OJMO

Open Journal of Mathematical Optimization

Olivier Fercoq

Quadratic error bound of the smoothed gap and the restarted averaged primal-dual hybrid gradient

Volume 4 (2023), article no. 6 (34 pages)

https://doi.org/10.5802/ojmo.26

Article submitted on May 31, 2022, revised on April 6, 2023,
accepted on July 3, 2023.

Open Journal of Mathematical Optimization is a member of the
Centre Mersenne for Open Scientific Publishing
http://www.centre-mersenne.org/
e-ISSN: 2777-5860

CENTRE MERSENNE

Quadratic error bound of the smoothed gap and the restarted averaged primal-dual hybrid gradient

Olivier Fercoq
olivier.fercoq@telecom-paris.fr
LTCI, Télécom Paris, Institut Polytechnique de Paris, France

Abstract

We study the linear convergence of the primal-dual hybrid gradient method. After a review of current analyses, we show that they do not explain properly the behavior of the algorithm, even on the most simple problems. We thus introduce the quadratic error bound of the smoothed gap, a new regularity assumption that holds for a wide class of optimization problems. Equipped with this tool, we manage to prove tighter convergence rates. Then, we show that averaging and restarting the primal-dual hybrid gradient allows us to leverage better the regularity constant. Numerical experiments on linear and quadratic programs, ridge regression and image denoising illustrate the findings of the paper.

Digital Object Identifier 10.5802/ojmo.26

Keywords linear convergence; primal-dual algorithm; error bound; restart.

Acknowledgments This work was supported by the Agence National de la Recherche grant ANR-20-CE40-0027, Optimal Primal-Dual Algorithms (APDO).

1 Introduction

Primal-dual algorithms are widely used for the resolution of optimization problems with constraints. Thanks to them, we can replace complex nonsmooth functions like those encoding the constraints by simpler, sometimes even separable functions, at the expense of solving a saddle point problem instead of an optimization problem. Then, this amounts to replacing a complex optimization problem by a sequence of simpler problems. In this paper, we shall consider more specifically

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

where f and g are convex with easily computable proximal operators, $A : \mathcal{X} \to \mathcal{Y}$ is a linear operator and f_2 and g_2^* are differentiable with L_f and L_{g^*} lipschitz gradients. Here, $g \,\square\, g_2(z) = \inf_y g(y) + g_2(z - y)$ is the infimal convolution of g. and g_2. To encode constraints, we just need to consider an indicator function for g. When using a primal-dual method, one is looking for a saddle point of the Lagrangian, which is given by

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

Of course, we shall assume throughout this paper that saddle points do exist, which can be guaranteed using conditions like Slater's constraint qualification condition [4].

A natural question is then: at what speed do primal-dual algorithms converge? This is trickier for saddle point problems than when we deal with a problem which is in primal form only. For instance, if we just assume convexity, methods like Primal-Dual Hybrid Gradient (PDHG) [6] or Alternating Directions Method of Multipliers (ADMM) [17] can be very slow, with a rate of convergence in the worst case in $O(1/\sqrt{k})$ [10]. Yet, if we average the iterates, we obtain an ergodic rate in $O(1/k)$. Nevertheless, it has been observed that, except for specially designed counter-examples, the averaged algorithms usually perform less well that the plain algorithm.

This is not unexpected. Indeed, the problem you are interested in has no reason to be the most difficult convex problem. In order to get a more positive answer, we should understand what makes a given problem easier to solve than another. In the case of gradient descent, strong convexity of the objective function implies a linear rate of convergence, and the more strongly convex the function, the faster is the algorithm. Strong convexity can be generalized to the objective quadratic error bound (QEB) and the Kurdyka–Łojasiewicz inequality in order to show improved rates for a large class of functions [5].

Before going further, let us discuss how one quantifies convergence speed for saddle point problems. Several measures of optimality have been considered in the literature. The most natural one is feasibility error and optimality gap. It directly fits the definition of the optimization problem at stake. However, one cannot compute the optimality gap before the problem is solved. Hence, in algorithms, we usually use the Karush–Kuhn–Tucker (KKT) error instead. It is a computable quantity and if the Lagrangian's gradient is metrically subregular [28], then a small KKT error implies that the current point is close to the set of saddle points. When the primal and dual domains are bounded, the duality gap is a very good way to measure optimality: it is often easily computable and it is an upper bound to the optimality gap. A generalization to unbounded domains has been proposed in [30]: the smoothed gap, based on the smoothing of nonsmooth functions [25], takes finite values even for constrained problems, unlike the duality gap. Moreover, if the smoothness parameter is small and the smoothed gap is small, this means that optimality gap and feasibility error are both small. In the present paper, we shall reuse this concept not only for showing a convergence speed but also to define a new regularity assumption that we believe is better suited to the study of primal-dual algorithms.

Regularity conditions for saddle point problems have been investigated more recently than for plain optimization problems. The most successful one is the metric subregularity of the Lagrangian's generalized gradient [22]. It holds among others for all linear-quadratic programs [21] and implies a linear convergence rate for PDHG and ADMM, as well as the proximal point algorithm [24]. One can also show linear convergence if the objective is smooth and strongly convex and the constraints are affine [2, 13, 29]. If the function defined as the maximum between objective gap and constraint error has the error bound property, then we can also show improved rates [23]. These result can also be extended to the coordinate descent case [1, 32], as well as the setup of distributed computations where doing less communication steps is an important matter [20]. The other assumptions look more restrictive because they require some form of strong convexity. Yet, we will see that for a problem that satisfies two assumptions, the rate predicted by each theory may be different.

Our contribution is as follows.

- In Section 2, we formally review the main regularity assumptions and do first comparisons.
- In order to do deeper comparisons, we analyze PDHG in detail in Sections 3 and 4 under each assumption. This choice is motivated by the self-containedness of the method, which does not require to solve any subproblem.
- In Section 5, we show that the present regularity assumptions may not reflect properly the behavior of PDHG, even on a very simple optimization problem.
- We introduce a new regularity assumption in Section 6: the quadratic error bound of the smoothed gap. We then show its advantages against previous approaches. The smoothed gap was introduced in [30] as a tool to analyse and design primal-dual algorithms. Here, we use it directly in the definition of the regularity assumption. We analyze PDHG under this assumption in Section 7
- We then present and analyze the Restarted Averaged Primal-Dual Hybrid Gradient (RAPDHG) in Section 8 and show that is some situations, it leads to a faster algorithm. An adaptive restart scheme is also presented for the cases where the regularity parameters are not known. This is a first step in leveraging our new understanding of saddle point problems to design more efficient algorithms.
- The theoretical results are illustrated in Section 9, devoted to numerical experiments.

We note striking similarities between this paper and the concurrent work of Applegate, Hinder, Lu and Lubin [3]. Although they focus on linear programs, the authors analyse PDHG and other first order methods thanks to the sharpness of the restricted duality. Indeed, in the case of linear programs, the restricted duality gap is a computable finite-valued measure of optimality and it is always sharp. The methodology is very similar except that the arguments are taylored to linear programs.

2 Regularity assumptions for saddle point problems

In this section, we define three regularity assumptions for saddle point problems from the literature. We will then present their application range.

2.1 Notation

We shall denote $\mathcal{X}$ the primal space and $\mathcal{Y}$ the dual space. We assume that thoses vector spaces are Hilbert spaces. Let us denote $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ the primal-dual space. Similarly for a primal vector x and a dual vector y, we shall denote $z = (x, y)$. This notation will be throughout the paper: for instance $\bar{x}$ and $\bar{y}$ will be the primal and dual parts of the vector $\bar{z}$. For $z = (x, y) \in \mathcal{Z}$, and $\tau, \sigma > 0$, we denote $\|z\|_V = (\frac{1}{\tau}\|x\|^2 + \frac{1}{\sigma}\|y\|^2)^{1/2}$ and $\langle z, z'\rangle_V = \frac{1}{\tau}\langle x, x'\rangle + \frac{1}{\sigma}\langle y, y'\rangle$. The proximal operator of a function f is given by $\mathrm{prox}_f(x) = \arg\min_{x'} f(x') + \frac{1}{2}\|x - x'\|^2$. For a set-value function $F : \mathcal{Z} \rightrightarrows \mathcal{Z}$, we can define $F^{-1} : \mathcal{Z} \rightrightarrows \mathcal{Z}$ by $w \in F(z) \Leftrightarrow z \in F^{-1}(w)$. We will make use of the convex indicator function

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

In order to ease reading of the paper, we shall use a blue font for results that use differentiable parts of the objective f_2 and g_2 and an orange font for results that use strong convexity.

2.2 Definitions

The simplest regularity assumption is strong convexity.

Definition 1. *A function $f : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ is* μ-strongly convex *if $f - \frac{\mu}{2}\|\cdot\|^2$ is convex.*

Assumption 2. The Lagrangian function is μ-strongly convex-concave, that is $(x \mapsto L(x, y))$ is μ-strongly convex for all y and $(y \mapsto L(x, y))$ is μ-strongly concave for all x.

This regularity assumption is used for instance in [6]. We can generalize strong convexity as follows.

Definition 3. *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 $\mathcal{R} \subseteq \mathcal{X}$ that contains $\arg\min f$ such that for all $x \in \mathcal{R}$,*

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

We shall use the acronym f has a η-QEB.

Although this is more general than strong convexity, the quadratic error bound is an assumption which is not general enough for saddle point problems. Indeed, for the fundamental class of problems with linear constraints $(y \mapsto L(x, y)$ is linear. Thus, it cannot satisfy a quadratic error bound in y. To resolve this issue, we may resort to metric regularity.

Definition 4. *A set-valued function $F : \mathcal{Z} \rightrightarrows \mathcal{Z}$ is* metrically subregular *at z for b if there exists $\eta > 0$ and a neighborhood $N(z)$ of z such that $\forall\, z' \in N(z)$,*

$$\operatorname{dist}(F(z'), b) \geq \eta \operatorname{dist}(z', F^{-1}(b))$$

We denote $C(z) = \partial f(x) \times \partial g^*(y)$ (where $\times$ denotes the Cartesian product), $B(z) = [\nabla f_2(x), \nabla g_2^*(y)]$ and $M(z) = [A^\top y, -Ax]$. The Lagrangian's subgradient is then $\widetilde{\partial} L(z) = (B + C + M)(z)$. We put a tilde to emphasize the fact that the dual component is the negative of the supergradient. We shall use the term *generalized gradient*.

We have $0 \in \widetilde{\partial} L(z^*)$ if and only if z^* is a saddle point of L. If $\widetilde{\partial} L$ is metrically sub-regular at z^* for 0, this means that we can measure the distance to the set of saddle points with the distance of the subgradient to 0.

Assumption 5. The Lagrangian's generalized gradient is metrically subregular, that is there exists η such that for all $z^* \in \mathcal{Z}^* = (\widetilde{\partial} L)^{-1}(0)$, $\widetilde{\partial} L$ is η-metrically subregular at z^* for 0.

This regularity assumption is used for instance in [22]. Another regularity assumption considered in the literature is as follows.

Assumption 6. The problem is a smooth strongly convex linearly constrained problem. Said otherwise, $f + f_2$ is strongly convex and differentiable, f and f_2 both have a Lipschitz continuous gradient, $g_2 = \iota_{\{0\}}$ and $g = \iota_{\{b\}}$, where $b \in \mathcal{Y}$.

This assumption is used for instance in [13]. The indicator functions encode the constraint $Ax = b$.

Table 1 Domain of applicability of each assumption. "Strongly convex & smooth" means that $g \square g_2$ is a differentiable function and $f + f_2$ is strongly convex.

Assumption	Strongly convex & smooth	Linear program	Quadratic program
Strongly convex-concave	Yes	No	No
Smooth strongly convex with linear constraints	Solve in primal space only	No	Strongly convex obj. & linear constraints
Error bound with inequality constraints	No	Yes	No
Metric sub-regularity	Yes	Yes	Yes

Assumption 7. Suppose that $g_2 = \iota_{\{0\}}$ and $g = \iota_{b+\mathbb{R}^m_-}$ and we encode the constraints $Ax - b \leq 0$. Denote x^* a minimizer of (1) and $\mathcal{X}^*$ the set of minimizers. The problem with inequality constraints satisfies the error bound if there exists $\mu > 0$ such that

$$F(x) = \max\Big(f(x) + f_2(x) - f(x^*) - f_2(x^*), \max_{1\leq j\leq m} (Ax - b)_j\Big) \geq \mu \operatorname{dist}(x, \mathcal{X}^*)$$

This regularity assumption is used to deal with functional inequality constraints in [23] but we restrict our study to linear inequalities to simplify the exposition of this paper. Yet, since it involves primal quantities only, it is not really adapted to a primal-dual algorithm and we will not discuss it much further in this paper.

The next two propositions show that for the minimization of a convex function, quadratic error bound of the objective is merely equivalent to metric subregularity of the subgradient.

Proposition 8 ([12, Theorem 3.3])**.** *Let f be a convex function such that $\forall\, x \in \mathcal{R}$, $f(x) \geq f(x^*) + \frac{\mu}{2} \operatorname{dist}(x, \mathcal{X}^*)^2$, where $\mathcal{X}^* = \arg\min f$ and $x^* \in \mathcal{X}^*$. Then $\forall\, x \in \mathcal{R}$, $\|\partial f(x)\|_0 = \inf_{g\in\partial f(x)} \|g\| \geq \frac{\mu}{2} \operatorname{dist}(x, \mathcal{X}^*)$.*

Proposition 9 ([12, Theorem 3.3])**.** *Let f be a convex function such that $f(x) \leq f_0$ implies $\|\partial f(x)\|_0 \geq \eta \operatorname{dist}(x, \mathcal{X}^*)$. Then $f(x) \geq f(x^*) + \frac{\eta}{2} \operatorname{dist}(x, \mathcal{X}^*)^2$ as soon as $f(x) \leq f_0$.*

For saddle point problems, we have the following result.

Proposition 10 ([21, Lemma 4.2])**.** *If L is μ-strongly convex-concave, then $\widetilde{\partial} L$ is μ-metrically sub-regular at z^* for 0 where z^* is the unique saddle point of L.*

In Table 1, we can see that the situation is more complex for saddle point problems than plain optimization problems. Indeed, the assumptions are not generalizations one of the other. Yet, metric subregularity seems to be the most general since it holds for more types of problems. In particular all linear programs and quadratic programs have a metrically subregular Lagrangian's generalized gradient [21].

3 Basic inequalities for the study of PDHG

Primal-Dual Hybrid Gradient (also known as asymmetric forward-backward-adjoint) is the algorithm defined by Algorithm 1. We shall use the definition of [21] rather than [8, 31] because we believe it simplifies the analysis.

Algorithm 1 Primal-Dual Hybrid Gradient (PDHG)

$$\begin{aligned}
\bar{x}_{k+1} &= \operatorname{prox}_{\tau f}(x_k - \tau\nabla f_2(x_k) - \tau A^\top y_k)\\
\bar{y}_{k+1} &= \operatorname{prox}_{\sigma g^*}(y_k - \sigma\nabla g_2^*(y_k) + \sigma A\bar{x}_{k+1})\\
x_{k+1} &= \bar{x}_{k+1} - \tau A^\top(\bar{y}_{k+1} - y_k)\\
y_{k+1} &= \bar{y}_{k+1}
\end{aligned}$$

Note that the algorithm of Chambolle and Pock [6] can be recovered in the case $f_2 = 0$ by taking $\bar{z}_{k+1}$ as a state variable instead of z_{k+1} and using $x_k = \bar{x}_k - \tau A^\top(y_k - y_{k-1}) = \bar{x}_k - \tau A^\top(\bar{y}_k - \bar{y}_{k-1})$:

$$\begin{aligned}
\bar{x}_{k+1} &= \operatorname{prox}_{\tau f}(\bar{x}_k - \tau A^\top(2\bar{y}_k - \bar{y}_{k-1}))\\
\bar{y}_{k+1} &= \operatorname{prox}_{\sigma g^*}(\bar{y}_k - \sigma\nabla g_2^*(\bar{y}_k) + \sigma A\bar{x}_{k+1})
\end{aligned}$$

PDHG is widely used for the resolution of large-dimensional convex-concave saddle point problems. Indeed, this algorithm only requires simple operations, namely matrix-vector multiplications, proximal operators and gradients, while keeping good convergence properties. We refer the reader to [9] for a review of variants of the algorithm and their analysis. As shown in [19], the proof techniques for all these variants share strong similarities and we believe that the results of the present paper could be easily adapted to them.

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

$$\begin{aligned} &\bar{x} = \mathrm{prox}_{\tau f}(x - \tau \nabla f_2(x) - \tau A^\top y) && \bar{y} = \mathrm{prox}_{\sigma g^*}(y - \sigma \nabla g_2^*(y) + \sigma A \bar{x}) \\ &x^+ = \bar{x} - \tau A^\top(\bar{y} - y) && y^+ = \bar{y} \\ &T(x,y) = (x^+, y^+) \end{aligned} \tag{3}$$

For $z = (x, y) \in \mathcal{Z}$, we denote $\|z\|_V = (\frac{1}{\tau}\|x\|^2 + \frac{1}{\sigma}\|y\|^2)^{1/2}$, $\gamma = \sigma\tau\|A\|^2$, $\alpha_f = \tau L_f/2$, $\alpha_g = \sigma L_{g^*}/2$ and

$$\begin{aligned} \widetilde{V}(z, z') &= \frac{1 - \tau L_f/2}{2\tau}\|\bar{x} - x - \bar{x}' + x'\|^2 + \left(\frac{1 - \sigma L_{g^*}/2}{2\sigma} - \frac{\tau\|A\|^2}{2}\right)\|\bar{y} - y - \bar{y}' + y'\|^2 \\ &= \frac{1 - \alpha_f}{2\tau}\|\bar{x} - x - \bar{x}' + x'\|^2 + \frac{1 - \alpha_g - \gamma}{2\sigma}\|\bar{y} - y - \bar{y}' + y'\|^2 \ . \end{aligned}$$

We will first show that the fixed point operator T is an averaged operator [4] in this norm. Then, we will give an upper bound on the Lagrangian's gap and a convergence result. All the results are small variations of already known facts so we defer the proofs to the appendix. Note that we may have adapted the results for our purpose.

Lemma 11 ([4, Proposition 12.26])**.** *Let $p = \mathrm{prox}_{\tau f}(x)$ and $p' = \mathrm{prox}_{\tau f}(x')$ where f is μ_f-strongly convex. For all x and x',*

$$\begin{aligned} &f(p) + \frac{1}{2\tau}\|p - x\|^2 \le f(x') + \frac{1}{2\tau}\|x' - x\|^2 - \frac{1 + \tau\mu_f}{2\tau}\|p - x'\|^2 \\ &(1 + 2\tau\mu_f)\|p - p'\|^2 \le \|x' - x\|^2 - \|p - x - p' + x'\|^2 \ . \end{aligned}$$

The following lemma can be mostly found in [21, Theorem 2.5]. In comparison, we write everything in the same norm $\|\cdot\|_V$ and we do not restrict to z' being a saddle point of the Lagrangian.

Lemma 12. *Let $T : \mathcal{X} \times \mathcal{Y} \to \mathcal{X} \times \mathcal{Y}$ be defined for any (x, y) by (3). Suppose that ∇f_2 is L_f-Lipschitz continuous and ∇g_2^* is L_{g^*}-Lipschitz continuous. If the step sizes satisfy $\gamma = \sigma\tau\|A\|^2 < 1$, $\alpha_f = \tau L_f/2 < 1$, $\alpha_g = \sigma L_{g^*}/2 < 1$ then T is nonexpansive in the norm $\|\cdot\|_V$,*

$$\|T(z) - T(z')\|_V^2 \le \|z - z'\|_V^2 - 2\widetilde{V}(z, z') \tag{4}$$

and T is $\frac{1}{1+\lambda}$-averaged where

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

which means for $z = (x, y)$ and $z' = (x', y')$

$$\|T(z) - T(z')\|_V^2 \le \|z - z'\|_V^2 - \lambda\|z - T(z) - z' + T(z')\|_V^2 \ . \tag{5}$$

As a consequence, (z_k) converges to a saddle point of the Lagrangian. Moreover, if $\sigma L_{g^*}/2 \le \alpha_f(1 - \sigma\tau\|A\|^2)$, then $\lambda \ge (1 - \sqrt{\gamma})(1 - \alpha_f)$.

A side result of independent interest proved within Lemma 12 is as follows.

Lemma 13. *For any $z^* \in \mathcal{Z}^*$, $\widetilde{V}$ satisfies*

$$\widetilde{V}(z_k, z^*) = \frac{1 - \alpha_f}{2\tau}\|\bar{x}_{k+1} - x_k\|^2 + \frac{1 - \alpha_g - \gamma}{2\sigma}\|\bar{y}_{k+1} - y_k\|^2 \ge \frac{\lambda}{2}\|z_{k+1} - z_k\|_V^2 \ .$$

As noted in [19], the case $\alpha_f > \frac{1}{2}$ is not covered by most of the results in the literature on convergence speed results. We propose here an extension of results in the proof of [6, Theorem 1] that allows the larger step size range $0 \le \alpha_f < 1$ where convergence is guaranteed.

Lemma 14. *Suppose that $\gamma = \sigma\tau\|A\|^2 < 1$, $\tau L_f/2 = \alpha_f < 1$, $\alpha_g = \sigma L_{g^*}/2 < 1$. For all $k \in \mathbb{N}$ and for all $z \in \mathcal{Z}$,*

$$L(\bar{x}_{k+1}, y) - L(x, \bar{y}_{k+1}) \leq \frac{1}{2}\|z - z_k\|_V^2 - \frac{1}{2}\|z - z_{k+1}\|_V^2 + a_2\widetilde{V}(z_k, z^*) \tag{6}$$

where $\widetilde{V}(z_k, z^*) = (\frac{1}{2\tau} - \frac{L_f}{2})\|\bar{x}_{k+1} - x_k\|^2 + (\frac{1}{2\sigma} - \frac{\tau\|A\|^2}{2} - \frac{L_{g^*}}{2})\|\bar{y}_{k+1} - y_k\|^2$ and $a_2 = \max(\frac{2\alpha_f - 1}{1-\alpha_f}, \frac{2\alpha_g - 1 + \gamma}{1 - \alpha_g - \gamma})$. $a_2 \geq -1$ may be positive or negative.

The next proposition is adapted from [6, Theorem 1]. We shall show in Section 8 how to generalize it to $\tau L_f < 2$.

