The Smoothed Duality Gap as a Stopping Criterion

Iyad Walwil, Olivier Fercoq

Noname manuscript No.
(will be inserted by the editor)

The Smoothed Duality Gap as a Stopping Criterion

Iyad Walwil · Olivier Fercoq

Received: date / Accepted: date

Abstract We optimize the running time of the primal-dual algorithms by optimizing their stopping criteria for solving convex optimization problems under affine equality constraints, which means terminating the algorithm earlier with fewer iterations. We study the relations between four stopping criteria and show under which conditions they are accurate to detect optimal solutions. The uncomputable one: *"Optimality gap and Feasibility error"*, and the computable ones: the *"Karush-Kuhn-Tucker error"*, the *"Projected Duality Gap"*, and the *"Smoothed Duality Gap"*. Assuming metric sub-regularity or quadratic error bound, we establish that all of the computable criteria provide practical upper bounds for the optimality gap, and approximate it effectively. Furthermore, we establish comparability between some of the computable criteria under certain conditions. Numerical experiments on basis pursuit, and quadratic programs with(out) non-negative weights corroborate these findings and show the superior stability of the smoothed duality gap over the rest.

Keywords Convex optimization · Stopping criteria · Optimality gap and Feasibility error · Karush-Kuhn-Tucker · Projected Duality Gap · Smoothed Duality Gap

Mathematics Subject Classification (2020) 65K05 · 90C25 · 90C46

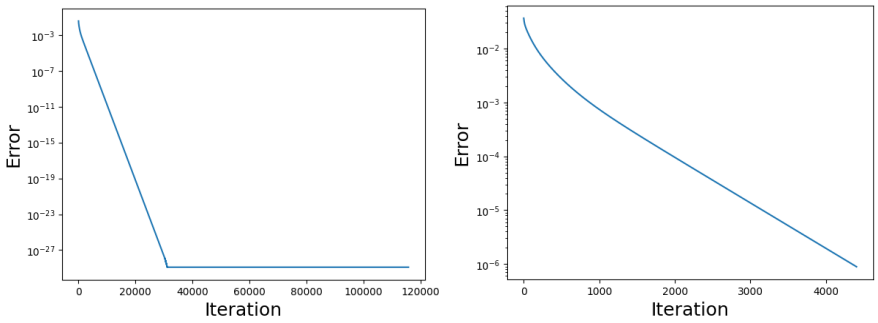
Iyad Walwil
LTCI, Télécom Paris, Institut Polytechnique de Paris, France
Department of Mathematics, Faculty of Sciences, An-Najah National University, Nablus, Palestine
E-mail: iyad.walwil@telecom-paris.fr

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

1 Introduction

How do we decide when to stop doing something when we don't really know when to stop? This is a non-trivial question in general that we can face in several events during our life. For instance, as researchers, deciding when to stop conducting more experiments to practically validate a new theoretical result is one of the hardest decisions to make. In this paper, we would like to answer a quite similar question: How do we decide when to stop an iterative algorithm seeking an $\varepsilon-$solution for an optimization problem?

When solving an optimization problem with an iterative method, one is looking for an $\varepsilon-$solution (i.e. a solution that's ε close to the optimal one). However, testing if a point is actually an $\varepsilon-$solution is not always easy, so we do not really know when to stop the algorithm. This can be detrimental to the running time of the method because of unnecessary iterations. In this work, we aim to address this problem by studying several stopping criteria and determine under which conditions they are accurate to detect $\varepsilon-$solutions.



(a) Fixed number of iterations, $\varepsilon_1 \approx 10^{-28}$ (b) Karush-Kuhn-Tucker error, $\varepsilon_2 \approx 10^{-6}$

Fig. 1 Gradient descent is employed to address an unconstrained Least-Squares problem, utilizing two distinct stopping criteria aimed at achieving an $\varepsilon = 10^{-5}$ solution

Figure 1 illustrates the significance of our study by demonstrating a substantial disparity observed when solving an unconstrained Least-Squares problem with the same algorithm but employing two distinct stopping criteria to achieve an $\varepsilon = 10^{-5}$ solution. At first glance, one might infer that sub-figure 1a outperforms sub-figure 1b. Indeed, this inference holds when solely considering the $y-$axis, where a smaller error indicates closer proximity to optimality. However, the $x-$axis conveys an alternate narrative regarding the requisite number of iterations and thereby the computational run-time. While our objective is to detect an $\varepsilon = 10^{-5}$ solution, sub-figure 1a has gone too far for, approximately, a 10^{-28} solution but at the expense of time. Although sub-figure 1a yields almost an optimal solution, it might become unfeasible to track in other scenarios involving high-dimensional problems for example.

We are interested in convex optimization problems under affine equality constraints. That is:

$$\min_{x\in\mathcal{X}} \; f(x) \qquad \text{subject to} \qquad Ax = b \tag{$\mathcal{P}$}$$

where $f\colon \mathcal{X} \to \mathbb{R}\cup\{+\infty\}$ is a proper, lower semi-continuous, and convex function with a computable proximal operator, $A\colon \mathcal{X}\to\mathcal{Y}$ is a linear operator, and $b\in\mathcal{Y}$. We will, simultaneously, solve the primal and dual problems by solving their associated saddle point problem:

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} \; \mathcal{L}(x,y) := f(x) + \langle Ax-b, y\rangle \tag{$\mathcal{SPP}$}$$

where $\mathcal{L}(x,y)$ is the associated Lagrangian with ($\mathcal{P}$) and $y\in\mathcal{Y}$ is the so-called Lagrange multiplier or dual variable. Throughout the paper, we assume the existence of a solution to ($\mathcal{SPP}$).

Problems of form ($\mathcal{SPP}$) are ubiquitous in operational research, signal processing, shape optimization, statistical learning, etc... For example,

- ❖ Method of framers [12] and Basis Pursuit [8] are techniques for decomposing a signal into an "optimal" superposition of dictionary elements by picking a solution whose coefficients have minimum ℓ_2, and ℓ_1 norms, respectively.
- ❖ Statistical learning covers wide range of (un)constrained optimization problems like LASSO [25,28], constrained least-squares [4,27], etc...
- ❖ In operational research: linear programs [11,24] stand out as the most renowned problems within a broader framework that encompasses the presence of inequality constraints, $\min_{x\geq 0}\max_{y}\langle c,x\rangle + \langle Ax-b,y\rangle$

Several measures of optimality have been considered in the literature. The first and most natural one is *"Optimality Gap and Feasibility error (OGFE)"*, which directly fits the definition of the optimization problem at stake. Indeed, the *Optimality Gap* represents the difference between the objective function value at the current solution (x_k) and the optimal solution ($x^\star$), while the *Feasibility Error* measures the constraints violation. The second traditional one is the *"Distance to the Optimal Solution Set (DOSS)"* that measures how far or close the solution is to be an optimal one. However, as both of those measures depend on the unknown point: $x^\star$, one cannot compute them before the problem is actually solved! Hence in algorithms, the *"Karush–Kuhn–Tucker error"* [22,23] is widely used [16,14,6], it is a computable quantity and serves as a first-order optimality measure for achieving optimality in non-linear programming problems. It accomplishes this by quantifying the error in feasibility and identifying, within the sub-differential of the associated Lagrangian, the element with the smallest norm. Moreover, if the Lagrangian's gradient is metrically sub-regular [21], 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 difference between the primal and dual optimal values that define the so-called *"Duality Gap (DG)"* [23] is a good way to measure optimality:

it is often easily computable, and it is an upper bound to the *optimality gap*. However, for unbounded domains, it's no longer informative, as it will consistently be infinity, except for the final iterations when the algorithm begins identifying feasible solutions. A first generalization to unbounded domains has been proposed in [26]: the *"Smoothed Duality Gap (SDG)"*, a new measure of optimality that is widely applicable but less well-studied than the other ones. It's based on the smoothing of non-smooth functions [20], and takes finite values for constrained problems, unlike the duality gap. Moreover, if the smoothness parameter is small and the smoothed duality gap is small, this means that the *optimality gap*, and the *feasibility error* are both small. A second one has been proposed for linear programs in [2], we have extended its definition within our framework ($\mathcal{SPP}$) and termed it the "Projected Duality Gap (PDG)." This concept involves calculating the duality gap at each iteration while simultaneously projecting the primal-dual solution onto their respective feasibility spaces. Another recent measure has been proposed for analyzing the primal-dual hybrid gradient algorithm in [18], the authors dub it: the *"Infimal Sub-differential Size (IDS)"*. It always has a finite value, easy to compute, and more importantly, it monotonically decays. IDS essentially measures the distance between 0 and the sub-differential of the objective function. Throughout the remainder of this paper, our attention will be directed towards four optimality measures: the optimality gap and feasibility error, the Karush–Kuhn–Tucker error, the projected duality gap, and the smoothed duality gap. On the one hand, we aim to demonstrate the conditions under which the computable measures (KKT error, PDG, and SDG) serve as upper bounds or approximations for the uncomputable measure, optimality gap and feasibility error. On the other hand, our objective is to assess and compare the performance of smoothed duality gap against both the Karush–Kuhn–Tucker error and the projected duality gap to determine its efficacy and stability.

1.1 Our contributions and Paper organization

The paper's contributions can be outlined as follows:

1. In section 2, we start by providing a comprehensive background.
2. Section 3 is dedicated to establishing a common understanding by precisely defining the different optimality measures, accompanied by a detailed analysis when necessary. Moreover, this section introduces our generalization of the existing measure: the projected duality gap, originally defined only for linear programs, to our framework ($\mathcal{SPP}$).
3. Section 4 is allocated to present and deeply study the novel measure, the smoothed duality gap, while also deriving some new properties.
4. Section 5 is devoted to the establishment of computable approximations for the uncomputable measure, the optimality gap, in terms of the other computable measures (the KKT error error, PDG, and SDG). The necessary regularity assumptions for this purpose are also introduced.

5. A more in-depth examination of the smoothed duality gap and its interconnections with both: the KKT error and PDG is presented in section 6. More precisely, we elucidate the conditions under which these measures can function as approximations to the smoothed duality gap, and vice versa.
6. Section 7 presents several numerical experiments illustrating our findings, while the technical proofs are relegated to the appendix.

