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

Olivier Fercoq

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

Olivier Fercoq†

May 18, 2021

**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.

## 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. 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.

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) [5] or Alternating Directions Method of Multipliers (ADMM) [14] can be very slow, with a rate of convergence in the worst case in $O(1/\sqrt{k})$ [7]. 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 work was supported by the Agence National de la Recherche grant ANR-20-CE40-0027, Optimal Primal-Dual Algorithms (APDO).

†LTCI, Télécom Paris, Institut Polytechnique de Paris, France

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 [4].

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 [23], 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 [24]: the smoothed gap, based on the smoothing of nonsmooth functions [21], takes finite values 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 [18]. It holds among others for all linear-quadratic programs [17] and implies a linear convergence rate for PDHG and ADMM, as well as the proximal point algorithm [20]. One can also show linear convergence if the objective is smooth and strongly convex and the constraints are affine [10, 2, 16]. 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 [19]. These result can also be extended to the coordinate descent [25, 1]. Metric subregularity holds for a wide range of problems that includes all piecewise linear-quadratic functions. 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 the 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 [24] 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 in Section 8 and show that is some situations, it leads to a faster algorithm. A heuristic 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.

## 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, $\mathcal{Y}$ the dual space and $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ 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 $\|z\|_V = (\frac{1}{\tau}\|x\|^2 + \frac{1}{\sigma}\|y\|^2)^{1/2}$. The proximal operator of a function $f$ is given by $\text{prox}_f(x) = \arg\min_{x'} f(x') + \frac{1}{2}\|x - x'\|^2$. 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 *$\mu$-strongly convex* if $f - \frac{\mu}{2}\|\cdot\|^2$ is convex.

**Assumption 1.** The Lagrangian function is $\mu$-strongly convex-concave, that is $(x \mapsto L(x, y))$ is $\mu$-strongly convex for all $y$ and $(y \mapsto L(x, y))$ is $\mu$-strongly concave for all $x$.

This regularity assumption is used for instance in [5]. We can generalize strong convexity as follows.

**Definition 2.** We say that a function $f : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ has a quadratic error bound if there exists $\eta$ 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 $\eta$-QEB.

Although this is more general than strong convexity, the quadratic error bound is not 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 3.** 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 $C(z) = [\partial f(x), \partial g^*(y)]$, $B(z) = [\nabla f_2(x), \nabla g_2^*(y)]$ and $M(z) = [A^\top y, -Ax]$. The Lagrangian's subgradient is then $\tilde{\partial} L(z) = (B + C + M)(z)$. We put a tilde to emphasize the fact that the dual component is the negative of the supergradient.

We have $0 \in \tilde{\partial} L(z^*)$ if and only if $z^*$ is a saddle point of $L$. If $\tilde{\partial} L$ 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 2.** The Lagrangian's generalized gradient is metrically subregular, that is there exists $\eta$ such that for all $z^* \in \mathcal{Z}^* = (\tilde{\partial} L)^{-1}(0)$, $\tilde{\partial} L$ is $\eta$-metrically subregular at $z^*$ for 0.

| 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 |

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.

This regularity assumption is used for instance in [18]. Another regularity assumption considered in the literature is as follows.

**Assumption 3.** 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 [10]. The indicator functions encode the constraint $Ax = b$.

**Assumption 4.** 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 [19] 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 1** ([9])**.** *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 2** ([9])**.** *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 3** ([17])**.** *If $L$ is $\mu$-strongly convex-concave, then $\tilde{\partial} L$ is $\mu$-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 [17].

# 3 Basic inequalities for the study of PDHG

Primal-Dual Hybrid Gradient is the algorithm defined by Algorithm 1. We shall use the definition of [17]

**Algorithm 1** Primal-Dual Hybrid Gradient (PDHG)

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

because we believe it simplifies the analysis. Note that the algorithm of [5] 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 $x_k = \bar{x}_k - \tau A^\top (y_k - y_{k-1}) = \bar{x}_k - \tau A^\top (\bar{y}_k - \bar{y}_{k-1})$:

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

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.

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 $\|z\|_V = (\frac{1}{\tau}\|x\|^2 + \frac{1}{\sigma}\|y\|^2)^{1/2}$. We will first show that this fixed point operator $T$ is an averaged operator [3] in this norm. Then, we will give an upper bound on the Lagrangian's gap and a convergence result. All the results are already known so we defer the proofs to the appendix.

**Lemma 1** ([3]). *Let $p = \mathrm{prox}_{\tau f}(x)$ and $p' = \mathrm{prox}_{\tau f}(x')$ where $f$ is $\mu_f$-strongly convex. For all $x$ and $x'$,*

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

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

**Lemma 2** ([17]). *Let $T : \mathcal{X} \times \mathcal{Y} \to \mathcal{X} \times \mathcal{Y}$ be defined for any $(x, y)$ by*

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

*If the step sizes satisfy $\gamma = \sigma\tau\|A\|^2 < 1$, $\tau L_f/2 \leq \alpha_f < 1$, $\alpha_g = \sigma L_{g^*}/2 \leq 1$ and $\sigma L_{g^*}/2 \leq \alpha_f(1 - \sigma\tau\|A\|^2)$ then $T$ is nonexpansive in the norm $\|\cdot\|_V$, and $T$ is $\frac{1}{1+\lambda}$-averaged where*

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

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

$$\|T(z) - T(z')\|_V^2 + 2\mu_f\|\bar{x} - \bar{x}'\|^2 + 2\mu_{g^*}\|\bar{y} - \bar{y}'\|^2 \leq \|z - z'\|_V^2 - \lambda\|z - T(z) - z' + T(z')\|^2\,.$$

*As a consequence, $(z_k)$ converges to a saddle point of the Lagrangian.*

**Lemma 3** ([5]). *For all $k \in \mathbb{N}$ and for all $z \in \mathcal{Z}$,*

$$L(\bar{x}_{k+1}, y) - L(x, \bar{y}_{k+1}) \leq \frac{1}{2}\|z - z_k\|_V^2 - \frac{1}{2}\|z - z_{k+1}\|_V^2 - \tilde{V}(\bar{z}_{k+1} - z_k) \tag{4}$$

*where* $\tilde{V}(\bar{z}_{k+1} - z_k) = (\frac{1}{2\tau} - \frac{L_f}{2})\|\bar{x}_{k+1} - x_k\|^2 + (\frac{1}{2\sigma} - \frac{\tau\|A\|^2}{2} - \frac{L_{g^*}}{2})\|\bar{y}_{k+1} - y_k\|^2$

**Lemma 4** ([5]). *$\tilde{V}$ satisfies*

$$\tilde{V}(\bar{z}_{k+1} - z_k) = (\frac{1}{2\tau} - \frac{L_f}{2})\|\bar{x}_{k+1} - x_k\|^2 + (\frac{1}{2\sigma} - \frac{\tau\|A\|^2}{2} - \frac{L_{g^*}}{2})\|\bar{y}_{k+1} - y_k\|^2 \geq \frac{(1-\alpha_f)(1-\sqrt{\gamma})}{2}\|z_{k+1} - z_k\|_V^2 .$$

*We shall denote $\Gamma = (1-\alpha_f)(1-\sqrt{\gamma})$ so that $\tilde{V}(\bar{z}_{k+1} - z_k) \geq \frac{\Gamma}{2}\|z_{k+1} - z_k\|_V^2$.*

**Proposition 4** ([5]). *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 $\tilde{z}_k = \frac{1}{k}\sum_{l=1}^k \bar{z}_l$ 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*

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

# 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.

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

**Lemma 5.** *For $0 < \alpha \leq 1$,*

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

*Proof.*

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

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 slight modification of [11, Theorem 4].

