{i,j=1}^{m,n} \sqrt{D_{i,j,1}^2 + D_{i,j,2}^2}$. The regularization term represents the isotropic TV norm, which is known to help in recovering sharp images by preserving discontinuities and reducing noise [Chambolle et al., 2010, Condat, 2017].

For our experiments, we vectorize the images of size 256×256 and transform D accordingly. We represent P_Ω as a matrix built by removing rows uniformly at random from the identity matrix, which removes 60% of pixels from the original image (sampling ratio 0.4). We perform parameter sweep for $\lambda \in$[`1e-4`, `1e`], $\beta \in$ [`1e-5`, `1e`] and $p \in$ [`1e-5`, `1e3`] (see Section 4.1 for details on p within PDHG) and settle for $\lambda =$ `1e-2`, $\beta =$ `1e-2` and $p =$ `1.64e2`. We remark that it was significantly more difficult to tune PDHG (which we did in 2 rounds), because the optimal range of p was very small and outside of it reconstruction quality abruptly dropped.

Experiment results are presented in Figure1, where we show the reconstructions, alongside the convergence plot and a comparison of the fixed stepsizes of PDHG with those of APDA. For each reconstruction we report the Peak Signal to Noise Ratio (PSNR) and the Structural Similarity Index Measure (SSIM), and in both cases higher values mark better reconstruction. Finally, note that our stepsizes are indeed better than the ones we found to be optimal for PDHG. Without retuning the parameters, we repeated the experiment for different subsampling ratios 0.3, 0.2, 0.1 and the relative performance of the algorithms remained the same.



(a) (b) (c) (d)

Figure 2: (a) APDA reconstruction, PSNR = 21.34, SSIM = 0.76. (b) PDHG reconstruction, PSNR = 20.56, SSIM = 0.70. (c) Convergence rate. (d) Primal stepsize comparison. The dual stepsize comparison is omitted for lack of space, but we report that $\sigma \approx 10^3$ (PDHG) whereas σ_k stabilizes similarly to τ_k around 1 (APDA).

5.2 Nonconvex phase retrieval

In this section we provide the results for applying our algorithm, heuristically, on the nonconvex least squares formulation of the phase retrieval (PR) problem. The phase-retrieval problem has attracted intense interest recently, due to its application is domains such as optical imaging [Walther, 1963], astronomy [Fienup and Dainty, 1987] and many others. Here, we consider the real counterpart of the original complex PR formulation for square images, where given $\{(A_i, b_i) \in \mathbb{R}^{n\times n} \times \mathbb{R}\}$ we want to recover $X \in \mathbb{R}^{n\times n}$ up to its sign, such that $b_i = \mathrm{Tr}(A_i^T x)^2$. To this end, we consider the following optimization objective:

$$\min_{X\in\mathbb{R}^{n\times n}} F(X) := \underbrace{\frac{1}{4m}\sum_{i=1}^{m}\left(b_i - \mathrm{Tr}(A_i^T X)^2\right)^2}_{f(X)} + \underbrace{\lambda\|DX\|_{2,1}}_{g(X)\equiv\|\cdot\|_{\mathrm{TV}}}. \tag{10}$$

We note a few things: first, objective (10) is nonconvex and not L-smooth. Secondly, Sun et al. [2018] have recently shown that given m i.i.d Gaussian measurements, the global geometry of $F(X)$ is 'benign' for $m > Cd\log(d)^3$, where d is the problem dimension. It is posed that in such cases iterative algorithms should, with high probability, find the minimizer without requiring special initialization as is needed for current state of the art solvers.

For our experiments we use 84×84 sized images and choose a smaller number of measurements than suggested above: $m = d\log(d) \approx 27,155$. We generate m sparse matrices $A_i \in R^{n\times n}$ with i.i.d standard Gaussian entries and sparsity coefficient 0.3 and corrupt 0.1 entries in b_i by setting them to 0. We perform parameter sweep for $\lambda \in [\texttt{1e-4}, \texttt{1e4}]$, $\beta \in [\texttt{1e-3}, \texttt{1e4}]$. Without guidelines for setting τ, σ for PDHG since f is not L-smooth, we search for the best $\tau \in [\texttt{1e-4}, \texttt{1e4}]$ and $p \in [\texttt{1e-2}, \texttt{1e2}]$ such that $\sigma = \frac{1}{p\tau\|A\|}$. We note that PDHG diverged for $32/40$ grid points, whereas our method converged for all instances.

The results are depicted in Figure 2, which contains the reconstructions and convergence plots. We tested several random seeds, and for some cases PDHG exhibited a strong oscillatory behavior (not depicted).