Proposition 15. *Let $z_0 \in \mathcal{Z}$ and let $R \subseteq \mathcal{Z}$. If $\sigma\tau\|A\|^2 + \sigma L_{g^*} \leq 1$ and $\tau L_f \leq 1$ then we have the stability*

$$\|z_k - z^*\|_V \leq \|z_0 - z^*\|_V$$

for all $z^* \in \mathcal{Z}^*$. Define $\widetilde{z}_k = \frac{1}{k}\sum_{l=1}^k \bar{z}_l$ and the restricted duality gap $G(\bar{z}, R) = \sup_{z\in R} L(\bar{x}, y) - L(x, \bar{y})$. We have the sublinear iteration complexity

$$G(\widetilde{z}_k, R) \leq \frac{1}{2k}\sup_{z\in R}\|z - z_0\|_V^2 .$$

4 Linear convergence of PDHG

In this section, we show that under the regularity assumptions stated in Section 2, the Primal-Dual Hybrid Gradient converges linearly. Most of the results were already known, we only improved slightly some constants. Hence, in this section also, we defer some of the proofs to Appendix B.

We begin with a technical lemma showing that $\bar{z}_{k+1}$ is close to z_{k+1}.

Lemma 16. *For $0 < \alpha \leq 1$,*

$$\mathrm{dist}_V(\bar{z}_{k+1}, \mathcal{Z}^*)^2 \geq (1-\alpha)\,\mathrm{dist}_V(z_{k+1}, \mathcal{Z}^*)^2 - (\alpha^{-1} - 1)\frac{1}{\sigma}\|y_{k+1} - y_k\|^2 .$$

Proof. We use the fact that for any z, $\|z_{k+1} - P_{\mathcal{Z}^*}(z)\|_V^2 \geq \mathrm{dist}_V(z_{k+1}, \mathcal{Z}^*)^2$ and Young's inequality to get

$$\begin{aligned}
\mathrm{dist}_V(\bar{z}_{k+1}, \mathcal{Z}^*)^2 &= \|\bar{z}_{k+1} - z_{k+1} + z_{k+1} - P_{\mathcal{Z}^*}(\bar{z}_{k+1})\|_V^2 \\
&= \|z_{k+1} - P_{\mathcal{Z}^*}(\bar{z}_{k+1})\|_V^2 + \|\bar{z}_{k+1} - z_{k+1}\|_V^2 + 2\langle z_{k+1} - P_{\mathcal{Z}^*}(\bar{z}_{k+1}), \bar{z}_{k+1} - z_{k+1}\rangle_V \\
&= \|z_{k+1} - P_{\mathcal{Z}^*}(\bar{z}_{k+1})\|_V^2 + \frac{1}{\tau}\|\bar{x}_{k+1} - x_{k+1}\|^2 + \frac{2}{\tau}\langle x_{k+1} - P_{\mathcal{X}^*}(\bar{x}_{k+1}), \bar{x}_{k+1} - x_{k+1}\rangle \\
&\geq \frac{1}{\sigma}\mathrm{dist}(y_{k+1}, \mathcal{Y}^*)^2 + \frac{1}{\tau}(1-\alpha)\,\mathrm{dist}(x_{k+1}, \mathcal{X}^*)^2 - \frac{1}{\tau}(\alpha^{-1} - 1)\|\bar{x}_{k+1} - x_{k+1}\|^2 \\
&\geq (1-\alpha)\,\mathrm{dist}_V(z_{k+1}, \mathcal{Z}^*)^2 - \frac{1}{\tau}(\alpha^{-1} - 1)\|\bar{x}_{k+1} - x_{k+1}\|^2
\end{aligned}$$

for all $\alpha \in (0, 1)$. Since $\frac{1}{\tau}\|\bar{x}_{k+1} - x_{k+1}\|^2 = \tau\|A^\top(y_{k+1} - y_k)\|^2 \leq \frac{1}{\sigma}\|y_{k+1} - y_k\|^2$, we get the result of the lemma. ◀

The next proposition is a modification of [14, Theorem 4] in order to allow $\alpha_f < 1$ instead of $\alpha_f \leq 1/2$. Here, we also concentrate on the deterministic version of PDHG. We put the proof in the main text because the proof of Theorem 28 in Section 7 will reuse some of the arguments.

Proposition 17. *If L is μ-strongly convex concave in the norm $\|\cdot\|_V$, then the iterates of PDHG satisfy for all k,*

$$\left(1 + \frac{\mu}{(2 + a_2)(1 + \mu/\lambda)}\right)\|z_{k+1} - z^*\|_V^2 \leq \|z_k - z^*\|_V^2$$

where z^* is the unique saddle point of L, $a_2 = \max(\frac{2\alpha_f - 1}{1-\alpha_f}, \frac{\gamma + 2\alpha_g - 1}{1 - \gamma - \alpha_g})$ and λ is defined in Lemma 12.

Proof. From Lemma 14 applied at $z = z^*$, we have

$$L(\bar{x}_{k+1}, y^*) - L(x^*, \bar{y}_{k+1}) \leq \frac{1}{2}\|z^* - z_k\|_V^2 - \frac{1}{2}\|z^* - z_{k+1}\|_V^2 + a_2 \widetilde{V}(\bar{z}_{k+1} - z_k) \;.$$

In order to deal with the case $a_2 \geq 0$, we add to this inequatity a times (4), where $a \geq 0$, $z = z_k$ and $z' = z^*$

$$L(\bar{x}_{k+1}, y^*) - L(x^*, \bar{y}_{k+1}) \leq \frac{1+a}{2}\|z^* - z_k\|_V^2 - \frac{1+a}{2}\|z^* - z_{k+1}\|_V^2 + (a_2 - a)\widetilde{V}(z_k, z^*) \;.$$

Since L is μ-strongly convex-concave, $(x \mapsto L(x, y^*))$ is minimized at x^* and $(y \mapsto L(x^*, y))$ is minimized at y^*, we have

$$L(\bar{x}_{k+1}, y^*) - L(x^*, \bar{y}_{k+1}) \geq \frac{\mu}{2}\|\bar{x}_{k+1} - x^*\|_{\tau^{-1}}^2 + \frac{\mu}{2}\|\bar{y}_{k+1} - y^*\|_{\sigma^{-1}}^2 \;.$$

We combine these two inequalities with Lemma 13 and Lemma 16 to get for all $\alpha \in (0,1)$ and $a \geq \max(0, a_2)$

$$(1{+}a + \mu(1-\alpha))\|z_{k+1} - z^*\|_V^2 \leq (1+a)\|z_k - z^*\|_V^2 + \frac{1}{\sigma}(\mu(\alpha^{-1} - 1) - \lambda(a_2 - a))\|y_{k+1} - y_k\|^2 \;.$$

We then choose $\alpha = \frac{\mu}{\lambda(a - a_2) + \mu}$ so that $\mu(\alpha^{-1} - 1) = \lambda(a - a_2)$ and we choose $a = a_2 + 1 \geq 0$. Thus

$$\left(2 + a_2 + \frac{\mu\lambda}{\mu + \lambda}\right)\|z_{k+1} - z^*\|_V^2 \leq (2 + a_2)\|z_k - z^*\|_V^2 \;. \qquad ◀$$

We next study the second case where some primal-dual methods have been proved to have a linear rate of convergence [2, Theorem 1], [13], [29, Theorem 6.2], that is, minimizing a strongly convex objective under affine equality constraints. Here also, we pay attention to allow $1/2 < \alpha_f < 1$ in our proof.

Proposition 18. *If $f + f_2$ has a $L'_f + L_f$-Lipschitz gradient and is μ_f-strongly convex, and $g + g_2 = \iota_{\{b\}}$, then PDHG converges linearly with rate*

$$\left(1 + \frac{\eta}{(2 + a_2)(1 + \eta/\lambda)}\right)\mathrm{dist}_V(z_{k+1}, \mathcal{Z}^*)^2 \leq \mathrm{dist}_V(z_k, \mathcal{Z}^*)^2$$

where $\eta = \min(\mu_f \tau, \frac{\sigma\tau\sigma_{\min}(A)^2}{\tau L_f + \tau L'_f + \frac{1}{\lambda}})$, λ is defined in Lemma 12 and $a_2 \geq -1$ is defined in Lemma 14.

Note that this does not contradict the lower bound of [27]. In [27], the authors consider the setup where the number of iterations is smaller than the dimension of the problem and showed that the convergence is necessarily sublinear in the worst case. On the other hand, our result becomes useful after a number of iterations that may be large for ill-conditioned problems but is more optimistic.

Finally, we will show that if the Lagrangian's generalized gradient is metrically sub-regular then PDHG converges linearly. Compared to [21, Theorem 5], we obtain a rate where the dependence in the norm is directly taken into account in the definition of metric sub-regularity and does not appear explicitly in the rate.

Proposition 19. *If $\widetilde{\partial} L$ is metrically subregular at z^* for 0 for all $z^* \in \mathcal{Z}^*$ with constant $\eta > 0$ in the norm $\|\cdot\|_V$, then $(I - T)$ is metrically subregular at z^* for 0 for all $z^* \in \mathcal{Z}^*$ with constant bounded below by $\frac{\eta}{\sqrt{3}\eta + (2 + 2\sqrt{3}\max(\alpha_f, \alpha_g))}$ and PDHG converges linearly with rate $1 - \frac{\eta^2\lambda}{(\sqrt{3}\eta + (2 + 2\sqrt{3}\max(\alpha_f, \alpha_g)))^2}$.*

5 Coarseness of the analysis

5.1 Strongly convex-concave Lagrangian

Suppose that f is μ_f strongly convex and that g^* is μ_{g^*} strongly convex. Then L is μ_L strongly convex in the norm $\|\cdot\|_V$ with $\mu_L = \min(\mu_f \tau, \mu_{g^*}\sigma)$. Note that in this case, the objective is the sum of the differentiable term $g(Ax)$ and the strongly convex proximable term $f(x)$. We have seen that this implies a linear rate of convergence for PDHG with rate $(1 - c\mu_L)$ with c close to 1. We may wonder what is the choice of τ and σ that leads to the best rate.

We need $\mu_L = \min(\mu_f \tau, \mu_{g^*}\sigma)$ the largest possible and $\sigma\tau\|A\|^2 \leq 1$. Hence, we take $\tau = \sqrt{\frac{\mu_{g^*}}{\mu_f}}\frac{1}{\|A\|}$ and $\sigma = \sqrt{\frac{\mu_f}{\mu_{g^*}}}\frac{1}{\|A\|}$. We do have $\sigma\tau\|A\|^2 \leq 1$ and also $\eta = \frac{\sqrt{\mu_f \mu_{g^*}}}{\|A\|}$. This rate is optimal for this class of problem [26], which is noticeable.

We have seen in Proposition 10 that having a strongly convex concave Lagrangian implies the metric sub-regularity of the Lagrangian's gradient. However, applying Proposition 19 with $\eta = \mu_L$ leads to a rate equal to $(1 - c\mu_L^2)$ which is much worse than what we can show using the more specialized assumption. This means that metric sub-regularity applies to more problems but is not a more general assumption because it leads to a coarser analysis.

5.2 Quadratic problem

We consider the toy problem

$$\min_{x \in \mathsf{R}} \frac{\mu}{2}x^2 \qquad ax = b$$

where $a, b \in \mathbb{R}$ and $\mu \geq 0$.

The Lagrangian is given by $L(x,y) = \frac{\mu}{2}x^2 + y(ax - b)$. Its gradient is $\nabla L(x,y) = [\mu x + ay, ax - b]$. Since ∇L is affine, we can see using an eigenvalue decomposition that ∇L is globally metrically sub-regular with constant $\frac{\sqrt{\mu^2\tau^2 + 4\sigma\tau a^2} - \mu\tau}{2}$ in the norm $\|\cdot\|_V$. We can also do a direct calculation. For all $\alpha > 0$ and the unique primal-dual optimal pair x^*, y^*,

$$\begin{aligned}
\|\nabla L(x,y)\|_{V*}^2 &= \tau\|\mu x + ay\|^2 + \sigma\|ax - b\|^2 = \tau\|\mu x - \mu x^* + ay - ay^*\|^2 + \sigma\|ax - ax^*\|^2 \\
&= (\tau\mu^2 + \sigma a^2)\|x - x^*\|^2 + \tau a^2\|y - y^*\|^2 + 2\tau\mu a\langle x - x^*, y - y^*\rangle \\
&\geq (\tau^2\mu^2 + \sigma\tau a^2 - \tau^2\mu a\alpha)\frac{1}{\tau}\|x - x^*\|^2 + (\sigma\tau a^2 - \mu\sigma\tau a\alpha^{-1})\frac{1}{\sigma}\|y - y^*\|^2 .
\end{aligned}$$

We choose $\alpha > 0$ such that $\tau^2\mu^2 + \sigma\tau a^2 - \tau^2\mu a\alpha = \sigma\tau a^2 - \mu\sigma\tau a\alpha^{-1}$, that is $\alpha = \frac{\tau\mu + \sqrt{\tau^2\mu^2 + 4\sigma\tau a^2}}{2\tau a}$, which leads to

$$\|\nabla L(x,y)\|_{V*}^2 \geq \left(\frac{\tau^2\mu^2}{2} + \sigma\tau a^2 - \frac{\tau\mu}{2}\sqrt{\tau^2\mu^2 + 4\sigma\tau a^2}\right)\|z - z^*\|^2 = \left(\frac{\sqrt{\mu^2\tau^2 + 4\sigma\tau a^2} - \mu\tau}{2}\right)^2 \|z - z^*\|^2 .$$

Let us now try to solve this (trivial) problem using PDHG:

$$\begin{aligned}
\bar{x}_{k+1} &= x_k - \tau(\mu x_k + ay_k) \\
\bar{y}_{k+1} &= y_k - \sigma(b - a\bar{x}_{k+1}) \\
x_{k+1} &= \bar{x}_{k+1} - \tau a(\bar{y}_{k+1} - y_k) \\
y_{k+1} &= \bar{y}_{k+1}
\end{aligned}$$

This can be written $z_{k+1} - z_* = R(z_k - z*)$ for

$$R = \begin{bmatrix} (1 - \sigma\tau a^2)(1 - \tau\mu) & -\tau a(1 - \sigma\tau a^2) \\ \sigma a(1 - \tau\mu) & (1 - \sigma\tau a^2) \end{bmatrix}$$

Hence, we can compute the exact rate of convergence, which is given by the largest eigenvalue of R different from 1.

We shall compare this actual rate with what is predicted by Proposition 19, that is $1 - \frac{\eta^2\lambda}{(\sqrt{3}\eta + (2 + 2\sqrt{3}\max(\alpha_f, \alpha_g)))^2}$ where λ, $\gamma = \sigma\tau a^2$, $\alpha_g = 0$, $\alpha_f = \tau\mu/2$ and $\eta = \frac{\sqrt{\mu^2\tau^2 + 4\sigma\tau a^2} - \mu\tau}{2}$ and what is predicted by Proposition 18, that is $\left(1 + \frac{\eta'}{(2 + a_2)(1 + \eta'/\lambda)}\right)^{-1}$ where $2 + a_2 = \frac{1}{1 - \tau\mu_f/2}$ and $\eta' = \min\left(\mu_f\tau, \frac{\sigma\tau\sigma_{\min}(A)^2}{\tau L_f + \tau L'_f + \frac{1}{\lambda}}\right)$. On Figure 1, we can see that there can be a large difference between what is predicted and what is observed, even for the simplest problem. Moreover, although the actual rate improves when μ increases, metric sub-regularity decreases, so that theory suggests the opposite of what is actually observed. On the other hand, using strong convexity explains the improvement of the rate when μ increases but does not manage to capture the linear convergence for $\mu = 0$.

6 Quadratic error bound of the smoothed gap

We now introduce a new regularity assumption that truly generalized strongly convex-concave Lagrangians and smooth strongly convex objectives with linear constraints and is as broadly applicable as metric subregularity of the Lagrangian's gradient.

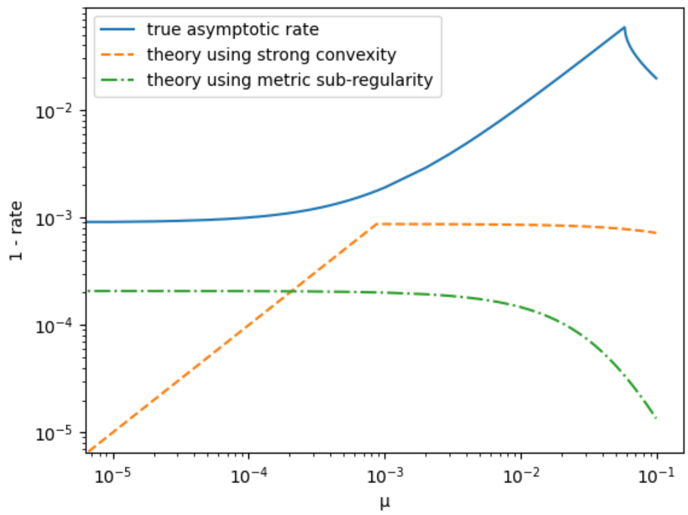

Figure 1 Comparison of the true rate (line above) and what is predicted by theory (2 lines below) for $a = 0.03$, $\tau = \sigma = 1$ and various values for μ.

6.1 Main assumption

Definition 20. *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*

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

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

Although the smooth 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 21. *Let $\beta \in [0, +\infty)^2$. If $z^* \in \mathcal{Z}^*$, then $z \in \mathcal{Z}^* \Leftrightarrow G_\beta(z; z^*) = 0$.*

Proof. We first remark that $G_0(z, z^*)$ is the usual duality gap and that $G_\infty(z; z^*) = L(x, y^*) - L(x^*, y) \geq 0$. Moreover, $G_0(z, z^*) \geq G_\beta(z, z^*) \geq G_\infty(z; z^*) \geq 0$. Since $z \in \mathcal{Z}^* \Rightarrow G_0(z; z^*) = 0$, we have the implication $z \in \mathcal{Z}^* \Rightarrow G_\beta(z; z^*) = 0$.

For the converse implication, we denote

$$\begin{aligned} y_\beta(x) &= \arg\max_{y'} L(x, y') - \frac{\beta_y}{2\sigma}\|y^* - y'\|^2 = \arg\max_{y'} \langle Ax, y'\rangle - g^*(y') - g_2^*(y') - \frac{\beta_y}{2\sigma}\|y^* - y'\|^2 \\ &= \mathrm{prox}_{\sigma/\beta_y(g^* + g_2^*)}\big(y^* + \frac{\sigma}{\beta} Ax\big) \end{aligned}$$

By the strong convexity of the problem defining $G_\beta(\,\cdot\,; z^*)$, we know that

$$\sup_{y'} L(x, y') - \frac{\beta_y}{2\sigma}\|y^* - y'\|^2 \geq L(x, y^*) - \frac{\beta_y}{2\sigma}\|y^* - y^*\|^2 + \frac{\beta_y}{2\sigma}\|y_\beta(x) - y^*\|^2 \geq L(x^*, y^*) + \frac{\beta_y}{2\sigma}\|y_\beta(x) - y^*\|^2 \,.$$

With a similar argument for $x_\beta(y)$, we get

$$G_\beta(z; z^*) \geq \frac{\beta_y}{2\sigma}\|y_\beta(x) - y^*\|^2 + \frac{\beta_x}{2\tau}\|x_\beta(y) - x^*\|^2 \,.$$

Thus, if $G_\beta(z; z^*) = 0$, then $y_\beta(x) = y^*$ and $x_\beta(y) = x^*$.

$$\begin{aligned} y_\beta(x) = y^* &\iff y^* = \mathrm{prox}_{\sigma/\beta_y(g^* + g_2^*)}\big(y^* + \frac{\sigma}{\beta_y} Ax\big) \\ &\iff 0 \in y^* - (y^* + \frac{\sigma}{\beta_y} Ax) + \frac{\sigma}{\beta_y}\partial g^*(y^*) + \frac{\sigma}{\beta_y}\nabla g_2^*(y^*) \\ &\iff 0 \in -Ax + \partial g^*(y^*) + \nabla g_2^*(y^*) \iff x \in \mathcal{X}^* \end{aligned}$$

and similarly $x_\beta(y) = x^* \Leftrightarrow y \in \mathcal{Y}^*$, which completes the proof of the proposition. ◀

Assumption 22. There exists $\beta = (\beta_x, \beta_y) \in]0, +\infty]^2$, $\eta > 0$ and a region $\mathcal{R} \subseteq \mathcal{Z}$ such that for all $z^* \in \mathcal{Z}^*$, $G_\beta(\,\cdot\,, z^*)$ has a quadratic error bound with constant η in the region $\mathcal{R}$ and with the norm $\|\cdot\|_V$. Said otherwise, for all $z \in \mathcal{R}$,

$$G_\beta(z; z^*) \geq \frac{\eta}{2} \operatorname{dist}_V(z, \mathcal{Z}^*)^2 \,.$$

The next proposition, which is a simple consequence of [16, Proposition 1] says that even though QEB is a local concept, it can be extended to any compact set at the expense of degrading the constant.

Proposition 23. *If $G_\beta(\,\cdot\,, z^*)$ has a η-QEB on $\{z : \operatorname{dist}(z, \mathcal{Z}^*)_V < a\}$ then for all $M > 1$, $G_\beta(\,\cdot\,, z^*)$ has a $\frac{\eta}{M}$-QEB on $\{z : \operatorname{dist}(z, \mathcal{Z}^*)_V < Ma\}$.*

6.2 Problems with strong convexity

We now give a few examples to show that Assumption 22 is often satisfied.

Proposition 24. *If L is μ-strongly convex-concave in the norm $\|\cdot\|_V$, then $\forall\, z \in \mathcal{Z}$, $G_\infty(z; z^*) \geq \frac{\mu}{2}\|z - z^*\|_V^2$.*

Proof. $G_\infty(z; z^*) = L(x, y^*) - L(x^*, y) \geq \frac{\mu}{2}\|z - z^*\|_V^2$. ◀

Proposition 25. *If $f + f_2$ has a $L_f' + L_f$-Lipschitz gradient, $g \,\square\, g_2 = \iota_{\{b\}}$, the primal function given by $(x \mapsto f(x) + f_2(x) + g \,\square\, g_2(Ax))$ has a $\bar{\mu}$-QEB and $f + f_2$ is μ_f-strongly convex, then the smoothed gap has a QEB:*

$$G_\beta(z, z^*) \geq \min\left(\max\left(\frac{\tau\mu_f}{2}, \frac{\bar{\mu}^2}{(L_f + L_f')^2}\frac{\sigma\tau\sigma_{\min}(A)^2}{16\beta_y}\right), \frac{\sigma\sigma_{\min}(A)^2}{2(L_f + L_f' + \beta_x/\tau)}\right) \operatorname{dist}_V(z, \mathcal{Z}^*)^2 \,.$$

Note that we require either $\mu_f > 0$ or $\bar{\mu} > 0$.

