Research Project
Primal-Dual Hybrid Gradient method with Adaptive Stepsizes

Rustem Islamov* Olivier Fercoq †

March 29, 2022

Abstract

The performance of gradient-type methods depends on the stepsize choice. In practice the stepsize is chosen randomly or as the best one in a set. However, such selection does not guarantee the optimality of the method. Besides, the best theoretical stepsizes are based on the parameters of the problem such as Lipschitz or strong convexity constants. As a consequence, the adaptive stepsize selection has been studied in many works. However, such works analyzing adaptive Primal-Dual Hybrid Gradient (PDHG) suffer from a luck of theory or can be considered only as heuristics. In this work we develop adaptive stepsize selection for PDHG based on Quadratic Error Bound of the smoothed gap in strongly convex-concave setting. Moreover, we demonstrate the empirical superiority of this approach in verious scenarious.

Contents

*Institut Polytechnique de Paris, Palaiseau, France
†Télécom Paris, Institut Polytechnique de Paris, Palaiseau, France.

1 Introduction

Primal-dual methods are widely used for solving optimization problems with constraints. Encoding such nonsmooth constrains we replace original problem with the problem of finding saddle points of the Lagrangian. More precisely we consider the problem of the form

$$\min_{x \in \mathcal{X}} \Big[f(x) + f_2(x) + g \square g_2(Ax) \Big]. \tag{1}$$

Here f and g are convex functions for which the proximal operators are easily computable; $A : \mathcal{X} \to \mathcal{Y}$ is a linear operator. We assume that we have an access to g^* and g_2^* Fenchel-Legendre conjugates of g and g_2 respectively. Finally, f_2 and g_2^* are convex functions with L_f and L_{g^*} Lipschitz gradients. As it was stated before, we consider primal-dual methods which are searching for a saddle point of of the Lagrangian, which has the following form

$$L(x, y) = f(x) + f_2(x) + \langle Ax, y \rangle - g^*(y) - g_2^*(y). \tag{2}$$

The point (x^*, y^*) is called a saddle point for the Lagrangian (2) if it satisfies

$$L(x, y^*) \leq L(x^*, y^*) \leq L(x^*, y) \quad \forall x \in \mathcal{X}, y \in \mathcal{Y}. \tag{3}$$

We assume throughout the paper that at least one saddle point exists. It can be guaranteed using Slater's constraint qualification condition.

There exist different ways to measure the convergence of primal-dual algorithms: duality gap, Karush-Kuhn-Tacker (KKT) error, metrical subregularity [Rockafellar and Wets, 1998], smoothed gap [Tran-Dinh et al., 2018]. Recently Quadratic Error Bound has been introduced in [Fercoq, 2021] which properly reflect the behaviour of PDHG. This regularity assumption holds for a wide range of problems such as strongly convex-concave problem or linear programming.

One of the most powerful methods to find a saddle point of the Lagrangian is Primal-Dual Hybrid Gradient (PDHG) [Chambolle and Pock, 2011, Zhu and Chan, 2008] method which is defined by Algorithm 1. The convergence of PDHG under different optimality measures has been recently studied [Liang et al., 2016b, Du and Hu, 2019, Fercoq, 2021] in order to guarantee the linear speed. Quadratic Error Bound (QEB) was introduced in [Fercoq, 2021] unifies existing optimilaty measures for PDHG analysis like strong convexity [Chambolle and Pock, 2011] and metrical subregularity [Liang et al., 2016a]. [Fercoq, 2021] shows linear convergence of PDHG under QEB regularity condition. This optimality measure properly reflects the behaviour of PDHG and holds for a wide range of problems such as strongly convex-concave problem or linear programming.

Algorithm 1 Primal-Dual Hybrid Gradient (PDHG)

Parameters: stepsizes τ, σ
Initialization: $x_0 \in \mathcal{X}, y_0 \in \mathcal{Y}$
for $k = 0, 1, 2, \ldots, K$ **do**
  $\bar{x}_{k+1} = \text{prox}_{\tau f}\left(x_k - \tau \nabla f_2(x^k) - \tau A^\top y_k\right)$
  $\bar{y}_{k+1} = \text{prox}_{\sigma g^*}\left(y_k - \sigma \nabla g_2(y_k) + \sigma A \bar{x}_{k+1}\right)$
  $x_{k+1} = \bar{x}_{k+1} - \tau A^\top (\bar{y}_{k+1} - y_k)$
  $y_{k+1} = \bar{y}_{k+1}$
end for
return (x_K, y_K)

2 Related works

One of the main problems that makes first-order methods impractical is the stepsize choice. Usually, the stepsize is selected arbitrarly and intuitively or using a grid search. However, such approaches do not guarantee that chosen stepsize will lead to optimal performance of a method.

Nevertheless, there are many works [Nesterov, 2013] that provide the choice of stepsizes such that the performance of a method cannot be improved. The issue of those works is that always the theoretical stepsizes are based on parameters of a problem (for example, Lipschitz or strong convexity constants). These parameters could be unknown in practice, or the cost of computing them is extremely high. On top of that, these parameters usually describe the global curvature of the objective function and could be overestimated. The afformentioned challenges lead to inefficient performance of gradient-type methods.