1.2 Notations

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. We assume that those vector spaces are Euclidean spaces. Similarly, for a primal vector x, and a dual vector y, we shall denote $z = (x, y)$. The set of saddle points will be denoted as $\mathcal{Z}^\star$. Let $\Gamma_0(\mathcal{X})$ denote the set of all proper, lower semi-continuous, and convex functions $f\colon \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$. The proximal operator of a function f and a step size $s > 0$ is given by:

$$\mathrm{Prox}_{sf}(x) = \arg\min_{x' \in \mathcal{X}} f(x') + \frac{1}{2s}\|x' - x\|^2$$

Let $\mathrm{dist}(z, \mathcal{Z}) = \min_{z' \in \mathcal{Z}} \|z - z'\|$ denote the distance between point z and set $\mathcal{Z}$. We will make use of the convex indicator function associated with the convex subset $\mathcal{C} \subset \mathcal{X}$:

$$\imath_{\mathcal{C}}(x) = \begin{cases} 0 & \text{if } x \in \mathcal{C} \\ +\infty & \text{Otherwise} \end{cases}$$

2 Background

Proposition 1 (Young's inequality)
For all vectors $\mathbf{u}$ *and* $\mathbf{v}$ *of an inner product space, and for any scalar* λ*. The following inequality holds:*

$$|\langle \mathbf{u}, \mathbf{v} \rangle| \leq \frac{\lambda^2}{2}\|\mathbf{u}\|^2 + \frac{1}{2\lambda^2}\|\mathbf{v}\|^2 \tag{1}$$

Lemma 1 *A function* $f\colon \mathcal{X} \to (-\infty, +\infty]$ *is* $\mu-$*strongly convex if for any* $x, y \in \mathrm{dom} f$ *and for any* $q \in \partial f(x)$*, the following inequality holds:*

$$f(y) \geq f(x) + \langle q, y - x \rangle + \frac{\mu}{2}\|y - x\|^2$$

Lemma 2 *Given* $f \in \Gamma_0(\mathcal{X})$*, if* $q_1 \in \partial f(x_1)$ *and* $q_2 \in \partial f(x_2)$*, then:*

$$\langle q_1 - q_2, x_1 - x_2 \rangle \geq 0$$

Lemma 3 *Let* $g \in \Gamma_0(\mathcal{X})$*, and denoting* $p = \mathrm{Prox}_{sg}(x)$*, then we have for all* $v \in \mathcal{X}$*:*

$$g(p) + \frac{1}{2s}\|x - p\|^2 \leq g(v) + \frac{1}{2s}\|x - v\|^2 - \frac{1}{2s}\|p - v\|^2$$

Proposition 2 (Projection properties)
Let $\mathcal{C} \subseteq \mathcal{X}$ be a nonempty, closed, and convex subset. Define $a := \mathrm{Proj}_{\mathcal{C}}(x)$ for $x \in \mathcal{X}$. Then for any $u \in \mathcal{C}$, the following holds:

$$\|a - x\|^2 \leq \|u - x\|^2 \tag{2}$$
$$\langle u - a, a - x \rangle \geq 0 \tag{3}$$

Definition 1 (Separable function)
We say that a function $\varphi\colon \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is separable if there exists n functions $\varphi_i\colon \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ such that $\forall x \in \mathbb{R}^n, \varphi(x) = \sum_{i=1}^{n} \varphi_i(x_i)$.

Proposition 3 (Properties of Separable function)
Let $\varphi\colon \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a separable function, then for any $x \in \mathbb{R}^n$,

$$\partial\varphi(x) = \partial\varphi_1(x_i) \times \cdots \times \partial\varphi_n(x_n) \tag{4}$$
$$\mathrm{Prox}_{\gamma\varphi}(x) = (\mathrm{Prox}_{\gamma\varphi_1}(x_1), \ldots, \mathrm{Prox}_{\gamma\varphi_n}(x_n)) \tag{5}$$
$$\sup_{x \in \mathbb{R}^n} \varphi(x) = \sup_{x_1 \in \mathbb{R}} \varphi_1(x_1) + \cdots + \sup_{x_n \in \mathbb{R}} \varphi_n(x_n) \tag{6}$$

Definition 2 (Fenchel-Legendre Conjugate)
Let $f\colon \mathcal{X} \to [-\infty, +\infty]$. The *Fenchel-Legendre conjugate* of f is the function $f^*\colon \mathcal{X} \to [-\infty, +\infty]$ defined by:

$$f^*(\phi) = \sup_{x \in \mathcal{X}} \langle \phi, x \rangle - f(x), \qquad \phi \in \mathcal{X}$$

Proposition 4 (Fenchel-Young's inequality)
Let $f\colon \mathcal{X} \to [-\infty, +\infty]$. For all $(x, \phi) \in \mathcal{X} \times \mathcal{X}$, the following inequality holds:

$$f(x) + f^*(\phi) \geq \langle \phi, x \rangle$$

*with **equality** if, and only if, $\phi \in \partial f(x)$.*

Proposition 5 *Let $f\colon \mathcal{X} \to (-\infty, +\infty]$ be proper, convex, and l.s.c. at some point $x \in \mathcal{X}$. Then,*

$$\phi \in \partial f(x) \iff x \in \partial f^*(\phi)$$

Proposition 6 (Moreau's identity)
Consider $f \in \Gamma_0(\mathcal{X})$ and $s > 0$. Then, for any $x \in \mathcal{X}$,

$$\mathrm{Prox}_{sf}(x) + s\mathrm{Prox}_{s^{-1}f^*}\left(\frac{x}{s}\right) = x$$

Definition 3 A set-valued function $F\colon \mathcal{Z} \rightrightarrows \mathcal{Z}$ is **metrically sub-regular** at z for v if there exists $\gamma > 0$ and a neighborhood $N(z)$ of z such that $\forall z' \in N(z)$,

$$\mathrm{dist}\left(F(z'), v\right) \geq \gamma \mathrm{dist}\left(z', F^{-1}(v)\right)$$

Definition 4 We say that a function $f\colon \mathcal{X} \to \mathbb{R}\cup\{+\infty\}$ has a *quadratic error bound* if there exists η and an open region $\mathcal{R} \subseteq \mathcal{X}$ that contains $\arg\min f$ such that for all $x \in \mathcal{R}$:

$$f(x) - \min f \geq \frac{\eta}{2}\mathrm{dist}(x, \arg\min f)^2$$

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

Proposition 7 (*Proposition 2 in [15]*)
Let f be a convex function such that $f(x) \leq f_0$ implies $\|\partial f(x)\|_0 \geq \eta \mathrm{dist}\,(x, \mathcal{X}^\star)$. Then, $f(x) \geq f(x^\star) + \frac{\eta}{2}\mathrm{dist}\,(x, \mathcal{X}^\star)$ as soon as $f(x) \leq f_0$.

Lemma 4 *Let $M \in \mathbb{R}^{m\times n}$ be a symmetric matrix, then for any vector $x \in \mathbb{R}^n$:*

$$\|Mx\| \geq |\lambda(M)|_{\min}\|x\|$$

where $|\lambda(M)|_{\min}$ is the smallest absolute value of the non-zero eigenvalues of M.

3 Optimality measures

Within this section, our objective is to establish a common understanding by defining the aforementioned optimality measures and the concept of $\varepsilon-$solution. Furthermore, where necessary, we provide a comprehensive analysis supporting these definitions.

Definition 5 ($\varepsilon-$Solution)
Given a target accuracy $\varepsilon > 0$. A point $\hat{x} \in \mathcal{X}$ is said to be an $\varepsilon-$Solution of ($\mathcal{P}$) if:

$$|f(\hat{x}) - f(x^\star)| \leq \varepsilon \qquad \text{and} \qquad \|A\hat{x} - b\| \leq \varepsilon \tag{7}$$

Definition 6 (Optimality Gap and Feasibility error (OGFE))
The *Optimality gap*, and the *Feasibility error* for ($\mathcal{SPP}$) at a point $\hat{x} \in \mathcal{X}$ are defined, respectively, as follows:

$$\mathcal{O}(\hat{x}) = \max\left(f(\hat{x}) - f^\star, 0\right) \qquad \mathcal{F}(\hat{x}) = \|A\hat{x} - b\| \tag{8}$$

One can observe that the definition of the optimality gap and feasibility error aligns precisely with the definition of $\varepsilon-$solutions. Consequently, any combination of the *optimality gap* and the *feasibility error* can be utilized to assess whether a provided solution constitutes an $\varepsilon-$solution. However, due to the uncomputability of the *optimality gap*, determining whether a solution truly qualifies as an ε solution becomes a challenging task.

Next, we introduce the *Karush–Kuhn–Tucker error*, which is quite similar to the outlined definition of the *Infimal Sub-differential Size* as presented in [18]. They employ the same definition as we do but with weighted norms. This definition draws inspiration from the Karush-Kuhn-Tucker conditions applied to ($\mathcal{SPP}$), where the saddle points are identified by having the sub-differential

of the associated Lagrangian equal to 0. Said otherwise, $(x^\star, y^\star)$ qualifies as a saddle point for ($\mathcal{SPP}$) if, and only if:

$$\begin{aligned} \partial_x \mathcal{L}(x^\star, y^\star) &= \partial f(x^\star) + A^T y^\star = 0 && \text{(Stationarity)} \\ \partial_y \mathcal{L}(x^\star, y^\star) &= Ax^\star \quad - b \quad = 0 && \text{(Primal-feasibility)} \end{aligned}$$

Therefore, employing any combination of these two conditions would provide insight into whether the current point functions as a saddle point or not. We utilize the squared norm of the vector that combines these two conditions.

Definition 7 (Karush–Kuhn–Tucker error (KKT))
The Karush-Kuhn-Tucker error for ($\mathcal{SPP}$) is defined as follows:

$$\mathcal{K}(z) := \left\|\partial f(x) + A^T y\right\|_0^2 + \|Ax - b\|^2 \tag{9}$$

where we define the "Infimal size" of a set $\mathcal{Q}$ as:

$$\|\mathcal{Q}\|_0 := \min\{\|q\| \mid q \in \mathcal{Q}\} \tag{10}$$

3.1 Projected Duality Gap

The optimality measure, which we termed as the *projected duality gap*, was initially introduced in [2] only for linear programs. This metric serves as the stopping criterion utilized in the `linprog` solver within SciPy for Python. In this work, we have extended its application to integrate with our ($\mathcal{SPP}$) framework. Essentially, our generalization operates quite similarly to the conventional duality gap. However, it differs in that it computes the duality gap at each iteration while simultaneously projecting the primal-dual solution onto their respective feasibility spaces. Consequently, this method always yields a finite value, unlike the conventional duality gap.