Proof. The proof is a generalization of Proposition 18 and reuses most of the argument.

$$\sup_{y' \in \mathcal{Y}} L(x, y') - \frac{\beta_y}{2\sigma}\|y' - y^*\|^2 = f(x) + f_2(x) + \langle y^*, Ax - b\rangle + \frac{\sigma}{2\beta_y}\|Ax - b\|^2 \,.$$

We decompose $x = x_A + x_{A^\perp}$ with $x_{A^\perp} = P_{\{x' : Ax' = b\}}(x)$ and $x_A = x - x_{A^\perp} \in (\ker A)^\perp$. We have $Ax - b = Ax_A$, so that $\|Ax - b\| \geq \sigma_{\min}(A)\|x_A\|$. Moreover by convexity of $f + f_2$ and optimality condition $\nabla f(x^*) + \nabla f_2(x^*) = -A^\top y^*$,

$$\begin{aligned}
&f(x) + f_2(x) + \langle y^*, Ax - b\rangle + \frac{\sigma}{2\beta_y}\|Ax - b\|^2 \\
&\quad\geq f(x_{A^\perp}) + f_2(x_{A^\perp}) + \langle \nabla(f + f_2)(x_{A^\perp}), x - x_{A^\perp}\rangle - \langle \nabla(f + f_2)(x^*), x - x_{A^\perp}\rangle + \frac{\sigma}{2\beta_y}\sigma_{\min}(A)^2\|x_A\|^2 \\
&\quad\geq f(x^*) + f_2(x^*) + \frac{\bar{\mu}}{2}\operatorname{dist}(x_{A^\perp}, \mathcal{X}^*)^2 - (L_f + L_f')\|x_{A^\perp} - x^*\|\|x_A\| + \frac{\sigma}{2\beta_y}\sigma_{\min}(A)^2\|x_A\|^2
\end{aligned}$$

where the last inequality comes from the assumption on the primal function and smoothness of $\nabla(f + f_2)$. We combine this with

$$f(x) + f_2(x) + \langle y^*, Ax - b\rangle \geq f(x^*) + f_2(x^*) + \frac{\mu_f}{2}\operatorname{dist}(x, \mathcal{X}^*)^2$$

to get for all $\lambda \in [0, 1]$ and $\alpha > 0$,

$$\begin{aligned}
&f(x) + f_2(x) + \langle y^*, Ax - b\rangle + \frac{\sigma}{2\beta_y}\|Ax - b\|^2 \\
&\quad\geq f(x^*) + f_2(x^*) + \left(\frac{\lambda\bar{\mu}}{2} - \frac{\lambda\alpha(L_f + L_f')}{2} + \frac{(1 - \lambda)\mu_f}{2}\right)\operatorname{dist}(x_{A^\perp}, \mathcal{X}^*)^2 \\
&\qquad + \left(\frac{\sigma}{2\beta_y}\sigma_{\min}(A)^2 - \frac{\lambda(L_f + L_f')}{2\alpha} + \frac{(1 - \lambda)\mu_f}{2}\right)\|x_A\|^2
\end{aligned}$$

We take $\alpha = \frac{\bar{\mu}}{2(L_f + L_f')}$, $\lambda = \frac{\bar{\mu}}{4(L_f + L_f')^2}\frac{\sigma\sigma_{\min}(A)^2}{\beta_y}$ to get

$$f(x) + f_2(x) + \langle y^*, Ax - b\rangle + \frac{\sigma}{2\beta_y}\|Ax-b\|^2 \\ \geq f(x^*) + f_2(x^*) + \max\Big(\frac{\mu_f}{2}, \frac{\bar\mu^2}{(L_f+L'_f)^2}\frac{\sigma\sigma_{\min}(A)^2}{16\beta_y}\Big)\,\mathrm{dist}(x, \mathcal{X}^*)^2 \ . \quad (7)$$

For the dual vector, we use the smoothness of the objective, the equality $\nabla f(x^*) + \nabla f_2(x^*) = -A^\top y^*$ and $Ax^* = b$.

$$\begin{aligned} -L(x', y) &= -f(x') - f_2(x') - \langle Ax' - b, y\rangle \\ &\geq -f(x^*) - f_2(x^*) - \langle \nabla f(x^*) - \nabla f_2(x^*), x' - x^*\rangle - \frac{L_f + L'_f}{2}\|x' - x^*\|^2 - \langle Ax' - b, y\rangle \\ &= -L(x^*, y^*) + \langle A^\top y^*, x' - x^*\rangle - \langle x' - x^*, A^\top y\rangle - \frac{L_f + L'_f}{2}\|x' - x^*\|^2 \end{aligned}$$

For $a \in \mathbb{R}$, we restrict ourselves to $x' = x^* + aA^\top(y^* - y)$ so that

$$\begin{aligned} \sup_{x'\in\mathcal{X}} -L(x', y) - \frac{\beta_x}{2\tau}\|x' - x^*\|^2 &\geq \sup_{a\in\mathsf{R}} -L(x^* + aA^\top(y^* - y), y) - \frac{\beta_x a^2}{2\tau}\|A^\top(y^* - y)\|^2 \\ &\geq \sup_{a\in\mathsf{R}} -L(x^*, y^*) + \Big(a - a^2\frac{L_f + L'_f + \beta_x/\tau}{2}\Big)\|A^\top(y - y^*)\|^2 \\ &= -L(x^*, y^*) + \frac{1}{2(L_f + L'_f + \beta_x/\tau)}\|A^\top(y - y^*)\|^2 \end{aligned}$$

Moreover, as in Proposition 18, we know that $\|A^\top y - A^\top y^*\| \geq \sigma_{\min(A)}\,\mathrm{dist}(y, \mathcal{Y}^*)$, where $\sigma_{\min(A)}$ is the smallest singular value of A.

Combining this with (7) yields the result of the proposition. ◀

Proposition 26. *Suppose that $\mathcal{X}$ and $\mathcal{Y}$ are finite-dimensional. Suppose that f, f_2, g, g_2 are convex piecewise linear-quadratic, which means that their domain is a union of polyhedra and on each of these polyhedra, they are quadratic functions. Then for all $\beta \in [0, +\infty[^2$, there exists $\eta(\beta)$ and $\mathcal{R}(\beta)$ such that $G_\beta(z; z^*) \geq \frac{\eta(\beta)}{2}\,\mathrm{dist}_V(z, \mathcal{Z}^*)^2$ for all $z \in \mathcal{R}(\beta)$ and $z^* \in \mathcal{Z}^*$.*

Proof. The proof follows the lines of [21]. The class of piecewise linear-quadratic functions is closed under scalar multiplication, addition, conjugation and Moreau envelope [28]. Hence for all $\beta \in [0, +\infty[^2$, $G_\beta(\,\cdot\,, z^*)$ is piecewise linear quadratic. As a consequence, its subgradient $\partial_z G_\beta(\,\cdot\,, z^*)$ is piecewise polyhedral and thus there exists $\eta > 0$ such that it satisfies metric sub-regularity with constant η at all $z^* \in \mathcal{Z}^*$ for 0 [11]. Since $G_\beta(\,\cdot\,, z^*)$ is a convex function, this implies the result by Proposition 9. ◀

6.3 Linear programs

In the rest of the section, we are going to show that linear programs do satisfy Assumption 22 and give the constant as a function of the Hoffman constant [18].

We consider the linear optimization problem

$$\begin{aligned} &\min_{x\in\mathsf{R}^n} c^\top x \qquad (8) \\ &A_{E,:}x = b_E \\ &A_{I,:}x \leq b_I \\ &x_N \geq 0 \end{aligned}$$

where A is a $m \times n$ matrix, $b \in \mathbb{R}^m$, E and I are disjoint sets of indices such that $E \cup I = \{1, \dots, m\}$ and N, F are disjoint sets of indices such that $N \cup F = \{1, \dots, n\}$.

A dual of this problem is given by

$$\begin{aligned} &\max_{y\in\mathsf{R}^m} -b^\top y \\ &(A_{:,F})^\top y + c_F = 0 \\ &(A_{:,N})^\top y + c_N \geq 0 \\ &y_I \geq 0 \end{aligned}$$

It happens that the set of primal-dual solution of an LP is characterized by a system of linear equalities and inequalities. This holds true because a feasible primal-dual pair with equal values is necessarily optimal. We get the following system

$$\begin{cases} c^\top x + b^\top y = 0 & (A_{:,F})^\top y + c_F = 0 \\ A_{E,:} x = b_E & (A_{:,N})^\top y + c_N \geq 0 \\ A_{I,:} x \leq b_I & y_I \geq 0 \\ x_N \geq 0 & \end{cases} \tag{9}$$

Let us denote the Hoffman constant [18] of this system by θ. This constant is the smallest positive number such that for all $z \in \mathbb{R}^{m+n}$

$$\mathrm{dist}(z, \mathcal{Z}^*) \leq \theta\Big(|c^\top x + b^\top y|^2 + \|A_{E,:}x - b_E\|^2 + \mathrm{dist}(A_{I,:}x - b_I, \mathbb{R}^I_-)^2 \\ + \mathrm{dist}(x_N, \mathbb{R}^N_+)^2 + \|(A_{:,F})^\top y + c_F\|^2 + \mathrm{dist}((A_{:,N})^\top y + c_N, \mathbb{R}^N_+)^2 + \mathrm{dist}(y_I, \mathbb{R}^I_+)^2\Big)^{1/2} \tag{10}$$

It is known that the Lagrangian's subgradient of an LP satisfies metric sub-regularity with a constant proportional to θ [24]. We shall show that the same holds for the QEB of the smoothed gap centered at z^*.

Proposition 27. *For any $\beta \geq 0$, $R > 0$ and $z^* \in \mathcal{Z}^*$, the linear program (8) satisfies the quadratic error bound: for all z such that $G_\beta(z; z^*) \leq R$, we have*

$$G_\beta(z; z^*) \geq \frac{\mathrm{dist}(z, \mathcal{Z}^*)^2}{\theta^2\Big(\sqrt{\frac{2\beta}{\tau}}(\sqrt{2} + \|x^*_F\| + \|x^*_N\|) + \sqrt{\frac{2\beta}{\sigma}}(\sqrt{2} + \|y^*_E\| + \|y^*_I\|) + 3\sqrt{R}\Big)^2} .$$

Hence, for R of the order of $\frac{1}{\theta}$, $G_{\frac{1}{\theta}}(\,\cdot\,, z^*)$ has a $\frac{c}{\theta}$-QEB with c independent of θ.

Proof. See Appendix C. ◀

7 Analysis of PDHG under quadratic error bound of the smoothed gap

In this section, we show that under the new regularity assumption, PDHG converges linearly. Moreover, we give an explicit value for the rate. This result is central to the paper because it shows that the quadratic error bound of the smoothed gap is a fruitful assumption: not only it is as broadly applicable as the metric subregularity of the Lagrangian's generalized gradient, but also the rates it predicts reach the state of the art in all subcases of interest.

Theorem 28. *Under Assumption 22, if $\mathcal{R}$ contains $\{z : \|z - P_{\mathcal{Z}^*}(z_0)\| \leq \mathrm{dist}_V(z_0, \mathcal{Z}^*)\}$, then PDHG converges linearly at a rate*

$$\Big(1 + \Lambda\frac{\eta}{1 + \eta/\lambda}\Big)\,\mathrm{dist}_V(z_{k+1}, \mathcal{Z}^*)^2 \leq \mathrm{dist}_V(z_k, \mathcal{Z}^*)^2$$

where

$$\Lambda = \frac{\lambda}{\max((1 + a_2)\lambda + 1/\beta_x, (2 + a_2)\lambda + 1/\beta_y),}$$

λ is defined in Lemma 12 and $a_2 = \max\big(\frac{2\alpha_f - 1}{1-\alpha_f}, \frac{2\alpha_g - 1 + \gamma}{1 - \alpha_g - \gamma}\big) \geq -1$ is defined in Lemma 14.

Proof. In this proof, we will use the notation $\beta \odot z = (\beta_x x, \beta_y y)$ and $\|z\|^2_{\beta V} = \frac{\beta_x}{\tau}\|x\|^2 + \frac{\beta_y}{\sigma}\|y\|^2$. By Lemma 14, we have

$$L(\bar{x}_{k+1}, y) - L(x, \bar{y}_{k+1}) \leq \frac{1}{2}\|z - z_k\|^2_V - \frac{1}{2}\|z - z_{k+1}\|^2_V + a_2\widetilde{V}(\bar{z}_k, z^*) .$$

For $z^* = P_{\mathcal{Z}^*}(z_k)$, the projection of z_k onto the set of saddle points using norm $\|\cdot\|_V$,

$$\begin{aligned} G_\beta(\bar{z}_{k+1}; z^*) &= \sup_x \sup_y L(\bar{x}_{k+1}, y) - \frac{\beta_y}{2}\|y - y^*\|^2_{\sigma^{-1}} - L(x, \bar{y}_{k+1}) - \frac{\beta_x}{2}\|x - x^*\|^2_{\tau^{-1}} \\ &\leq \sup_z \frac{1}{2}\|z - z_k\|^2_V - \frac{1}{2}\|z - z_{k+1}\|^2_V - \frac{1}{2}\|z - z^*\|^2_{\beta V} + a_2\widetilde{V}(\bar{z}_k, z^*) \end{aligned}$$

For the right hand side, we are looking for z such that $\beta \odot (z - z^*) + (z - z_{k+1}) - (z - z_k) = 0$ so that $\beta \odot z = \beta \odot z^* + z_{k+1} - z_k$ and

$$\begin{aligned}
&\frac{1}{2}\|z - z_k\|_V^2 - \frac{1}{2}\|z - z_{k+1}\|_V^2 - \frac{1}{2}\|z - z^*\|_{\beta V}^2 \\
&\qquad = \frac{1}{2}\|z^* - z_k\|_V^2 - \frac{1}{2}\|z^* - z_{k+1}\|_V^2 + \frac{1}{2}\|z_{k+1} - z_k\|_{\beta^{-1}V}^2 \\
&\qquad \leq \frac{1}{2}\operatorname{dist}_V(z_k, \mathcal{Z}^*)^2 - \frac{1}{2}\operatorname{dist}_V(z_{k+1}, \mathcal{Z}^*)^2 + \frac{1}{2}\|z_{k+1} - z_k\|_{\beta^{-1}V}^2
\end{aligned}$$

where the last inequality comes from our choice of z^*. We also have by Lemma 12

$$\frac{1}{2}\operatorname{dist}_V(z_k, \mathcal{Z}^*)^2 - \frac{1}{2}\operatorname{dist}_V(z_{k+1}, \mathcal{Z}^*)^2 - \widetilde{V}(z_k, z^*) \geq \frac{1}{2}\|z^* - z_k\|_V^2 - \frac{1}{2}\|z^* - z_{k+1}\|_V^2 - \widetilde{V}(z_k, z^*) \geq 0 \ .$$

Using Assumption 22, this leads to: $\forall\, \Lambda \in [0, 1]$,

$$\begin{aligned}
&\frac{1}{2}\operatorname{dist}_V(z_k, \mathcal{Z}^*)^2 - \frac{1}{2}\operatorname{dist}_V(z_{k+1}, \mathcal{Z}^*)^2 + \frac{\Lambda}{2}\|z_k - z_{k+1}\|_{\beta^{-1}V}^2 + (\Lambda a_2 - (1-\Lambda))\widetilde{V}(z_k, z^*) \\
&\qquad\qquad \geq \frac{\Lambda\eta}{2}\operatorname{dist}_V(\bar{z}_{k+1}, \mathcal{Z}^*)^2 \ .
\end{aligned}$$

Using Lemma 16 and Lemma 13, we get, as soon as $\Lambda a_2 - (1 - \Lambda) \leq 0$,

$$\begin{aligned}
&\frac{1}{2}\operatorname{dist}_V(z_k, \mathcal{Z}^*)^2 - \frac{1}{2}\operatorname{dist}_V(z_{k+1}, \mathcal{Z}^*)^2 + \Big(\frac{\Lambda}{\beta_x} + (\Lambda a_2 - (1-\Lambda))\lambda\Big)\frac{1}{2\tau}\|x_k - x_{k+1}\|^2 \\
&\qquad + \Big(\frac{\Lambda}{\beta_y} + (\alpha^{-1} - 1)\Lambda\eta + (\Lambda a_2 - (1-\Lambda))\lambda\Big)\frac{1}{2\sigma}\|y_k - y_{k+1}\|^2 \\
&\quad \geq \frac{(1-\alpha)\Lambda\eta}{2}\operatorname{dist}_V(z_{k+1}, \mathcal{Z}^*)^2
\end{aligned}$$

So, taking $\alpha = \frac{\eta}{\lambda + \eta}$ and $\Lambda = \frac{\lambda}{\max((1+a_2)\lambda + 1/\beta_x, (2+a_2)\lambda + 1/\beta_y)} \leq 1$ leads to $\frac{\Lambda}{\beta_y} + (\alpha^{-1} - 1)\Lambda\eta + (\Lambda a_2 - (1-\Lambda))\lambda = \frac{\Lambda}{\beta_y} + \lambda\Lambda + (a_2 + 1)\lambda\Lambda - \lambda \leq 0$ and $\frac{\Lambda}{\beta_x} + (\Lambda a_2 - (1-\Lambda))\lambda \leq 0$, so that

$$\operatorname{dist}_V(z_k, \mathcal{Z}^*)^2 \geq \Big(1 + \Lambda\frac{\eta}{1 + \eta/\lambda}\Big)\operatorname{dist}_V(z_{k+1}, \mathcal{Z}^*)^2$$

and thus the algorithm enjoys a linear rate of convergence. ◀

Strongly convex-concave Lagrangian

If the Lagrangian is strongly convex concave, then we can take $\beta = (+\infty, +\infty)$ and $\eta = \mu$ (Proposition 24), so that we recover the rate of Proposition 17.

Note that in that case, the rate of order $1 - c\mu$ given by Proposition 17, and so by its generalized version Theorem 28, is much better than what Proposition 19 tells us: a rate of order $1 - c\mu^2$. Hence, we can see that for this important particular case, the rate predicted using the quadratic error bound of the smoothed gap is more informative than using the metric subregularity of the Lagrangian's gradient. Moreover, the new assumption applies to all piecewise-linear quadratic problems, making it at the same time accurate and general.

Back to the toy problem

We consider again the linearly constrained 1D problem $\min_{x \in \mathbb{R}}\{\frac{\mu}{2}x^2 \ : \ ax = b\}$ where $a, b \in \mathbb{R}$ and $\mu \geq 0$ introduced in Section 5.2 and we calculate the quadratic error bound of the smoothed gap.

$$\begin{aligned}
G_\beta(\bar{z}, z^*) &= \sup_y \frac{\mu}{2}\bar{x}^2 + y(a\bar{x} - b) - \frac{\beta_y}{2\sigma}(y - y^*)^2 + \sup_x -\frac{\mu}{2}x^2 - \bar{y}(ax - b) - \frac{\beta_x}{2\tau}(x - x^*)^2 \\
&= \frac{\mu}{2}\bar{x}^2 + y^*(a\bar{x} - b) + \frac{\sigma}{2\beta_y}(a\bar{x} - b)^2 + b\bar{y} + \frac{1}{2(\frac{\beta_x}{\tau} + \mu)}(\frac{\beta_x}{\tau}x^* - a\bar{y})^2 - \frac{\beta_x}{2\tau}(x^*)^2 \\
&\geq \frac{\mu\tau + \frac{\sigma\tau a^2}{\beta_y}}{2\tau}(\bar{x} - x^*)^2 + \frac{\sigma\tau a^2}{2\sigma(\beta_x + \mu\tau)}(\bar{y} - y^*)^2 \\
&\geq \frac{1}{2}\min\Big(\mu\tau + \frac{\sigma\tau a^2}{\beta_y}, \frac{\sigma\tau a^2}{\beta_x + \mu\tau}\Big)\|\bar{z} - z^*\|_V^2
\end{aligned}$$

As we have seen in Proposition 25, we can leverage the strong convexity of the objective. But also the smoothed gap may enjoy a quadratic error bound even if the objective is not strongly convex.

According to Theorem 28, since $2 + a_2 = \frac{1}{1-\tau\mu_f/2}$, the rate is $(1+\rho)^{-1}$ where

$$\rho = \Lambda\frac{\eta}{1+\eta/\lambda} = \frac{\lambda}{\max(\lambda\frac{\tau\mu_f}{1-\tau\mu_f/2} + 1/\beta_x, \lambda\frac{1+\tau\mu_f/2}{1-\tau\mu_f/2} + 1/\beta_y)} \frac{\min\Big(\mu\tau + \frac{\sigma\tau a^2}{\beta_y}, \frac{\sigma\tau a^2}{\beta_x+\mu\tau}\Big)}{1+\min\Big(\mu\tau + \frac{\sigma\tau a^2}{\beta_y}, \frac{\sigma\tau a^2}{\beta_x+\mu\tau}\Big)/\lambda} .$$

with $\lambda = (1-\mu\tau/2)(1-\sqrt{\sigma\tau a^2})$. Since the algorithm does not depend on β_x or β_y we can choose them so that they minimize the rate (or maximize ρ). On Figure 2, we can see that the rate of convergence explained using the quadratic error bound of the smoothed gap is as good as the rate using strong convexity (Assumption 6) when μ is large and does not vanish when μ goes to 0. On top of this, for small values of μ, we obtain a much better rate than what is predicted using metric sub-regularity.

In Appendix D, Proposition 34, we derive a finer analysis in the case where we solve a linearly constrained problem whose objective function is strongly convex. Indeed, we can show that the largest singular value of the matrix R described in Section 5.2 is $1-\gamma$. Yet, its spectral radius is much smaller. This implies that a contraction on $\mathrm{dist}_V(z_k - z^*)^2$ is not enough to account for the actual rate. We propose to combine it with a contraction on $\|z_{k+1} - z_k\|_V^2$. The rationale for this addition is that for large strong convexity parameters, the primal sequence will behave as if it were tracking $\arg\min_{x'} L(x', y_k)$. This is a kind of slow-fast system where the dual variable is slowly varying and the primal variable is fast.