One of the solutions is to choose stepsizes adaptively based on local curvature of the objective. [Malitsky and Mishchenko, 2020, Vladarean et al., 2021] analyze first-order methods in the setting when the Lipschitz constant is unknown. They propose a novel idea how to estimate local smoothness parameter and choose the stepsize according to it. Their theory covers the case when the objective is not globally smooth. Their adaptive strategy achieves the same convergence rate $\mathcal{O}(k^{-1})$ for convex functions. [Aujol et al., 2021] propose adaptive restart scheme for FISTA [Beck and Teboulle, 2009] that allows to get the optimal convergence rate when the objective function satisfies quadratic growth condition with unknown constant. There is an alernative scheme for accelerated gradient type methods proposed by [Fercoq and Qu, 2019], but it requires the initial estimate of quadratic growth constant.

To the best of our knowledge, there is no work propose adaptive stepsize choice for PDHG that has strong theory. There are several papers [Goldstein et al., 2015, Yokota and Hontani, 2017] that investigate PDHG with adaptive stepsizes, however, they do not provide theoretical guarantees for their schemes. The strategy proposed by [Goldstein et al., 2015] relies on the assumption that cannot be checked in practice. The scheme of [Yokota and Hontani, 2017] has significant issue that it does not ensure the necessary convergence condition of PDHG, which is that the stepsizes must satisfy the inequality $\tau\sigma\|A\|^2 \leq 1$, holds. As a consequence, this scheme could lead to divergence in some practical scenarious.

3 Contributions

We propose a new adaptive stepsizes selection strategies for PDHG method. Our theory covers strongly convex-concave case. When both $f + f_2$ and $g^* + g_2^*$ are strongly convex, and one of the strong convexity constants is known. This holds in the case when, for example, $f + f_2$ is ℓ_2 or elastic net regularizations. We introduce iterative scheme how to choose the stepsizes for primal and dual spaces. The idea is close to that by [Fercoq and Qu, 2019] where the authors deduce the inequality that involves the unknown quadratic growth constant and the rest can be computed in each iteration. We derive the inequality of the same form based on QEB condition that allows to check if the convergence rate is fast enough. If the theoretical speed is faster than the practical one, then we decrease the estimate of unknwon strong convexity parameter. Such iteration scheme eventually makes possible to find true or better approximate of strong convexity parameter.

4 Preliminaries

In this section we introduce necessary definitions and notation which will be used throughout the paper.

4.1 Notation

We denote $\mathcal{X}$ the primal space, $\mathcal{Y}$ the dual space. The proximal operator of a function f is given by $\operatorname{prox}_f(x') = \arg\min_{x'} \left[f(x') + \frac{1}{2}\|x - x'\|^2 \right]$. We will use the indicator function ι_C of a convex set C which is defined as follows

$$\iota_C := \begin{cases} 0 & \text{if } x \in C \\ +\infty & \text{if } x \notin C. \end{cases}$$

Besides, Fenchel-Legendre conjugate f^* of a function f is defined by

$$f^*(y) = \sup_{x \in \mathcal{X}} \left[\langle x, y \rangle - f(x) \right].$$

4.2 Important Definitions

First we define the epigraph of a function f.

Definition 4.1. *Let $f : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$. The epigraph of f, denoted by epi f, is the subset of $\mathcal{X} \times \mathbb{R}$ defined by*

$$epi\ f = \{(x, t) \in \mathcal{X} \times \mathbb{R} : t \geq f(x)\}.$$

Knowing what the epigraph is, we can define the definition of convex function.

Definition 4.2. *A function $f : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ is convex if its epigraph is a convex set.*

More restricted class of convex functions is a class of so called strongly-convex functions.

Definition 4.3. *A function $f : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ is μ-strongly convex if $f - \frac{\mu}{2}\|x\|^2$ is convex.*

Now we define what strongly convex-concave Lagrangian function means.

Definition 4.4. *The Lagrangian function is μ-strongly convex-concave, if $x \mapsto L(x, y)$ is μ-strongly convex for all y, and $y \mapsto -L(x, y)$ is μ-strongly convex for all x.*

Moreover, we will work with L-smooth functions.

Definition 4.5. *Let $f : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ is L-smooth, if it is continuously differentiable and $\|\nabla f(x) - \nabla f(x')\| \leq L\|x - x'\|$ for all x, x'.*

Definition 4.6. *We say that a function $f : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ has a quadratic error bound if there exists η and an open region $\mathcal{R}$ that contains* $\arg\min f$ *such that for all $x \in \mathcal{R}$,*

$$f(x) \geq \min f + \frac{\eta}{2}\mathrm{dist}(x, \arg\min f)^2. \tag{4}$$

We shall use the acronym f has η-QEB.

Although this is more general than strong convexity, the quadratic error bound is not enough for saddle point problems. For example, for a large class of problems with linear constraints (i.e. $y \to L(x, y)$ is linear) QEB is not satisfied in y. To resolve this issue, we may resort to metric regularity.

Definition 4.7. *A set-valued function $F : \mathcal{Z} \rightrightarrows \mathcal{Z}$ is metrically subregular at z for b if there exists $\eta > 0$ and a neighbourhood $N(z)$ of z such that $\forall\ z' \in N(z)$,*

$$\mathrm{dist}_V(F(z', b)) \geq \eta \mathrm{dist}_V(z', F^{-1}(b)). \tag{5}$$

Definition 4.8. *Given $\beta = (\beta_x, \beta_y) \in [0, +\infty]^2, z \in \mathcal{Z}$ and $\dot{z} \in \mathcal{Z}$, the smoothed gap G_β is the function defined by*

$$G_\beta(z, \dot{z}) = \sup_{z' \in \mathcal{Z}} L(x, y') - L(x', y) - \frac{\beta_x}{2\tau}\left\|x' - \dot{x}\right\|^2 - \frac{\beta_y}{2\sigma}\left\|y' - \dot{y}\right\|^2. \tag{6}$$

We call the function $z \to G_\beta(z, \dot{z})$ the smoothed gap centered at $\dot{z}$.