Definition 8 (Projected Duality Gap (PDG))
The Projected Duality Gap for ($\mathcal{SPP}$) is defined as follows:

$$\mathcal{D}(z) := |f(x) + f^*(a) + \langle b, y\rangle|^2 + \left\|a + A^T y\right\|^2 + \|Ax - b\|^2 \tag{11}$$

$$a := \mathrm{Proj}_{\mathrm{dom} f^*}\left(-A^T y\right) \tag{12}$$

This definition is inspired by the aforementioned saddle point problem ($\mathcal{SPP}$):

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \; \mathcal{L}(x, y) := f(x) + \langle Ax - b, y\rangle$$

Thanks to the Fenchel-Legendre transform, we can express the associated primal and dual problems equivalently in terms of it. That is:

$$\begin{aligned} &\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} f(x) + \langle Ax, y\rangle - \langle b, y\rangle \\ &\quad \equiv \min_{x\in\mathcal{X}} f(x) + \imath_{\{b\}}(Ax) \end{aligned}$$

$$\equiv \max_{y \in \mathcal{Y}} -f^*\left(-A^T y\right) - \langle b, y \rangle$$

Therefore, (x, y) is a saddle point if the following conditions hold:

$$\begin{cases} |f(x) + \imath_{\{b\}}(Ax) + f^*\left(-A^T y\right) + \langle b, y \rangle| = 0 \\ Ax \in \mathrm{dom}(\imath_{\{b\}}) \equiv Ax = b \\ -A^T y \in \mathrm{dom}(f^*) \end{cases}$$

Hence, by considering any combination of these conditions, an optimality measure is obtained. We take the squared norm of the vector combining the three conditions.

The fourth measure under consideration is entirely novel and less well-studied compared to the aforementioned ones. This is an area where we invest more time and effort in our study. Consequently, we allocate a dedicated section to comprehensively introduce it, along with deriving some new properties.

4 Smoothed Duality Gap

In this section, we present the last optimality measure along with some new properties. The *smoothed duality gap*, initially introduced in [26], represents a novel measure of optimality that is widely applicable but remains less studied compared to the previously discussed ones.

Definition 9 (*Definition 4 in [15]*) .

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

$$\mathcal{G}_\beta(z; \dot{z}) = \sup_{z' \in \mathcal{Z}} \mathcal{L}(x, y') - \mathcal{L}(x', y) - \frac{\beta_x}{2}\|x' - \dot{x}\|^2 - \frac{\beta_y}{2}\|y' - \dot{y}\|^2 \qquad (13)$$

When the smoothness parameter $\beta = 0$, we recover the conventional duality gap. The smoothed duality gap concept involves smoothing the duality gap through a proximity function [20], thereby ensuring that the smoothed duality gap attains finite values for constrained problems, unlike its conventional counterpart. Additionally, when the smoothness parameter is small and the smoothed duality gap is small, it signifies that both the optimality gap and the feasibility error are also small.

Moreover, the author in [15] has found that the smoothed duality gap offers a robust outcome. Independently of any unknown or uncomputable variables, it serves as a valid optimality measure. Therefore, it could be utilized as a stopping criterion.

Definition 10 (*Definition 5 in [15]*)

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

Theorem 1 (***Proposition 15 in [15]***)
The self-centered smoothed gap is a measure of optimality. Indeed, $\forall z \in \mathcal{Z}, \forall \beta \in [0, +\infty[^2$*: (i)* $\mathcal{G}_\beta(z,z) \geq 0$ *and (ii)* $\mathcal{G}_\beta(z,z) = 0 \iff z \in \mathcal{Z}^\star$

An obstacle in the definition of the smoothed duality gap lies in it constituting an optimization problem in itself, thereby adding complexity. However, for the previously mentioned ($\mathcal{SPP}$), we have managed to derive a closed-form expression for the smoothed duality gap.

Proposition 8 *The self-centered smoothed gap for ($\mathcal{SPP}$) can be computed as follows:*

$$\mathcal{G}_\beta(z) := f(x) - f(p) + \langle A(x-p), y\rangle - \frac{\beta_x}{2}\|p-x\|^2 + \frac{1}{2\beta_y}\|Ax-b\|^2 \tag{14}$$

$$p := \mathrm{Prox}_{\beta_x^{-1} f}\left(x - \frac{1}{\beta_x} A^T y\right) \tag{15}$$

Proof We start with the definition of the smoothed duality gap:

$$\begin{aligned}
\mathcal{G}_\beta(z) &= f(x) + \langle b, y\rangle + \max_{x'} -f(x') - \langle Ax', y\rangle - \frac{\beta_x}{2}\|x'-x\|^2 + \max_{y'}\langle Ax-b, y'\rangle \\
&\qquad\qquad - \frac{\beta_y}{2}\|y'-y\|^2 \\
&= f(x) + \langle b, y\rangle - f(p) - \langle Ap, y\rangle - \frac{\beta_x}{2}\|p-x\|^2 + \langle Ax-b, y\rangle + \frac{1}{2\beta_y}\|Ax-b\|^2 \\
&= f(x) - f(p) + \langle A(x-p), y\rangle - \frac{\beta_x}{2}\|p-x\|^2 + \frac{1}{2\beta_y}\|Ax-b\|^2
\end{aligned}$$

where

$$\begin{aligned}
p &= \arg\max_{x'} -f(x') - \langle Ax', y\rangle - \frac{\beta_x}{2}\|x'-x\|^2 \\
&= \arg\min_{x'} f(x') + \langle x', A^T y\rangle + \frac{\beta_x}{2}\left\|x' - \left(x - \frac{1}{\beta_x}A^T y\right) - \frac{1}{\beta_x}A^T y\right\|^2 \\
&= \arg\min_{x'} f(x') + \langle x', A^T y\rangle + \frac{\beta_x}{2}\left\|x' - \left(x - \frac{1}{\beta_x}A^T y\right)\right\|^2 + \frac{1}{2\beta_x}\|A^T y\|^2 \\
&\qquad\qquad - \langle x', A^T y\rangle + \langle x - \frac{1}{\beta_x}A^T y, A^T y\rangle \\
&= \mathrm{Prox}_{\beta_x^{-1} f}\left(x - \frac{1}{\beta_x}A^T y\right)
\end{aligned}$$

□

Where the red terms cancel, and the green ones are free of x'.

Throughout the remainder of this section, we outline several properties of the self-centered smoothed gap, which will significantly contribute to justifying our subsequent findings. Additionally, we will call $\mathcal{G}_\beta(x^\star, y; x, y^\star)$ and $\mathcal{G}_\beta(x, y^\star; x^\star, y)$ the *outer-saddle* and the *inner-saddle* smoothed gaps, respectively.

Lemma 5 *Given $\beta = (\beta_x, \beta_y) \in [0, +\infty[^2$, then the proximal point, p, defined in (15) satisfies:*

$$p = \text{Prox}_{\beta_x^{-1} f}\left(x - \frac{1}{\beta_x} A^T y\right) \iff \beta_x(x - p) \in \partial f(p) + A^T y$$

Proof Direct implication of Fermat's rule.

Lemma 6 *Given $\beta = (\beta_x, \beta_y) \in [0, +\infty[^2$, then for all $z \in \mathcal{Z}$ the self-centered smoothed gap satisfies:*

$$\mathcal{G}_\beta(z) \geq \frac{\beta_x}{2}\|x - p\|^2 + \frac{1}{2\beta_y}\|Ax - b\|^2$$

Proof By Lemma 3, we know that:

$$p = \text{Prox}_{\beta_x^{-1} f}\left(x - \frac{1}{\beta_x} A^T y\right) \iff \forall v \in \mathcal{X},$$
$$\frac{1}{\beta_x} f(p) + \frac{1}{2}\left\|p - \left(x - \frac{1}{\beta_x} A^T y\right)\right\|^2 \leq \frac{1}{\beta_x} f(v) + \frac{1}{2}\left\|v - \left(x - \frac{1}{\beta_x} A^T y\right)\right\|^2 - \frac{1}{2}\|v - p\|^2$$

Taking $v = x$, we obtain:

$$f(x) - f(p) - \frac{\beta_x}{2}\|x - p\|^2 \geq \frac{\beta_x}{2}\left\|p - \left(x - \frac{1}{\beta_x} A^T y\right)\right\|^2 - \frac{1}{2\beta_x}\left\|A^T y\right\|^2 \quad (16)$$

Thus,

$$\begin{aligned}
\mathcal{G}_\beta(z) &= f(x) - f(p) - \frac{\beta_x}{2}\|p - x\|^2 + \langle A(x - p), y\rangle + \frac{1}{2\beta_y}\|Ax - b\|^2 \\
&\overset{(16)}{\geq} \frac{\beta_x}{2}\left\|p - \left(x - \frac{1}{\beta_x} A^T y\right)\right\|^2 - \frac{1}{2\beta_x}\left\|A^T y\right\|^2 + \langle A(x - p), y\rangle \\
&\qquad + \frac{1}{2\beta_y}\|Ax - b\|^2 \\
&= \frac{\beta_x}{2}\|x - p\|^2 + \frac{1}{2\beta_y}\|Ax - b\|^2
\end{aligned}$$

□

Corollary 1 *Given $\beta = (\beta_x, \beta_y) \in [0, +\infty]^2$. Then, for any $z \in \mathcal{Z}$, the feasibility error defined in (8) could be approximated in terms of the self-centered smoothed duality gap defined in (14) for ($\mathcal{SPP}$). More precisely, for all $z \in \mathcal{Z}$,*

$$\|Ax - b\| \leq \sqrt{2\beta_y \mathcal{G}_\beta(z)} \quad (17)$$

Proof Direct implication of Lemma 6.