When we plot the curve of the rate as a function of μ_f (with the legend "slow-fast double concentration rate") we can see that this more complex analysis manages to explain the improvement of the rate for an increasing strong convexity parameter, together with its degradations when the parameter becomes too large.

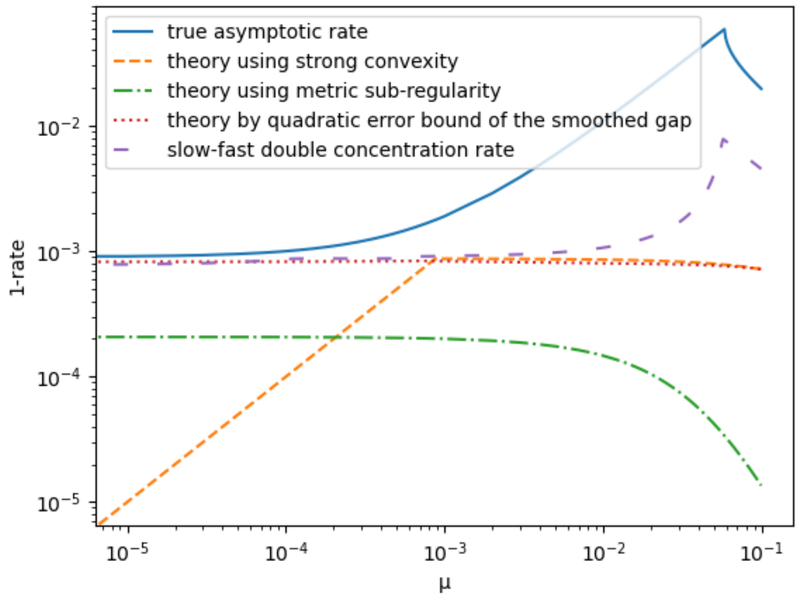

Figure 2 Comparison of the true rate ρ (line above), what is predicted by theory using previous theories and what is predicted by using quadratic error bound of the smoothed gap for $a = 0.03$, $\tau = \sigma = 1$ and various values for μ. We plot $1-\rho$ in logarithmic scale.

8 Restarted averaged primal-dual hybrid gradient

8.1 Restarted Averaged Primal-Dual Hybrid Gradient

In this section we will see how our new understanding of the rate of convergence of PDHG can help us design a faster algorithm.

Algorithm 2 Averaged Primal Dual Hybrid Gradient – APDHG(x_0, y_0, K)

For $k \in \{0, \dots, K-1\}$:

$$\bar{x}_{k+1} = \mathrm{prox}_{\tau f}(x_k - \tau \nabla f_2(x_k) - \tau A^\top y_k)$$
$$\bar{y}_{k+1} = \mathrm{prox}_{\sigma g^*}(y_k - \sigma \nabla g_2^*(y_k) + \sigma A \bar{x}_{k+1})$$
$$x_{k+1} = \bar{x}_{k+1} - \tau A^\top (\bar{y}_{k+1} - y_k)$$
$$y_{k+1} = \bar{y}_{k+1}$$
$$\widetilde{x}_{k+1} = \tfrac{1}{k+1} \sum_{l=0}^{k} \bar{x}_{l+1} \qquad \widetilde{y}_{k+1} = \tfrac{1}{k+1} \sum_{l=0}^{k} \bar{y}_{l+1}$$

Return $(\widetilde{x}_K, \widetilde{y}_K)$

Let averaged PDHG be given by Algorithm 2. On the class of convex functions, averaged PDHG has an improved convergence speed in $O(1/k)$ in the worst case while PDHG has a convergence in $O(1/\sqrt{k})$ [10].

However, when averaging, we loose the linear convergence for well behaved problems. We thus propose to restart the algorithm as in Algorithm 3. The following proposition shows that RAPDHG enjoys an improved rate of convergence where the product $\beta\eta$ is replaced by $\max(\beta, \eta)$. Hence for problems where $\eta(\beta)$ is a decreasing function of β, like linear programs, we will expect an improved convergence rate by averaging and restarting.

Algorithm 3 Restarted Averaged Primal Dual Hybrid Gradient – RAPDHG(x_0, y_0)

Let $K \in \mathbb{N}$ and $z_0 = (x_0, y_0)$. For $s \geq 0$:

$$z_{s+1} = \mathrm{APDHG}(z_s, K)$$

Proposition 29. *Under Assumption 22 with $\beta_x = \beta_y = \beta$, if the restart frequency K satisfies $K\beta \geq 2$ and $K\eta \geq 2(2 + a_2^+)/\eta$, where $a_2^+ = \max(0, a_2)$ and a_2 is defined in Lemma 14, then RAPDHG converges linearly at a rate $2^{-1/K}$. Moreover, if $K = \lceil \max(2/\beta, 2(2 + a_2^+)/\eta/\eta) \rceil$, then the rate is $\exp\Big(-\frac{1}{\lceil \max(2/\beta, 2(2+a_2^+)/\eta/\eta) \rceil} \ln(2)\Big) \approx \exp\big(-\min(\beta/2, \eta/(2(2 + a_2^+))) \ln(2)\big)$.*

Proof. Let us denote by $(z_s^R)_{s \in \mathbf{N}}$ the iterates of RAPDHG. We keep the notation $z_k, \bar{z}_k$ for the iterates of the inner loop.

Consider $z^* \in \mathcal{Z}^*$ and denote $a_2^+ = \max(0, a_2)$. We combine (6) with $a_2^+/2$ times (4) to get

$$L(\bar{x}_{k+1}, y) - L(x, \bar{y}_{k+1}) \\ \leq \frac{1}{2}\|z - z_k\|_V^2 - \frac{1}{2}\|z - z_{k+1}\|_V^2 + \frac{a_2^+}{2}\|z^* - z_k\|_V^2 - \frac{a_2^+}{2}\|z^* - z_{k+1}\|_V^2 + (a_2 - a_2^+)\widetilde{V}(z_k, z^*) \ .$$

Summing this inequality for k between 0 and $K - 1$, using the fact that the Lagrangian is convex-concave, and that $a_2 - a_2^+ \leq 0$, we get

$$L(\widetilde{x}_K, y) - L(x, \widetilde{y}_K) \leq \frac{1}{2K}\|z - z_0\|_V^2 - \frac{1}{2K}\|z - z_K\|_V^2 + \frac{a_2^+}{2K}\|z^* - z_0\|_V^2 - \frac{a_2^+}{2K}\|z^* - z_K\|_V^2$$

which leads to

$$L(\widetilde{x}_K, y) - L(x, \widetilde{y}_K) - \frac{\beta}{2}\|z - z^*\|_V^2 \leq \frac{1}{2K}\|z - z_0\|_V^2 - \frac{\beta}{2}\|z - z^*\|_V^2 + \frac{a_2^+}{2K}\|z^* - z_0\|_V^2$$

and so, as soon as $K\beta > 1$, since the maximum of the right hand side is attained at $z = \frac{K\beta z^* - z_0}{K\beta - 1}$,

$$G_\beta(\widetilde{z}_K, z^*) \leq \frac{1}{2K}\left(\frac{K\beta}{K\beta - 1} + a_2^+\right)\|z^* - z_0\|_V^2 \ .$$

We now use Assumption 22 to get

$$\frac{1}{K}\left(\frac{K\beta}{K\beta - 1} + a_2^+\right)\|z^* - z_0\|_V^2 \geq \eta\|z^* - \widetilde{z}_K\|^2 \ .$$

We choose $z^* = P_{\mathcal{Z}^*}(z_0)$ and K such that $K\beta \geq 2$ and $K\eta \geq 2(2 + a_2^+)$ in order to get

$$\mathrm{dist}_V(z_1^R, \mathcal{Z}^*)^2 = \mathrm{dist}_V(\tilde{z}_K, \mathcal{Z}^*)^2 \leq \frac{1}{2}\,\mathrm{dist}_V(z_0, \mathcal{Z}^*)^2 .$$

If we choose $K = \lceil \max(2/\beta, 2(2 + a_2^+)/\eta) \rceil$ we thus get a linear convergence

$$\begin{aligned}\mathrm{dist}_V(z_s^R, \mathcal{Z}^*)^2 &\leq \frac{1}{2^s}\,\mathrm{dist}_V(\tilde{z}_0^R, \mathcal{Z}^*)^2 \\ &\leq \exp\Big(-\frac{1}{\lceil \max(2/\beta, 2(2 + a_2^+)/\eta) \rceil}\ln(2)\Big)^{sK} \mathrm{dist}_V(z_0, \mathcal{Z}^*)^2\end{aligned}$$

where sK is the total number of iterations. ◀

The rate of convergence of RAPDHG has two nice features as compared to plain PDHG. Indeed, there is a factor Λ in Theorem 28 in front of the quadratic error bound constant η, which is of order $\lambda\beta$ when β is small. On the other hand, the rate of RAPDHG has no direct dependence on λ, which means that it will behave well even if $\sigma\tau\|A\|^2$ is close to 1. Moreover, it replaces $\beta\eta$ by $\min(\beta, \eta)$, which will be orders of magnitude better in the case of linear programs where $\eta = O(\beta)$ for $\beta = 1/\theta$ (Proposition 27)

8.2 Self-centered smoothed gap

In this paper, we have shown that the smoothed gap is a convenient quantity for the analysis of PDHG and that assuming that it satisfies a quadratic error bound condition explains well its behaviour. However, since computing it requires the knowledge of a saddle point, one cannot use the smoothed gap for algorithmic design, and in particular for the tuning of RAPDHG.

We thus propose the following approximation, that we call the self-centered smoothed gap.

Definition 30. *Given $\beta = (\beta_x, \beta_x) \in [0, +\infty[^2$, and $z \in \mathcal{Z}$, the self-centered smoothed gap is given by $G_\beta(z, z)$.*

The motivation for this definition is the following lemma.

Lemma 31. *For all $z, \dot{z} \in \mathcal{Z}$ and z^* equal to the projection of $\dot{z}$ onto $\mathcal{Z}^*$,*

$$G_\beta(z, \dot{z}) \geq G_{2\beta}(z, z^*) - \beta\,\mathrm{dist}_V(\dot{z}, \mathcal{Z}^*)_V^2 . \tag{11}$$

Proof.

$$\begin{aligned}G_\beta(z, \dot{z}) &= \max_{z'} L(x, y') - L(x', y) - \frac{\beta}{2}\|\dot{z} - z'\|_V^2 \\ &\geq \max_{z'} L(x, y') - L(x', y) - \beta\|z^* - z'\|_V^2 - \beta\|\dot{z} - z^*\|_V^2 \\ &= G_{2\beta}(z, z^*) - \beta\|\dot{z} - z^*\|_V^2 = G_{2\beta}(z, z^*) - \beta\,\mathrm{dist}_V(\dot{z}, \mathcal{Z}^*)_V^2\end{aligned}$$ ◀

This shows that $G_\beta(z, \dot{z})$ is a good approximation to the measure of optimality $G_{2\beta}(z, z^*)$ as soon as β is small enough or $\dot{z}$ is close enough to z^*. It happens that for $\dot{z} = z$, we can prove even more.

Proposition 32. *The self-centered smoothed gap is a measure of optimality. Indeed, $\forall\, z \in \mathcal{Z}$, $\forall\, \beta \in [0, +\infty[^2$:*

- **i.** $G_\beta(z, z) \geq 0$.
- **ii.** $G_\beta(z, z) = 0 \Leftrightarrow z \in \mathcal{Z}^*$.
- **iii.** *For $z^* = P_{\mathcal{Z}^*}(z) \in \mathcal{Z}^*$, if $G_\beta(z, z^*) \geq \frac{\eta}{2}\mathrm{dist}_V(z, \mathcal{Z}^*)^2$, then we have $G_{\beta'}(z, z) \geq \frac{\eta'}{2}\mathrm{dist}_V(z, \mathcal{Z}^*)^2$ where $\beta' = \min(\beta/2, \eta/4)$ and $\eta' = \eta/2$.*

Proof. The function $\Phi : z' \mapsto L(x, y') - L(x', y) - \frac{\beta}{2}\|z - z'\|_V^2$ is β-strongly concave in the norm $\|\cdot\|_V$ so for $z_\beta^*(z) = \arg\max \Phi$, we have

$$G_\beta(z, z) = \max_{z'} \Phi(z') \geq \Phi(z) + \frac{\beta}{2}\|z_\beta^*(z) - z\|_V^2 .$$

Using the fact that $\Phi(z) = 0$ gives point **i**.

For the second point, it is clear by Proposition 21 that $G_\beta(z^*, z^*) = 0$. For the converse implication, we shall do the proof only for $\beta > 0$ because $G_0(z, z)$ is the usual duality gap.

$$G_\beta(z, z) = 0 \;\Rightarrow\; \frac{\beta}{2}\|z_\beta^*(z) - z\|_V^2 = 0 \;\Rightarrow\; z_\beta^*(z) = z \;\Rightarrow\; \begin{cases} 0 \in -\partial_x L(x, y) - \frac{\beta}{\tau}(x - x) \\ 0 \in -\partial_y(-L)(x, y) - \frac{\beta}{\sigma}(y - y) \end{cases} \;\Rightarrow z \in \mathcal{Z}^*$$

so that point **ii** holds.

Finally, suppose that $G_\beta(z,z^*) \geq \frac{\eta}{2}\operatorname{dist}_V(z,\mathcal{Z}^*)^2$. Since $\beta' = \min(\beta/2, \eta(\beta)/4) \leq \beta/2$, we have $G_{2\beta'}(z,z^*) \geq G_\beta(z,z^*)$. Using Lemma 31, we have

$$\begin{aligned}G_{\beta'}(z,z) &\geq G_{2\beta'}(z,z^*) - \beta' \operatorname{dist}_V(z,\mathcal{Z}^*)^2 \geq G_\beta(z,z^*) - \beta' \operatorname{dist}_V(z,\mathcal{Z}^*)^2 \geq \big(\frac{\eta}{2} - \beta'\big) \operatorname{dist}_V(z,\mathcal{Z}^*)^2 \\ &\geq \frac{\eta}{4}\operatorname{dist}_V(z,\mathcal{Z}^*)^2 \ .\end{aligned}$$ ◀

In the numerical experiment section, we shall use the self-centered smoothed gap as a stopping criterion with $\beta = (0, \delta)$ where δ is the dual infeasibility.

8.3 Adaptive restart

We now modify RAPDHG so that instead of using unknown quantities β and η to set the restart period K, we monitor the self-centered smoothed gap and restart when this quantity has been halved. In order to take into account cases where averaging is detrimental, we then compare $\tilde{z}_k$ and $\bar{z}_k$ and restart at the best of these in terms of smoothed gap. This adaptive restart is formalized in Algorithm 4 and justified by the following proposition.

Proposition 33. *Suppose that Assumption 22 holds, i.e., there exists β, η such that for all $z^* \in \mathcal{Z}^*$ and z verifying $\operatorname{dist}_V(z,\mathcal{Z}^*) \leq \operatorname{dist}_V(z_0,\mathcal{Z}^*)$ we have $G_\beta(z;z^*) \leq \frac{\eta}{2}\operatorname{dist}_V(z,\mathcal{Z}^*)$. Denote $\eta'(\beta') = 0$ if $\beta' \geq \min(\beta/2, \eta/4)$ and $\eta'(\beta') = \eta$ otherwise. Then, as soon as $\beta_s \leq \min(\beta/2,\eta/4)$ the iterates of Algorithm 4 satisfy for all $\beta' \in]0, +\infty[$,*

$$G_{\beta'}(\tilde{z}_k, \tilde{z}_k) \leq \frac{2}{(k-s)\eta'(\beta_s)}\Big(2 + a_2^+ + \frac{2}{(k-s)\beta'}\Big) G_{\beta_s}(z_s, z_s) \ .$$

where $a_2^+ = \max(0, a_2)$ and a_2 is defined in Lemma 14.

Proof. As in Proposition 29, we have $\forall\, z$,

$$L(\tilde{x}_k, y) - L(x, \tilde{y}_k) \leq \frac{1}{2(k-s)}\|z - z_s\|_V^2 - \frac{1}{2(k-s)}\|z - z_k\|_V^2 + \frac{a_2^+}{2(k-s)}\|z^* - z_s\|_V^2 - \frac{a_2^+}{2(k-s)}\|z^* - z_k\|_V^2 \ .$$

Summing (6) for l between s and $k-1$ and using the fact that the Lagrangian is convex-concave, we get for all z, We go on with

$$\begin{aligned}L(\tilde{x}_k, y) - L(x, \tilde{y}_k) - \frac{\beta'}{2}\|z - \tilde{z}_k\|_V^2 &\leq \frac{1}{2(k-s)}\|z - z_s\|_V^2 - \frac{1}{2(k-s)}\|z - z_k\|_V^2 - \frac{\beta'}{2}\|z - \tilde{z}_k\|_V^2 \\ &\quad + \frac{a_2^+}{2(k-s)}\|z^* - z_s\|_V^2 - \frac{a_2^+}{2(k-s)}\|z^* - z_k\|_V^2 \\ G_{\beta'}(\tilde{z}_k, \tilde{z}_k) &\leq \sup_z \frac{1}{2(k-s)}\|z - z_s\|_V^2 - \frac{1}{2(k-s)}\|z - z_k\|_V^2 - \frac{\beta'}{2}\|z - \tilde{z}_k\|_V^2 \\ &\quad + \frac{a_2^+}{2(k-s)}\|z^* - z_s\|_V^2 - \frac{a_2^+}{2(k-s)}\|z^* - z_k\|_V^2\end{aligned}$$

This supremum is attained at $z = \widetilde{z}_k + \frac{1}{\beta'(k-s)}(z_k - z_s)$ so that, denoting $z^* = P_{\mathcal{Z}^*}(z_s)$,

$$\begin{aligned}
G_{\beta'}(\widetilde{z}_k, \widetilde{z}_k) &\leq \frac{1}{2(k-s)}\Big\langle z_k - z_s, 2\widetilde{z}_k + \frac{1}{\beta'(k-s)}(z_k - z_s) - z_k - z_s\Big\rangle_V - \frac{1}{2\beta'(k-s)^2}\|z_k - z_s\|_V^2 \\
&\qquad + \frac{a_2^+}{2(k-s)}\|z^* - z_s\|_V^2 - \frac{a_2^+}{2(k-s)}\|z^* - z_k\|_V^2 \\
&\leq \frac{1}{2(k-s)}\|\widetilde{z}_k - z_s\|_V^2 - \frac{1}{2(k-s)}\|\widetilde{z}_k - z_k\|_V^2 + \frac{1}{2\beta'(k-s)^2}\|z_k - z_s\|_V^2 \\
&\qquad + \frac{a_2^+}{2(k-s)}\|z^* - z_s\|_V^2 - \frac{a_2^+}{2(k-s)}\|z^* - z_k\|_V^2 \\
&\leq \frac{1}{k-s}\|\widetilde{z}_k - z^*\|_V^2 + \frac{1}{k-s}\|z_s - z^*\|_V^2 - 0 + \frac{1}{\beta'(k-s)^2}\|z_k - z^*\|_V^2 + \frac{1}{\beta'(k-s)^2}\|z_s - z^*\|_V^2 \\
&\qquad + \frac{a_2^+}{2(k-s)}\|z^* - z_s\|_V^2 - \frac{a_2^+}{2(k-s)}\|z^* - z_k\|_V^2 \\
&\leq \frac{1}{k-s}\|\widetilde{z}_k - z^*\|_V^2 + \left(\frac{1}{k-s} + \frac{1}{\beta'(k-s)^2} + \frac{a_2^+}{2(k-s)}\right)\|z_s - z^*\|_V^2 \\
&\qquad + \left(\frac{1}{\beta'(k-s)^2} - \frac{a_2^+}{2(k-s)}\right)\|z_k - z^*\|_V^2 \\
&\leq \frac{1}{k-s}\left(2 + a_2^+ + \frac{2}{\beta'(k-s)}\right)\mathrm{dist}_V(z_s, \mathcal{Z}^*)^2
\end{aligned}$$

because Lemma 12 implies that $\|z_k - z^*\| \leq \|z_s - z^*\|$ for all $k \geq s$, and thus also $\|\widetilde{z}_k - z^*\| \leq \|z_s - z^*\|$. We now use the quadratic error bound of the self-centered smoothed gap, which holds thanks to Proposition 32.

$$G_{\beta'}(\widetilde{z}_k, \widetilde{z}_k) \leq \frac{2}{\eta'(\beta_s)(k-s)}\left(2 + a_2^+ + \frac{2}{\beta'(k-s)}\right)G_{\beta_s}(z_s, z_s)\ . \qquad ◀$$

Hence, choosing $\beta' = \frac{1}{k-s}$, as soon as $k - s \geq \frac{4(4+a_2^+)}{\eta'(\beta_s)}$, we have $G_{\beta'}(\widetilde{z}_k, \widetilde{z}_k) \leq 0.5 G_{\beta_s}(z_s, z_s)$. We have added additional safeguards – $\beta' = \min(\frac{1}{k-s+1}, \beta_s)$ and $G_{\beta_s}(z_s, z_s) \leq 0.01 \min(G_{\beta'}(\widetilde{z}_{k+1}, \widetilde{z}_{k+1}), G_{\beta'}(\bar{z}_{k+1}, \bar{z}_{k+1}))$ – for cases where a precipitous restart may lead to $\beta' > \min(\beta/2, \eta/4)$ and thus slow down the algorithm afterwards because we have lost control on $\eta(\beta')$.