Although the smoothed gap can be defined for any center $\dot{z}$, the next proposition shows that if $\dot{z} = z^* \in \mathcal{Z}^*$, then the smoothed gap is a measure of optimality.

Proposition 4.9. *Let $\beta \in [0, +\infty]^2$. If $z^* \in \mathcal{Z}^*$, then $z \in \mathcal{Z}^* \leftrightarrow G_\beta(z, z^*) = 0$.*

5 Proposed Adaptive Scheme

In this section we introduce our adaptive schemes for stepsize selection of PDHG. Before that we state useful proprositions from [Fercoq, 2021]. We work with strongly convex-concave Lagrangian function. Besides, we assume that μ_{g^*} is known, but μ_f not (the opposite is also possible). Such problem statement allows to adaptively change the approximation $\hat{\mu}_f$ of the strong convexity parameter μ_f throughout the iteration process.

Now we are ready to claim all necessary propositions and theoretical statements in order to prove our adaptive scheme.

Proposition 5.1. *If L is μ-strongly convex-concave, then $\tilde{\partial} L$ us μ-metrically sub-regular at z^* for 0, where z^* is the unique saddle point of L.*

PDHG can be conveniently seen as a fixed point algorithm $z_{k+1} = T(z_k)$ where T is defined as follows

$$\begin{aligned} \bar{x} &= \operatorname{prox}_{\tau f}(x - \tau A^\top y), & \bar{y} &= \operatorname{prox}_{\sigma g^*}(y + \sigma A \bar{x}) \\ x^+ &= \bar{x} - \tau A^\top(\bar{y} - y), & y^+ &= \bar{y} \\ T(x, y) &= (x^+, y^+). \end{aligned} \tag{7}$$

In strongly convex-concave case we are able to prove linear convergence of PDHG in the norm $\|\cdot\|_V$.

Proposition 5.2. *If L is μ-strongly convex-concave in the norm $\|\cdot\|_V$, then the iterates of PDHG satisfy for all k,*

$$\|z_{k+1} - z^*\|_V^2 \leq \left(1 + \frac{\mu}{1 + \mu/\Gamma}\right)^{-1} \|z_k - z^*\|_V^2, \tag{8}$$

where z^* is the unique saddle point of L and $\Gamma = (1 - \alpha_f)(1 - \sqrt{\gamma})$.

[Fercoq, 2021] additionally provides the stepsizes that allow to get the optimal convergence rate of PDHG. They have the following form

$$\tau = \sqrt{\frac{\mu_{g^*}}{\mu_f}} \frac{1}{\|A\|}, \quad \sigma = \sqrt{\frac{\mu_f}{\mu_{g^*}}} \frac{1}{\|A\|}. \tag{9}$$

For $z = (x, y) \in \mathcal{Z}$, we denote $\|z\|_V^2 = (\tau^{-1}\|x\|^2 + \sigma^{-1}\|y\|^2)^{1/2}$. Let stepsizes satisfy $\gamma = \sigma\tau\|A\|^2 < 1$, $\tau L_f/2 \leq \alpha_f < 1$, $\alpha_g = \sigma L_{g^*}/2 \leq 1$, and $\sigma L_{g^*}/2 \leq \alpha_f(1 - \sigma\tau\|A\|^2)$. Using Proposition 5.1 we get another Proposition.

Proposition 5.3. *If $\tilde{\partial} L$ is metrically sub-regular at z^* for 0 and for all $z^* \in \mathcal{Z}^*$ with constant $\eta > 0$, then $(I - T)$ is metrically sub-regular at z^* for 0 and for all $z^* \in \mathcal{Z}^*$ with constant $\frac{\eta}{\sqrt{3}\eta + 2 + 2\sqrt{3}\max\{\alpha_g, \alpha_f\}}$, and PDHG converges linearly with rate $\left(1 - \frac{\eta^2(1-\alpha_f)(1-\sqrt{\gamma})}{\left(\sqrt{3}\eta + 2 + 2\sqrt{3}\max\{\alpha_f, \alpha_g\}\right)^2}\right)$.*

Let us assume that f and g^* are strongly convex function, but we do not know the strong convexity parameter of f. In this case L is strongly convex-concave. By Propositions 5.3 and 5.1 we get that $\tilde{\partial} L$ is μ-metrically sub-regular at z^* for 0, and $(I - T)$ is η-metrically sub-regular, where

$$\eta = \frac{\mu}{\sqrt{3}\mu + 2 + 2\sqrt{3}\max\{\alpha_g, \alpha_f\}}. \tag{10}$$