Lemma 7 *Given $\beta = (\beta_x, \beta_y) \in [0, +\infty]^2$, then for all $z \in \mathcal{Z}$, the self-centered smoothed duality gap satisfies:*

$$\mathcal{G}_\beta(z) \geq \frac{\beta_x}{2}\|x_\beta(y) - x\|^2 + \frac{\beta_y}{2}\|y_\beta(x) - y\|^2 \tag{18}$$

$$x_\beta(y) := \arg\max_{x'} -f(x') - \langle Ax' - b, y\rangle - \frac{\beta_x}{2}\|x' - x\|^2 \tag{19}$$

$$y_\beta(x) := \arg\max_{y'} \langle Ax - b, y'\rangle - \frac{\beta_y}{2}\|y' - y\|^2 \tag{20}$$

Proof

$$\begin{aligned}\mathcal{G}_\beta(z) &= f(x) + \sup_{x'} -f(x') - \langle Ax' - b, y\rangle - \frac{\beta_x}{2}\|x' - x\|^2 + \sup_{y'} \langle Ax - b, y'\rangle \\ &\qquad - \frac{\beta_y}{2}\|y' - y\|^2 \\ &= f(x) + F_1(x_\beta(y)) + F_2(y_\beta(x))\end{aligned}$$

Where

$$\begin{aligned}F_1(\mu) &= -f(\mu) - \langle A\mu - b, y\rangle - \frac{\beta_x}{2}\|\mu - x\|^2 \\ F_2(\nu) &= \langle Ax - b, \nu\rangle - \frac{\beta_y}{2}\|\nu - y\|^2\end{aligned}$$

One can observe that F_1 and F_2 are $\frac{\beta_x}{2}$ and $\frac{\beta_y}{2}$−strongly concave, respectively. Hence, by Lemma 1:

$$\begin{aligned}F_1(x_\beta(y)) &\geq F_1(x) + \frac{\beta_x}{2}\|x_\beta(y) - x\|^2 \\ F_2(y_\beta(x)) &\geq F_2(y) + \frac{\beta_y}{2}\|y_\beta(x) - y\|^2\end{aligned} \tag{21}$$

Thus,

$$\begin{aligned}\mathcal{G}_\beta(z) &= f(x) + F_1(x_\beta(y)) + F_2(y_\beta(x)) \\ &\overset{(21)}{\geq} f(x) - f(x) - \langle Ax - b, y\rangle + \frac{\beta_x}{2}\|x_\beta(y) - x\|^2 + \langle Ax - b, y\rangle + \frac{\beta_y}{2}\|y_\beta(x) - y\|^2 \\ &= \frac{\beta_x}{2}\|x_\beta(y) - x\|^2 + \frac{\beta_y}{2}\|y_\beta(x) - y\|^2 \qquad \square\end{aligned}$$

Corollary 2 *Given $\beta = (\beta_x, \beta_y) \in [0, +\infty]^2, z \in \mathcal{Z}$, and $z^\star \in \mathcal{Z}^\star$, then the outer-saddle smoothed gap satisfies:*

$$\mathcal{G}_\beta(x^\star, y; x, y^\star) \geq -\frac{\beta_x}{2}\|x - x^\star\|^2 + \frac{\beta_x}{2}\|x_\beta(y) - x^\star\|^2$$

$$x_\beta(y) = \arg\max_{x'} -f(x') - \langle Ax' - b, y\rangle - \frac{\beta_x}{2}\|x' - x\|^2$$

Proof

$$\begin{aligned}\mathcal{G}_\beta(x^\star, y; x, y^\star) &= f(x^\star) + \sup_{x'} \left(-f(x') - \langle Ax' - b, y\rangle - \frac{\beta_x}{2}\|x' - x\|^2 \right) \\ &\qquad + \sup_{y'} -\frac{\beta_y}{2}\|y' - y^\star\|^2 \\ &= f(x^\star) - f(x_\beta(y)) - \langle Ax_\beta(y) - b, y\rangle - \frac{\beta_x}{2}\|x_\beta(y) - x\|^2 + 0 \\ &= f(x^\star) + F(x_\beta(y))\end{aligned}$$

Where

$$F(\mu) = -f(\mu) - \langle A\mu - b, y\rangle - \frac{\beta_x}{2}\|\mu - x\|^2$$

Since F is $\frac{\beta_x}{2}-$Strongly-concave and its maximum is attained at $x_\beta(y)$, then by Lemma 1:

$$F(x_\beta(y)) \geq F(x^\star) + \frac{\beta_x}{2}\|x_\beta(y) - x^\star\|^2$$

Therefore,

$$\begin{aligned}\mathcal{G}_\beta(x^\star, y; x, y^\star) &\geq f(x^\star) + F(x^\star) + \frac{\beta_x}{2}\|x_\beta(y) - x^\star\|^2 \\ &= f(x^\star) - f(x^\star) - \langle Ax^\star - b, y\rangle - \frac{\beta_x}{2}\|x^\star - x\|^2 + \frac{\beta_x}{2}\|x_\beta(y) - x\|^2 \\ &= -\frac{\beta_x}{2}\|x - x^\star\|^2 + \frac{\beta_x}{2}\|x_\beta(y) - x\|^2 \qquad \square\end{aligned}$$

Lemma 8 *Let $z^\star \in \mathcal{Z}^\star$, then for any $z \in \mathcal{Z}$, the self-centered smoothed duality gap can be decomposed in terms of the outer and inner saddle smoothed gaps as:*

$$\mathcal{G}_\beta(z) = \mathcal{G}_\beta(x, y^\star; x^\star, y) + \mathcal{G}_\beta(x^\star, y; x, y^\star)$$

Proof By using the definition of SDG twice, first:

$$\begin{aligned}\mathcal{G}_\beta(x, y^\star; x^\star, y) &= f(x) + \sup_{x'} -f(x') - \langle Ax' - b, y^\star\rangle - \frac{\beta_x}{2}\|x' - x^\star\|^2 \\ &\qquad + \sup_{y'} \langle Ax - b, y'\rangle - \frac{\beta_y}{2}\|y' - y\|^2 \\ &= f(x) - f(x^\star) + \sup_{y'} \langle Ax - b, y'\rangle - \frac{\beta_y}{2}\|y' - y\|^2\end{aligned}$$

Second:

$$\begin{aligned}\mathcal{G}_\beta(x^\star, y; x, y^\star) &= f(x^\star) + \sup_{x'} -f(x') - \langle Ax' - b, y\rangle - \frac{\beta_x}{2}\|x' - x\|^2 \\ &\qquad + \sup_{y'} -\frac{\beta_y}{2}\|y' - y^\star\|^2 \\ &= f(x^\star) + \sup_{x'} -f(x') - \langle Ax' - b, y\rangle - \frac{\beta_x}{2}\|x' - x\|^2 + 0 \qquad \square\end{aligned}$$

Summing the two terms implies the result.

Lemma 9 *Given $\beta = (\beta_x, \beta_y) \in [0, +\infty]^2, z \in \mathcal{Z}$, and $z^\star \in \mathcal{Z}^\star$, then the outer-saddle smoothed gap satisfies:*

$$\mathcal{G}_\beta(x^\star, y; x, y^\star) \geq -2\sqrt{\beta_x \mathcal{G}_\beta(z)}\|x - x^\star\|$$

Proof By corollary 2, we know that:

$$\begin{aligned}
\mathcal{G}_\beta(x^\star, y; x, y^\star) &\geq -\frac{\beta_x}{2}\|x - x^\star\|^2 + \frac{\beta_x}{2}\|x_\beta(y) - x + x - x^\star\|^2 \\
&= \frac{\beta_x}{2}\|x_\beta(y) - x\|^2 + \beta_x \langle x_\beta(y) - x, x - x^\star \rangle \\
&\geq \beta_x \langle x_\beta(y) - x, x - x^\star \rangle \\
&\overset{(CS)}{\geq} -\beta_x \|x_\beta(y) - x\| \|x - x^\star\| \\
&\overset{(18)}{\geq} -\sqrt{2\beta_x \mathcal{G}_\beta(z)}\|x - x^\star\|
\end{aligned}$$

□

This last lemma will play a crucial role in establishing an upper bound for the optimality gap in terms of the smoothed duality gap as we will elucidate in the subsequent section.

5 Optimality Gap Bounds

In our earlier discussion regarding the definition of an ε-solution, a solution is considered as an ε-solution if both the optimality gap and the feasibility error are below ε. While the computation of the feasibility error is straightforward, the same cannot be said for the optimality gap. Consequently, determining whether the optimality gap is indeed less than ε or not poses a more complicated challenge.

This section aims to establish computable approximations for the uncomputable measure, *optimality gap*, by setting upper bounds in terms of the aforementioned computable ones: the KKT error, PDG, and SDG. The initial approach involves attempting to set an upper bound on the optimality gap in terms of the KKT error defined in equation (9). That is:

$$\mathcal{O}(x) = f(x) - f^\star \overset{?}{\leq} \mathcal{W}(\mathcal{K}(z))$$

For instance, this $\mathcal{W}$ could be $\mathcal{K}^2$ or $c\mathcal{K}$ for any scalar c, and so forth. Consequently, if $\mathcal{W}(\mathcal{K}(z)) \leq \varepsilon$, it follows that $\mathcal{O}(x) \leq \varepsilon$ as well. Therefore, we will attain an ε_g-solution with $\varepsilon \leq \varepsilon_g$, depending on the tightness of our subsequent bounds.

One possible starting point is by taking a vector that satisfies the stationarity property of the Lagrangian:

$$\begin{aligned}
q \in \partial f(x) + A^T y &\iff \forall u \in \mathcal{X},\ f(u) \geq f(x) + \langle q - A^T y, u - x \rangle \\
&\overset{u=x^\star}{\implies} f(x) - f(x^\star) \leq \langle q - A^T y, x - x^\star \rangle
\end{aligned}$$

$$\iff f(x) - f(x^\star) \leq \langle q, x - x^\star \rangle + \langle -y, A(x - x^\star) \rangle$$
$$\overset{Ax^\star = b}{\iff} f(x) - f(x^\star) \leq \underbrace{\|q\|\|x - x^\star\| + \|y\|\|Ax - b\|}_{\text{Initial bound}} \quad (22)$$

Yet, our efforts have resulted in an expression that remains depending on the unknown quantity, $x^\star$. This implies that unless we can eliminate this problematic term, our progress will not surpass the limitations of the optimality gap itself. Furthermore, when we attempted a similar approach with the *projected duality gap* and the *smoothed duality gap*, we encountered an identical issue:

❖ Projected duality gap

$$f(x) - f^\star \leq |f(x) + f^*(a) + \langle b, y \rangle| + \|x^\star\| \left\|a + A^T y\right\| \leq (1 + \|x^\star\|) \sqrt{\mathcal{D}(z)}$$

❖ Smoothed duality gap

$$\begin{aligned} f(x) - f^\star &= f(x) + \langle Ax - b, y \rangle + \frac{1}{2\beta_y} \|Ax - b\|^2 - f^\star - \langle Ax - b, y \rangle \\ &\qquad - \frac{1}{2\beta_y} \|Ax - b\|^2 \\ &= \mathcal{G}_\beta(z) - \mathcal{G}_\beta(x^\star, y; x, y^\star) - \langle Ax - b, y \rangle - \frac{1}{2\beta_y} \|Ax - b\|^2 \\ &\leq \mathcal{G}_\beta(z) + \sqrt{2\beta_x \mathcal{G}_\beta(z)} \|x - x^\star\| - \ldots\ldots \end{aligned}$$

In each scenario, we continuously encountered the presence of the blue annoying unknown term. Nevertheless, we successfully transformed the initial bounds of the optimality gap into final ones by introducing additional assumptions that effectively eliminate the aforementioned blue annoying term.

Just to remark, more detailed proofs of each step of what we did in our initial bounds will be presented later on in this section.