Algorithm 4 RAPDHG with adaptive restart

$s = 0$, $\beta_0 > 0$
for $k \in \mathbb{N}$ **do**
  $z_{k+1} = T(z_k)$ – *see* (3)
  $\widetilde{z}_{k+1} = \frac{1}{k-s+1}\sum_{l=s+1}^{k+1} \bar{z}_l$
  $\beta' = \min(\frac{1}{k-s+1}, \beta_s)$
  $G^{curr} = \min(G_{\beta'}(\widetilde{z}_{k+1}, \widetilde{z}_{k+1}), G_{\beta'}(\bar{z}_{k+1}, \bar{z}_{k+1}))$
  if $G^{curr} \leq 0.5\ G_{\beta_s}(z_s, z_s)$ **or** $G_{\beta_s}(z_s, z_s) \leq 0.01\ G^{curr}$ **then**
    if $G_{\beta'}(\widetilde{z}_{k+1}, \widetilde{z}_{k+1}) \leq G_{\beta'}(\bar{z}_{k+1}, \bar{z}_{k+1})$ **then**
      Reassign $z_{k+1} \leftarrow \widetilde{z}_{k+1}$
    else
      Keep current iterate
    $z_s = z_{k+1}$
    $\beta_s = \beta'$
    $s = k$

9 Numerical experiments

In the last section, we present some numerical experiments to illustrate the linear convergence behaviour of PDHG and RAPDHG[1]. We will first look at a two linear program to show that the linear rate of RAPDHG can

[1] The code is available on `https://perso.telecom-paristech.fr/ofercoq/Software.html`

β	$\widehat{\eta}(\beta)$
1	0.00018
0.1	0.00183
0.01	0.01829
0.001	0.14474

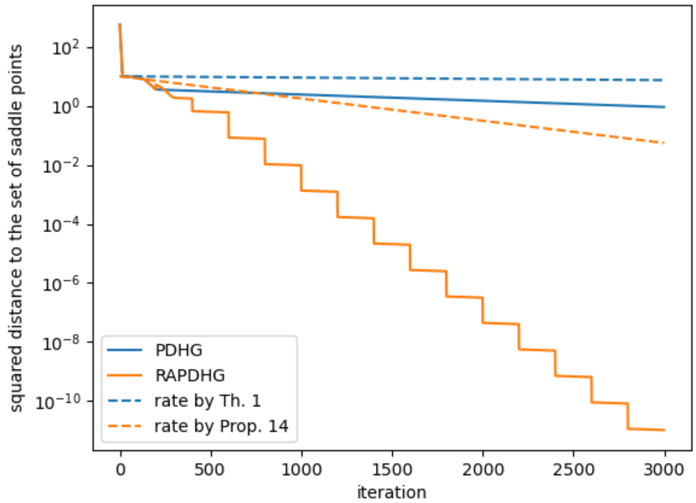

Figure 3 Table: Estimates of the quadratic error bound of the smoothed gap for several smoothing parameters. Figure: Comparison of PDHG and RAPDHG on the small linear program. The restart period of 200 was chosen because for $\beta = 1/100$, we have $\widehat{\eta}(\beta) \approx 2/100$, so that $K = \lceil \max(2/\beta, 4/\eta) \rceil = 200$.

be much faster than PDHG's. Then, we will exemplify the limits of the methods with a ridge regression problem where restarted averaging does not help and a non-polyhedral problem where we do not observe a linear rate of convergence.

9.1 Small linear program

The first experiment is on a small LP where the dual optimal set is known:

$$\begin{aligned}\min_{x \in \mathbb{R}^4, x \geq 0} \quad & -7x_1 - 9x_2 - 18x_3 - 17x_4 \\ & 2x_1 + 4x_2 + 6x_3 + 7x_4 \leq 41 \\ & x_1 + x_2 + 2x_3 + 2x_4 \leq 17 \\ & x_1 + 2x_2 + 3x_3 + 3x_4 \leq 24\end{aligned}$$

To give an estimate the quadratic error bound constant, we compute for several values of β the quantity $\widehat{\eta}(\beta) = \min_k \frac{G_\beta(z_k; z^*)}{0.5 \operatorname{dist}(z_k, \mathcal{Z}^*)^2}$. We can do it because $\mathcal{Z}^*$ is known for this small problem. Using a similar idea we can also get an estimate of the metric subregularity constant of the Lagrangian's gradient, here $\eta \approx 0.0187$.

On Figure 3, we can see that the actual rate of convergence is rather close to what is predicted by theory. Moreover, RAPDHG is much faster than PDHG. Yet, note that thousands of iterations for a LP with 4 variables and 3 constraints is not competitive with the state of the art.

9.2 Larger polyhedral problem

We then run an experiment on a more realistic problem. We run PDHG and RAPDHG with adaptive restart on the following sparse SVM problem:

$$\min_{w \in \mathbb{R}^d} \sum_{i=1}^n \max(0, 1 - y_i x_{i,:} w) + \|w\|_1$$

where $(y_i, x_{i,:})_{1 \leq i \leq n}$ are the data points from the a1a dataset [7] ($d = 119$ and $n = 1,605$). We normalized the data matrix so that $\|x_{:,j}\|_2 = 1$.

The convergence profile is given in Figure 4. The behaviour of the algorithms is similar to what was seen in the small size problem. Here however, we can see clearly two phases. In the beginning, we observe a sublinear convergence, where restart and averaging does not help. Then the linear rate kicks in after a nonnegligible time. We believe that it comes from something related to the condition $G_\beta(z; z^*) \leq R$ in Proposition 27. Note that this cold start phase is quite long. On our laptop computer with 4 Intel(R) Core(TM) i5-7200U CPU @ 2.50GHz it took 5.7s while the adaptive proximal point method of [24] took 0.93s to solve the problem.

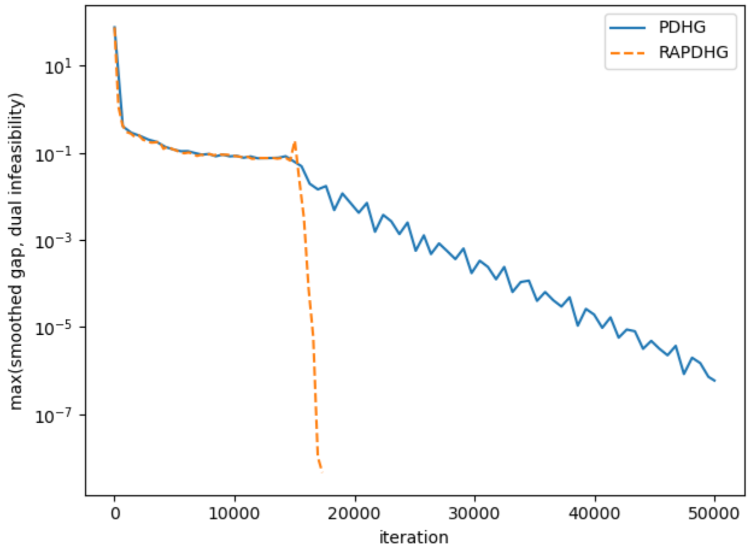

Figure 4 Comparison of PDHG and RAPDHG: sparse SVM on the a1a dataset. We are plotting the optimality measure for the last iterate

9.3 Ridge regression

In this experiment, we test on a problem where restarting does not help. We consider least squares with ℓ_2 regularization

$$\min_x \frac{1}{2}\|Ax - b\|^2 + 50\|x\|^2$$

where A and b are given by the real-sim dataset [7]. Since we know the strong convexity-concavity parameter of the Lagrangian, we choose the step sizes σ and τ as in Section 5.1. As a consequence, PDHG has a convergence rate that matches the theoretical lower bound for this class problem and cannot be improved.

We can see on Figure 5 that, as expected, restart and averaging does not help: $\bar{z}_k$ is consistently better than $\widetilde{z}_k$ so that RAPDHG with adaptive restart selects the same sequence as PDHG and the two curves match. We added a comparison with restarted-FISTA [15] to show that the choice of step sizes indeed suffices to get an algorithm with accelerated rate.

9.4 TV-L1

We consider the minimization of the following non-polyhedral function

$$\min_x \lambda\|x - I\|_1 + \|Dx\|_{2,1}$$

where I is the Cameraman image, D is the 2D discrete gradient, $\|z\|_{2,1} = \sum_{p\in P}\sqrt{z_{p,1}^2 + z_{p,2}^2}$ and $\lambda = 1.9$. This problem is not piecewise linear-quadratic, so that our linear convergence result does not hold. Yet is rather structured: it is equivalent to a second order cone program. We can see in Figure 6 that this is a difficult problem for PDHG but that RAPDHG does improve the convergence speed significantly. The solution we obtain is shown in Figure 7.

10 Conclusion

In this paper, we have tried to understand the linear rate of convergence of primal-dual hybrid gradient. Even on a very simple problem, we have seen that current regularity assumptions are not sufficient to explain the

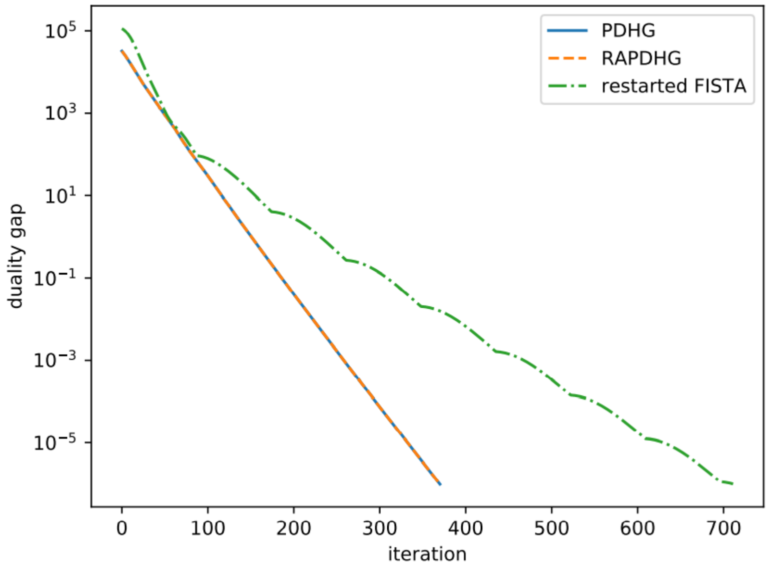

Figure 5 Solving ℓ_2 regularized least squares on the real-sim dataset.

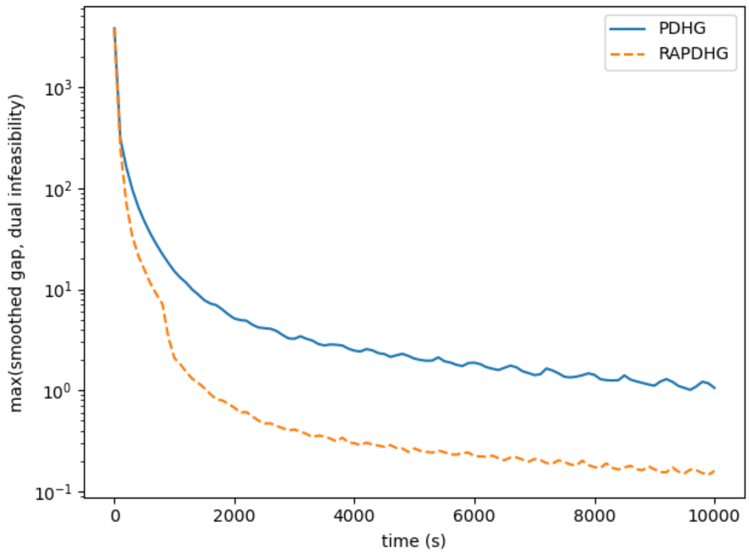

Figure 6 Comparison of PDHG and RAPDHG on the ℓ_1 ROF problem.

behavior of the algorithm. We have then introduced the quadratic error bound of the smoothed gap and argue that this new condition is more widely applicable and more precise than previous ones. Finally, we showed how this new knowledge can be used to improve the algorithm.

Figure 7 Left:original image – Right: solution, 59% of the pixels are unchanged

This work opens several perspectives:

- Can the quadratic error bound of the smooth gap be used to understand better the convergence rate of other primal-dual algorithms? Interesting cases would be the ADMM, the augmented Lagrangian method and coordinate update methods to cite a few.
- We have seen in (11) that the smoothed gap at a non-optimal point can approximate the smoothed gap at an optimal point. Considering it as a stopping criterion would be an alternative to the KKT error, which implicitly requires metric sub-regularity to make sense, and duality gap, which is $+\infty$ nearly everywhere for linearly constrained problems.
- Our first attempt for the design of a primal-dual algorithm with an improved linear rate of convergence has shown the usefulness of our regularity assumption. Would we be able to design an optimal algorithm for the class of problems with a given quadratic error bound of the smoothed gap function?

A Proofs of Section 3

Lemma 11. Let $p = \mathrm{prox}_{\tau f}(x)$ and $p' = \mathrm{prox}_{\tau f}(x')$ where f is μ_f-strongly convex. For all x and x',

$$f(p) + \frac{1}{2\tau}\|p - x\|^2 \leq f(x') + \frac{1}{2\tau}\|x' - x\|^2 - \frac{1 + \tau\mu_f}{2\tau}\|p - x'\|^2$$
$$(1 + 2\tau\mu_f)\|p - p'\|^2 \leq \|x' - x\|^2 - \|p - x - p' + x'\|^2 \ .$$

Proof. $p = \arg\min_z f(z) + \frac{1}{2\tau}\|z - x\|^2$

Yet, $h : z \mapsto f(z) + \frac{1}{2\tau}\|z - x\|^2 - \frac{1+\tau\mu_f}{2\tau}\|p - z\|^2$ is convex and $0 \in \partial h(p)$. This implies the first inequality by Fermat's rule.

We now apply the first inequality at (x, p') and at (x', p) and then sum.

$$f(p) + \frac{1}{2\tau}\|p - x\|^2 + f(p') + \frac{1}{2\tau}\|p' - x'\|^2$$
$$\leq f(p') + \frac{1}{2\tau}\|p' - x\|^2 - \frac{1 + \tau\mu_f}{2\tau}\|p - p'\|^2 + f(p) + \frac{1}{2\tau}\|p - x'\|^2 - \frac{1 + \tau\mu_f}{2\tau}\|p' - p\|^2$$

Rearranging the squared norm terms we get

$$(1 + \tau\mu_f)\|p' - p\|^2 \leq \langle p - p', x - x'\rangle$$
$$\|p - x - p' + x'\|^2 = \|p - p'\|^2 + \|x - x'\|^2 - 2\langle p - p', x - x'\rangle \leq \|x - x'\|^2 - (1 + 2\tau\mu_f)\|p - p'\|^2$$ ◀

Lemma 12. Let $T : \mathcal{X} \times \mathcal{Y} \to \mathcal{X} \times \mathcal{Y}$ be defined for any (x, y) by (3). Suppose that ∇f_2 is L_f-Lipschitz continuous and ∇g_2^* is L_{g^*}-Lipschitz continuous. If the step sizes satisfy $\gamma = \sigma\tau\|A\|^2 < 1$, $\alpha_f = \tau L_f/2 < 1$, $\alpha_g = \sigma L_{g^*}/2 < 1$ then T is nonexpansive in the norm $\|\cdot\|_V$,

$$\|T(z) - T(z')\|_V^2 \leq \|z - z'\|_V^2 - 2\tilde{V}(z, z') \tag{4}$$

and T is $\frac{1}{1+\lambda}$-averaged where

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

which means for $z = (x, y)$ and $z' = (x', y')$

$$\|T(z) - T(z')\|_V^2 \leq \|z - z'\|_V^2 - \lambda\|z - T(z) - z' + T(z')\|_V^2\,. \tag{5}$$

As a consequence, (z_k) converges to a saddle point of the Lagrangian. Moreover, if $\sigma L_{g^*}/2 \leq \alpha_f(1 - \sigma\tau\|A\|^2)$, then $\lambda \geq (1-\sqrt{\gamma})(1-\alpha_f)$.

Proof. In the appendix, we will improve slightly the result in the case where f or g^* is strongly convex. Note that all what follows works even if $\mu_f = \mu_{g^*} = 0$.

Since the proximal operator of a convex function is firmly nonexpansive, for (x, y), $(x', y') \in \mathcal{Z}$,

$$\begin{aligned}
(1 + 2\mu_f\tau)\|\bar{x} - \bar{x}'\|^2 &\leq \|x - \tau\nabla f_2(x) - \tau A^\top y - x' + \tau\nabla f_2(x') + \tau A^\top y'\|^2 \\
&\quad - \|x - \tau\nabla f_2(x) - \tau A^\top y - \bar{x} - x' + \tau\nabla f_2(x') + \tau A^\top y' + \bar{x}'\|^2 \\
&= \|x - \tau\nabla f_2(x) - x' + \tau\nabla f_2(x')\|^2 + \tau^2\|A^\top(y - y')\|^2 \\
&\quad - 2\tau\langle x - \tau\nabla f_2(x) - x' + \tau\nabla f_2(x'), A^\top(y - y')\rangle \\
&\quad - \|x - \tau\nabla f_2(x) - \bar{x} - x' + \tau\nabla f_2(x') + \bar{x}'\|^2 - \tau^2\|A^\top(y - y')\|^2 \\
&\quad + 2\tau\langle x - \tau\nabla f_2(x) - \bar{x} - x' + \tau\nabla f_2(x') + \bar{x}', A^\top(y - y')\rangle \\
&= \|x - \tau\nabla f_2(x) - x' + \tau\nabla f_2(x')\|^2 - \|x - \tau\nabla f_2(x) - \bar{x} - x' + \tau\nabla f_2(x') + \bar{x}'\|^2 \\
&\quad - 2\tau\langle \bar{x} - \bar{x}', A^\top(y - y')\rangle
\end{aligned}$$

We also have

$$\begin{aligned}
\|x - \tau\nabla f_2(x) - x' + \tau\nabla f_2(x')\|^2 &= \|x - x'\|^2 + \tau^2\|\nabla f_2(x) - \nabla f_2(x')\|^2 - 2\tau\langle\nabla f_2(x) - \nabla f_2(x'), x - x'\rangle \\
&\leq \|x - x'\|^2 - \Big(\frac{2\tau}{L_f} - \tau^2\Big)\|\nabla f_2(x) - \nabla f_2(x')\|^2
\end{aligned}$$

$$\begin{aligned}
&\|x - \tau\nabla f_2(x) - \bar{x} - x' + \tau\nabla f_2(x') + \bar{x}'\|^2 \\
&\qquad = \|x - \bar{x} - x' + \bar{x}'\|^2 + \tau^2\|\nabla f_2(x) - \nabla f_2(x')\|^2 - 2\tau\langle\nabla f_2(x) - \nabla f_2(x'), x - x' - \bar{x} + \bar{x}'\rangle \\
&\qquad \geq (1 - \alpha_f)\|x - \bar{x} - x' - \bar{x}'\|^2 + \tau^2(1 - \alpha_f^{-1})\|\nabla f_2(x) - \nabla f_2(x')\|^2
\end{aligned}$$

for all $\alpha_f > 0$. Hence,

$$\begin{aligned}
(1 + 2\mu_f\tau)\|\bar{x} - \bar{x}'\|^2 \leq \|x - x'\|^2 - (1 - \alpha_f)\|x - \bar{x} - x' + \bar{x}'\|^2 - 2\tau\langle\bar{x} - \bar{x}', A^\top(y - y')\rangle \\
- \Big(\frac{2\tau}{L_f} - \alpha_f^{-1}\tau^2\Big)\|\nabla f_2(x) - \nabla f_2(x')\|^2
\end{aligned}$$

Similarly,

$$\begin{aligned}
(1 + 2\mu_{g^*}\sigma)\|\bar{y} - \bar{y}'\|^2 \leq \|y - y'\|^2 - (1 - \alpha_g)\|y - \bar{y} - y' + \bar{y}'\|^2 + 2\sigma\langle\bar{y} - \bar{y}', A(\bar{x} - \bar{x}')\rangle \\
- \Big(\frac{2\sigma}{L_{g^*}} - \alpha_g^{-1}\sigma^2\Big)\|\nabla g_2(y) - \nabla g_2(y')\|^2
\end{aligned}$$

We then proceed to

$$\begin{aligned}
\|T(x,y)-T(x',y')\|_V^2 &= \frac{1}{\tau}\|\bar{x}-\tau A^\top(\bar{y}-y)-\bar{x}'+\tau A^\top(\bar{y}'-y')\|^2+\frac{1}{\sigma}\|\bar{y}-\bar{y}'\|^2\\
&= \frac{1}{\tau}\|\bar{x}-\bar{x}'\|^2+\tau\|A^\top(\bar{y}-y)-A^\top(\bar{y}'-y')\|^2\\
&\qquad -2\langle\bar{x}-\bar{x}', A^\top(\bar{y}-y)-A^\top(\bar{y}'-y')\rangle+\frac{1}{\sigma}\|\bar{y}-\bar{y}'\|^2\\
&\le \frac{1}{\tau}\|x-x'\|^2-\frac{1-\alpha_f}{\tau}\|x-\bar{x}-x'+\bar{x}'\|^2-2\langle\bar{x}-\bar{x}', A^\top(y-y')\rangle\\
&\qquad +\tau\|A^\top(\bar{y}-y-\bar{y}'+y')\|^2-2\langle\bar{x}-\bar{x}', A^\top(\bar{y}-y)-A^\top(\bar{y}'-y')\rangle\\
&\qquad +\frac{1}{\sigma}\|y-y'\|^2-\frac{1-\alpha_g}{\sigma}\|y-\bar{y}-y'+\bar{y}'\|^2+2\langle\bar{y}-\bar{y}', A(\bar{x}-\bar{x}')\rangle\\
&\qquad -\left(\frac{2\tau}{L_f}-\alpha_f^{-1}\tau^2\right)\|\nabla f_2(x)-\nabla f_2(x')\|^2-2\mu_f\|\bar{x}-\bar{x}'\|^2\\
&\qquad -\left(\frac{2\sigma}{L_{g^*}}-\alpha_g^{-1}\sigma^2\right)\|\nabla g_2(y)-\nabla g_2(y')\|^2-2\mu_{g^*}\|\bar{y}-\bar{y}'\|^2
\end{aligned}$$

We choose $\alpha_f=\tau L_f/2<1$ and $\alpha_g=\sigma L_{g^*}/2<1$ and we note that

$$-2\langle\bar{x}-\bar{x}', A^\top(y-y')\rangle-2\langle\bar{x}-\bar{x}', A^\top(\bar{y}-y)-A^\top(\bar{y}'-y')\rangle+2\langle\bar{y}-\bar{y}', A(\bar{x}-\bar{x}')\rangle=0\ .$$