**Proposition 5.** *If $L$ is $\mu$-strongly convex concave in the norm $\|\cdot\|_V$, then the iterates of PDHG satisfy for all $k$,*

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

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

*Proof.* From Lemma 3 applied at $z=z^*$, we have

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

Since $L$ is $\mu$-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 4 and Lemma 5 to get for all $\alpha\in(0,1)$

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

We just need to choose $\alpha=\frac{\mu}{\Gamma+\mu}$ so that $\mu(\alpha^{-1}-1)=\Gamma$ to conclude. □

We next study the second case where some primal-dual methods have been proved to have a linear rate of convergence [10, 2, 16].

**Proposition 6.** *If $f+f_2$ has a $L_f'+L_f$-Lipschitz gradient and is $\mu_f$-strongly convex, and $g+g_2=\iota_{\{b\}}$, then PDGH converges linearly with rate*

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

*where $\eta=\min(\mu_f\tau,\frac{\sigma\tau\sigma_{\min}(A)^2}{\tau L_f+\tau L_f'+\frac{1}{\Gamma}})$, $\Gamma=(1-\alpha_f)(1-\sqrt{\gamma})$.*

*Proof.* We know by Lemmas 3 and 4 that for all $z=(x,y)$,

$$\begin{aligned}L(\bar{x}_{k+1},y)-L(x,\bar{y}_{k+1}) &\leq \frac{1}{2}\|z-z_k\|_V^2-\frac{1}{2}\|z-z_{k+1}\|_2^2-\tilde{V}(\bar{z}_{k+1}-z_k)\\ &\leq \frac{1}{2}\|z-z_k\|_V^2-\frac{1}{2}\|z-z_{k+1}\|_2^2-\frac{\Gamma}{2}\|z_{k+1}-z_k\|_V^2\end{aligned}$$

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 $Ax^*=b$.

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

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

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

Combining the three inequalities, we obtain

$$\begin{aligned}
&\frac{1}{2}\|z^* - z_k\|^2 - \frac{1}{2}\|z^* - z_{k+1}\|^2 - \frac{\Gamma}{2}\|y_{k+1} - y_k\|^2_{\sigma^{-1}} \\
&\qquad\qquad \geq \frac{\mu_f}{2}\|\bar{x}_{k+1} - x^*\|^2 + \Big(a - a^2\frac{L_f + L'_f}{2} - a^2\frac{1}{2\tau\Gamma}\Big)\|A^\top(\bar{y}_{k+1} - y^*)\|^2
\end{aligned}$$

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

$$\begin{aligned}
&\frac{1}{2}\|z^* - z_k\|^2 - \frac{1}{2}\|z^* - z_{k+1}\|^2 - \frac{\Gamma}{2}\|y_{k+1} - y_k\|^2_{\sigma^{-1}} \\
&\qquad\qquad \geq \frac{\mu_f\tau}{2}\|\bar{x}_{k+1} - x^*\|^2_{\tau^{-1}} + \frac{\sigma\tau\sigma_{\min}(A)^2/2}{\tau L_f + \tau L'_f + \frac{1}{\Gamma}}\|\bar{y}_{k+1} - y^*\|^2_{\sigma^{-1}} .
\end{aligned}$$

Denote $\eta = \min(\mu_f\tau, \frac{\sigma\tau\sigma_{\min}(A)^2}{\tau L_f + \tau L'_f + \frac{1}{\Gamma}})$. Lemma 5 with $\alpha = \frac{\eta}{\Gamma+\eta}$ chosen such that $\eta(\alpha^{-1} - 1) = \Gamma$ allows us to conclude. □

Finally, we will show that if the Lagrangian's generalized gradient is metrically sub-regular then PDHG converges linearly. Compared to [18] and [17], 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.

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

**Proposition 7.** *If $\tilde{\partial}L$ is metrically subregular at $z^*$ for 0 for all $z^* \in \mathcal{Z}^*$ with constant $\eta > 0$, then $(I-T)$ is metrically subregular at $z^*$ for 0 for all $z^* \in \mathcal{Z}^*$ with constant $\frac{\eta}{\sqrt{3}\eta + (2+2\sqrt{3}\max(\alpha_f,\alpha_g))}$ and PDHG converges linearly with rate $\left(1 - \frac{\eta^2(1-\alpha_f)(1-\sqrt{\gamma})}{\left(\sqrt{3}\eta + \left(2+2\sqrt{3}\max(\alpha_f,\alpha_g)\right)\right)^2}\right)$.*

*Proof.* First we remark that

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

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

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

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

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

so that

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

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

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

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

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

Gathering these three inequalities gives

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

Finally, we remark that

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

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

$$\|T(z) - z^*\|_V^2 \leq \|z - z^*\|_V^2 - (1-\alpha_f)(1-\sqrt{\gamma})\|(I-T)(z)\|_V^2$$

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

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

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

$$\begin{aligned}\operatorname{dist}_V(T(z), \mathcal{Z}^*)^2 &\leq \|T(z) - P_{\mathcal{Z}^*}(z)\|_V^2 \\ &\leq \left(1 - \frac{\eta^2(1-\alpha_f)(1-\sqrt{\gamma})}{\left(\sqrt{3}\eta + \left(2 + 2\sqrt{3}\max(\alpha_f, \alpha_g)\right)\right)^2}\right) \operatorname{dist}_V(z, \mathcal{Z}^*)^2\end{aligned}$$

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

# 5 Coarseness of the analysis

## 5.1 Strongly convex-concave Lagrangian

Suppose that $f$ is $\mu_f$ strongly convex and that $g^*$ is $\mu_{g^*}$ strongly convex. Then $L$ is $\mu_L$ 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 $\tau$ and $\sigma$ 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 $\sigma = \sqrt{\frac{\mu_{g^*}}{\mu_f}}\frac{1}{\|A\|}$. 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 [22], which is noticeable.

We have seen in Proposition 3 that metric sub-regularity of the Lagrangian's gradient is a more general assumption than being strongly convex-concave. However, applying Proposition 7 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

$$\begin{aligned}\min_{x\in\mathbb{R}} \ & \frac{\mu}{2}x^2 \\ & ax = b\end{aligned}$$

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 $\nabla L$ is affine, it is easy to see that $\nabla L$ is globally metrically sub-regular with constant $\frac{\sqrt{\mu^2\tau^2 + 4\sigma\tau a^2} - \mu\tau}{2}$ in the norm $\|\cdot\|_V$.

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

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

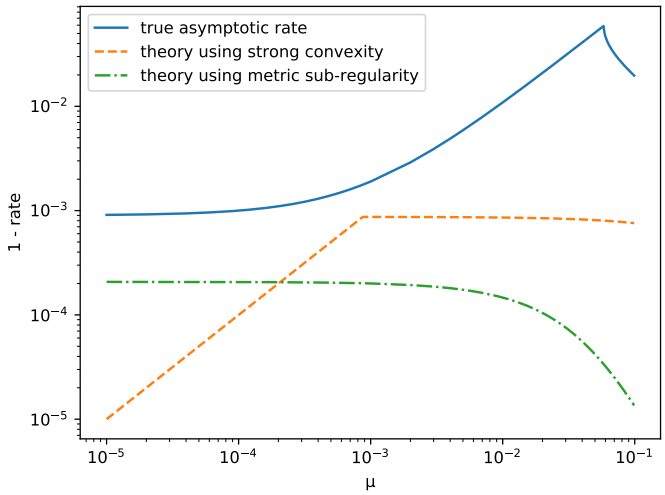


Figure 1: Comparison of the true rate (line above) and what is predicted by theory (2 lines below) for $a = 0.03$, $\tau = \sigma = 1$ and various values for $\mu$.