This implies the following

$$\|T(z) - z\|_V^2 \geq \eta^2 \|z - z^*\|_V^2. \tag{11}$$

Moreover, from Lemma 2 of [Fercoq, 2021] we get for $z' = z^*$ (note that z^* is a fixed point of T)

$$\begin{aligned} \lambda \|z - T(z) - z^* + T(z^*)\|_V^2 &\leq \|z - z^*\|_V^2 - \|T(z) - T(z^*)\|_V^2 - 2\mu_f \|\bar{x} - \bar{x}^*\|^2 - 2\mu_{g^*} \|\bar{y} - \bar{y}^*\|^2 \\ \lambda \|z - T(z)\|_V^2 &\leq \|z - z^*\|_V^2, \end{aligned} \tag{12}$$

where

$$\lambda = 1 - \alpha_f - \frac{\alpha_g - (1 - \gamma)\alpha_f}{2} - \sqrt{(1 - \alpha_f)^2\gamma + \left((1 - \gamma)\alpha_f - \alpha_g\right)^2/4}.$$

5.1 Update Rule #1

Using (12) in the Proposition 5.2 we get

$$\lambda \|z_{k+1} - T(z_{k+1})\|_V^2 \leq \|z_{k+1} - z^*\|_V^2 \leq \left(1 + \frac{\mu}{1+\mu/\Gamma}\right)^{-1} \|z_k - z^*\|_V^2$$
$$\leq \eta^{-2} \left(1 + \frac{\mu}{1+\mu/\Gamma}\right)^{-1} \|T(z_k) - z_k\|_V^2 . \tag{13}$$

Note that the last inequality should hold throughout optimization process. That's why we can obtain a tighter inequality in order to detect faster the true strong convexity parameter.

$$\boxed{\lambda \|z_{k+1} - T(z_{k+1})\|_V^2 \leq \min_{l \leq k} \eta^{-2} \left(1 + \frac{\mu}{1+\mu/\Gamma}\right)^{-(l+1)} \|T(z_{k-l}) - z_{k-l}\|_V^2 .} \tag{14}$$

We may use (14) as criterion for restarted PDHG because both sides of (13) are computable. If this inequality does not hold, then we As a consequence, we deduce the method summaraized as Algorithm 2. We would like to point out that in (14) there is η^{-2} in the right-hand side. However, in [Fercoq and Qu, 2019] the rate depends on this parameter as η^{-1} which makes the adaptive scheme faster. This is because in practice η is small, and (14) holds for a huge number of iterations before detection that $\hat{\mu}_f$ is too optimistic.

Algorithm 2 Primal-Dual Hybrid Gradient (PDHG) with Adaptive Stepsizes

Initialization: $x_0 \in \mathcal{X}, y_0 \in \mathcal{Y}$, true value of μ_{g^*}, estimate of the strong convexity parameter $\hat{\mu}_f$ of f
compute τ_0, σ_0 according to (9) with $\hat{\mu}_f$ and μ_{g^*}
for $k = 0, 1, 2, \ldots, K$ **do**
  $\bar{x}_{k+1} = \text{prox}_{\tau_k f}\left(x_k - \tau_k \nabla f_2(x^k) - \tau_k A^\top y_k\right)$
  $\bar{y}_{k+1} = \text{prox}_{\sigma_k g^*}\left(y_k - \sigma_k \nabla g_2(y_k) + \sigma_k A \bar{x}_{k+1}\right)$
  $x_{k+1} = \bar{x}_{k+1} - \tau A^\top (\bar{y}_{k+1} - y_k)$
  $y_{k+1} = \bar{y}_{k+1}$
  compute $\mu = \min\{\tau_k \mu_{g^*}, \sigma_k \hat{\mu}_f\}$
  If $\lambda \|z_{k+1} - T(z_{k+1})\|_V^2 > \min\limits_{l \leq k} \eta^{-2} \left(1 + \frac{\mu}{1+\mu/\Gamma}\right)^{-(l+1)} \|T(z_{k-l}) - z_{k-l}\|_V^2$
    then $\hat{\mu}_f \leftarrow \hat{\mu}_f/2$, compute σ_{k+1} and τ_{k+1} according to (9) with updated $\hat{\mu}_f$ and μ_{g^*}
end for
return (x_K, y_K)

5.2 Update Rule #2

[Fercoq, 2021] provide the heuristic adaptive restart of PDHG. They derive the following key inequality

$$G_\beta(z, \dot{z}) = \max_{z' \in \mathcal{Z}} L(x, y') - L(x', y) - \frac{\beta}{2} \left\|z' - \dot{z}\right\|_V^2$$
$$\geq \max_{z' \in \mathcal{Z}} L(x, y') - L(x', y) - \beta \left\|z' - z^*\right\|_V^2 - \beta \|\dot{z} - z^*\|_V^2$$
$$= G_{2\beta}(z, z^*) - \beta \|\dot{z} - z^*\|_V^2$$
$$\geq \frac{\eta(2\beta)}{2} \|z - z^*\|_V^2 - \beta \|\dot{z} - z^*\|_V^2 . \tag{15}$$

Using the above with $z = z_0, \dot{z} = z_k$ we derive

$$\|z_0 - z^*\|_V^2 \leq \frac{2}{\eta(\beta)} \left[G_{\beta/2}(z_0, z_k) + \frac{\beta}{2} \|z_k - z^*\|_V^2\right].$$

From Proposition 5.2 we also know

$$\|z_k - z^*\|_V^2 \leq \left(1 + \frac{\mu}{1 + \mu/\Gamma}\right)^{-k} \|z_0 - z^*\|_V^2.$$