This leads to

$$\begin{aligned}
\|T(x,y)-T(x',y')\|_V^2\le\|z-z'\|_V^2-\frac{1-\alpha_f}{\tau}\|x-\bar{x}-x'+\bar{x}'\|^2-\frac{1-\alpha_g-\tau\sigma\|A\|^2}{\sigma}\|y-\bar{y}-y'+\bar{y}'\|^2\\
-2\mu_f\|\bar{x}-\bar{x}'\|^2-2\mu_{g^*}\|\bar{y}-\bar{y}'\|^2
\end{aligned}$$

which proves (4). Now, we shall prove that $V(z,z')\ge\frac{\lambda}{2}\|z-T(z)-z'+T(z')\|_V^2$. For any $\lambda\in[0,1-\alpha_f]$ and $\alpha>0$,

$$\begin{aligned}
\|T(x,y)-T(x',y')\|_V^2\le{}&\frac{1}{\tau}\|x-x'\|^2-\frac{1-\alpha_f-\lambda}{\tau}\|x-\bar{x}-x'+\bar{x}'\|^2\\
&-\frac{\lambda}{\tau}\|x-\bar{x}+\tau A^\top(\bar{y}-y)-x'+\bar{x}'-\tau A^\top(\bar{y}'-y')\|^2\\
&+\lambda\tau\|A^\top(\bar{y}-y-\bar{y}'+y')\|^2\\
&+2\lambda\langle x-\bar{x}-x'+\bar{x}', A^\top(\bar{y}-y)-A^\top(\bar{y}'-y')\rangle\\
&+\frac{1}{\sigma}\|y-y'\|^2-\frac{1-\alpha_g-\sigma\tau\|A\|^2}{\sigma}\|y-\bar{y}-y'+\bar{y}'\|^2\\
&-2\mu_f\|\bar{x}-\bar{x}'\|^2-2\mu_{g^*}\|\bar{y}-\bar{y}'\|^2
\end{aligned}$$

$$\begin{aligned}
\|T(x,y)-T(x',y')\|_V^2\le{}&\frac{1}{\tau}\|x-x'\|^2+\frac{1}{\sigma}\|y-y'\|^2\\
&-\frac{\lambda}{\tau}\|x-\bar{x}+\tau A^\top(\bar{y}-y)-x'+\bar{x}'-\tau A^\top(\bar{y}'-y')\|^2\\
&-\frac{\lambda}{\sigma}\|y-\bar{y}-y'+\bar{y}'\|^2+\left(\frac{\lambda}{\tau\alpha}-\frac{1-\alpha_f-\lambda}{\tau}\right)\|x-\bar{x}-x'+\bar{x}'\|^2\\
&+\left((1+\lambda+\lambda\alpha)\tau\|A\|^2-\frac{1-\alpha_g-\lambda}{\sigma}\right)\|(\bar{y}-y-\bar{y}'+y')\|^2\\
&-2\mu_f\|\bar{x}-\bar{x}'\|^2-2\mu_{g^*}\|\bar{y}-\bar{y}'\|^2
\end{aligned}$$

where $\lambda\in[0,1-\alpha_f]$ and $\alpha>0$ are arbitrary. We choose λ and α such that

$$\begin{aligned}
&\frac{\lambda}{\alpha}=1-\alpha_f-\lambda\\
&(1+\lambda+\lambda\alpha)\gamma=1-\alpha_g-\lambda
\end{aligned}$$

that is $\lambda=1-\sqrt{\gamma}$ and $\alpha=\frac{\lambda}{1-\lambda}=\frac{1-\sqrt{\gamma}}{\sqrt{\gamma}}$ when $f_2=0$ and $g_2=0$. In the case f_2 and g_2 non zero, we take

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

Note that as soon as $\alpha_g \leq (1-\gamma)\alpha_f$, we have $(1-\alpha_f)(1-\sqrt{\gamma}) \leq \lambda \leq 1-\alpha_f$. We continue as

$$\|T(x,y) - T(x',y')\|_V^2 \leq \frac{1}{\tau}\|x-x'\|^2 + \frac{1}{\sigma}\|y-y'\|^2 - \frac{\lambda}{\tau}\|x - \bar{x} + \tau A^\top(\bar{y} - y) - x' + \bar{x}' - \tau A^\top(\bar{y}' - y')\|^2$$
$$- \frac{\lambda}{\sigma}\|y - \bar{y} - y' + \bar{y}'\|^2 - 2\mu_f\|\bar{x} - \bar{x}'\|^2 - 2\mu_{g^*}\|\bar{y} - \bar{y}'\|^2 \; .$$

We get that T is β-averaged with $\frac{1-\beta}{\beta} = \lambda$, that is $\beta = \frac{1}{\lambda+1}$.

For the convergence, we use Krasnosels'kii Mann theorem [4]. ◀

Lemma 13. For any $z^* \in \mathcal{Z}^*$, $\widetilde{V}$ satisfies

$$\widetilde{V}(z_k, z^*) = \frac{1-\alpha_f}{2\tau}\|\bar{x}_{k+1} - x_k\|^2 + \frac{1-\alpha_g-\gamma}{2\sigma}\|\bar{y}_{k+1} - y_k\|^2 \geq \frac{\lambda}{2}\|z_{k+1} - z_k\|_V^2 \; .$$

Proof. The last part of the proof of Lemma 12 shows that for any $z, z' \in \mathcal{Z}$,

$$V(z,z') \geq \frac{\lambda}{2}\|z - T(z) - z' + T(z')\|_V^2$$

Since $T(z^*) = z^*, T(z_k) = z_{k+1}$, we get the desired result. ◀

Lemma 14. Suppose that $\gamma = \sigma\tau\|A\|^2 < 1$, $\tau L_f/2 = \alpha_f < 1$, $\alpha_g = \sigma L_{g^*}/2 < 1$. For all $k \in \mathbb{N}$ and for all $z \in \mathcal{Z}$,

$$L(\bar{x}_{k+1}, y) - L(x, \bar{y}_{k+1}) \leq \frac{1}{2}\|z - z_k\|_V^2 - \frac{1}{2}\|z - z_{k+1}\|_V^2 + a_2\widetilde{V}(z_k, z^*) \tag{6}$$

where $\widetilde{V}(z_k, z^*) = (\frac{1}{2\tau} - \frac{L_f}{2})\|\bar{x}_{k+1} - x_k\|^2 + (\frac{1}{2\sigma} - \frac{\tau\|A\|^2}{2} - \frac{L_{g^*}}{2})\|\bar{y}_{k+1} - y_k\|^2$ and $a_2 = \max(\frac{2\alpha_f - 1}{1-\alpha_f}, \frac{2\alpha_g - 1 + \gamma}{1-\alpha_g-\gamma})$. $a_2 \geq -1$ may be positive or negative.

Proof. By Taylor–Lagrange inequality and convexity of f_2 and g_2^*,

$$\begin{aligned}
f_2(\bar{x}_{k+1}) &\leq f_2(x_k) + \langle \nabla f_2(x_k), \bar{x}_{k+1} - x_k\rangle + \frac{L_f}{2}\|\bar{x}_{k+1} - x_k\|^2 \\
&\leq f_2(x) + \langle \nabla f_2(x_k), \bar{x}_{k+1} - x\rangle + \frac{L_f}{2}\|\bar{x}_{k+1} - x_k\|^2 - \frac{\tau\mu_{f_2}}{\tau}\|x_k - x\|^2 \\
g_2^*(\bar{y}_{k+1}) &\leq g_2^*(y_k) + \langle \nabla g_2^*(y_k), \bar{y}_{k+1} - y_k\rangle + \frac{L_{g^*}}{2}\|\bar{y}_{k+1} - y_k\|^2 \\
&\leq g_2^*(y) + \langle \nabla g_2^*(y_k), \bar{y}_{k+1} - y\rangle + \frac{L_{g^*}}{2}\|\bar{y}_{k+1} - y_k\|^2 - \frac{\sigma\mu_{g_2^*}}{\sigma}\|y_k - y\|^2
\end{aligned}$$

By definitions of $\bar{x}_{k+1}$ and $\bar{y}_{k+1}$, for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, we have:

$$f(\bar{x}_{k+1}) \leq f(x) + \langle \nabla f_2(x_k) + A^\top y_k, x - \bar{x}_{k+1}\rangle + \frac{1}{2\tau}\|x - x_k\|^2 - \frac{1+\tau\mu_f}{2\tau}\|x - \bar{x}_{k+1}\|^2 - \frac{1}{2\tau}\|\bar{x}_{k+1} - x_k\|^2$$
$$g^*(\bar{y}_{k+1}) \leq g^*(y) + \langle \nabla g_2^*(y_k) - A\bar{x}_{k+1}, y - \bar{y}_{k+1}\rangle + \frac{1}{2\sigma}\|y - y_k\|^2 - \frac{1+\sigma\mu_{g^*}}{2\sigma}\|y - \bar{y}_{k+1}\|^2 - \frac{1}{2\sigma}\|\bar{y}_{k+1} - y_k\|^2$$

Summing these inequalities and using the relations $x_{k+1} = \bar{x}_{k+1} - \tau A^\top(\bar{y}_{k+1} - y_k)$ and $y_{k+1} = \bar{y}_{k+1}$ yields

$$
\begin{aligned}
&L(\bar{x}_{k+1}, y) - L(x, \bar{y}_{k+1}) \\
&\quad = f(\bar{x}_{k+1}) + f_2(\bar{x}_{k+1}) + \langle A\bar{x}_{k+1}, y\rangle - g^*(y) - g_2^*(y) - f(x) - f_2(x) - \langle Ax, \bar{y}_{k+1}\rangle + g^*(\bar{y}_{k+1}) + g_2^*(\bar{y}_{k+1}) \\
&\quad \leq \frac{1-\tau\mu_{f_2}}{2\tau}\|x - x_k\|^2 + \frac{1-\sigma\mu_{g_2^*}}{2\sigma}\|y - y_k\|^2 - \frac{1}{2\tau}\|x - x_{k+1}\|^2 - \frac{1}{2\sigma}\|y - y_{k+1}\|^2 \\
&\qquad - \frac{1}{2\tau}\|x_{k+1} - \bar{x}_{k+1}\|^2 - \frac{1}{\tau}\langle x - x_{k+1}, x_{k+1} - \bar{x}_{k+1}\rangle \\
&\qquad + \langle A\bar{x}_{k+1}, y\rangle - \langle Ax, \bar{y}_{k+1}\rangle + \langle A^\top y_k, x - \bar{x}_{k+1}\rangle - \langle A\bar{x}_{k+1}, y - \bar{y}_{k+1}\rangle \\
&\qquad - \frac{1}{2\tau}\|\bar{x}_{k+1} - x_k\|^2 + \frac{1}{2\sigma}\|\bar{y}_{k+1} - y_k\|^2 + \frac{L_f}{2}\|\bar{x}_{k+1} - x_k\|^2 + \frac{L_{g^*}}{2}\|\bar{y}_{k+1} - y_k\|^2 \\
&\qquad - \frac{\tau\mu_f}{2\tau}\|\bar{x}_{k+1} - x\|^2 - \frac{\sigma\mu_{g^*}}{2\sigma}\|\bar{y}_{k+1} - y\|^2 \\
&\quad = \frac{1}{2}\|z - z_k\|^2_{V-\mu_2} - \frac{1}{2}\|z - z_{k+1}\|^2_V - \frac{\tau}{2}\|A^\top(\bar{y}_{k+1} - y_k)\|^2 \\
&\qquad + \langle x - \bar{x}_{k+1} + \tau A^\top(\bar{y}_{k+1} - y_k), A^\top(\bar{y}_{k+1} - y_k)\rangle + \langle A(\bar{x}_{k+1} - x), \bar{y}_{k+1} - y\rangle \\
&\qquad - \frac{1}{2}\|\bar{z}_{k+1} - z_k\|^2_V + \frac{L_f}{2}\|\bar{x}_{k+1} - x_k\|^2 + \frac{L_{g^*}}{2}\|\bar{y}_{k+1} - y_k\|^2 - \frac{1}{2}\|\bar{z}_{k+1} - z\|^2_\mu \\
&\quad = \frac{1}{2}\|z - z_k\|^2_V - \frac{1}{2}\|z - z_{k+1}\|^2_V + \frac{\tau}{2}\|A^\top(\bar{y}_{k+1} - y_k)\|^2 - \frac{1}{2}\|\bar{z}_{k+1} - z_k\|^2_V \\
&\qquad + \frac{L_f}{2}\|\bar{x}_{k+1} - x_k\|^2 + \frac{L_{g^*}}{2}\|\bar{y}_{k+1} - y_k\|^2 - \frac{1}{2}\|\bar{z}_{k+1} - z\|^2_\mu
\end{aligned}
$$

Since $\widetilde{V}(z_k, z^*) = \frac{1-\alpha_f}{2\tau}\|\bar{x}_{k+1} - x_k\|^2 + (\frac{1-\alpha_g-\gamma}{2\sigma})\|\bar{y}_{k+1} - y_k\|^2$, $\alpha_f \geq \frac{\tau L_f}{2}$ and $\alpha_g = \frac{\sigma L_{g^*}}{2}$, we can write

$$
\begin{aligned}
&\frac{\tau}{2}\|A^\top(\bar{y}_{k+1} - y_k)\|^2 - \frac{1}{2}\|\bar{z}_{k+1} - z_k\|^2_V + \frac{L_f}{2}\|\bar{x}_{k+1} - x_k\|^2 + \frac{L_{g^*}}{2}\|\bar{y}_{k+1} - y_k\|^2 \\
&\quad \leq \frac{1}{2\tau}(2\alpha_f - 1)\|\bar{x}_{k+1} - x_k\|^2 + \frac{1}{2\sigma}(\gamma + 2\alpha_g - 1)\|\bar{y}_{k+1} - y_k\|^2 \leq \max\left(\frac{2\alpha_f - 1}{1-\alpha_f}, \frac{\gamma + 2\alpha_g - 1}{1 - \gamma - \alpha_g}\right)\widetilde{V}(z_k, z^*)
\end{aligned}
$$

Hence,

$$
L(\bar{x}_{k+1}, y) - L(x, \bar{y}_{k+1}) \leq \frac{1}{2}\|z - z_k\|^2_{V-\mu_2} - \frac{1}{2}\|z - z_{k+1}\|^2_V + a_2\widetilde{V}(z_k, z^*) - \frac{1}{2}\|\bar{z}_{k+1} - z\|^2_\mu
$$

where $a_2 = \max\left(\frac{2\alpha_f-1}{1-\alpha_f}, \frac{\gamma+2\alpha_g-1}{1-\gamma-\alpha_g}\right) \geq -1$ may be negative or positive. ◀

Proposition 15. Let $z_0 \in \mathcal{Z}$ and let $R \subseteq \mathcal{Z}$. If $\sigma\tau\|A\|^2 + \sigma L_{g^*} \leq 1$ and $\tau L_f \leq 1$ then we have the stability

$$
\|z_k - z^*\|_V \leq \|z_0 - z^*\|_V
$$

for all $z^* \in \mathcal{Z}^*$. Define $\widetilde{z}_k = \frac{1}{k}\sum_{l=1}^k \bar{z}_l$ and the restricted duality gap $G(\bar{z}, R) = \sup_{z\in R} L(\bar{x}, y) - L(x, \bar{y})$. We have the sublinear iteration complexity

$$
G(\widetilde{z}_k, R) \leq \frac{1}{2k}\sup_{z\in R}\|z - z_0\|^2_V \ .
$$

Proof. For any $z^* \in \mathcal{Z}^*$, $L(\bar{x}_{k+1}, y^*) - L(x^*, \bar{y}_{k+1}) \geq 0$ which implies by Lemma 14 the stability inequality, since $a_2 \leq 0$ in the case $\alpha_f \leq \frac{1}{2}$ and $2\alpha_g + \gamma \leq 1$.

$$
\frac{1}{2}\|z^* - z_{k+1}\|^2_V \leq \frac{1}{2}\|z^* - z_k\|^2_V \leq \frac{1}{2}\|z^* - z_0\|^2_V \ .
$$

We then sum (6) for k between 0 and $K-1$ and use convexity in x and concavity in y of the Lagrangian:

$$
K\big(L(\widetilde{x}_K, y) - L(x, \widetilde{y}_K)\big) \leq \sum_{k=0}^{K-1} L(\bar{x}_{k+1}, y) - L(x, \bar{y}_{k+1}) \leq \frac{1}{2}\|z - z_0\|^2_V - \frac{1}{2}\|z - z_K\|^2_V - \sum_{k=0}^{K-1}\widetilde{V}(\bar{z}_{k+1} - z_k)
$$

In particular,

$$
G((\widetilde{x}_K, \widetilde{y}_K), R) \leq \frac{1}{2K}\sup_{z\in R}\|z - z_0\|^2_V - \|z - z_K\|^2_V \ .
$$

◀

B Proofs of Section 4

Proposition 18. If $f + f_2$ has a $L'_f + L_f$-Lipschitz gradient and is μ_f-strongly convex, and $g + g_2 = \iota_{\{b\}}$, then PDHG converges linearly with rate

$$\Big(1 + \frac{\eta}{(2+a_2)(1+\eta/\lambda)}\Big) \operatorname{dist}_V(z_{k+1}, \mathcal{Z}^*)^2 \leq \operatorname{dist}_V(z_k, \mathcal{Z}^*)^2$$

where $\eta = \min(\mu_f \tau, \frac{\sigma\tau\sigma_{\min}(A)^2}{\tau L_f + \tau L'_f + \frac{1}{\lambda}})$, λ is defined in Lemma 12 and $a_2 \geq -1$ is defined in Lemma 14.

Proof. We know by Lemmas 14 and 13 that for all $z = (x, y)$,

$$L(\bar{x}_{k+1}, y) - L(x, \bar{y}_{k+1}) \leq \frac{1}{2}\|z - z_k\|_V^2 - \frac{1}{2}\|z - z_{k+1}\|_2^2 + a_2\widetilde{V}(z_k, z^*) \ .$$

We shall choose $y = y^* \in \mathcal{Y}^*$. By strong convexity of $f + f_2$,

$$L(\bar{x}_{k+1}, y^*) \geq L(x^*, y^*) + \frac{\mu_f}{2}\|\bar{x}_{k+1} - x^*\|^2 \ .$$

For the dual vector, we use the smoothness of the objective, the equality $\nabla f(x^*) + \nabla f_2(x^*) = -A^\top y^*$ and $Ax^* = b$.

$$\begin{aligned}
-L(x, \bar{y}_{k+1}) &= -f(x) - f_2(x) - \langle Ax - b, \bar{y}_{k+1}\rangle \\
&\geq -f(x^*) - f_2(x^*) - \langle \nabla f(x^*) - \nabla f_2(x^*), x - x^*\rangle - \frac{L_f + L'_f}{2}\|x - x^*\|^2 - \langle Ax - b, \bar{y}_{k+1}\rangle \\
&= -L(x^*, y^*) + \langle A^\top y^*, x - x^*\rangle - \langle x - x^*, A^\top \bar{y}_{k+1}\rangle - \frac{L_f + L'_f}{2}\|x - x^*\|^2
\end{aligned}$$

For $a \in \mathbb{R}$, we choose $x = x^* + aA^\top(y^* - \bar{y}_{k+1})$ so that

$$-L(x^* + aA^\top(y^* - \bar{y}_{k+1}), \bar{y}_{k+1}) \geq -L(x^*, y^*) + (a - a^2\frac{L_f + L'_f}{2})\|A^\top(\bar{y}_{k+1} - y^*)\|^2 \ .$$

Moreover, we can show that $\|A^\top \bar{y} - A^\top y^*\| \geq \sigma_{\min(A)} \operatorname{dist}(\bar{y}, \mathcal{Y}^*)$, where $\sigma_{\min(A)}$ is the smallest singular value of A. Indeed, $\mathcal{Y}^* = \{y : A^\top y = -\nabla(f + f_2)(x^*)\} = P_{\mathcal{Y}^*}(\bar{y}) + \ker A^\top$ is an affine space. Here, we denoted by $P_{\mathcal{Y}^*}$ the orthogonal projection on $\mathcal{Y}^*$. We can then decompose $\bar{y}$ as $\bar{y} = P_{\mathcal{Y}^*}(\bar{y}) + z$ where $z \in \ker A^\top = (\operatorname{Im} A)^\perp$. This leads to $\|A^\top \bar{y} - A^\top y^*\| = \|A^\top P_{\mathcal{Y}^*}(\bar{y}) - A^\top y^*\| \geq \sigma_{\min(A)}\|P_{\mathcal{Y}^*}(\bar{y}) - y^*\|$ because $P_{\mathcal{Y}^*}(\bar{y}) - y^* \in (\ker A^\top)^\perp$.

We now develop

$$\begin{aligned}
&\frac{1}{2\tau}\|x^* + aA^\top(y^* - \bar{y}_{k+1}) - x_k\|^2 - \frac{1}{2\tau}\|x^* + aA^\top(y^* - \bar{y}_{k+1}) - x_{k+1}\|^2 \\
&\quad = \frac{1}{2\tau}\|x^* - x_k\|^2 - \frac{1}{2\tau}\|x^* - x_{k+1}\|^2 + \frac{a}{\tau}\langle x_k - x_{k+1}, A^\top(y^* - \bar{y}_{k+1})\rangle \\
&\quad \leq \frac{1}{2\tau}\|x^* - x_k\|^2 - \frac{1}{2\tau}\|x^* - x_{k+1}\|^2 + \frac{\lambda}{2\tau}\|x_k - x_{k+1}\|^2 + \frac{a^2}{2\tau\lambda}\|A^\top(y^* - \bar{y}_{k+1})\|^2
\end{aligned}$$

Combining the three inequalities, we obtain

$$\frac{1}{2}\|z^* - z_k\|^2 - \frac{1}{2}\|z^* - z_{k+1}\|^2 + a_2\widetilde{V}(z_k, z^*) \geq \frac{\mu_f}{2}\|\bar{x}_{k+1} - x^*\|^2 + \Big(a - a^2\frac{L_f + L'_f}{2} - a^2\frac{1}{2\tau\Gamma}\Big)\|A^\top(\bar{y}_{k+1} - y^*)\|^2 \ .$$

We choose $a = \frac{\tau}{\tau L_f + \tau L'_f + \frac{1}{\lambda}}$ and we use $\|A^\top \bar{y} - A^\top y^*\| \geq \sigma_{\min(A)} \operatorname{dist}(\bar{y}, \mathcal{Y}^*)$ to get

$$\frac{1}{2}\|z^* - z_k\|^2 - \frac{1}{2}\|z^* - z_{k+1}\|^2 + a_2\widetilde{V}(z_k, z^*) \geq \frac{\mu_f\tau}{2}\|\bar{x}_{k+1} - x^*\|_{\tau^{-1}}^2 + \frac{\sigma\tau\sigma_{\min}(A)^2/2}{\tau L_f + \tau L'_f + \frac{1}{\lambda}}\|\bar{y}_{k+1} - y^*\|_{\sigma^{-1}}^2 \ .$$

Denote $\eta = \min\Big(\mu_f\tau, \frac{\sigma\tau\sigma_{\min}(A)^2}{\tau L_f + \tau L'_f + \frac{1}{\lambda}}\Big)$. We then add $\frac{1}{2}(a_2 + 1)$ times (4) and use Lemma 16 to get

$$\frac{2 + a_2}{2}\|z^* - z_k\|^2 - \frac{2 + a_2}{2}\|z^* - z_{k+1}\|^2 - \widetilde{V}(z_k, z^*) + \frac{\eta(\alpha^{-1} - 1)}{2\sigma}\|y_{k+1} - y_k\|^2 \geq \frac{\eta(1 - \alpha)}{2}\|z_{k+1} - z^*\|_V^2 \ .$$

Taking $\alpha = \frac{\eta}{\lambda + \eta}$ chosen such that $\eta(\alpha^{-1} - 1) = \lambda$ and using Lemma 13 allows us to conclude. ◀

Proposition 19. If $\widetilde{\partial} L$ is metrically subregular at z^* for 0 for all $z^* \in \mathcal{Z}^*$ with constant $\eta > 0$ in the norm $\|\cdot\|_V$, then $(I-T)$ is metrically subregular at z^* for 0 for all $z^* \in \mathcal{Z}^*$ with constant bounded below by $\frac{\eta}{\sqrt{3}\eta+(2+2\sqrt{3}\max(\alpha_f,\alpha_g))}$ and PDHG converges linearly with rate $1-\frac{\eta^2\lambda}{(\sqrt{3}\eta+(2+2\sqrt{3}\max(\alpha_f,\alpha_g)))^2}$.