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 7, that is $\left(1 - \frac{\eta^2\Gamma}{\left(\sqrt{3}\eta + \left(2+2\sqrt{3}\max(\alpha_f,\alpha_g)\right)\right)^2}\right)$ where $\Gamma = (1-\alpha_f)(1-\sqrt{\gamma})$, $\gamma = \sigma\tau a^2$, $\alpha_g = 0$, $\alpha_f = \mu/2$ and $\eta = \frac{\sqrt{\mu^2\tau^2+4\sigma\tau a^2}-\mu\tau}{2}$ and what is predicted by Proposition 6, that is $(1 + \frac{\mu}{1+\mu/\Gamma})^{-1}$. On Figure 2, 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 $\mu$ 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 $\mu$ 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.

## 6.1 Main assumption

**Definition 4.** Given $\beta = (\beta_x, \beta_y) \in [0, +\infty]^2$, $z \in \mathcal{Z}$ and $\dot{z} \in \mathcal{Z}$, the smoothed gap $G_\beta$ 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 8.** *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

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

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

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

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 $x_\beta(y) = x^*$.

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

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

**Assumption 5.** 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 $\eta$ 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 [13, Prop. 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 9.** *If $G_\beta(\cdot, z^*)$ has a $\eta$-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 $\{z : \operatorname{dist}(z, \mathcal{Z}^*)_V < Ma\}$*

We now give a few examples to show that this assumption is often satisfied.

**Proposition 10.** *If $L$ is $\mu$-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 11.** *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 $G_\beta(z;z^*) \geq \frac{\eta(\beta)}{2}\operatorname{dist}_V(z,\mathcal{Z}^*)^2$ for all $z \in \mathcal{R}(\beta)$ and $z^* \in \mathcal{Z}^*$.*

*Proof.* The proof follows the lines of [17]. The class of piecewise linear-quadratic functions is closed under scalar multiplication, addition, conjugation and Moreau envelope [23]. 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 $\eta$ at all $z^* \in \mathcal{Z}^*$ for 0 [8]. Since $G_\beta(\cdot,z^*)$ is a convex function, this implies the result by Proposition 2. □

## 6.2 Linear programs

In the rest of the section, we are going to show that linear programs do satisfy Assumption 5 and give the constant as a function of a Hoffman constant [15].

We consider the linear optimization problem

$$\begin{aligned} &\min_{x\in\mathbb{R}^n} c^\top x \\ &A_{E,:}x = b_E \\ &A_{I,:}x \leq b_I \\ &x_N \geq 0 \end{aligned} \tag{5}$$

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

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

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

$$\begin{cases} c^\top x + b^\top y = 0 \\ A_{E,:}x = b_E \\ A_{I,:}x \leq b_I \\ x_N \geq 0 \\ (A_{:,F})^\top y + c_F = 0 \\ (A_{:,N})^\top y + c_N \geq 0 \\ y_I \geq 0 \end{cases} \tag{6}$$

Let us denote the Hoffman constant [15] of this system by $\theta$. This constant verifies

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

It is known that the Lagrangian's subgradient of an LP satisfies metric sub-regularity with a constant proportional to $\theta$ [20] . We shall show that the same holds for the QEB of the smoothed gap centered at $z^*$.

**Proposition 12.** *For any $\beta \geq 0$, $R > 0$ and $z^* \in \mathcal{Z}^*$, the linear program (5) satisfies the quadratic error bound: for all $z$ such that $G_\beta(z; z^*) \leq R$, we have*

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

*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 $\theta$.*

*Proof.* First of all, we calculate the smoothed gap for (5).

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

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

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

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

Suppose that $S^P_\beta(x, y^*) \leq \epsilon$. Then each of these terms is smaller than $\epsilon$. The most complex term is the last one. We shall consider separately 2 sub cases: $I_- = \{j \in I : y_j^* + \frac{\sigma}{\beta}(A_{j,:}x - b_j) \leq 0\}$, 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 $S^P_\beta(x, y^*) \leq \epsilon$, 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 $\langle c, x\rangle + \langle b, y^*\rangle = \langle c + A^\top y^*, x\rangle + \langle b - Ax, y^*\rangle$. $S^P_\beta(x, y^*) \le \epsilon$ implies $0 \le \langle c + A^\top y^*, x\rangle \le \epsilon$. 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 $S^P_\beta(x, y^*) \le \epsilon$ implies $\|A_{E,:}x - b_E\|^2 \le 2\beta\epsilon/\sigma$, we get

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

For $I_+$, $|\sum_{j\in I_+} y^*_j(b_j - A_{j,:}x)| \le \|y^*_{I_+}\|\|b_{I_+} - A_{I_+,:}x\| \le \|y^*_I\|\sqrt{2\beta\epsilon/\sigma}$.

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

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

Combining the three cases, we get

$$\sqrt{2\beta\epsilon/\sigma}(\|y^*_E\| + \|y^*_I\|) \le \langle c, x\rangle + \langle b, y^*\rangle \le \sqrt{2\beta\epsilon/\sigma}(\|y^*_E\| + \|y^*_I\|) + 3\epsilon$$

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

$$\begin{aligned}
&\left(|c^\top x + b^\top y^*|^2 + \|A_{E,:}x - b_E\|^2 + \operatorname{dist}(A_{I,:}x - b_I, \mathbb{R}^I_-)^2 + \operatorname{dist}(x_N, \mathbb{R}^N_+)^2\right)^{1/2} \\
&\qquad \le \left(\left(\sqrt{\frac{2\beta\epsilon}{\sigma}}(\|y^*_E\| + \|y^*_I\|) + 3\epsilon\right)^2 + \frac{2\beta\epsilon}{\sigma} + \frac{2\beta\epsilon}{\sigma}\right)^{1/2} \\
&\qquad \le \sqrt{\frac{2\beta\epsilon}{\sigma}}(\|y^*_E\| + \|y^*_I\|) + 3\epsilon + 2\sqrt{\frac{\beta\epsilon}{\sigma}}
\end{aligned}$$

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

$$\begin{aligned}
\operatorname{dist}(z, \mathcal{Z}^*) \le \theta\Big(&\sqrt{\frac{2\beta}{\tau}}(\sqrt{2} + \|x^*_F\| + \|x^*_N\|)\sqrt{G_\beta(z; z^*)} \\
&+ \sqrt{\frac{2\beta}{\sigma}}(\sqrt{2} + \|y^*_E\| + \|y^*_I\|)\sqrt{G_\beta(z; z^*)} + 3G_\beta(z; z^*)\Big) \,.
\end{aligned}$$

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

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

□

# 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 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 1.** *Under Assumption 5, if $\mathcal{R}$ contains $\{z \ : \ \|z - P_{\mathcal{Z}^*}(z_0)\| \leq \operatorname{dist}_V(z_0, \mathcal{Z}^*)\}$, then PDHG converges linearly at a rate*

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

*where $\Gamma = (1-\alpha_f)(1-\sqrt{\gamma})$ and $\lambda = \frac{\Gamma}{\max(1/\beta_x, \Gamma + 1/\beta_y)}$.*

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

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

so, for $z^* = P_{\mathcal{Z}^*}(z_0)$, we get the stability of the set $\{z \ : \ \|z - P_{\mathcal{Z}^*}(z_0)\| \leq \operatorname{dist}_V(z_0, \mathcal{Z}^*)\}$.