5.1 Regularity assumptions

In this subsection, we present the regularity assumptions that we employ to eliminate the blue annoying term from our initial bounds of the optimality gap. The first assumption is the *metric sub-regularity* of the sub-differential of the Lagrangian, it was first introduced in the works of [17]. The approach involves applying the definition of metric sub-regularity *(Definition 3)* to the sub-differential of the Lagrangian of ($\mathcal{SPP}$) (*i.e.* $F = \partial\mathcal{L}$ and $v = 0$). Formally, this can be expressed as follows:

Assumption 1 (MSRSDL, [17])
The metric sub-regularity of the sub-differential of the Lagrangian (MSRSDL) assumes that: there exists $\gamma > 0$ such that

$$\|\partial_x \mathcal{L}(x, y)\|_0 + \|\nabla_y \mathcal{L}(x, y)\| \geq \gamma \text{dist}(x, \mathcal{X}^\star) + \gamma \text{dist}(y, \mathcal{Y}^\star) \quad (23)$$

As we can observe, assuming the metric sub-regularity of the sub-differential of the Lagrangian provides an upper bound for the earlier identified blue annoying term in terms of the KKT error.

The second assumption we introduce is referred to as the *quadratic error bound of the smoothed gap*. While the *quadratic error bound* is a widely recognized and commonly used assumption in general contexts, its application to the smoothed duality gap is relatively recent, first introduced in [15]. It's as broadly applicable as the metric sub-regularity of the Lagrangian's sub-differential, and it also serves as an upper bound for the blue annoying term in terms of the smoothed duality gap.

Assumption 2 (QEBSG, *Proposition 15 (iii)* in [15])
The quadratic error bound of the smoothed gap (QEBSG) assumes that: there exists $\beta = (\beta_x, \beta_y) \in]0, +\infty]^2, \eta > 0$ and a region $\mathcal{R} \subseteq \mathcal{Z}$ such that $\mathcal{G}_\beta$ has a *quadratic error bound* with constant η in the region $\mathcal{R}$. Said otherwise, for all $z \in \mathcal{R}$:

$$\mathcal{G}_\beta(z) \geq \frac{\eta}{2} \operatorname{dist}(z, \mathcal{Z}^\star)^2 \tag{24}$$

5.2 Final bounds

In the subsection, we present our final bounds of the optimality gap by assuming the aforementioned regularity assumptions.

Theorem 2 *Let $f \in \Gamma_0(\mathcal{X}), x^\star = \operatorname{Proj}_{\mathcal{X}^\star}(x)$, and $q = \left\|\partial f(x) + A^T y\right\|_0$. Then, under Assumption 1, the optimality gap defined in (8) could be approximated in terms of the Karush–Kuhn–Tucker error defined in (9) for ($\mathcal{SPP}$). More precisely, for all $z \in \mathcal{Z}$,*

$$f(x) - f^\star \leq \frac{2}{\gamma}\mathcal{K}(z) + \|y\|\sqrt{\mathcal{K}(z)} \tag{25}$$

Proof As we have seen in our initial bound (eqn:22), taking $q \in \partial f(x) + A^T y$ yields:

$$\begin{aligned}
f(x) - f^\star &\leq \|q\|\|x - x^\star\| + \|y\|\|Ax - b\| \\
&\overset{(23)}{\leq} \left\|\partial f(x) + A^T y\right\|_0 \frac{1}{\gamma}\left(\|\partial f(x) + A^T y\|_0 + \|Ax - b\|\right) + \|y\|\|Ax - b\| \\
&= \frac{1}{\gamma}\left\|\partial f(x) + A^T y\right\|_0^2 + \frac{1}{\gamma}\left\|\partial f(x) + A^T y\right\|_0 \|Ax - b\| + \|y\|\|Ax - b\| \\
&\overset{(9)}{\leq} \frac{1}{\gamma}\mathcal{K}(z) + \frac{1}{\gamma}\sqrt{\mathcal{K}(z)}\sqrt{\mathcal{K}(z)} + \|y\|\sqrt{\mathcal{K}(z)} \\
&= \frac{2}{\gamma}\mathcal{K}(z) + \|y\|\sqrt{\mathcal{K}(z)}
\end{aligned}$$

□

Remark 1 An interesting observation in this first finding pertains to the term $\left(\frac{2}{\gamma}\mathcal{K}(z)\right)$ that depends on γ. The metric sub-regularity constant, γ, typically

takes very small values, such as 10^{-8}, 10^{-10}, or even smaller. So, having $\frac{1}{\gamma}$ multiplied with $\mathcal{K}(z)$ rather than $\sqrt{\mathcal{K}(z)}$ yields a tighter and more efficient bound where the algorithm will require fewer iterations to beat $\frac{1}{\gamma}$ before identifying an $\varepsilon-$solution.

Counter-example 1 In Theorem 2, we derived an upper bound for the optimality gap based on the KKT error. It is important to note, however, that the reverse relationship may not always hold. The following counterexample illustrates this point:

Let $\varepsilon > 0$, and consider the unconstrained optimization problem $\min_{x\in\mathbb{R}} f(x)$ where $f\colon \mathbb{R} \to \mathbb{R}$ is defined as:

$$f(x) = \begin{cases} x & \text{If } x > \varepsilon \\ \frac{x^2}{2\varepsilon} + \frac{\varepsilon}{2} & \text{If } x \leq \varepsilon \end{cases}$$

Then,

$$f'(x) = \begin{cases} 1 & \text{If } x > \varepsilon \\ \frac{x}{\varepsilon} & \text{If } x \leq \varepsilon \end{cases}$$

One can observe that, $\forall \varepsilon > 0$:

$$f'(\varepsilon) = 1 \qquad \textbf{but} \qquad f(\varepsilon) - f(0) = \frac{\varepsilon^2}{2\varepsilon} + \frac{\varepsilon}{2} - \frac{\varepsilon}{2} = \frac{\varepsilon}{2} \xrightarrow[\varepsilon\to 0]{} 0 \qquad \square$$

which concludes the example.

Theorem 3 *Let $f \in \Gamma_0(\mathcal{X}), x^\star = \mathrm{Proj}_{\mathcal{X}^\star}(x)$, and $\beta = (\beta_x, \beta_y) \in (0, +\infty)^2$. Then, under Assumption 2, the optimality gap defined in (8) could be approximated in terms of the self-centered smoothed duality gap defined in (14) for (SPP). More precisely, for all $z \in \mathcal{Z}$,*

$$f(x) - f^\star \leq \left(1 + \sqrt{\frac{2\beta_x}{\eta}}\right) \mathcal{G}_\beta(z) + \sqrt{2\beta_y}\|y\|\sqrt{\mathcal{G}_\beta(z)} \tag{26}$$

Proof We start by rewriting the optimality gap in a decomposed way:

$$\begin{aligned} f(x) - f^\star &= f(x) + \langle Ax - b, y\rangle + \frac{1}{2\beta_y}\|Ax - b\|^2 - f(x^\star) - \langle Ax - b, y\rangle \\ &\qquad - \frac{1}{2\beta_y}\|Ax - b\|^2 \\ &= \mathcal{G}_\beta(x, y^\star; x^\star, y) - \langle Ax - b, y\rangle - \frac{1}{2\beta_y}\|Ax - b\|^2 \\ &\overset{(8)}{=} \mathcal{G}_\beta(z) - \mathcal{G}_\beta(x^\star, y; x, y^\star) - \langle Ax - b, y\rangle - \frac{1}{2\beta_y}\|Ax - b\|^2 \\ &\overset{(9)}{\leq} \mathcal{G}_\beta(z) + \sqrt{2\beta_x \mathcal{G}_\beta(z)}\|x - x^\star\| - \langle Ax - b, y\rangle - \frac{1}{2\beta_y}\|Ax - b\|^2 \end{aligned}$$

$$\overset{(24)}{\leq} \mathcal{G}_\beta(z) + \sqrt{2\beta_x \mathcal{G}_\beta(z)}\sqrt{\frac{2}{\eta}\mathcal{G}_\beta(z)} - \langle Ax - b, y\rangle - 0$$

$$\leq \left(1 + 2\sqrt{\frac{\beta_x}{\eta}}\right)\mathcal{G}_\beta(z) + \|y\|\|Ax - b\|$$

$$\overset{(17)}{\leq} \left(1 + 2\sqrt{\frac{\beta_x}{\eta}}\right)\mathcal{G}_\beta(z) + \|y\|\sqrt{2\beta_y \mathcal{G}_\beta(z)} \qquad \square$$

Remark 2 The quadratic error bound constant, η, exhibits a similar characteristic of taking very small values, akin to the metric sub-regularity constant. In Theorem 3, we managed to establish an upper bound that depends on the square root of η, while maintaining its multiplication with $\mathcal{G}_\beta(z)$ instead of $\sqrt{\mathcal{G}_\beta(z)}$. Nevertheless, it is noteworthy that an alternative bound could be formulated directly in terms of η itself by using Corollary 2.

Theorem 4 *Let $f \in \Gamma_0(\mathcal{X}), x^\star = \mathrm{Proj}_{\mathcal{X}^\star}(x)$, $\beta = (\beta_x, \beta_y) \in (0, +\infty)^2$, and $\beta_{\min} = \min(\beta_x, \beta_y)$. Then, under Assumption 2, the optimality gap defined in (8) could be approximated in terms of the projected duality gap defined in (11) for (SPP). More precisely, for all $z \in \mathcal{Z}$,*

$$f(x) - f^\star \leq \left(1 + \|x\| + \sqrt{\frac{2}{\eta}}\sqrt{(1 + \|x\| + \|y\|)\sqrt{\mathcal{D}(z)} + \frac{1}{2\beta_{\min}}\mathcal{D}(z)}\right)\sqrt{\mathcal{D}(z)} \tag{27}$$

In the proof of this theorem, we will use some arguments that will come later in this paper. So, we will provide its proof later on as well.

Remark 3 Designing a new regularity assumption that inherently fits PDG is out of the scope of the paper. Instead, we take advantage of assuming QEBSG to derive our result.

6 Comparability Bounds

Our subsequent objective is motivated by the belief that the newly proposed optimality measure, the smoothed duality gap, might serve as a more appropriate stopping criterion compared to the others. This belief comes to light from the smoothed duality gap's definition; whereas the Karush–Kuhn–Tucker error relies on the sub-differential of the objective function to assess optimality, the smoothed duality gap is more directly based on the objective function itself. Hence, in this section, our goal is to conduct a comparative analysis between the smoothed duality gap, the Karush–Kuhn–Tucker error, and the projected duality gap. We aim to investigate and identify the conditions under which these measures could serve as approximations for the smoothed duality gap and vice versa.