Proof. We denote $D(z)=[\tau x,\sigma y]$, $C(z)=\partial f(x)\times\partial g^*(y)$, $B(z)=[\nabla f_2(x),\nabla g_2^*(y)]$, $M(z)=[A^\top y,-Ax]$ and $H(z)=[\tau^{-1}x,\sigma^{-1}y-Ax]$. This will help us decompose the operator T.

First we remark that

$$\widetilde{\partial} L(z)=(B+C+M)(z)\;.$$

We continue with

$$\begin{aligned}
&T(z)=z^+=DH\bar{z}+(I-DH)z\\
&x-\tau\nabla f_2(x)-\tau A^\top y-\bar{x}\in\tau\partial f(\bar{x})\\
&y-\sigma\nabla g_2^*(y)+\sigma A\bar{x}-\bar{y}\in\sigma\partial g^*(\bar{y})
\end{aligned}$$

so that using the fact that $(H-M)(z)=[\tau^{-1}x-A^\top y,\sigma^{-1}y]$,

$$\bar{z}=(C+H)^{-1}(H-M-B)(z)\;.$$

Thus

$$T(z)=DH(C+H)^{-1}(H-M-B)(z)+(I-DH)z$$

$$(I-T)(z)=DH(I-(C+H)^{-1}(H-M-B))(z)=DH(z-\bar{z})\;.$$

$$\begin{aligned}
&\widetilde{\partial} L(\bar{z})=(B+C+M)(\bar{z})=B(\bar{z})+(C+H)(\bar{z})+(M-H)(\bar{z})\\
&B(\bar{z})+(H-B-M)(z)+(M-H)(\bar{z})\in\widetilde{\partial} L(\bar{z})
\end{aligned}$$

so that

$$(H-B-M)(z-\bar{z})=(H-B-M)(DH)^{-1}(I-T)(z)\in\widetilde{\partial} L(\bar{z})\;.$$

Using the fact that B is Lipschitz-continuous with constant $2\max(\alpha_f,\alpha_g)$ in the norm $\|\cdot\|_V$ and that $\|z\|_V=\|D^{-1/2}z\|$, this leads to

$$\begin{aligned}
\eta\,\mathrm{dist}_V(\bar{z},\mathcal{Z}^*)&\le\|(H-B-M)(z-\bar{z})\|_{V*}\\
&\le\|(H-M)(z-\bar{z})\|_{V*}+\|B(z-\bar{z})\|_{V*}\\
&\le\big(\|(H-M)(DH)^{-1}\|_{V*,V}+2\max(\alpha_f,\alpha_g)\big)\times\|(DH)^{-1}\|_V\|(I-T)(z)\|_V\\
&=\big(\|D^{1/2}(H-M)H^{-1}D^{-1}D^{1/2}\|+2\max(\alpha_f,\alpha_g)\|D^{-1/2}H^{-1}D^{-1}D^{1/2}\|\big)\|(I-T)(z)\|_V\\
&=\big(\|I-D^{1/2}MH^{-1}D^{-1/2}\|+2\max(\alpha_f,\alpha_g)\|D^{-1/2}H^{-1}D^{-1/2}\|\big)\|(I-T)(z)\|_V
\end{aligned}$$

Moreover, $\|D^{-1/2}H^{-1}D^{-1/2}z\|^2\le\|x\|^2+2\sigma\tau\|A\|^2\|x\|^2+2\|y\|^2\le 3\|z\|^2$ and

$$\begin{aligned}
\|I-D^{1/2}MH^{-1}D^{-1/2}z\|^2&=\|x-\sigma\tau A^\top Ax+\sigma^{1/2}\tau^{1/2}A^\top y\|^2+\|-\tau^{1/2}\sigma^{1/2}Ax+y\|^2\\
&\le 2(\|I-\sigma\tau A^\top A\|^2\|x\|^2+\sigma\tau\|A\|^2\|y\|^2)+2(\tau\sigma\|A\|^2\|x\|^2+\|y\|^2)\\
&\le 4\|z\|^2
\end{aligned}$$

Gathering these three inequalities gives

$$\|z-P_{\mathcal{Z}^*}(\bar{z})\|_V=\mathrm{dist}_V(\bar{z},\mathcal{Z}^*)\le\eta^{-1}\big(2+2\max(\alpha_f,\alpha_g)\sqrt{3}\big)\|(I-T)(z)\|_V\;.$$

Finally, we remark that

$$\begin{aligned}
\mathrm{dist}_V(z,\mathcal{Z}^*)=\|z-P_{\mathcal{Z}^*}(z)\|_V&\le\|z-P_{\mathcal{Z}^*}(\bar{z})\|_V\le\|\bar{z}-P_{\mathcal{Z}^*}(\bar{z})\|_V+\|z-\bar{z}\|_V\\
&\le\eta^{-1}\big(2+2\max(\alpha_f,\alpha_g)\sqrt{3}\big)\|(I-T)(z)\|_V+\|(DH)^{-1}\|_V\|(I-T)(z)\|_V\\
&\le(\sqrt{3}+\eta^{-1}(2+2\sqrt{3}\max(\alpha_f,\alpha_g)))\|(I-T)(z)\|_V
\end{aligned}$$

Then, to prove the linear rate of convergence, we recall that for all $z^* \in \mathcal{Z}^*$,

$$\|T(z) - z^*\|_V^2 \le \|z - z^*\|_V^2 - \lambda\|(I-T)(z)\|_V^2 \,.$$

Combined with the metric sub-regularity of $(I-T)$, we get

$$\|T(z) - z^*\|_V^2 \le \|z - z^*\|_V^2 - \frac{\eta^2\lambda}{\big(\sqrt{3}\eta + \big(2 + 2\sqrt{3}\max(\alpha_f, \alpha_g)\big)\big)^2}\operatorname{dist}_V(z, \mathcal{Z}^*)^2 \,.$$

Choosing $z^* = P_{\mathcal{Z}^*}(z)$ leads to

$$\operatorname{dist}_V(T(z), \mathcal{Z}^*)^2 \le \|T(z) - P_{\mathcal{Z}^*}(z)\|_V^2 \le \left(1 - \frac{\eta^2\lambda}{\big(\sqrt{3}\eta + \big(2 + 2\sqrt{3}\max(\alpha_f, \alpha_g)\big)\big)^2}\right)\operatorname{dist}_V(z, \mathcal{Z}^*)^2$$

and thus the linear rate of PDHG follows directly from this contraction property of operator T. ◀

C Proof of Proposition 27

Proposition 27. For any $\beta \ge 0$, $R > 0$ and $z^* \in \mathcal{Z}^*$, the linear program (8) satisfies the quadratic error bound: for all z such that $G_\beta(z; z^*) \le R$, we have

$$G_\beta(z; z^*) \ge \frac{\operatorname{dist}(z, \mathcal{Z}^*)^2}{\theta^2\Big(\sqrt{\frac{2\beta}{\tau}}(\sqrt{2} + \|x_F^*\| + \|x_N^*\|) + \sqrt{\frac{2\beta}{\sigma}}(\sqrt{2} + \|y_E^*\| + \|y_I^*\|) + 3\sqrt{R}\Big)^2} \,.$$

Hence, for R of the order of $\frac{1}{\theta}$, $G_{\frac{1}{\theta}}(\,\cdot\,, z^*)$ has a $\frac{c}{\theta}$-QEB with c independent of θ.

Proof. First of all, we calculate the smoothed gap for (8).

$$\begin{aligned}
G_\beta(z; z^*) &= \sup_{z' \in \mathbb{R}^{n+m}} \langle c, x\rangle + I_{\mathbb{R}^N_+}(x_N) + \langle Ax, y'\rangle - \langle b, y'\rangle - I_{\mathbb{R}^I_+}(y'_I) - \frac{\beta}{2\sigma}\|y' - y^*\|^2 \\
&\qquad - \langle c, x'\rangle - I_{\mathbb{R}^N_+}(x'_N) - \langle Ax', y\rangle + \langle b, y\rangle + I_{\mathbb{R}^I_+}(y_I) - \frac{\beta}{2\tau}\|x' - x^*\|^2 \\
&= \langle c, x\rangle + I_{\mathbb{R}^N_+}(x_N) + \langle A_{E,:}x - b_E, y_E^*\rangle + \frac{\sigma}{2\beta}\|A_{E,:} - b_E\|^2 \\
&\qquad + \frac{\beta}{2\sigma}\|\max\Big(0, y_I^* + \frac{\sigma}{\beta}(A_{I,:}x - b_I)\Big)\|^2 - \frac{\beta}{2\sigma}\|y_I^*\|^2 + \langle b, y\rangle \\
&\qquad + I_{\mathbb{R}^I_+}(y_I) - \langle (A_{:,F})^\top y + c_F, x_F^*\rangle + \frac{\tau}{2\beta}\|(A_{:,F})^\top y + c_F\|^2 \\
&\qquad + \frac{\beta}{2\tau}\|\max\Big(0, x_N^* - \frac{\tau}{\beta}((A_{:,N})^\top y + c_N)\Big)\|^2 - \frac{\tau}{2\sigma}\|x_N^*\|^2
\end{aligned}$$

Let us denote $S_\beta^P(x, y^*) = G_\beta((x, y^*); z^*)$ and $S_\beta^D(y, x^*) = G_\beta((x^*, y); z^*)$ so that $G_\beta(z; z^*) = S_\beta^P(x, y^*) + S_\beta^D(y, x^*)$. We know that $\operatorname{dist}(x, \mathcal{X}^*) \le \theta\big(|c^\top x + b^\top y^*|^2 + \|A_{E,:}x - b_E\|^2 + \operatorname{dist}(A_{I,:}x - b_I, \mathbb{R}^I_-)^2 + \operatorname{dist}(x_N, \mathbb{R}^N_+)^2\big)^{1/2}$ thanks to (10). Our goal is to upper bound this by a function of $S_\beta^P(x, y^*)$.

First, we note that $S_\beta^P(x, y^*) = \langle c, x\rangle + I_{\mathbb{R}^N_+}(x_N) + \langle A_{E,:}x - b_E, y_E^*\rangle + \frac{\sigma}{2\beta}\|A_{E,:}x - b_E\|^2 + \frac{\beta}{2\sigma}\|\max\big(0, y_I^* + \frac{\sigma}{\beta}(A_{I,:}x - b_I)\big)\|^2 - \frac{\beta}{2\sigma}\|y_I^*\|^2 + \langle b, y^*\rangle$ is the sum of many nonnegative terms:

$$\begin{aligned}
&(A_{:,i}^\top y^* + c_i)x_i = 0 && \forall\, i \in F \\
&(A_{:,i}^\top y^* + c_i)x_i \ge 0 && \forall\, i \in N \\
&I_{\mathbb{R}_+}(x_i) \ge 0 && \forall\, i \in N \\
&\frac{\sigma}{2\beta}(A_{j,:}x - b_j)^2 \ge 0 && \forall\, j \in E \\
&\frac{\beta}{2\sigma}\max\big(0, y_j^* + \frac{\sigma}{\beta}(A_{j,:}x - b_j)\big)^2 - \frac{\beta}{2\sigma}(y_j^*)^2 - (A_{j,:}x - b_j)y_j^* \ge 0 && \forall\, j \in I
\end{aligned}$$

Suppose that $S_\beta^P(x, y^*) \le \epsilon$. Then each of these terms is smaller than ϵ. The most complex term is the last one. We shall consider separately 2 sub cases: $I_- = \{j \in I : y_j^* + \frac{\sigma}{\beta}(A_{j,:}x - b_j) \le 0\}$, and $I_+ = \{j \in I : y_j^* + \frac{\sigma}{\beta}(A_{j,:}x - b_j) > 0\}$.

If $j \in I_+$, then

$$\frac{\beta}{2\sigma}\max\big(0, y_j^* + \frac{\sigma}{\beta}(A_{j,:}x - b_j)\big)^2 - \frac{\beta}{2\sigma}(y_j^*)^2 - (A_{j,:}x - b_j)y_j^* = \frac{\sigma}{2\beta}(A_{j,:}x - b_j)^2 \,.$$

Hence, if $S_\beta^P(x, y^*) \leq \epsilon$, then $\sum_{j\in I_+} \max(0, A_{j,:}x - b_j)^2 \leq \sum_{j\in I_+}(A_{j,:}x - b_j)^2 \leq 2\beta\epsilon/\sigma$

If $j \in I_-$, then $-(A_{j,:}x - b_j) \geq \frac{\beta}{\sigma} y_j^*$, so that $(A_{j,:}x - b_j) \leq 0$.

Combining both cases, $\sum_{j\in I} \max(0, A_{j,:}x - b_j)^2 = \sum_{j\in I_+} \max(0, A_{j,:}x - b_j)^2 \leq 2\beta\epsilon/\sigma$.

We now look at $\langle c, x\rangle + \langle b, y^*\rangle = \langle c + A^\top y^*, x\rangle + \langle b - Ax, y^*\rangle$. $S_\beta^P(x, y^*) \leq \epsilon$ implies $0 \leq \langle c + A^\top y^*, x\rangle \leq \epsilon$. Then we need to focus on the complementary slackness $\langle b - Ax, y^*\rangle = \langle b_E - A_{E,:}x, y_E^*\rangle + \langle b_I - A_{I,:}x, y_I^*\rangle$.

Since $S_\beta^P(x, y^*) \leq \epsilon$ implies $\|A_{E,:}x - b_E\|^2 \leq 2\beta\epsilon/\sigma$, we get

$$|\langle b_E - A_{E,:}x, y_E^*\rangle| \leq \|y_E\|\|A_{E,:}x - b_E\| \leq \sqrt{2\beta\epsilon/\sigma}\|y_E\| \,.$$

For I_+, $|\sum_{j\in I_+} y_j^*(b_j - A_{j,:}x)| \leq \|y_{I_+}^*\|\|b_{I_+} - A_{I_+,:}x\| \leq \|y_I^*\|\sqrt{2\beta\epsilon/\sigma}$.

For I_-, since $-\frac{\beta}{2\sigma}(y_j^*)^2 \geq \frac{1}{2}(A_{j,:}x - b_j)y_j^*$,

$$\begin{aligned}\epsilon &\geq \sum_{j\in I_-} \frac{\beta}{2\sigma}\max\big(0, y_j^* + \frac{\sigma}{\beta}(A_{j,:}x - b_j)\big)^2 - \frac{\beta}{2\sigma}(y_j^*)^2 - (A_{j,:}x - b_j)y_j^* \\ &= \sum_{j\in I_-} -\frac{\beta}{2\sigma}(y_j^*)^2 - (A_{j,:}x - b_j)y_j^* \geq \sum_{j\in I_-} -\frac{1}{2}(A_{j,:}x - b_j)y_j^* \geq 0\end{aligned}$$

Combining the three cases, we get

$$\sqrt{2\beta\epsilon/\sigma}(\|y_E^*\| + \|y_I^*\|) \leq \langle c, x\rangle + \langle b, y^*\rangle \leq \sqrt{2\beta\epsilon/\sigma}(\|y_E^*\| + \|y_I^*\|) + 3\epsilon \,.$$

Finally, for x such that $x_N \geq 0$,

$$\begin{aligned}&\big(|c^\top x + b^\top y^*|^2 + \|A_{E,:}x - b_E\|^2 + \mathrm{dist}(A_{I,:}x - b_I, \mathbb{R}_-^I)^2 + \mathrm{dist}(x_N, \mathbb{R}_+^N)^2\big)^{1/2} \\ &\qquad \leq \Bigg(\Bigg(\sqrt{\frac{2\beta\epsilon}{\sigma}}(\|y_E^*\| + \|y_I^*\|) + 3\epsilon\Bigg)^2 + \frac{2\beta\epsilon}{\sigma} + \frac{2\beta\epsilon}{\sigma}\Bigg)^{1/2} \leq \sqrt{\frac{2\beta\epsilon}{\sigma}}(\|y_E^*\| + \|y_I^*\|) + 3\epsilon + 2\sqrt{\frac{\beta\epsilon}{\sigma}}\end{aligned}$$

The argument for the dual problem is exactly the same. Hence

$$\mathrm{dist}(z, \mathcal{Z}^*) \leq \theta\Bigg(\sqrt{\frac{2\beta}{\tau}}(\sqrt{2}+\|x_F^*\|+\|x_N^*\|)\sqrt{G_\beta(z;z^*)}+\sqrt{\frac{2\beta}{\sigma}}(\sqrt{2}+\|y_E^*\|+\|y_I^*\|)\sqrt{G_\beta(z;z^*)}+3G_\beta(z;z^*)\Bigg).$$

If $G_\beta(z;z^*) \leq R$, we get the quadratic error bound

$$G_\beta(z;z^*) \geq \frac{\mathrm{dist}(z, \mathcal{Z}^*)^2}{\theta^2\Big(\sqrt{\frac{2\beta}{\tau}}(\sqrt{2} + \|x_F^*\| + \|x_N^*\|) + \sqrt{\frac{2\beta}{\sigma}}(\sqrt{2} + \|y_E^*\| + \|y_I^*\|) + 3\sqrt{R}\Big)^2} \,.$$ ◀

D Idea to take profit of strong convexity

Proposition 34. *Suppose that $\mu_f > 0$, $g = \iota_{\{b\}}$ and $G_\beta(\,\cdot\,, z^*)$ has a η-QEB where $\frac{1}{\beta_x} \geq \frac{1}{\beta_y} + \sqrt{\eta_x} - \eta_x$. Then, for all $C > 0$,*

$$(1 + \lambda_4)\,\mathrm{dist}_V(z_{k+1} - z^*)^2 + \lambda_1\|z_{k+1} - z_k\|_V^2 \leq \rho\Big((1 + \lambda_4)\,\mathrm{dist}_V(z_k - z^*)^2 + \lambda_1\|z_k - z_{k-1}\|_V^2\Big)$$

where, denoting $\alpha_1 = \frac{2\mu_f\sigma\tau}{2\mu_f\sigma\tau + \Gamma}$:

- *if $2\mu_f\tau(1 - \alpha_1) \leq C\eta_x$, then $\lambda_1 = 0$, $\lambda_4 = \frac{1}{\beta_x\Gamma} - 1$ and*

$$\rho = \max\Bigg(\Big(1 + \frac{C\eta_x\beta_x}{\Gamma}\Big)^{-1}, \Big(1 + \frac{\eta_y\beta_x}{\Gamma}\Big)^{-1}\Bigg) \,;$$

- *if*

$$2\mu_f\tau(1-\alpha_1) > C\eta_x \quad \textit{and} \quad \frac{\frac{1}{\beta_x}-\Gamma}{2\mu_f(1-\alpha_1)-C\eta_x} > \frac{-\frac{1}{\beta_y}+\frac{(1-\sqrt{\eta_x}-C)\eta_x}{2\gamma(1-\sqrt{\eta_x})}-C\sqrt{\eta_x}+\frac{1}{\beta_x}}{2\mu_f(1-\alpha_1)},$$

then we take

$$\lambda_1 = \frac{-\frac{1}{\beta_y}+\frac{(1-\sqrt{\eta_x}-C)\eta_x}{2\gamma(1-\sqrt{\eta_x})}-C\sqrt{\eta_x}+\frac{1}{\beta_x}}{2\mu_f\tau(1-\alpha_1)}, \quad \lambda_4 = \frac{\frac{1}{\beta_x}-\lambda_1(2\mu_f\tau(1-\alpha_1)-C\eta_x)}{\Gamma}-1$$

and we have

$$\rho = \left(1+\frac{\min(C\eta_x,\eta_y)\Gamma}{\frac{1}{\beta_x}-\frac{2\mu_f\tau(1-\alpha_1)-C\eta_x}{2\mu_f\tau(1-\alpha_1)}(-\frac{1}{\beta_y}+\frac{(1-\sqrt{\eta_x}-C)\eta_x}{2\gamma(1-\sqrt{\eta_x})}-C\sqrt{\eta_x}+\frac{1}{\beta_x})}\right)^{-1}$$

- *if*

$$2\mu_f\tau(1-\alpha_1) > C\eta_x \quad \textit{and} \quad \frac{\frac{1}{\beta_x}-\Gamma}{2\mu_f\tau(1-\alpha_1)-C\eta_x} \leq \frac{-\frac{1}{\beta_y}+\frac{(1-\sqrt{\eta_x}-C)\eta_x}{2\gamma(1-\sqrt{\eta_x})}-C\sqrt{\eta_x}+\frac{1}{\beta_x}}{2\mu_f\tau(1-\alpha_1)},$$

then

$$\lambda_4 = 0, \quad \lambda_1 = \frac{\frac{1}{\beta_x}-\Gamma}{2\mu_f\tau(1-\alpha_1)-C\eta_x} \quad \textit{and} \quad \rho = \max\big((1+C\eta_x)^{-1},(1+\eta_y)^{-1}\big)$$

In order to use this proposition, we shall compute ρ for a grid of values of C and select the best one.