For $z^* = P_{\mathcal{Z}^*}(z_k)$,

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

For the right hand side, $\beta \odot (z - z^*) + (z - z_{k+1}) - (z - z_k) = 0$ 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

$$\begin{aligned} &\frac{1}{2}\operatorname{dist}_V(z_k, \mathcal{Z}^*)^2 - \frac{1}{2}\operatorname{dist}_V(z_{k+1}, \mathcal{Z}^*)^2 - \tilde{V}(\bar{z}_{k+1} - z_k) \\ &\qquad\qquad \geq \frac{1}{2}\|z^* - z_k\|_V^2 - \frac{1}{2}\|z^* - z_{k+1}\|_V^2 - \tilde{V}(\bar{z}_{k+1} - z_k) \geq 0 \end{aligned}$$

Using the assumption, this leads to: $\forall \lambda \in [0, 1]$,

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

Using Lemma 5 we get

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

Moreover, $\tilde{V}(\bar{z}_{k+1} - z_k) \geq \frac{\Gamma}{2}\|z_{k+1} - z_k\|_V^2$ (Lemma 4), so taking $\alpha = \frac{\eta}{\Gamma + \eta}$ and $\lambda = \frac{\Gamma}{\max(1/\beta_x, \Gamma + 1/\beta_y)} \leq 1$ leads to $\frac{\lambda}{2\beta_y} + \frac{(\alpha^{-1}-1)\lambda\eta}{2} = \frac{\lambda}{2\beta_y} + \frac{\lambda\Gamma}{2} = \frac{\lambda}{2}(\Gamma + 1/\beta_y) \leq \frac{\Gamma}{2}$ and $\frac{\lambda}{2\beta_x} \leq \frac{\Gamma}{2}$, so that

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

and thus 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 10), so that we recover the rate of Proposition 5.

**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.

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

According to Theorem 1, the rate is thus $(1 + \rho)^{-1}$ where

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

with $\Gamma = (1 - \mu\tau/2)(1 - \sqrt{\sigma\tau a^2})$. Since the algorithm does not depend on $\beta_x$ or $\beta_y$ we can choose them so that they minimize the rate (or maximize $\rho$). On Figure 1, 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 3) when $\mu$ is large and does not vanish when $\mu$ goes to 0. On top of this, for small values of $\mu$, we obtain a much better rate than what is predicted using metric sub-regularity.

With the analysis including a $\|z_{k+1} - z_k\|_V^2$ term (available in Appendix B, Proposition 14), we can explain an even better rate. When we plot the curve of the rate as a function of $\mu_f$ (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.

# 8 Restarted averaged primal-dual hybrid gradient

## 8.1 Presentation of RAPDHG

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

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 $O(1/\sqrt{k})$ [7].

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 $\beta$, like linear programs, we will expect an improved convergence rate by averaging and restarting.

**Proposition 13.** *Under Assumption 5 with $\beta_x = \beta_y = \beta$, if the restart frequency $K$ satisfies $K\beta \geq 2$ and $K\eta \geq 4$, then RAPDHG converges linearly at a rate $2^{-1/K}$. Moreover, if $K = \lceil\max(2/\beta, 4/\eta)\rceil$, then the rate is $\exp\Big(-\frac{1}{\lceil\max(2/\beta, 4/\eta)\rceil}\ln(2)\Big) \approx \exp\Big(-\min(\beta/2, \eta/4)\ln(2)\Big)$.*

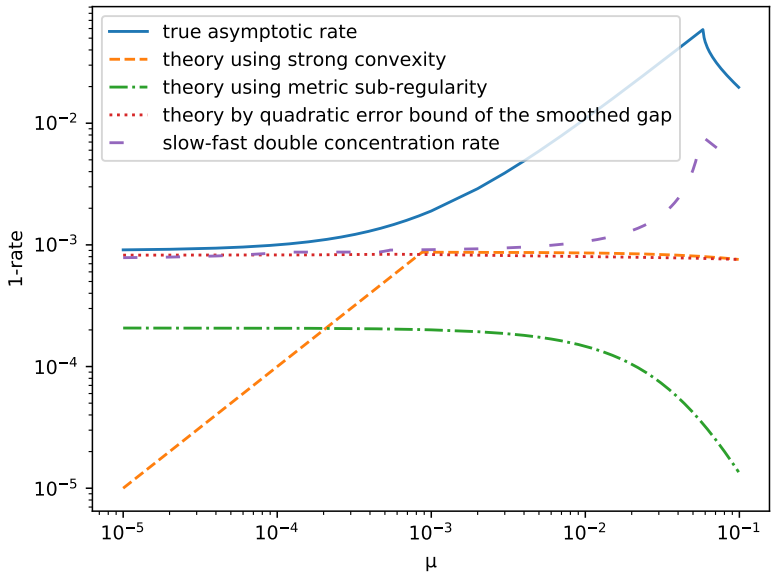


Figure 2: Comparison of the true rate (line above), what is predicted by theory using previous theories and what is predicted by using quadratic error bound of the smoothed gap for $a = 0.03$, $\tau = \sigma = 1$ and various values for $\mu$.

*Proof.* Summing (4) for $k$ between 0 and $K-1$ and using the fact that the Lagrangian is convex-concave, we get

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

which leads to

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

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

$$G_\beta(\tilde{z}_K, z^*) \leq \frac{1}{2K}\frac{K\beta}{K\beta - 1}\|z^* - z_0\|_V^2$$

We now use Assumption 5 to get

$$\frac{1}{K}\frac{K\beta}{K\beta - 1}\|z^* - z_0\|_V^2 \geq \eta\|z^* - \tilde{z}_K\|^2$$

We choose $z^* = P_{\mathcal{Z}^*}(z_0)$ and $K$ such that $K\beta \geq 2$ and $K\eta \geq 4$ in order to get

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

If we choose $K = \lceil \max(2/\beta, 4/\eta) \rceil$ we thus get a linear convergence

$$\begin{aligned}\mathrm{dist}_V(\tilde{z}_K, \mathcal{Z}^*)^2 &\leq \frac{1}{2^s}\,\mathrm{dist}_V(\tilde{z}_0, \mathcal{Z}^*)^2 \\ &\leq \exp\Big(-\frac{1}{\lceil \max(2/\beta, 4/\eta) \rceil}\ln(2)\Big)^{sK}\,\mathrm{dist}_V(\tilde{z}_0, \mathcal{Z}^*)^2\end{aligned}$$

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

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

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

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

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

Let $K = C/\eta$ and $z_0 = (x_0, y_0)$.
For $s \geq 0$:

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

where $sK$ is the total number of iterations. □

## 8.2 Heuristic adaptive restart

In general, we do not know the set of saddle points, so that computing the smoothed gap with a saddle point as reference point is not possible. We propose the following approximation. For $z^*$ equal to the projection of $z$ onto $\mathcal{Z}^*$, we have:

$$\begin{aligned}
G_\beta(z, \dot{z}) &= \max_{z'} L(x, y') - L(x', y) - \frac{\beta}{2} \|\dot{z} - z'\|_V^2 \qquad (7) \\
&\geq \max_{z'} L(x, y') - L(x', y) - \beta \|z^* - z'\|_V^2 - \beta \|\dot{z} - z^*\|_V^2 \\
&= G_{2\beta}(z, z^*) - \beta \|\dot{z} - z^*\|_V^2 \geq \frac{\eta(2\beta)}{2} \|z - z^*\|_V^2 - \beta \|\dot{z} - z^*\|_V^2 \qquad (8)
\end{aligned}$$

and thus $G_\beta(z, \dot{z})$ is a good approximation to the measure of optimality $G_{2\beta}(z, z^*)$ as soon as $\beta$ is small enough (and $\dot{z}$ is closer to $z^*$ than $z$). In the numerical experiment section, we will use it as a stopping criterion with $\beta = (0, \delta)$ where $\delta$ is the dual infeasibility and $\dot{z} = z$.