6.1 SDG – KKT bounds

In this subsection, we start our comparative analysis between the smoothed duality gap and the Karush–Kuhn–Tucker error. Initially, we present the upper bound obtained for SDG in terms of the KKT error. Subsequently, we demonstrate the reverse relationship.

Theorem 5 *Given $\beta = (\beta_x, \beta_y) \in (0, +\infty)^2$, $z \in \mathcal{Z}$, and a function $f \in \Gamma_0(\mathcal{X})$. Assume $\|q\| = \left\|\partial f(x) + A^T y\right\|_0$. Then, for the Karush–Kuhn–Tucker error and the smoothed duality gap defined, respectively, in (9) and (14) we have:*

$$\mathcal{G}_\beta(z) \leq \underline{\beta}\mathcal{K}(z) \tag{28}$$

$$\underline{\beta} = \max\left(\frac{1}{\beta_x}, \frac{1}{2\beta_y}\right) \tag{29}$$

Proof First of all, let us observe the following:

❖ Using the definition of the sub-differential, we obtain:

$$\begin{aligned} q \in \partial f(x) + A^T y &\iff \forall u \in \mathbb{R}^n,\ f(u) \geq f(x) + \langle q - A^T y, u - x\rangle \\ &\overset{u=p}{\implies} f(x) - f(p) \leq \langle q - A^T y, x - p\rangle \end{aligned} \tag{30}$$

❖ From Lemma 5, we know that

$$p = \mathrm{Prox}_{\beta_x^{-1} f}\left(x - \frac{1}{\beta_x} A^T y\right) \iff \beta_x(x - p) \in \partial f(p) + A^T y$$

Hence, by Lemma 2, for any $q \in \partial f(x) + A^T y$, we get:

$$\begin{aligned} \langle \beta_x(x - p) - q, p - x\rangle \geq 0 &\iff \langle q, x - p\rangle \geq \beta_x \|p - x\|^2 \\ &\iff \|q\| \geq \beta_x \|p - x\| \end{aligned} \tag{31}$$

Now, from the definition of the SDG:

$$\begin{aligned} \mathcal{G}_\beta(z) &= f(x) - f(p) + \langle A(x - p), y\rangle - \frac{\beta_x}{2}\|p - x\|^2 + \frac{1}{2\beta_y}\|Ax - b\|^2 \\ &\overset{(30)}{\leq} \langle q - A^T y, x - p\rangle + \langle A(x - p), y\rangle - \frac{\beta_x}{2}\|p - x\|^2 + \frac{1}{2\beta_y}\|Ax - b\|^2 \\ &\leq \langle q, x - p\rangle + \frac{1}{2\beta_y}\|Ax - b\|^2 \\ &\leq \|q\|\|x - p\| + \frac{1}{2\beta_y}\|Ax - b\|^2 \\ &\overset{(31)}{\leq} \|q\|\frac{1}{\beta_x}\|q\| + \frac{1}{2\beta_y}\|Ax - b\|^2 \\ &\overset{(29)}{\leq} \underline{\beta}\left[\left\|\partial f(x) + A^T y\right\|^2 + \|Ax - b\|^2\right] \\ &= \underline{\beta}\mathcal{K}(z) \end{aligned}$$

□

Theorem 6 *Given $\beta = (\beta_x, \beta_y) \in (0, +\infty)^2$, $z \in \mathcal{Z}$, and a differentiable function $f \in \Gamma_0(\mathcal{X})$ that has an $L-$Lipschitz gradient. Then, for the Karush–Kuhn–Tucker error and the smoothed duality gap defined, respectively, in (9) and (14) we have:*

$$\mathcal{K}(z) \leq \bar{\beta}_L \mathcal{G}_\beta(z) \tag{32}$$

$$\bar{\beta}_L = \max\left(\frac{2(L+\beta_x)^2}{\beta_x}, 2\beta_y\right) \tag{33}$$

Proof We start by observing the following:

❖ From Lemma 5, we know that

$$p = \mathrm{Prox}_{\beta_x^{-1} f}\left(x - \frac{1}{\beta_x} A^T y\right) \iff \beta_x(x-p) \in \partial f(p) + A^T y$$

Hence, keeping in mind that f is differentiable, we get:

$$\beta_x \|x-p\| = \left\|\nabla f(p) + A^T y\right\| \tag{34}$$

❖ Since the gradient of f is $L-$Lipschitz, then by the reverse triangle inequality, we get:

$$\begin{aligned}\left\|\nabla f(x) + A^T y\right\| - \left\|\nabla f(p) + A^T y\right\| &\leq \left\|\left(\nabla f(x) + A^T y\right) - \left(\nabla f(p) + A^T y\right)\right\| \\ &\leq L\|x-p\|\end{aligned}$$

Hence,

$$\begin{aligned}\left\|\nabla f(x) + A^T y\right\| &\leq L\|x-p\| + \left\|\nabla f(p) + A^T y\right\| \\ &\overset{(34)}{\leq} L\|x-p\| + \beta_x\|x-p\| \\ &= (L+\beta_x)\|x-p\|\end{aligned} \tag{35}$$

Therefore, starting from Lemma 6, we get:

$$\begin{aligned}\mathcal{G}_\beta(z) &\geq \frac{\beta_x}{2}\|x-p\|^2 + \frac{1}{2\beta_y}\|Ax-b\|^2 \\ &\overset{(35)}{\geq} \frac{\beta_x}{2}\frac{\|\nabla f(x) + A^T y\|^2}{(L+\beta_x)^2} + \frac{1}{2\beta_y}\|Ax-b\|^2 \\ &\geq \min\left(\frac{\beta_x}{2(L+\beta_x)^2}, \frac{1}{2\beta_y}\right)\left[\|\nabla f(x) + A^T y\|^2 + \|Ax-b\|^2\right]\end{aligned}$$

which implies the result:

$$\mathcal{K}(z) \leq \bar{\beta}_L \mathcal{G}_\beta(z) \qquad \square$$

Counter-example 2 Theorem 6 necessitates the assumption that the objective function is differentiable and has an L-Lipschitz gradient. Without these conditions, Theorem 6 may not hold. This is exemplified in this counterexample.

Let $\beta = (\beta_x, \beta_y) = (1, 1)$, and consider the unconstrained optimization problem $\min_{x\in\mathbb{R}} |x|$.

Then, the associated Lagrangian is $\mathcal{L}(z) = \mathcal{L}(x) = |x|$. Hence,

❖ The smoothed duality gap is defined as:

$$\begin{aligned}\mathcal{G}_\beta(z) = \mathcal{G}_\beta(x) &= |x| - \min_{x'} |x'| + \frac{1}{2}(x' - x)^2 \\ &= |x| - \left|\text{Prox}_{|.|}(x)\right| - \frac{1}{2}\left(\text{Prox}_{|.|}(x) - x\right)^2 \\ &= |x| - \left|[|x| - 1]_+ \text{sgn}(x)\right| - \frac{1}{2}\left([|x| - 1]_+ \text{sgn}(x) - x\right)^2 \\ &= |x| - [|x| - 1]_+ - \frac{1}{2}\left([|x| - 1]_+ \text{sgn}(x) - x\right)^2\end{aligned}$$

❖ The KKT error is defined as:

$$\mathcal{K}(z) = \mathcal{K}(x) = \|\partial f(x)\|_0^2 \text{ with } \partial f(x) = \begin{cases} -1 & \text{If } x < 0 \\ [-1, 1] & \text{If } x = 0 \\ 1 & \text{If } x > 0 \end{cases}$$

Then,

$$\lim_{x\to 0} \mathcal{K}(x) = 1 \qquad \textbf{but} \qquad \lim_{x\to 0} \mathcal{G}_\beta(x) = \lim_{x\to 0} |x| - \frac{1}{2}x^2 = 0 \qquad \square$$

which concludes the example.

6.2 SDG – PDG bounds

In this subsection, we proceed with our comparative analysis, this time between the smoothed duality gap and the projected duality gap. Firstly, we present the upper bound acquired for the smoothed duality gap in terms of the projected duality gap. Following this, we illustrate the converse relationships. Furthermore, we will revisit Theorem 4 and derive its proof.

Theorem 7 *Given $\beta = (\beta_x, \beta_y) \in (0, +\infty)^2$, $z \in \mathcal{Z}$, and a function $f \in \Gamma_0(\mathcal{X})$. Let $\beta_{\min} = \min(\beta_x, \beta_y)$. Then, for the projected duality gap and the smoothed duality gap defined, respectively, in (11) and (14) we have:*

$$\mathcal{G}_\beta(z) \leq (1 + \|x\| + \|y\|)\sqrt{\mathcal{D}(z)} + \frac{1}{2\beta_{\min}}\mathcal{D}(z) \tag{36}$$

Proof By the Fenchel-conjugate definition, we have:

$$f^*(a) = \sup_x \ \langle a, x\rangle - f(x) \overset{x=p}{\geq} \langle a, p\rangle - f(p) \Longrightarrow -f(p) \leq f^*(a) - \langle p, a\rangle \quad (37)$$

Thus,

$$\begin{aligned}
\mathcal{G}_\beta(z) &= f(x) - f(p) + \langle A(x-p), y\rangle - \frac{\beta_x}{2}\|x-p\|^2 + \frac{1}{2\beta_y}\|Ax-b\|^2 \\
&\overset{(37)}{\leq} f(x) + f^*(a) + \langle Ax, y\rangle - \left\langle p, a + A^T y\right\rangle - \frac{\beta_x}{2}\|x-p\|^2 + \frac{1}{2\beta_y}\|Ax-b\|^2 \\
&= f(x) + f^*(a) + \langle Ax, y\rangle - \left\langle x, a + A^T y\right\rangle + \left\langle x-p, a + A^T y\right\rangle \\
&\qquad - \frac{\beta_x}{2}\|x-p\|^2 + \frac{1}{2\beta_y}\|Ax-b\|^2
\end{aligned}$$

Now, by making use of Young's inequality *(Proposition 1)* with $\mathbf{u} = x - p$, $\mathbf{v} = a + A^T y$ and $\lambda = \sqrt{\beta_x}$, we get:

$$\left\langle x-p, a + A^T y\right\rangle \leq \frac{\beta_x}{2}\|x-p\|^2 + \frac{1}{2\beta_x}\left\|a + A^T y\right\|^2 \quad (38)$$