Proof. We shall write the proof for $\mu_g > 0$, even though we state the proposition for $\mu_g = +\infty$ only. We apply Lemma 12 to $z = z_k$ and $z' = z_{k-1}$ so that $T(z) = z_{k+1}$ and $T(z') = z_k$. Note that we apply the appendix version of Lemma 12 in order to leverage the most of strong convexity.

$$\|z_{k+1}-z_k\|_V^2 + 2\mu_f\|\bar{x}_{k+1}-\bar{x}_k\|^2 \leq \|z_k-z_{k-1}\|_V^2 - \lambda\|z_k-z_{k+1}-z_{k-1}+z_k\|_V^2 \ .$$

$$\begin{aligned}\|\bar{x}_{k+1}-\bar{x}_k\|^2 &= \|x_{k+1}+\tau A^\top(y_{k+1}-y_k)-x_k-\tau A^\top(y_k-y_{k-1})\|^2 \\ &\geq (1-\alpha_1)\|x_{k+1}-x_k\|^2 - (\alpha_1^{-1}-1)\tau\|A^\top(y_{k+1}-y_k-y_k-y_{k-1})\|^2\end{aligned}$$

We choose α_1 such that $2\mu_f(\alpha_1^{-1}-1)\tau = \frac{\lambda}{\sigma}$, i.e. $\alpha_1 = (1+\frac{\lambda}{2\mu_f\sigma\tau})^{-1} \in O(\mu_f)$, which leads to

$$\|z_{k+1}-z_k\|_V^2 + 2\mu_f(1-\alpha_1)\|x_{k+1}-x_k\|^2 \leq \|z_k-z_{k-1}\|_V^2$$

We also have

$$\begin{aligned}&\frac{\eta_x}{2}\|\bar{x}_{k+1}-x^*\|_{\tau^{-1}}^2 + \frac{\eta_y}{2}\|\bar{y}_{k+1}-y^*\|_{\sigma^{-1}}^2 \leq G_\beta(\bar{z}_{k+1},z^*) \\ &\quad\leq \frac{1}{2}\|z_k-z^*\|_V^2 - \frac{1}{2}\|z_{k+1}-z^*\|_V^2 + \frac{1}{2\beta_x}\|x_{k+1}-x_k\|_{\tau^{-1}}^2 + \frac{1}{2\beta_y}\|y_{k+1}-y_k\|_{\sigma^{-1}}^2 + a_2\widetilde{V}(\bar{z}_{k+1}-z_k)\end{aligned}$$

Moreover, since $0 \in \partial g(y_{k+1}) + \nabla g_2(y_k) + A\bar{x}_{k+1} + \frac{1}{\sigma}(y_{k+1}-y_k)$,

$$\begin{aligned}\|y_{k+1}-y_k\|_{\sigma^{-1}} &\leq \sqrt{\sigma}(\|A\bar{x}_{k+1}-Ax^*\| + \frac{1}{\mu_g}\|y_{k+1}-y^*\| + L_{g_2^*}\|y_k-y^*\|) \\ &\leq \sqrt{\gamma}\|\bar{x}_{k+1}-x^*\|_{\tau^{-1}} + \frac{\sigma}{\mu_g}\|y_{k+1}-y^*\|_{\sigma^{-1}} + \sigma L_{g_2^*}\|y_k-y^*\|_{\sigma^{-1}} \\ \|y_{k+1}-y_k\|_{\sigma^{-1}}^2 &\leq 2\gamma\|\bar{x}_{k+1}-x^*\|_{\tau^{-1}}^2 + 4\frac{\sigma}{\mu_g}\|y_{k+1}-y^*\|_{\sigma^{-1}}^2 + 4\sigma L_{g_2^*}\|y_k-y^*\|_{\sigma^{-1}}^2\end{aligned}$$

We then sum the three inequalities with factors $\lambda_i \geq 0$, $i \in \{1,2,3\}$.

$$\begin{aligned}&\Big(\frac{\lambda_2\eta_x}{2}-\lambda_3\gamma\Big)\|\bar{x}_{k+1}-x^*\|_{\tau^{-1}}^2 + \Big(\frac{\lambda_2\eta_y}{2}-\frac{2\lambda_3\sigma}{\mu_g}\Big)\|\bar{y}_{k+1}-y^*\|_{\sigma^{-1}}^2 + \frac{\lambda_2}{2}\|z_{k+1}-z^*\|_V^2 \\ &\quad + \Big(\frac{\lambda_1}{2}+\lambda_1\mu_f\tau(1-\alpha_1)-\frac{\lambda_2}{2\beta_x}\Big)\|x_{k+1}-x_k\|_{\tau^{-1}}^2 + \Big(\frac{\lambda_1}{2}-\frac{\lambda_2}{2\beta_y}+\frac{\lambda_3}{2}\Big)\|y_{k+1}-y_k\|_{\sigma^{-1}}^2 \\ &\quad - \lambda_2 a_2\widetilde{V}(\bar{z}_{k+1}-z_k) \\ &\leq \frac{\lambda_2}{2}\|z_k-z^*\|_V^2 + \frac{\lambda_1}{2}\|z_k-z_{k-1}\|_V^2 + 2\lambda_3\sigma L_{g_2^*}\|y_k-y^*\|_{\sigma^{-1}}^2\end{aligned}$$

We combine with

$$\begin{aligned}\|\bar{x}_{k+1}-x^*\|^2_{\tau^{-1}} &\geq (1-\alpha_2)\|x_{k+1}-x^*\|^2_{\tau^{-1}} - (\alpha_2^{-1}-1)\|\bar{x}_{k+1}-x_{k+1}\|^2_{\tau^{-1}} \\ &\geq (1-\alpha_2)\|x_{k+1}-x^*\|^2_{\tau^{-1}} - (\alpha_2^{-1}-1)\|y_{k+1}-y_k\|^2_{\sigma^{-1}}\end{aligned}$$

and

$$\frac{1}{2}\|z_{k+1}-z^*\|^2_V \leq \frac{1}{2}\|z_k-z^*\|^2_V - \widetilde{V}(\bar{z}_{k+1}-z_k)$$

to get

$$\begin{aligned}&\left(\left(\frac{\lambda_2\eta_x}{2}-\lambda_3\gamma\right)(1-\alpha_2)+\frac{\lambda_2}{2}+\frac{\lambda_4}{2}\right)\|x_{k+1}-x^*\|^2_{\tau^{-1}} + \left(\frac{\lambda_2\eta_y}{2}-\frac{2\lambda_3\sigma}{\mu_g}+\frac{\lambda_2}{2}+\frac{\lambda_4}{2}\right)\|y_{k+1}-y^*\|^2_{\sigma^{-1}} \\ &\quad + \left(\frac{\lambda_1}{2}+\lambda_1\mu_f\tau(1-\alpha_1)-\frac{\lambda_2}{2\beta_x}+(\lambda_4-\lambda_2 a_2)\frac{\lambda}{2}\right)\|x_{k+1}-x_k\|^2_{\tau^{-1}} \\ &\quad + \left(\frac{\lambda_1}{2}-\frac{\lambda_2}{2\beta_y}+\frac{\lambda_3}{2}-\left(\frac{\lambda_2\eta_x}{2}-\lambda_3\sqrt{\gamma}\right)(\alpha_2^{-1}-1)+(\lambda_4-\lambda_2 a_2)\frac{\lambda}{2}\right)\|y_{k+1}-y_k\|^2_{\sigma^{-1}} \\ &\leq \frac{\lambda_2+\lambda_4}{2}\|z_k-z^*\|^2_V + \frac{\lambda_1}{2}\|z_k-z_{k-1}\|^2_V + 2\lambda_3\sigma L_{g_2^*}\|y_k-y^*\|^2_{\sigma^{-1}}\end{aligned}$$

To get the rate, we then need

$$\begin{aligned}&\rho\big((\lambda_2\eta_x-2\lambda_3\gamma)(1-\alpha_2)+\lambda_2+\lambda_4\big) \geq \lambda_2+\lambda_4 \\ &\rho\left(\lambda_2\eta_y-\frac{4\lambda_3\sigma}{\mu_g}+\lambda_2+\lambda_4\right) \geq \lambda_2+\lambda_4+4\lambda_3\sigma L_{g_2^*} \\ &\rho\left(\lambda_1+2\lambda_1\mu_f\tau(1-\alpha_1)-\frac{\lambda_2}{\beta_x}+(\lambda_4-\lambda_2 a_2)\lambda)\right) \geq \lambda_1 \\ &\rho\left(\lambda_1-\frac{\lambda_2}{\beta_y}+\lambda_3-\big(\lambda_2\eta_x-2\lambda_3\gamma\big)(\alpha_2^{-1}-1)+(\lambda_4-\lambda_2 a_2)\lambda\right) \geq \lambda_1\end{aligned}$$

We choose $\alpha_2=\sqrt{\eta_x}$, $\lambda_3=\frac{(1-\alpha_2-C)\eta_x}{2\gamma(1-\alpha_2)}$ and $\lambda_2=1$. We shall let the choice of $C\in[0,1-\alpha_2]$ for a 1D grid search since the rate will depend a lot on its value. This yields $\big(\lambda_2\eta_x-2\lambda_3\gamma\big)(1-\alpha_2)=C\eta_x$.

We assume that $\frac{1}{\beta_x}\geq\frac{1}{\beta_y}+\eta_x(\alpha_2^{-1}-1)$.

Case 1. If $2\mu_f\tau(1-\alpha_1)\leq C\eta_x$, we choose $\lambda_1=0$ and $\lambda_4=\frac{1}{\beta_x\lambda}+a_2$. This leads to

$$\begin{aligned}&\rho\big(1+\lambda_4+C\eta_x\big)\geq 1+\lambda_4 \\ &\rho\left(1+\lambda_4+\eta_y-\frac{4\lambda_3\sigma}{\mu_g}\right)\geq 1+\lambda_4+4\lambda_3\sigma L_{g_2^*} \\ &-\frac{1}{\beta_x}+(\lambda_4-a_2)\lambda=0\geq 0 \\ &-\frac{1}{\beta_y}+\frac{(1-\alpha_2-C)\eta_x}{2\gamma(1-\alpha_2)}-\frac{C\eta_x}{1-\alpha_2}(\alpha_2^{-1}-1)+\frac{1}{\beta_x} \\ &\qquad\geq \frac{(1-\alpha_2-C)\eta_x}{2\gamma(1-\alpha_2)}-C\eta_x\alpha_2^{-1}+\eta_x(\alpha_2^{-1}-1)\geq\eta_x(\alpha_2^{-1}-1)-(1-\alpha_2)\alpha_2^{-1}\eta_x=0\end{aligned}$$

where the last inequality uses $C\leq 1-\alpha_2$. Supposing that $\mu_g=+\infty$ and $L_{g_2^*}=0$, we get a rate

$$\rho=\max\left(\left(1+\frac{C\eta_x}{1+a_2+1/(\lambda\beta_x)}\right)^{-1},\left(1+\frac{\eta_y}{1+a_2+1/(\lambda\beta_x)}\right)^{-1}\right).$$

Case 2. If

$$2\mu_f\tau(1-\alpha_1)>C\eta_x \quad\text{and}\quad \frac{\frac{1}{\beta_x}+a_2\lambda}{2\mu_f\tau(1-\alpha_1)-C\eta_x}>\frac{-\frac{1}{\beta_y}+\frac{(1-\alpha_2-C)\eta_x}{2\gamma(1-\alpha_2)}-C\eta_x\alpha_2^{-1}+\frac{1}{\beta_x}}{2\mu_f\tau(1-\alpha_1)},$$

we choose

$$\lambda_1=\frac{-\frac{1}{\beta_y}+\lambda_3-C\eta_x\alpha_2^{-1}+\frac{1}{\beta_x}}{2\mu_f\tau(1-\alpha_1)} \quad\text{and}\quad \lambda_4=\frac{\frac{1}{\beta_x}-\lambda_1(2\mu_f\tau(1-\alpha_1)-C\eta_x)}{\lambda}+a_2.$$

We get $2\lambda_1\mu_f\tau(1-\alpha_1) - \frac{\lambda_2}{\beta_x} + (\lambda_4 - \lambda_2 a_2)\lambda = 2\lambda_1\mu_f\tau(1-\alpha_1) - \frac{1}{\beta_x} + \frac{1}{\beta_x} - 2\lambda_1\mu_f\tau(1-\alpha_1) + \lambda_1 C\eta_x = \lambda_1 C\eta_x$ and $-\frac{\lambda_2}{\beta_y} + \lambda_3 - \big(\lambda_2\eta_x - 2\lambda_3\gamma\big)(\alpha_2^{-1} - 1) + (\lambda_4 - \lambda_2 a_2)\lambda = -\frac{1}{\beta_y} + \lambda_3 - C\eta_x\alpha_2^{-1} + \frac{1}{\beta_x} - \lambda_1 2\mu_f\tau(1-\alpha_1) + \lambda_1 C\eta_x = \lambda_1 C\eta_x$. Hence,

$$\begin{aligned}
&\rho\Big(1 + \lambda_4 + C\eta_x\Big) \geq 1 + \lambda_4 \\
&\rho\Big(1 + \lambda_4 + \eta_y - \frac{4\lambda_3\sigma}{\mu_g}\Big) \geq 1 + \lambda_4 + 4\lambda_3\sigma L_{g_2^*} \\
&\rho\Big(\lambda_1 + C\eta_x\lambda_1\Big) \geq \lambda_1 \\
&\rho\Big(\lambda_1 + C\eta_x\lambda_1\Big) \geq \lambda_1
\end{aligned}$$

Supposing that $\mu_g = +\infty$ and $L_{g_2^*} = 0$, we get a rate

$$\begin{aligned}
\rho &= \max\bigg(\Big(1 + \frac{C\eta_x}{1+\lambda_4}\Big)^{-1}, \Big(1 + \frac{\eta_y}{1+\lambda_4}\Big)^{-1}\bigg) \\
&= \left(1 + \frac{\min(C\eta_x, \eta_y)\lambda}{\frac{1}{\beta_x} - \frac{2\mu_f\tau(1-\alpha_1) - C\eta_x}{2\mu_f\tau(1-\alpha_1)}\big(-\frac{1}{\beta_y} + \frac{(1-\alpha_2-C)\eta_x}{2\gamma(1-\alpha_2)} - C\eta_x\alpha_2^{-1} + \frac{1}{\beta_x}\big) + a_2\lambda}\right)^{-1}.
\end{aligned}$$

Case 3. If

$$2\mu_f\tau(1-\alpha_1) > C\eta_x \quad \text{and} \quad \frac{\frac{1}{\beta_x} + a_2\lambda}{2\mu_f\tau(1-\alpha_1) - C\eta_x} \leq \frac{-\frac{1}{\beta_y} + \frac{(1-\alpha_2-C)\eta_x}{2\gamma(1-\alpha_2)} - C\eta_x\alpha_2^{-1} + \frac{1}{\beta_x}}{2\mu_f\tau(1-\alpha_1)},$$

we choose $\lambda_4 = 0$ and $\lambda_1 = \frac{\frac{1}{\beta_x} + a_2\lambda}{2\mu_f\tau(1-\alpha_1) - C\eta_x}$. We get

$$\begin{aligned}
-\frac{1}{\beta_y} + \frac{(1-\alpha_2-C)\eta_x}{2\gamma(1-\alpha_2)} - C\eta_x\alpha_2^{-1} - a_2\lambda &\geq -\frac{1}{\beta_x} - a_2\lambda + 2\mu_f\tau(1-\alpha_1)\frac{\frac{1}{\beta_x} + a_2\lambda}{2\mu_f\tau(1-\alpha_1) - C\eta_x} \\
&= \lambda_1(-2\mu_f\tau(1-\alpha_1) + C\eta_x + 2\mu_f\tau(1-\alpha_1)) = C\eta_x\lambda_1.
\end{aligned}$$

Hence,

$$\begin{aligned}
&\rho\Big(1 + C\eta_x\Big) \geq 1 \\
&\rho\Big(1 + \eta_y - \frac{4\lambda_3\sigma}{\mu_g}\Big) \geq 1 + 4\lambda_3\sigma L_{g_2^*} \\
&\rho\Big(\lambda_1 + C\eta_x\lambda_1\Big) \geq \lambda_1 \\
&\rho\Big(\lambda_1 - \frac{1}{\beta_y} + \frac{(1-\alpha_2-C)\eta_x}{2\gamma(1-\alpha_2)} - C\eta_x\alpha_2^{-1} - a_2\lambda\Big) \geq \rho\Big(\lambda_1 + C\eta_x\lambda_1\Big) \geq \lambda_1
\end{aligned}$$

Supposing that $\mu_g = +\infty$ and $L_{g_2^*} = 0$, we get a rate $\rho = \max((1 + C\eta_x)^{-1}, (1 + \eta_y)^{-1})$. We finally combine the results and use the fact that $\alpha_2 = \sqrt{\eta_x}$. ◀

References

1 Ahmet Alacaoglu, Olivier Fercoq, and Volkan Cevher. On the convergence of stochastic primal-dual hybrid gradient. `https://arxiv.org/abs/1911.00799`, 2019.

2 Sulaiman A. Alghunaim and Ali H. Sayed. Linear convergence of primal–dual gradient methods and their performance in distributed optimization. *Automatica*, 117, 2020.

3 David Applegate, Oliver Hinder, Haihao Lu, and Miles Lubin. Faster First-Order Primal-Dual Methods for Linear Programming using Restarts and Sharpness. `https://arxiv.org/abs/2105.12715`, 2021.

4 Heinz H. Bauschke, Patrick L. Combettes, et al. *Convex analysis and monotone operator theory in Hilbert spaces*, volume 408. Springer, 2011.

5 Jérôme Bolte, Trong Phong Nguyen, Juan Peypouquet, and Bruce W. Suter. From error bounds to the complexity of first-order descent methods for convex functions. *Math. Program.*, 165(2):471–507, 2017.

6 Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vis.*, 40(1):120–145, 2011.

7 Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3):1–27, 2011.

8 Laurent Condat. A primal–dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms. *J. Optim. Theory Appl.*, 158(2):460–479, 2013.

9 Laurent Condat, Daichi Kitahara, Andrés Contreras, and Akira Hirabayashi. Proximal splitting algorithms: A tour of recent advances, with new twists. `https://arxiv.org/abs/1912.00137`, 2019.

10 Damek Davis and Wotao Yin. Convergence rate analysis of several splitting schemes. In *Splitting methods in communication, imaging, science, and engineering*, Scientific Computation, pages 115–163. Springer, 2016.

11 Asen L. Dontchev and R. Tyrrell Rockafellar. *Implicit functions and solution mappings. A view from variational analysis.* Springer Monographs in Mathematics. Springer, 2009.

12 Dmitriy Drusvyatskiy and Adrian S. Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *Math. Oper. Res.*, 43(3):919–948, 2018.

13 Simon S. Du and Wei Hu. Linear convergence of the primal-dual gradient method for convex-concave saddle point problems without strong convexity. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 196–205. PMLR, 2019.

14 Olivier Fercoq and Pascal Bianchi. A coordinate-descent primal-dual algorithm with large step size and possibly nonseparable functions. *SIAM J. Optim.*, 29(1):100–134, 2019.

15 Olivier Fercoq and Zheng Qu. Adaptive restart of accelerated gradient methods under local quadratic growth condition. *IMA J. Numer. Anal.*, 39(4):2069–2095, 2019.

16 Olivier Fercoq and Zheng Qu. Restarting the accelerated coordinate descent method with a rough strong convexity estimate. *Comput. Optim. Appl.*, 75(1):63–91, 2020.

17 Daniel Gabay and Bertrand Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Comput. Math. Appl.*, 2(1):17–40, 1976.

18 Alan J. Hoffman. On Approximate Solutions of Systems of Linear Inequalities. *Journal of Research of the National Bureau of Standards*, 49(4):263, 1952.

19 Fan Jiang, Zhongming Wu, Xingju Cai, and Hongchao Zhang. Unified linear convergence of first-order primal-dual algorithms for saddle point problems. *Optim. Lett.*, 16(6):1675–1700, 2022.

20 Dmitry Kovalev, Adil Salim, and Peter Richtárik. Optimal and practical algorithms for smooth and strongly convex decentralized optimization. *Adv. Neural Inf. Process. Syst.*, 33, 2020.

21 Puya Latafat, Nikolaos M. Freris, and Panagiotis Patrinos. A new randomized block-coordinate primal-dual proximal algorithm for distributed optimization. *IEEE Trans. Autom. Control*, 64(10):4050–4065, 2019.

22 Jingwei Liang, Jalal Fadili, and Gabriel Peyré. Convergence rates with inexact non-expansive operators. *Math. Program.*, 159(1-2):403–434, 2016.

23 Qihang Lin, Runchao Ma, Selvaprabu Nadarajah, and Negar Soheili. First-Order Methods for Convex Constrained Optimization under Error Bound Conditions with Unknown Growth Parameters. `https://arxiv.org/abs/2010.15267`, 2020.

24 Meng Lu and Zheng Qu. An adaptive proximal point algorithm framework and application to large-scale optimization. `https://arxiv.org/abs/2008.08784`, 2020.

25 Yurii Nesterov. Smooth minimization of non-smooth functions. *Math. Program.*, 103(1):127–152, 2005.

26 Yurii Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

27 Yuyuan Ouyang and Yangyang Xu. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. *Math. Program.*, 185(1-2):1–35, 2021.

28 R. Tyrrell Rockafellar and Roger J.-B. Wets. *Variational analysis*, volume 317. Springer, 2009.

29 Adil Salim, Laurent Condat, Konstantin Mishchenko, and Peter Richtárik. Dualize, split, randomize: Toward fast nonsmooth optimization algorithms. *J. Optim. Theory Appl.*, 195(1):102–130, 2022.

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

31 Bang Công Vũ. A splitting algorithm for dual monotone inclusions involving cocoercive operators. *Adv. Comput. Math.*, 38(3):667–681, 2013.

32 Daoli Zhu and Lei Zhao. Linear Convergence of Randomized Primal-Dual Coordinate Method for Large-scale Linear Constrained Convex Programming. In *International Conference on Machine Learning*, pages 11619–11628. PMLR, 2020.