Combining the last two inequalities we obtain

$$\|z_k - z^*\|_V^2 \leq \frac{2}{\eta(\beta)} \left(1 + \frac{\mu}{1 + \mu/\Gamma}\right)^{-k} \left[G_{\beta/2}(z_0, z_k) + \frac{\beta}{2} \|z_k - z^*\|_V^2\right].$$

The inequality above allows to transform (13) into the next one

$$\boxed{\lambda \|T(z_k) - z_k\|_V^2 \leq \frac{\frac{2}{\eta(\beta)}(1 + \frac{\mu}{1+\mu/\Gamma})^{-k} G_{\beta/2}(z_0, z_k)}{1 - (1 + \frac{\mu}{1+\mu/\Gamma})^{-k} \frac{\beta}{\eta(\beta)}}.} \tag{16}$$

One can notice that now we don't have the dependency on η^{-2} that is a drawback of the previous update rule #1. Nevertheless, this update rule is also has its disadvantage. The denominator of right-hand side could be negative, thus, we need to wait until it turns out to be positive in order to be able to use (16).

5.3 Convergence Theorem

In this section we establish the convergence theorem for the update rule #1. We denote as $\hat{\mu}_s, s = \{0, 1, \ldots\}$ the value of the strong convexity parameter estimator of the Lagrangian after s decreasing steps in line 9 of Algorithm 2. Besides, we denote by $\|T(z_k) - z_k\|_{V_s}^2$ and η_s the V-norm of the difference and η (10) when stepsizes τ and σ are computed with respect to $\hat{\mu}_s$.

Theorem 5.4. *Let $\hat{\mu}_f$ be the initial approximation of unknown strong convexity parameter μ_f, and μ_{g^*} be known strong convexity parameter of g^*. Let $\bar{l} := \lceil \log_2 \hat{\mu}_f - \mu_f \rceil$. Then Algorithm 2 requires*

$$\sum_{s=0}^{\bar{l}} K_s$$

steps to find true value μ_f, where K_s is defined as follows

$$K_s := \frac{\log\left(\frac{\|T(z_0) - z_0\|_{V_s}^2}{\eta_s^2 \lambda \varepsilon}\right)}{\log\left(1 + \frac{\mu_s}{1+\mu_s/\Gamma}\right)}.$$

Proof. Let $\hat{\mu}_f$ be an apporximation of unknown strong convexity parameter μ_f and τ, σ be the stepsizes chosen according to (9) with known μ_{g^*} and approximation $\hat{\mu}_f$. Then the strong convexity parameter μ of the Lagrangian is

$$\begin{aligned}
\mu &= \min\{\tau\mu_f, \sigma\mu_{g^*}\} \\
&= \frac{1}{\|A\|} \min\left\{\sqrt{\frac{\mu_{g^*}}{\hat{\mu}_f}}\mu_f, \sqrt{\frac{\hat{\mu}_f}{\mu_{g^*}}}\mu_{g^*}\right\} \\
&= \frac{\sqrt{\mu_f \mu_{g^*}}}{\|A\|} \min\left\{\sqrt{\frac{\mu_f}{\hat{\mu}_f}}, \sqrt{\frac{\hat{\mu}_f}{\mu_f}}\right\} \\
&= \frac{\sqrt{\mu_f \mu_{g^*}}}{\|A\|} \sqrt{\frac{\mu_f}{\hat{\mu}_f}} = \mu_* \sqrt{\frac{\mu_f}{\hat{\mu}_f}},
\end{aligned}$$

where we use the fact that $\hat{\mu}_f \geq \mu_f$. Let $\bar{l} = \lceil \log_2 \hat{\mu}_f - \mu_f \rceil$. Thus, after $\bar{l}$ decreasing steps the method reaches the true value μ_f. Note that

$$\varepsilon < \|T(z_k) - z_k\|_{V_s} \leq \eta_s^{-2} \left(1 + \frac{\mu_s}{1 + \mu_s/\Gamma}\right)^{-k} \|T(z_0) - z_0\|_{V_s}^2$$

cannot hold for k satisfying

$$\lambda^{-1}\eta_s^{-2} \left(1 + \frac{\mu_s}{1 + \mu_s/\Gamma}\right)^{-k} \|T(z_0) - z_0\|_{V}^2 \leq \varepsilon.$$

This means that we can run the adaptive scheme with μ_s for at most

$$K_s = \frac{\log\left(\frac{\|T(z_0)-z_0\|_{V_s}^2}{\eta_s^2 \lambda \varepsilon}\right)}{\log\left(1 + \frac{\mu_s}{1+\mu_s/\Gamma}\right)}.$$

By summing up $K_s, s \in \{0, 1, \ldots, \bar{l}\}$ we finish the proof. □

6 Optimization Problems

6.1 Ridge Regression

The first problem we consider is Ridge regression. The problem has the following form

$$\min_{x \in \mathcal{X}} \left[\frac{1}{2} \|Ax - b\|^2 + \frac{\lambda}{2}\|x\|^2\right],$$

where λ is positive regularization parameter. Such regularization is used when linear system $Ax = b$ has infinite number of solutions. We choose the solution with the smallest norm. Usually ℓ_2 is applied if data suffers from multicollinearity. Referring to (1), we set $g(Ax) = \frac{1}{2}\|Ax - b\|^2$, i.e. $g(z) = \frac{1}{2}\|z - b\|^2$, and $f(x) = \frac{\lambda}{2}\|x\|^2$. Other functions are zero.