5.3 Limitations of APDA

We have illustrated the performance of APDA for two imaging problems, whose common trait is that they do not impose hard constraints on the primal variables. As a consequence, we are able to take plain gradient steps in the primal domain. However, for instances such as Poisson linear inverse problems [Bertero et al., 2009], the iterates need to reside in $\mathbb{R}^n_+$ as the primal objective contains log functions. APDA cannot handle such cases, as any constraints imposed on the primal variables will only be satisfies asymptotically. We consider such scenarios as a future research direction.

References

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

Mario Bertero, Patrizia Boccacci, Gabriele Desiderà, and Giuseppe Vicidomini. Image deblurring with poisson data: from cells to galaxies. *Inverse Problems*, 25(12):123006, 2009.

Antonin Chambolle. An algorithm for total variation minimization and applications. *Journal of Mathematical imaging and vision*, 20(1):89–97, 2004.

Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of mathematical imaging and vision*, 40(1):120–145, 2011.

Antonin Chambolle and Thomas Pock. On the ergodic convergence rates of a first-order primal–dual algorithm. *Mathematical Programming*, 159(1):253–287, 2016a.

Antonin Chambolle and Thomas Pock. An introduction to continuous optimization for imaging. *Acta Numerica*, 25:161–319, 2016b.

Antonin Chambolle, Vicent Caselles, Daniel Cremers, Matteo Novaga, and Thomas Pock. An introduction to total variation for image analysis. *Theoretical foundations and numerical methods for sparse recovery*, 9(263-340):227, 2010.

Xiaokai Chang, Junfeng Yang, and Hongchao Zhang. Golden ratio primal-dual algorithm with linesearch. *arXiv preprint arXiv:2105.07108*, 2021.

Patrick L Combettes and Jean-Christophe Pesquet. Proximal splitting methods in signal processing. In *Fixed-point algorithms for inverse problems in science and engineering*, pages 185–212. Springer, 2011.

Laurent Condat. A primal–dual splitting method for convex optimization involving lipschitzian, proximable and linear composite terms. *Journal of optimization theory and applications*, 158(2): 460–479, 2013.

Laurent Condat. Discrete total variation: New definition and minimization. *SIAM Journal on Imaging Sciences*, 10(3):1258–1290, 2017.

C Fienup and J Dainty. Phase retrieval and image reconstruction for astronomy. *Image recovery: theory and application*, 231:275, 1987.

Tom Goldstein, Min Li, Xiaoming Yuan, Ernie Esser, and Richard Baraniuk. Adaptive primal-dual hybrid gradient methods for saddle-point problems. *arXiv preprint arXiv:1305.0546*, 2013.

Tom Goldstein, Christoph Studer, and Richard Baraniuk. A field guide to forward-backward splitting with a fasta implementation. *arXiv preprint arXiv:1411.3406*, 2014.

Tom Goldstein, Min Li, and Xiaoming Yuan. Adaptive primal-dual splitting methods for statistical learning and image processing. In *Advances in Neural Information Processing Systems*, pages 2089–2097, 2015.

Nikos Komodakis and Jean-Christophe Pesquet. Playing with duality: An overview of recent primal? dual approaches for solving large-scale optimization problems. *IEEE Signal Processing Magazine*, 32(6):31–54, 2015.

Yura Malitsky and Konstantin Mishchenko. Adaptive gradient descent without descent. In *International Conference on Machine Learning*, pages 6702–6712. PMLR, 2020.

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

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

Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in optimization*, 1(3): 127–239, 2014.

Fabian Pedregosa and Gauthier Gidel. Adaptive three operator splitting. In *International Conference on Machine Learning*, pages 4085–4094. PMLR, 2018.

Boris Teodorovich Polyak. Minimization of nonsmooth functionals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 9(3):509–521, 1969.

Ju Sun, Qing Qu, and John Wright. A geometric analysis of phase retrieval. *Foundations of Computational Mathematics*, 18(5):1131–1198, 2018.

Bang Công Vũ. A splitting algorithm for dual monotone inclusions involving cocoercive operators. *Advances in Computational Mathematics*, 38(3):667–681, 2013.

Adriaan Walther. The question of phase retrieval in optics. *Optica Acta: International Journal of Optics*, 10(1):41–49, 1963.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Yes, Section 5.3
   (c) Did you discuss any potential negative societal impacts of your work? [No] This is a theoretical work with no foreseeable societal consequences.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See the Section 3.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Supplementary material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Section 5
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [No]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] Code.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] The image used for the experiments is a widely used/known for testing imaging algorithms.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]