For RAPDHG, we do not know either the value of the quadratic error bound of the smoothed gap. We propose the following heuristic to adaptively restart the algorithm. Let $\bar{z}_s$ be the primal-dual point at the last restart. We restart when $G_{\frac{2}{k-s+1}}(\tilde{z}_k, \tilde{z}_k) \leq 0.5\, G_{\frac{2}{k-s+1}}(\bar{z}_s, \tilde{z}_k)$. We then compare $\tilde{z}_k$ and $\bar{z}_k$ and restart at the best of these in terms of smoothed gap. Note that $\text{dist}_V(\tilde{z}_k, \mathcal{Z}^*) \leq \text{dist}_V(\bar{z}_s, \mathcal{Z}^*)$. This adaptive restart is formalized in Algorithm 4. We added an additional safeguard for cases where the smoothed gap is increasing in the first phase of the algorithm.

**Algorithm 4** RAPDHG with adaptive restart

$s = 0$
**for** $k \in \{0, \dots, K-1\}$ **do**
  $z_{k+1} = T(z_k)$ – *see* (3)
  $\tilde{z}_{k+1} = \frac{1}{k-s+1} \sum_{l=s+1}^{k+1} \bar{z}_l$
  **if** $G_{\frac{2}{k-s+1}}(\tilde{z}_{k+1}, \tilde{z}_{k+1}) \leq 0.5\, G_{\frac{2}{k-s+1}}(\bar{z}_{s+1}, \tilde{z}_{k+1})$
      **or** $G_{\frac{2}{k-s+1}}(\tilde{z}_{k+1}, \tilde{z}_{k+1}) > 10\, G_{\frac{2}{k-s+1}}(\bar{z}_{s+1}, \tilde{z}_{k+1})$ **then**
    $s = k$
    **if** $G_{\frac{2}{k-s+1}}(\tilde{z}_{k+1}, \tilde{z}_{k+1}) \leq G_{\frac{2}{k-s+1}}(\bar{z}_{k+1}, \bar{z}_{k+1})$ **then**
      Reassign $z_{k+1} \leftarrow \tilde{z}_{k+1}$
    **else**
      Keep current iterate

# 9 Numerical experiments

In the last section, we present some numerical experiments to illustrate the linear convergence behaviour of PDHG and RAPDHG[1].

## 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 $\beta$ the quantity $\hat{\eta}(\beta) = \min_k \frac{G_\beta(z_k; z^*)}{0.5 \operatorname{dist}(z_k, \mathcal{Z}^*)^2}$. 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 rate of convergence matches 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 [6] ($d = 119$ 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

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

| $\beta$ | $\hat{\eta}(\beta)$ |
|---|---|
| 1 | 0.00018 |
| 0.1 | 0.00183 |
| 0.01 | 0.01829 |
| 0.001 | 0.14474 |

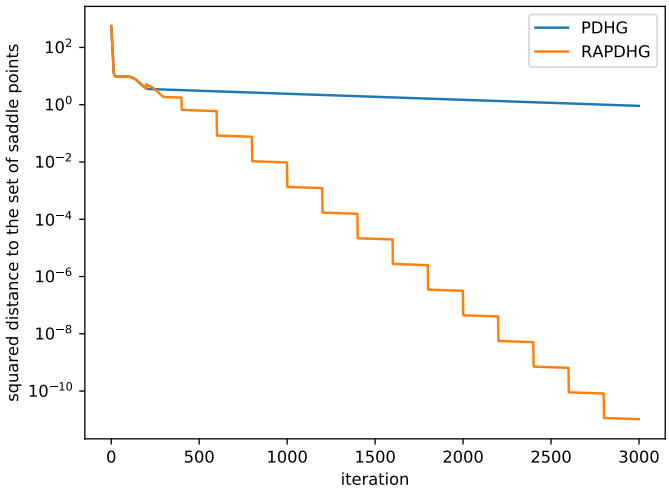


Figure 3: Table: Estimates of the quadratic error bound of the smoothed gap for several smoothing parameters. Figure: Comparison of PDHG and RAPDHG on the small linear program. The restart period of 200 was chosen because for $\beta = 1/100$, we have $\hat{\eta}(\beta) \approx 2/100$, so that $K = \lceil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

Proposition 12. 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 [20] took 0.93s to solve the problem.

### 9.3 Ridge regression

In this experiment, we test on a problem where restarting does not help. We consider least squares with $\ell_2$ regularization

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

where $A$ and $b$ are given by the real-sim dataset [6]. Since we know the strong convexity-concavity parameter of the Lagrangian, we choose the step sizes $\sigma$ and $\tau$ as in Section 5.1.

We can see on Figure 5 that, as expected, restart and averaging does not help: $\bar{z}_k$ is consistently better than $\tilde{z}_k$ so that the curves for PDHG and RAPDHG with adaptive restart match. We added a comparison with restarted FISTA [12] 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 but is rather structured. Indeed, 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.

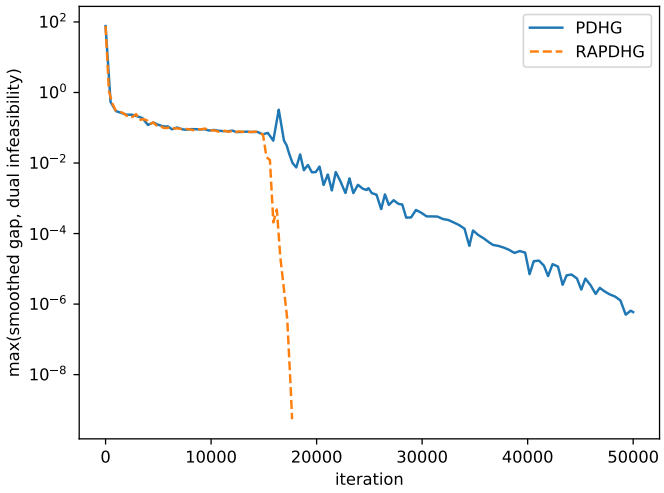


Figure 4: Comparison of PDHG and RAPDHG: sparse SVM on the a1a dataset.

# 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 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.

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 (8) 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 1** *Let $p = \operatorname{prox}_{\tau f}(x)$ and $p' = \operatorname{prox}_{\tau f}(x')$ where $f$ is $\mu_f$-strongly convex. For all $x$ and $x'$,*

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

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

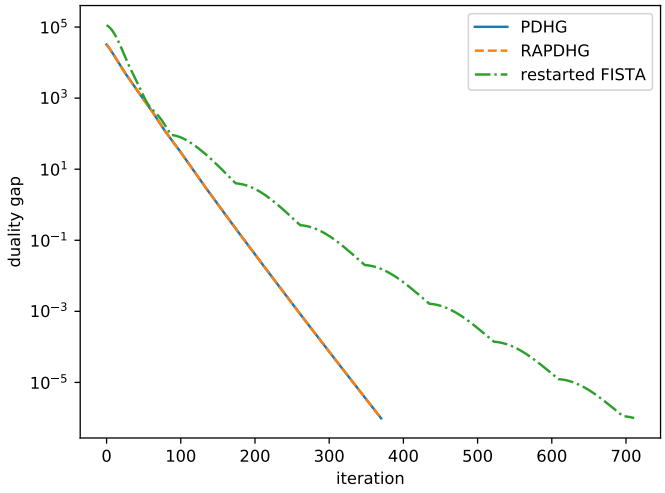


Figure 5: Solving $\ell_2$ regularized least squares on the real-sim dataset.