First, we need to find g^*.

Lemma 6.1. *The Fenchel-Legendre conjugate of $g = \frac{1}{2}\|x - b\|^2$ is given by*

$$g^*(y) = \frac{1}{2}\|y\|^2 + \langle y, b\rangle.$$

Proof. We write the definition of the Fenchel-Legendre conjugate

$$g^*(y) = \sup_x \left[\langle y, x\rangle - \frac{1}{2}\|x - b\|^2\right].$$

This is the strongly convex problem, thus the solution is unique. By the Fermat rule we get

$$y - x + b = 0 \Rightarrow x = y + b.$$

Finally, putting the above into the definition of Fenchel-Legendre conjugate we derive

$$\begin{aligned} g^*(y) &= \langle y, y + b\rangle - \frac{1}{2}\|y + b - b\|^2 \\ &= \frac{1}{2}\|y\|^2 + \langle y, b\rangle. \end{aligned}$$

□

Now we need to find the explicit form of proximal operators for f and g^*.

Lemma 6.2. *The proximal operator of τf, where $f = \frac{\lambda}{2}\|x\|^2$, is given by*

$$\text{prox}_{\tau f}(x) = \frac{x}{1+\tau\lambda}.$$

Proof. We write the definition of a proximal operator

$$\text{prox}_{\tau f}(x) = \arg\min_{x'}\left[\frac{\tau\lambda}{2}\|x'\|^2 + \frac{1}{2}\left\|x'-x\right\|^2\right].$$

By the Fermat rule we obtain

$$\tau\lambda x' + x' - x = 0 \Rightarrow x' = \frac{x}{1+\tau\lambda}.$$

□

Lemma 6.3. *The proximal operator of σg^*, where $g^*(y) = \frac{1}{2}\|y\|^2 + \langle y, b\rangle$, is given by*

$$\text{prox}_{\sigma g^*}(x) = \frac{x-\sigma b}{1+\sigma}.$$

Proof. We write the definition of a proximal operator

$$\text{prox}_{\sigma g^*}(x) = \arg\min_{x'}\left[\frac{\sigma}{2}\|x'\|^2 + \sigma\left\langle x', b\right\rangle + \frac{1}{2}\left\|x'-x\right\|^2\right].$$

By the Fermat rule we get

$$\sigma x' + \sigma b + x' - x = 0 \Rightarrow x' = \frac{x-\sigma b}{1+\sigma}.$$

□

The Lagrangian function of Ridge regression problem has the following problem

$$L(x,y) = \frac{\lambda}{2}\|x\|^2 + \langle Ax, y\rangle - \frac{1}{2}\|y\|^2 - \langle y, b\rangle.$$

It is λ-strongly convex in x and 1-strongly concave in y. Finally, f is L-smooth with $L = \lambda_{\max}(A^\top A)$, where $\lambda_{\max}(M)$ denotes the largest eigenvalue of M.

6.2 Elastic Net Regression

Now we consider Elastic net regression problem of the form

$$\min_{x\in\mathcal{X}}\left[\frac{1}{2}\|Ax-b\|^2 + \lambda_1\|x\|_1 + \frac{\lambda_2}{2}\|x\|^2\right],$$

where λ_1, λ_2 are positive regularization constants. Use of ℓ_2 regularization has the same meaning as for Ridge regression. Besides, we also add thresholding via ℓ_1 regularization. This means that we select only important features corresponding to large values of x and throw away others. In this case we use the following notation: $f(x) = \lambda_1\|x\|_1$, $f_2(x) = \frac{\lambda_2}{2}\|x\|^2$, and $g(z) = \frac{1}{2}\|z-b\|^2$, i.e. $g(Ax) = \frac{1}{2}\|Ax-b\|^2$.

We already now the explicit form of the Fenchel-Legendre conjugate of g (see Lemma 6.1). Moreover, we don't have to use proximal operator of f_2, but f_2 is L-smooth with $L = \lambda_2$. The only thing that is still unknown is the proximal operator of $f = \|\cdot\|_1$.

Lemma 6.4. *The proximal operator of $\tau f(x)$, where $f(x) = \lambda_1 \|x\|_1$, is given by*

$$\left[\text{prox}_{\tau f}(x)\right]_i = \begin{cases} x_i - \tau\lambda_1 & \text{if } x_i > \tau\lambda_1 \\ 0 & \text{if } x_i \in [-\tau\lambda_1, \tau\lambda_1] \\ x_i + \tau\lambda_1 & \text{if } x_i < -\tau\lambda_1 \end{cases}.$$

Proof. Recall that the subdifferntial of $|x|$ can be given in the following way

$$\partial|x| = \begin{cases} 1 & \text{if } x > 0 \\ [-1, 1] & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{cases}.$$

Now we write the definition of proximal operator of τf

$$\text{prox}_{\tau f}(x) = \arg\min_{x'} \left[\tau\lambda_1 \|x'\|_1 + \frac{1}{2}\left\|x' - x\right\|^2\right].$$

Note, that the subproblem inside $\arg\min$ is separable, thus the solution can be found for each component separately

$$\left[\text{prox}_{\tau f}(x)\right]_i = \arg\min_{x'_i} \left[\tau\lambda_1 |x'_i| + \frac{1}{2}(x'_i - x_i)^2\right].$$