Therefore,

$$\begin{aligned}
\mathcal{G}_\beta(z) &= f(x) + f^*(a) + \langle Ax, y\rangle - \left\langle x, a + A^T y\right\rangle + \left\langle x-p, a + A^T y\right\rangle \\
&\qquad - \frac{\beta_x}{2}\|x-p\|^2 + \frac{1}{2\beta_y}\|Ax-b\|^2 \\
&\overset{(38)}{\leq} f(x) + f^*(a) + \langle Ax, y\rangle - \left\langle x, a + A^T y\right\rangle + \frac{\beta_x}{2}\|x-p\|^2 \\
&\qquad + \frac{1}{2\beta_x}\left\|a + A^T y\right\|^2 - \frac{\beta_x}{2}\|x-p\|^2 + \frac{1}{2\beta_y}\|Ax-b\|^2 \\
&= f(x) + f^*(a) + \langle b, y\rangle + \langle Ax - b, y\rangle - \left\langle x, a + A^T y\right\rangle + \frac{1}{2\beta_x}\left\|a + A^T y\right\|^2 \\
&\qquad + \frac{1}{2\beta_y}\|Ax-b\|^2 \\
&\leq |f(x) + f^*(a) + \langle b, y\rangle| + \|Ax-b\|\,\|y\| + \|x\|\left\|a + A^T y\right\| \\
&\qquad + \frac{1}{2\beta_x}\left\|a + A^T y\right\|^2 + \frac{1}{2\beta_y}\|Ax-b\|^2 \\
&\leq |f(x) + f^*(a) + \langle b, y\rangle| + \|Ax-b\|\,\|y\| + \|x\|\left\|a + A^T y\right\| \\
&\qquad + \frac{1}{2\beta_{\min}}\left(\left\|a + A^T y\right\|^2 + \|Ax-b\|^2\right) \\
&\overset{(11)}{\leq} (1 + \|x\| + \|y\|)\sqrt{\mathcal{D}(z)} + \frac{1}{2\beta_{\min}}\mathcal{D}(z)
\end{aligned}$$

□

Now, we will derive Theorem 4. You'll notice in the proof that at a certain stage, we will establish a bound in terms of the smoothed duality gap. Therefore, it was necessary to acquire an approximation for the smoothed duality

gap in terms of the projected duality gap before completing the proof. With the assistance of Theorem 7, we now possess this approximation and can proceed to derive Theorem 4.

Proof Starting with the definition of the Fenchel-Conjugate function at an optimal solution, we get:

$$\begin{aligned} f(x^\star) = \sup_{\mu\in\mathcal{X}} \langle \mu, x^\star\rangle - f^*(\mu) &\overset{\mu=a}{\geq} \langle a, x^\star\rangle - f^*(a) \\ &= \left\langle x^\star, -A^T y\right\rangle - f^*(a) + \left\langle x^\star, a + A^T y\right\rangle \\ &\geq -\langle b, y\rangle - f^*(a) - \|x^\star\| \left\|a + A^T y\right\| \end{aligned}$$

Thus,

$$\begin{aligned} f(x) - f^\star &\leq |f(x) + f^*(a) + \langle b, y\rangle| + \|x^\star\| \left\|a + A^T y\right\| \\ &\overset{(11)}{\leq} (1 + \|x^\star\|)\sqrt{\mathcal{D}(z)} \\ &\leq (1 + \|x\| + \|x - x^\star\|)\sqrt{\mathcal{D}(z)} \\ &\overset{(24)}{\leq} \left(1 + \|x\| + \sqrt{\frac{2\mathcal{G}_\beta(z)}{\eta}}\right)\sqrt{\mathcal{D}(z)} \\ &\overset{(36)}{\leq} \left(1 + \|x\| + \sqrt{\frac{2}{\eta}}\sqrt{(1 + \|x\| + \|y\|)\sqrt{\mathcal{D}(z)} + \frac{1}{2\beta_{\min}}\mathcal{D}(z)}\right)\sqrt{\mathcal{D}(z)} \end{aligned}$$

□

Our latest finding establishes the upper bound of the projected duality gap in terms of the smoothed duality gap. This last part is quite technical and incorporates manifold and convex optimization concepts. Therefore, we outline the main result here, deferring the detailed proofs to Appendix B.

Theorem 8 *Given $\beta = (\beta_x, \beta_y) \in (0, +\infty)^2$, $z \in \mathcal{Z}$, and a function $f \in \Gamma_0(\mathcal{X})$. Then, under the following set of assumptions (we denote it $\mathcal{E}$):*

- *The Fenchel-Conjugate of the objective function, f^*, could be written in a separable way:*

$$f^*(\mu) = f_1^*(\mu_1) + f_2^*(\mu_2), \qquad \mu \in \mathcal{X}$$

- *f_1^* is $L_{f_1^*}-$Lipschitz on its domain, $\mathrm{dom} f_1^*$.*
- *The domain of f_2^* is a non-empty affine space.*
- *Let $\mu_0 \in \mathrm{dom} f_2^*$, then $\forall \mu_2 \in \mathrm{dom} f_2^*$, we define*

$$g(\lambda) = f_2^*\left(\mu_0 + \phi^{-1}(\lambda)\right) = f_2^*(\mu_2)$$

  where ϕ is the diffeomorphism defined in Lemma 16 (eqn:66) in Appendix B.
- *The function g is differentiable and has an L_g-Lipschitz gradient.*

The projected duality gap and the smoothed duality gap defined, respectively, in (11) and (14) satisfy:

$$\mathcal{D}(z) \leq \left((3 + \beta_x L_g)\, \mathcal{G}_\beta(z) + \left(\sqrt{2\beta_x}\, (2\|x\| + L_{f_1^*}) + \sqrt{2\beta_y}\|y\| \right) \sqrt{\mathcal{G}_\beta(z)} \right)^2 + 2\beta_{\max} \mathcal{G}_\beta(z) \tag{39}$$

$$a = \mathrm{Proj}_{\mathrm{dom} f^*}\left(-A^T y\right) \qquad p = \mathrm{Prox}_{\beta_x^{-1} f}\left(x - \frac{1}{\beta_x} A^T y\right) \tag{40}$$

Next, we will illustrate our theoretical findings through numerical experiments. The directed graph depicted in Figure 2 summarizes our theoretical findings and highlights their main assumptions.

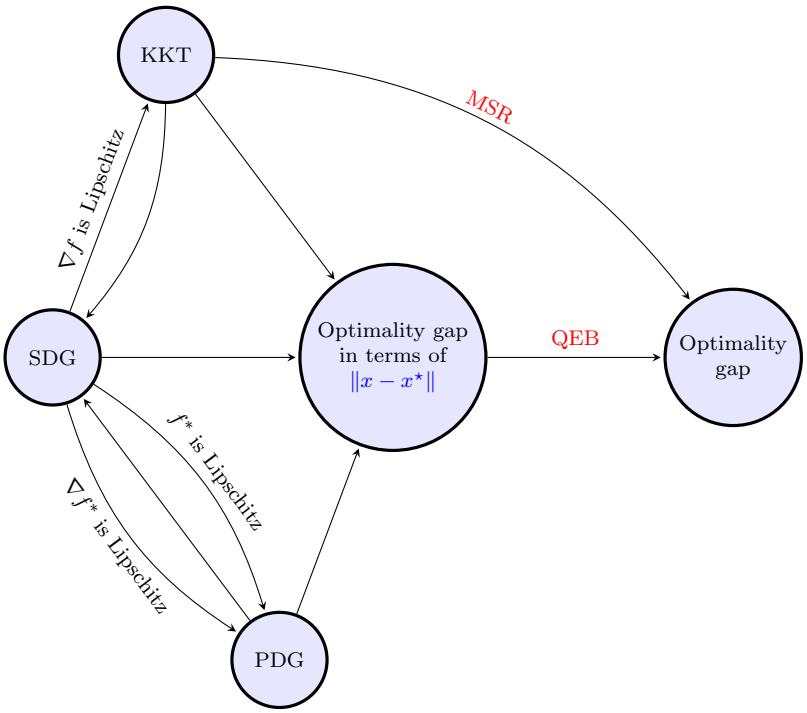


Fig. 2 Summary of our findings, the representation: $\mathcal{M}_1 \xrightarrow{\mathcal{S}} \mathcal{M}_2$ means that the measure ($\mathcal{M}_1$) provides an upper-bound for the measure $\mathcal{M}_2$ under the assumption $\mathcal{S}$

7 Numerical experiments

In the last section, we present a series of numerical experiments aimed at illustrating and validating our findings and their tightness. Initially, we inves-

tigate linearly-constrained least-squares problems where the objective is convex and smooth. Subsequently, we extend our findings beyond affine equality constraints, demonstrating their applicability to more generalized scenarios involving inequality constraints. Finally, we explore a non-smooth problem that showcases the superior stability of the smoothed duality gap compared to the Karush-Kuhn-Tucker error.

One could observe that while the measures of optimality themselves are not complex to implement programmatically, each one necessitates specific information that may be tricky. For example, computing the sub-differential of the objective function, crucial for the Karush–Kuhn–Tucker error, can be exceptionally complicated for certain functions. Similarly, while the smoothed duality gap prefers an easily computable proximal operator of the objective, the projected duality gap favors an objective with a tractable Fenchel-conjugate, allowing projection onto its domain. Moreover, our results rely on additional assumptions such as the metric sub-regularity of the sub-differential of the Lagrangian and quadratic error bound of the smoothed gap, requiring the determination of associated constants, a task often non-trivial in itself. Hence, for each experiment, we will provide all non-trivial computations along with detailed steps wherever necessary.

Primal-Dual Hybrid Gradient: we implement the *Primal-Dual Hybrid Gradient* (PDHG) algorithm [3,7,10,1] to tackle our experiments. We chose it because it's famous for its efficiency in handling large-scale problems due to its low cost per iteration, the PDHG algorithm enjoys linear convergence on the problems under consideration, which makes it more insightful for showing the numerical performance of the measures we consider. Additionally, it provides primal-dual solutions at each iteration.

Algorithm: Primal-Dual Hybrid Gradient (PDHG)

$$
\begin{aligned}
\bar{x}_{k+1} &= \mathrm{Prox}_{\tau f}\left(x_k - \tau A^T y_k\right) \\
\bar{y}_{k+1} &= y_k + \sigma\left(A\bar{x}_{k+1} - b\right) \\
x_{k+1} &= \bar{x}_{k+1} - \tau A^T\left(\bar{y}_{k+1} - y_k\right) \\
y_{k+1} &= \bar{y}_{k+1}
\end{aligned}
$$

The convergence of the algorithm is guaranteed when the step sizes τ and σ satisfy the inequality:

$$\tau\sigma\|A\|^2 < 1$$

Therefore, we choose the following step sizes:

$$\tau = \frac{0.95}{\|A\|} \qquad\qquad \sigma = \frac{1}{\|A\|}$$

Smoothing parameter selection: to determine the smoothing parameter $\beta \in]0, +\infty[^2$, we adopt the following approach:

❖ We set the same primal-dual smoothing parameters, i.e., $\beta_x = \beta_y$.