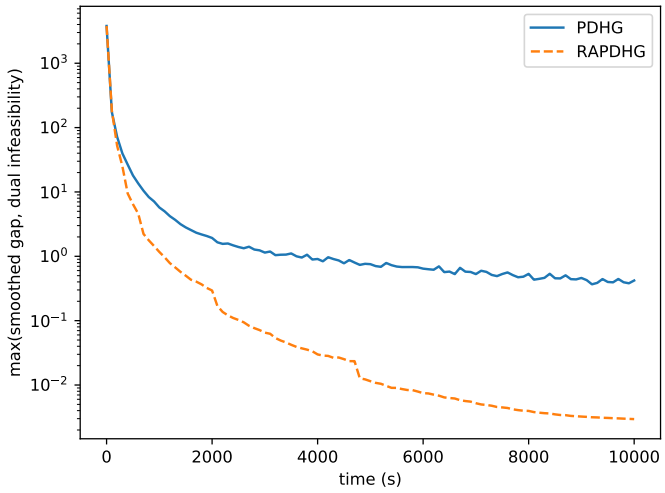


Figure 6: Comparison of PDHG and RAPDHG on the $\ell_1$ ROF problem.

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

*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 2** *Let $T : \mathcal{X}\times\mathcal{Y} \to \mathcal{X}\times\mathcal{Y}$ be defined for any $(x, y)$ by*

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

*If $\gamma = \sigma\tau\|A\|^2 < 1$, $\tau L_f/2 \leq \alpha_f < 1$, $\alpha_g = \sigma L_{g^*}/2 \leq 1$ and $\sigma L_{g^*}/2 \leq \alpha_f(1-\sigma\tau\|A\|^2)$ then $T$ is nonexpansive in the norm $\|\cdot\|_V$, and $T$ is $\frac{1}{1+\lambda}$-averaged where*

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

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

$$\|T(z) - T(z')\|_V^2 + 2\mu_f\|\bar{x}-\bar{x}'\|^2 + 2\mu_{g^*}\|\bar{y}-\bar{y}'\|^2 \leq \|z-z'\|_V^2 - \lambda\|z - T(z) - z' + T(z')\|^2\,.$$

*As a consequence, $(z_k)$ converges to a saddle point of the Lagrangian.*

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

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

We also have

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

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

for all $\alpha_f > 0$. Hence,

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

Similarly,

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

We then proceed to

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

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

$$\begin{aligned}\|T(x,y)-T(x',y')\|_V^2 \leq \frac{1}{\tau}\|x-x'\|^2 + \frac{1}{\sigma}\|y-y'\|^2 \\ - \frac{\lambda}{\tau}\|x-\bar{x}+\tau A^\top(\bar{y}-y)-x'+\bar{x}'-\tau A^\top(\bar{y}'-y')\|^2 \\ - \frac{\lambda}{\sigma}\|y-\bar{y}-y'+\bar{y}'\|^2 + \Big(\frac{\lambda}{\tau\alpha}-\frac{1-\alpha_f-\lambda}{\tau}\Big)\|x-\bar{x}-x'+\bar{x}'\|^2 \\ + \Big((1+\lambda+\lambda\alpha)\tau\|A\|^2 - \frac{1-\alpha_g-\lambda}{\sigma}\Big)\|(\bar{y}-y-\bar{y}'+y')\|^2 \\ -\Big(\frac{2\tau}{L_f}-\alpha_f^{-1}\tau^2\Big)\|\nabla f_2(x)-\nabla f_2(x')\|^2 - 2\mu_f\|\bar{x}-\bar{x}'\|^2 \\ -\Big(\frac{2\sigma}{L_{g^*}}-\alpha_g^{-1}\sigma^2\Big)\|\nabla g_2(y)-\nabla g_2(y')\|^2 - 2\mu_{g^*}\|\bar{y}-\bar{y}'\|^2\end{aligned}$$

where $\lambda \in [0, 1-\alpha_f]$ and $\alpha > 0$ are arbitrary. We choose $\alpha_f = \tau L_f/2 < 1$ and $\alpha_g = \sigma L_{g^*}/2 < 1$. We choose $\lambda$ and $\alpha$ 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

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

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

For the convergence, we use Krasnosels'kii Mann theorem [3]. □

**Lemma 3** *For all $k \in \mathbb{N}$ and for all $z \in \mathcal{Z}$,*

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

*where* $\tilde{V}(\bar{z}_{k+1}-z_k) = (\frac{1}{2\tau}-\frac{L_f}{2})\|\bar{x}_{k+1}-x_k\|^2 + (\frac{1}{2\sigma}-\frac{\tau\|A\|^2}{2}-\frac{L_{g^*}}{2})\|\bar{y}_{k+1}-y_k\|^2$

*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:

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

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

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

We can then write

$$\begin{aligned}
\frac{\tau}{2}\|A^\top(\bar{y}_{k+1} - y_k)\|^2 - \frac{1}{2}\|\bar{z}_{k+1} - z_k\|_V^2 + \frac{L_f}{2}\|\bar{x}_{k+1} - x_k\|^2 + \frac{L_{g^*}}{2}\|\bar{y}_{k+1} - y_k\|^2 \\
\leq (\frac{L_f}{2} - \frac{1}{2\tau})\|\bar{x}_{k+1} - x_k\|^2 + (\frac{\tau\|A\|^2}{2} + \frac{L_{g^*}}{2} - \frac{1}{2\sigma})\|\bar{y}_{k+1} - y_k\|^2 \\
= -\tilde{V}(\bar{z}_{k+1} - z_k)
\end{aligned}$$

where $\tilde{V}(z) \geq 0$ as soon as $\gamma \leq 1$. So

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

□

**Lemma 4** *$\tilde{V}$ satisfies*

$$\begin{aligned}\tilde{V}(\bar{z}_{k+1} - z_k) &= (\frac{1}{2\tau} - \frac{L_f}{2})\|\bar{x}_{k+1} - x_k\|^2 + (\frac{1}{2\sigma} - \frac{\tau\|A\|^2}{2} - \frac{L_{g^*}}{2})\|\bar{y}_{k+1} - y_k\|^2 \\ &\geq \frac{(1-\alpha_f)(1-\sqrt{\gamma})}{2}\|z_{k+1} - z_k\|_V^2 \ .\end{aligned}$$