Since both functions in $\arg\min$ are convex with domain $\mathbb{R}$, then the Fermat rule can be written as follows

$$0 \in \tau\lambda_1 \partial|x'_i| + x'_i - x_i.$$

Now we consider all three cases. If $x'_i > 0$, then

$$0 = \tau\lambda_1 + x'_i - x_i \Rightarrow x'_i = x_i - \tau\lambda_1.$$

We get that this case could be realized if $x_i > \tau\lambda_1$. Now let be $x_i < 0$, then

$$0 = -\tau\lambda_1 + x'_i - x_i \Rightarrow x'_i = x_i + \tau\lambda_1.$$

This case is realized, if $x_i < -\tau\lambda_1$. Finally, if $x_i = 0$, then we obtain

$$0 \in [-\tau\lambda_1, \tau\lambda_1] + 0 - x_i \Rightarrow x_i \in [-\tau\lambda_1, \tau\lambda_1].$$

Combining all the above we derive

$$x'_i = \begin{cases} x_i - \tau\lambda_1 & \text{if } x_i > \tau\lambda_1 \\ 0 & \text{if } x_i \in [-\tau\lambda_1, \tau\lambda_1] \\ x_i + \tau\lambda_1 & \text{if } x_i < -\tau\lambda_1 \end{cases},$$

that concludes the proof. □

Note that the result above can be written as follows:

$$\text{prox}_{\tau\lambda_1\|x\|_1}(x) = \text{sign}(x) \max\{|x| - \tau\lambda_1, \mathbf{0}\},$$

where all functions work element-wise. For example, for $x \in \mathbb{R}^d$ we have

$$\begin{aligned} &\text{sign} : \mathbb{R}^d \to \mathbb{R}^d, \quad [\text{sign}(x)]_i = \text{sign}(x_i)\ \forall i \in [d], \\ &|\cdot| : \mathbb{R}^d \to \mathbb{R}^d, \quad [|x|]_i = |x_i|, \\ &\max : \mathbb{R}^d \to \mathbb{R}^d, \quad [\max\{x, \mathbf{0}\}]_i = \max\{x_i, 0\}, \end{aligned}$$

where $\mathbf{0}$ is a vector of zeros. Such explicit form allows efficient implementation of this proximal operator in practice.

Finally, we write the explicit of form of the Lagrangian function of this problem

$$L(x, y) = \lambda_1 \|x\|_1 + \frac{\lambda_2}{2}\|x\|^2 + \langle Ax, y\rangle - \frac{1}{2}\|y\|^2 - \langle y, b\rangle\,.$$

This function is λ_2-strongly convex in x and 1-strongly concave in y.

7 Experiments

In this section we demonstrate empirical performance of PDHG and PDHG with adaptive stepsize choice. We test PDHG with optimal stepsizes(9), PDHG with two proposed update rules, and adaptive PDHG of [Goldstein et al., 2015]. The performance is checked on a1a data set from LibSVM library[Chang and Lin, 2011]. We do preprocessing by deleting all zero columns. Thus, the final data matrix is of the size 1600×113. The perforamnce of all algorithms is tested on Ridge Regression and Elastic Net Regression problems described in Section 6.

For this two problems we assume that the strong convexity parameter in y is known (it is 1 following the results from Section 6), and the strong convexity parameter, i.e. regularization parameter λ is oppositely unknown.