*Proof.* For all $\alpha \in ]0, 1[$,

$$\begin{aligned}\tilde{V}(\bar{z}_{k+1} - z_k) &= (\frac{1}{2\tau} - \frac{L_f}{2})\|\bar{x}_{k+1} - x_k\|^2 + (\frac{1}{2\sigma} - \frac{\tau\|A\|^2}{2} - \frac{L_{g^*}}{2})\|\bar{y}_{k+1} - y_k\|^2 \\ &= (\frac{1}{2\tau} - \frac{L_f}{2})\|x_{k+1} - x_k + \tau A^\top(y_{k+1} - y_k)\|^2 + (\frac{1}{2\sigma} - \frac{\tau\|A\|^2}{2} - \frac{L_{g^*}}{2})\|y_{k+1} - y_k\|^2 \\ &\geq \frac{1}{2\tau}(1-\alpha_f)\Big((1-\alpha)\|x_{k+1} - x_k\|^2 + (1-\alpha^{-1})\tau^2\|A^\top(y_{k+1} - y_k)\|^2\Big) \\ &\qquad + (\frac{1}{2\sigma} - \frac{\tau\|A\|^2}{2} - \frac{L_{g^*}}{2})\|y_{k+1} - y_k\|^2 \\ &\geq \frac{1}{2\tau}(1-\alpha_f)(1-\alpha)\|x_{k+1} - x_k\|_{\tau^{-1}}^2 \\ &\qquad + \Big(\frac{1}{2} - (1 + (1-\alpha_f)(\alpha^{-1} - 1))\frac{\sigma\tau\|A\|^2}{2} - \frac{\sigma L_{g^*}}{2}\Big)\|y_{k+1} - y_k\|_{\sigma^{-1}}^2\end{aligned}$$

We have $\gamma = \sigma\tau\|A\|^2 < 1$, $\tau L_f/2 \leq \alpha_f < 1$, $\alpha_g = \sigma L_{g^*}/2 \leq 1$ and $\sigma L_{g^*}/2 \leq \alpha_f(1 - \sigma\tau\|A\|^2)$, so

$$\begin{aligned}\tilde{V}(\bar{z}_{k+1} - z_k) &\geq \frac{1}{2}(1-\alpha_f)(1-\alpha)\|x_{k+1} - x_k\|_{\tau^{-1}}^2 \\ &\qquad + \frac{1}{2}\Big(1 - \alpha_g - \gamma - (1-\alpha_f)(\alpha^{-1} - 1)\gamma\Big)\|y_{k+1} - y_k\|_{\sigma^{-1}}^2\end{aligned}$$

We want

$$\begin{aligned}(1-\alpha_f)(1-\alpha) &= 1 - \gamma - \alpha_g - (1-\alpha_f)(\alpha^{-1} - 1)\gamma \\ (1-\alpha) &= \frac{1 - \gamma - \alpha_g}{1-\alpha_f} - (\alpha^{-1} - 1)\gamma \\ \alpha &= \gamma/\alpha - \frac{\alpha_f(1-\gamma) - \alpha_g}{1-\alpha_f} \\ \alpha^2 &+ \alpha\frac{\alpha_f(1-\gamma) - \alpha_g}{1-\alpha_f} - \gamma = 0 \\ \alpha &= \frac{1}{2}(-r + \sqrt{r^2 + 4\gamma}) \leq \sqrt{\gamma}\end{aligned}$$

where $r = \frac{\alpha_f(1-\gamma) - \alpha_g}{1-\alpha_f} \geq 0$. With this value of $\alpha$, we obtain

$$\begin{aligned}\tilde{V}(\bar{z}_{k+1} - z_k) &\geq \frac{1}{2}(1-\alpha_f)\frac{1}{2}(2 + r - \sqrt{r^2 + 4\gamma})\|z_{k+1} - z_k\|_V^2 \\ &\geq \frac{1}{2}(1-\alpha_f)(1-\sqrt{\gamma})\|z_{k+1} - z_k\|_V^2\end{aligned}$$

And if $L_f = L_{g^*} = 0$, we get $\tilde{V}(\bar{z}_{k+1} - z_k) \geq \frac{1-\sqrt{\gamma}}{2}\|z_{k+1} - z_k\|_V^2$ □

**Proposition 4** *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 $\tilde{z}_k = \frac{1}{k}\sum_{l=1}^{k} \bar{z}_l$ 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*

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

*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 3 the stability inequality

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

We then sum (4) for $k$ between 0 and $K-1$ and use convexity in $x$ and concavity in $y$ of the Lagrangian:

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

In particular,

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

□

# B Idea to take profit of strong convexity

The goal of this section is to derive a finer analysis in the case where we solve a linearly constrained problem whose objective function is strongly convex. In the toy problem of Section 5.2, we can show that the largest singular value of the matrix $R$ 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 here 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.

**Proposition 14.** *Suppose that $\mu_f > 0$, $g = \iota_{\{b\}}$ and $G_\beta(\cdot, z^*)$ has a $\eta$-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*

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

- *if* $2\mu_f\tau(1-\alpha_1) > C\eta_x$ *and* $\frac{\frac{1}{\beta_x}-\Gamma}{2\mu_f(1-\alpha_1)-C\eta_x} > \frac{-\frac{1}{\beta_y}+\frac{(1-\alpha_2-C)\eta_x}{2\gamma(1-\alpha_2)}-C\eta_x(1-\alpha_2)(\alpha_2^{-1}-1)+\frac{1}{\beta_x}}{2\mu_f(1-\alpha_1)}$, *then we take* $\lambda_1 = \frac{-\frac{1}{\beta_y}+\frac{(1-\alpha_2-C)\eta_x}{2\gamma(1-\alpha_2)}-C\eta_x(1-\alpha_2)(\alpha_2^{-1}-1)+\frac{1}{\beta_x}}{2\mu_f\tau(1-\alpha_1)}$, $\lambda_4 = \frac{\frac{1}{\beta_x}-\lambda_1(2\mu_f\tau(1-\alpha_1)-C\eta_x)}{\Gamma} - 1$ *and we have*

$$\rho = \Big(1 + \frac{\min(C\eta_x, \eta_y)\Gamma}{\frac{1}{\beta_x} - \frac{2\mu_f\tau(1-\alpha_1)-C\eta_x}{2\mu_f\tau(1-\alpha_1)}(-\frac{1}{\beta_y} + \frac{(1-\alpha_2-C)\eta_x}{2\gamma(1-\alpha_2)} - C\eta_x(1-\alpha_2)(\alpha_2^{-1}-1) + \frac{1}{\beta_x})}\Big)^{-1}$$

- *if* $2\mu_f\tau(1-\alpha_1) > C\eta_x$ *and* $\frac{\frac{1}{\beta_x}-\Gamma}{2\mu_f\tau(1-\alpha_1)-C\eta_x} \leq \frac{-\frac{1}{\beta_y}+\frac{(1-\alpha_2-C)\eta_x}{2\gamma(1-\alpha_2)}-C\eta_x(1-\alpha_2)(\alpha_2^{-1}-1)+\frac{1}{\beta_x}}{2\mu_f\tau(1-\alpha_1)}$, *then* $\lambda_4 = 0$, $\lambda_1 = \frac{\frac{1}{\beta_x}-\Gamma}{2\mu_f\tau(1-\alpha_1)-C\eta_x}$ *and*

$$\rho = \max\big((1+C\eta_x)^{-1}, (1+\eta_y)^{-1}\big)$$

In order to use this proposition, we shall compute $\rho$ 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 2 to $z = z_k$ and $z' = z_{k-1}$ so that $T(z) = z_{k+1}$ and $T(z') = z_k$:

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

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

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

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

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 $\lambda_i$, $i \in \{1,2,3\}$.

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

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

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

to get

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

To get the rate, we then need

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

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 $C$ for a 1D grid search since the rate will depend a lot on its value.

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 $\lambda_4=\frac{1}{\beta_x\Gamma}-1$. this leads to

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

Supposing that $\mu_g=+\infty$ and $L_{g_2^*}=0$, we get a rate $\rho=\max((1+\frac{C\eta_x\beta_x}{\Gamma})^{-1},(1+\frac{\eta_y\beta_x}{\Gamma})^{-1})$.

Case 2: if $2\mu_f\tau(1-\alpha_1)>C\eta_x$ and $\frac{\frac{1}{\beta_x}-\Gamma}{2\mu_f(1-\alpha_1)-C\eta_x}>\frac{-\frac{1}{\beta_y}+\frac{(1-\alpha_2-C)\eta_x}{2\gamma(1-\alpha_2)}-C\eta_x(1-\alpha_2)(\alpha_2^{-1}-1)+\frac{1}{\beta_x}}{2\mu_f(1-\alpha_1)}$

We choose $\lambda_1 = \frac{-\frac{1}{\beta_y}+\frac{(1-\alpha_2-C)\eta_x}{2\gamma(1-\alpha_2)}-C\eta_x(1-\alpha_2)(\alpha_2^{-1}-1)+\frac{1}{\beta_x}}{2\mu_f\tau(1-\alpha_1)}$ and $\lambda_4 = \frac{\frac{1}{\beta_x}-\lambda_1(2\mu_f\tau(1-\alpha_1)-C\eta_x)}{\Gamma} - 1$. We get

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

Supposing that $\mu_g = +\infty$ and $L_{g_2^*} = 0$, we get a rate $\rho = \max((1+\frac{C\eta_x}{1+\lambda_4})^{-1}, (1+\frac{\eta_y}{1+\lambda_4})^{-1}) = (1+\frac{\min(C\eta_x,\eta_y)\Gamma}{\frac{1}{\beta_x}-\frac{2\mu_f\tau(1-\alpha_1)-C\eta_x}{2\mu_f\tau(1-\alpha_1)}(-\frac{1}{\beta_y}+\frac{(1-\alpha_2-C)\eta_x}{2\gamma(1-\alpha_2)}-C\eta_x(1-\alpha_2)(\alpha_2^{-1}-1)+\frac{1}{\beta_x})})^{-1}$.

Case 3: if $2\mu_f\tau(1-\alpha_1) > C\eta_x$ and $\frac{\frac{1}{\beta_x}-\Gamma}{2\mu_f\tau(1-\alpha_1)-C\eta_x} \leq \frac{-\frac{1}{\beta_y}+\frac{(1-\alpha_2-C)\eta_x}{2\gamma(1-\alpha_2)}-C\eta_x(1-\alpha_2)(\alpha_2^{-1}-1)+\frac{1}{\beta_x}}{2\mu_f\tau(1-\alpha_1)}$

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

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

where the last inequality holds because $-\frac{1}{\beta_y}+\frac{(1-\alpha_2-C)\eta_x}{2\gamma(1-\alpha_2)}-C\eta_x(1-\alpha_2)(\alpha_2^{-1}-1)+\Gamma \geq 2\mu_f\tau(1-\alpha_1)\frac{\frac{1}{\beta_x}-\Gamma}{2\mu_f\tau(1-\alpha_1)-C\eta_x}-\frac{1}{\beta_x}+\Gamma = C\eta_x\lambda_1$

Supposing that $\mu_g = +\infty$ and $L_{g_2^*} = 0$, we get a rate $\rho = \max((1+C\eta_x)^{-1}, (1+\eta_y)^{-1})$. □

# References


[1] Alacaoglu, A., Fercoq, O., Cevher, V.: On the convergence of stochastic primal-dual hybrid gradient. arXiv preprint arXiv:1911.00799 (2019)

[2] Alghunaim, S.A., Sayed, A.H.: Linear convergence of primal–dual gradient methods and their performance in distributed optimization. Automatica **117**, 109003 (2020)

[3] Bauschke, H.H., Combettes, P.L.: Convex analysis and monotone operator theory in Hilbert spaces, vol. 408. Springer (2011)

[4] Bolte, J., Nguyen, T.P., Peypouquet, J., Suter, B.W.: From error bounds to the complexity of first-order descent methods for convex functions. Mathematical Programming **165**(2), 471–507 (2017)

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

[6] Chang, C.C., Lin, C.J.: Libsvm: a library for support vector machines. ACM transactions on intelligent systems and technology (TIST) **2**(3), 1–27 (2011)

[7] Davis, D., Yin, W.: Convergence rate analysis of several splitting schemes. In: Splitting methods in communication, imaging, science, and engineering, pp. 115–163. Springer (2016)

[8] Dontchev, A.L., Rockafellar, R.T.: Implicit functions and solution mappings, vol. 543. Springer (2009)

[9] Drusvyatskiy, D., Lewis, A.S.: Error bounds, quadratic growth, and linear convergence of proximal methods. Mathematics of Operations Research **43**(3), 919–948 (2018)

[10] Du, S.S., Hu, W.: 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, pp. 196–205. PMLR (2019)

[11] Fercoq, O., Bianchi, P.: A coordinate-descent primal-dual algorithm with large step size and possibly nonseparable functions. SIAM Journal on Optimization **29**(1), 100–134 (2019)

[12] Fercoq, O., Qu, Z.: Adaptive restart of accelerated gradient methods under local quadratic growth condition. IMA Journal of Numerical Analysis **39**(4), 2069–2095 (2019)

[13] Fercoq, O., Qu, Z.: Restarting the accelerated coordinate descent method with a rough strong convexity estimate. Computational Optimization and Applications **75**(1), 63–91 (2020)

[14] Gabay, D., Mercier, B.: A dual algorithm for the solution of nonlinear variational problems via finite element approximation. Computers & mathematics with applications **2**(1), 17–40 (1976)

[15] Hoffman, A.J.: On approximate solutions of systems of linear inequalities'. Journal of Research of the National Bureau of Standards **49**(4), 263 (1952)

[16] Kovalev, D., Salim, A., Richtárik, P.: Optimal and practical algorithms for smooth and strongly convex decentralized optimization. arXiv preprint arXiv:2006.11773 (2020)

[17] Latafat, P., Freris, N.M., Patrinos, P.: A new randomized block-coordinate primal-dual proximal algorithm for distributed optimization. IEEE Transactions on Automatic Control **64**(10), 4050–4065 (2019)

[18] Liang, J., Fadili, J., Peyré, G.: Convergence rates with inexact non-expansive operators. Mathematical Programming **159**(1-2), 403–434 (2016)

[19] Lin, Q., Ma, R., Nadarajah, S., Soheili, N.: First-order methods for convex constrained optimization under error bound conditions with unknown growth parameters. arXiv preprint arXiv:2010.15267 (2020)

[20] Lu, M., Qu, Z.: An adaptive proximal point algorithm framework and application to large-scale optimization. arXiv preprint arXiv:2008.08784 (2020)

[21] Nesterov, Y.: Smooth minimization of non-smooth functions. Mathematical programming **103**(1), 127–152 (2005)

[22] Nesterov, Y.: Lectures on convex optimization, vol. 137. Springer (2018)

[23] Rockafellar, R.T., Wets, R.J.B.: Variational analysis, vol. 317. Springer Science & Business Media (2009)

[24] Tran-Dinh, Q., Fercoq, O., Cevher, V.: A smooth primal-dual optimization framework for nonsmooth composite convex minimization. SIAM Journal on Optimization **28**(1), 96–134 (2018)

[25] Zhu, D., Zhao, L.: Linear convergence of randomized primal-dual coordinate method for large-scale linear constrained convex programming. In: International Conference on Machine Learning, pp. 11619–11628. PMLR (2020)