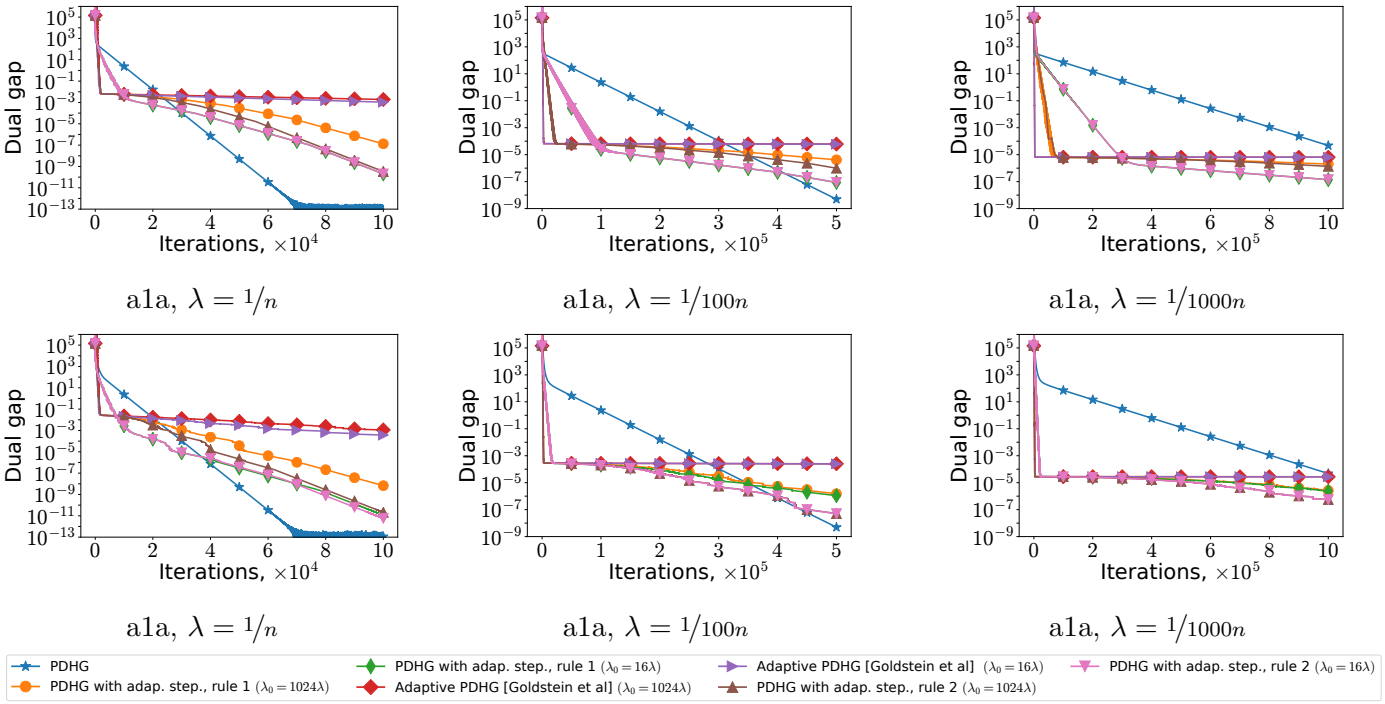

Figure 1: The comparison of PDHG, proposed adaptive PDHG with 2 types of update rules, and adaptive PDHG of [Goldstein et al., 2015]. **First row:** Ridge Regression, **Second row:** Elastic Net Regression.

According to the results presented in Figure 7, we observe that PDHG with optimal stepsizes is the best one for large enough regularization parameter $\lambda = 1/n$. Besides, the smaller λ, the slower PDHG with optimal stepsizes. However, proposed adaptive schemes become better than PDHG with optimal stepsizes when λ is decreasing. The adaptive schemes in all cases end with the regularization parameter estimator that is 2-4 times larger than the true one. Moreover, we clearly see that adaptive PDHG of [Goldstein et al., 2015] in all cases has fast dual gap decreasing in the beginning, and then a stognation. This means that the selected stepsizes by this adaptive schemes are far from the optimal ones, and are not suitable for these two class of problems.

Another statement we can make based on the results is that for $\lambda = 1/n$ (first column) the performance of PDHG with update rule #2 with initial λ estimator does not play a huge role. Indeed, the perfomance with

initial approximations $\lambda_0 = 16\lambda$ and $\lambda_0 = 1024\lambda$ is almost the same, but this is not the case for update rule #1. This means that the update rule #2 is more robust to the initial value of λ estimator. This makes the update rule #2 more practical.

8 Discussions and Future Work

We propose two adaptive schemes of stepsizes selection for PDHG. They provably work in practice and better than PDHG with optimal stepsizes in some cases.

In future works we plan to improve proposed adaptive schemes to get even faster methods. Besides, we would like to extend the theory to the case when both strong convexity parameters are unknown which is usually the case in practice.

References

Jean-François Aujol, Charles H Dossal, Hippolyte Labarrière, and Aude Rondepierre. FISTA restart using an automatic estimation of the growth parameter. 2021. working paper or preprint.

Amir Beck and Marc Teboulle. A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. *IMA Journal of Numerical Analysis 39(4), 2069–2095*, 2009.

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.

Chih-Chung Chang and Chih-Jen Lin. LibSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):1–27, 2011.

Simon S. Du and Wei Hu. Linear convergence of the primal-dual gradient method for convex-concave saddle point problems without strong convexity. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 196–205. PMLR, 16–18 Apr 2019.

Olivier Fercoq. Quadratic error bound of the smoothed gap and the restarted averaged primal-dual hybrid gradient. 2021. URL `https://hal.archives-ouvertes.fr/hal-03228252`.

Olivier Fercoq and Zheng Qu. Adaptive restart of accelerated gradient methods under local quadratic growth condition. *SIAM Journal on Imaging Sciences, 2(1), 183–202*, 2019.

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*, 2015.

Jingwei Liang, Jalal Fadili, and Gabriel Peyré. Convergence rates with inexact non-expansive operators. *Mathematical Programming 159(1-2), 403–434*, 2016a.

Jingwei Liang, Mohamed-Jalal Fadili, and Gabriel Peyré. Convergence rates with inexact non-expansive operators. *Mathematical Programming*, 159:403–434, 2016b.

Yura Malitsky and Konstantin Mishchenko. Adaptive gradient descent without descent. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6702–6712. PMLR, 13–18 Jul 2020.

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

R. Tyrrell Rockafellar and Roger J.-B. Wets. *Variational Analysis*. Springer Verlag, Heidelberg, Berlin, New York, 1998.

Quoc Tran-Dinh, Olivier Fercoq, and Volkan Cevher. A smooth primal-dual optimization framework for nonsmooth composite convex minimization. *SIAM Journal on Optimization 28(1), 96–134*, 2018.

Maria-Luiza Vladarean, Yura Malitsky, and Volkan Cevher. A first-order primal-dual method with adaptivity to local smoothness. In *Proceedings of the 34th Advances in Neural Information Processing Systems 34 pre-proceedings (NeurIPS 2021)*, 2021.

Tatsuya Yokota and Hidekata Hontani. An efficient method for adapting step-size parameters of primal-dual hybrid gradient method in application to total variation regularization. In *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 973–979, 2017.

Mingqiang Zhu and Tony Chan. An efficient primal-dual hybrid gradient algorithm for total variation image restoration. 2008.