❖ We construct a predefined geometric list of values for β, denoted by $\mathcal{I}$, with a total of 41 values. This list includes the feasibility error at the current iteration $\|Ax_k - b\|$, along with 40 equally-divided values in logarithmic scale ranging from 10^{-8} to 100:

$$\mathcal{I} := \left\{10^{-8}, 1.91 \times 10^{-7}, \ldots, 5.25 \times 10^{-1}, 100, \|Ax - b\|\right\}$$

❖ Then, for each iteration $k \in \mathbb{N}$, we assign a value for β from $\mathcal{I}$ according to the following criteria:
- For results of the form: $g(z_k) \leq h_\beta(z_k)$, we choose:

$$\tilde{\beta} = \arg\min_{\beta \in \mathcal{I}} h_\beta(z_k)$$

- For results of the form: $g_\beta(z_k) \leq h_\beta(z_k)$, we choose:

$$\tilde{\beta} = \arg\min_{\beta \in \mathcal{I}} \frac{h_\beta(z_k)}{g_\beta(z_k)}$$

Computer configurations: our numerical results were generated on macOS using an Apple M1 Pro processor with 16 GB of RAM. We employed Python 3.10.9 software for the computations.

7.1 Linearly-Constrained Least-Squares (LC-LS)

We are interested in illustrating several instances of the following Least-Squares problem [4,27]:

$$\min_{x \in \mathbb{R}^n} \ \frac{1}{2}\|Qx - c\|^2 \qquad \text{subject to} \qquad Ax = b \qquad \text{(LC-LS)}$$

or, equivalently, the associated saddle point problem:

$$\min_{x \in \mathbb{R}^n} \max_{y \in \mathbb{R}^m} \mathcal{L}(x, y) = \frac{1}{2}\|Qx - c\|^2 + \langle Ax - b, y \rangle$$

where the objective function $f(x) = \frac{1}{2}\|Qx - c\|^2$ is smooth $\left(\nabla f(x) = Q^T (Qx - c)\right)$, and convex $\left(\nabla^2 f(x) = Q^T Q \succcurlyeq 0\right)$. The matrix $A \in \mathbb{R}^{m \times n}$ and the vectors c & $b \in \mathbb{R}^m$ are given. Before getting started with the instances, let's analyze the problem.

7.1.1 Problem Analysis

1. **The gradient**, ∇f, is $L-$Lipschitz with $L = \left\|Q^T Q\right\|_{op} = \lambda_{\max}\left(Q^T Q\right)$

2. **The proximal operator is:**

$$\begin{aligned}\bar{u} = \mathrm{Prox}_{sf}(x) = \arg\min_{u\in\mathbb{R}^n} \frac{1}{2}\|Qu-c\|^2 + \frac{1}{2s}\|u-x\|^2 \\ \iff 0 \in \partial\left(\frac{1}{2}\|Q.-c\|^2 + \frac{1}{2}\|.-x\|^2\right)(\bar{u}) \\ \iff \bar{u} = \left(Q^TQ + \frac{1}{s}\mathrm{Id}\right)^{-1}\left(Q^Tc + \frac{1}{s}x\right) \qquad (41)\end{aligned}$$

3. **The Fenchel-Conjugate is**: $f^*(\mu) = \sup_{x\in\mathbb{R}^n} \langle\mu, x\rangle - \frac{1}{2}\|Qx-c\|^2$, let:

$$\begin{aligned}\bar{x} = \arg\max_{x\in\mathbb{R}^n} \langle\mu, x\rangle - \frac{1}{2}\|Qx-c\|^2 \\ \iff 0 \in \partial\left(\langle\mu,.\rangle - \frac{1}{2}\|Q.-c\|^2\right)(\bar{x}) \\ \iff Q^TQ\bar{x} = Q^Tc + \mu\end{aligned}$$

There are two cases:

❖ If $\mu \notin \mathrm{Ran}\left(Q^T\right)$, which means that $\mu = \mu_1 + \mu_2$ such that $\mu_1 \in \mathrm{Ran}\left(Q^T\right)$ and $\mu_2 \in \ker(Q)$, then one can find a maximizing sequence defined as:

$$x_n = n\mu_2, \qquad \forall n \in \mathbb{N}$$

Then,

$$\begin{aligned}f^*(x_n) &= \sup_{n\in\mathbb{N}} \langle\mu, x_n\rangle - \frac{1}{2}\|Qx_n-c\|^2 \\ &= \sup_{n\in\mathbb{N}} n\|\mu_2\|^2 - \frac{1}{2}\|c\|^2 \xrightarrow[n\to+\infty]{} +\infty\end{aligned}$$

❖ If $\mu \in \mathrm{Ran}\left(Q^T\right)$, then

$$\bar{x} = \left(Q^TQ\right)^\dagger\left(Q^Tc + \mu\right)$$

and f^* could be simplified to get:

$$f^*(\mu) = \frac{1}{2}\left\langle\mu, \left(Q^TQ\right)^\dagger\mu\right\rangle + \left\langle\mu, \left(Q^TQ\right)^\dagger Q^Tc\right\rangle - \frac{1}{2}\left\|Q\left(Q^TQ\right)^\dagger Q^Tc - c\right\|^2$$

Thus,

$$f^*(\mu) = \frac{1}{2}\left\langle\mu, \left(Q^TQ\right)^\dagger\mu\right\rangle + \left\langle\mu, \left(Q^TQ\right)^\dagger Q^Tc\right\rangle - \frac{1}{2}\left\|Q\left(Q^TQ\right)^\dagger Q^Tc - c\right\|^2 + \imath_{\mathrm{Ran}(Q^T)}(\mu) \qquad (42)$$

4. Rewriting $f^*(\mu) = f_1^*(\mu)$ and $f_2^* = 0$ satisfies the set of assumptions, $\mathcal{E}$, of Theorem 8. Thus, we define:

$$\begin{aligned} g(\lambda) &= f^*(\mu_0 + \phi^{-1}(\lambda)) \\ &= \frac{1}{2}\left\langle \mu, \left(Q^TQ\right)^\dagger \mu \right\rangle + \left\langle \mu, \left(Q^TQ\right)^\dagger Q^Tc \right\rangle - \frac{1}{2}\left\| Q\left(Q^TQ\right)^\dagger Q^Tc - c \right\|^2 \\ \lambda &= \phi(\mu) \qquad\qquad \forall \mu \in \mathrm{dom} f^* \end{aligned}$$

Hence, g is **differentiable** with **Lipschitz constant**

$$L_g = \left\| \left(Q^TQ\right)^\dagger \right\|_{op} = \lambda_{\max}\left(\left(Q^TQ\right)^\dagger\right)$$

5. **Projection** onto $\mathrm{dom} f^* = \mathrm{Ran}\left(Q^T\right)$

$$\begin{aligned} \mathrm{Proj}_{\mathrm{Ran}(Q^T)}(\mu) &= \arg\min_{u \in \mathrm{Ran}(Q^T)} \|u - \mu\|^2 \\ &\overset{u=Q^Tv}{=} Q^T \arg\min_{v \in \mathbb{R}^m} \left\|Q^Tv - \mu\right\|^2 \\ &= Q^T\left(QQ^T\right)^\dagger Q\mu \end{aligned}$$

We will show in the sequel that finding the values of the QEBSG and MSR constants η and γ, respectively, is tractable for this problem.

6. **Metric sub-regularity constant**, γ.

Lemma 10 *For the LC-LS problem, let $z^\star = \mathrm{Proj}_{\mathcal{Z}^\star}(z)$. Then, for any $z \in \mathcal{Z}$ the Lagrangian's sub-differential satisfies:*

$$\|\partial_x \mathcal{L}(x,y)\| + \|\nabla_y \mathcal{L}(x,y)\| \geq |\lambda(\mathcal{M})|_{\min} \,\mathrm{dist}\left(z, \mathcal{Z}^\star\right) \tag{43}$$

$$\mathcal{M} = \begin{bmatrix} Q^TQ & A^T \\ A & 0 \end{bmatrix} \tag{44}$$

where $|\lambda(\mathcal{M})|_{\min}$ is the smallest absolute value of the non-zero eigenvalues of $\mathcal{M}$.
Proof. See Appendix A

Thanks to this Lemma, one can take $\gamma = |\lambda(\mathcal{M})|_{\min}$

7. **Quadratic error bound of the smoothed gap constant**, η.

Lemma 11 *For the LC-LS problem, the self-centered smoothed duality gap could be reformulated into a quadratic form. That is, for any $z \in \mathcal{Z}$:*

$$\mathcal{G}_\beta(z) = z^T\mathcal{H}z + \langle z, v\rangle + cst \tag{45}$$

where,

$$v_{(n+m)} = \begin{bmatrix} v_x \\ v_y \end{bmatrix} \qquad \mathcal{H}_{(n+m)\times(n+m)} = \begin{bmatrix} M_{xx} & \frac{1}{2}M_{xy} \\ \frac{1}{2}M_{xy}^T & M_{yy} \end{bmatrix} \tag{46}$$

Proof. See Appendix A

Vector - Matrix	Size
$v_x = \frac{\beta_x}{\tau} B^{-1} Q^T c - Q^T c$	$n \times 1$
$v_y = -AB^{-1} Q^T c$	$m \times 1$
$B = Q^T Q + \frac{\beta_x}{\tau} \mathrm{Id}_n$	$n \times n$
$M_{xx} = \frac{1}{2} Q^T Q + \frac{\sigma}{2\beta_y} A^T A + \frac{\beta_x^2}{2\tau^2} B^{-1} - \frac{\beta_x}{2\tau} \mathrm{Id}_n$	$n \times n$
$M_{xy} = A^T - \frac{\beta_x}{\tau} B^{-1} A^T$	$n \times m$
$M_{yy} = \frac{1}{2} AB^{-1} A^T$	$m \times m$

The advantage of the lemma is that, for LC-LS, the self-centered smoothed duality gap is convex and $\lambda_{\min}(\mathcal{H})-$metrically sub-regular, with $\lambda_{\min}(\mathcal{H})$ being the smallest positive eigenvalue of the positive semi-definite matrix $\mathcal{H}$. Hence, by Proposition 7, we conclude that $\mathcal{G}_\beta(z)$ has a $\eta = \frac{\lambda_{\min}(\mathcal{H})}{2}-$QEB.

7.1.2 Problem instances

In this part, we explore various instances of the LC-LS problem. We start by examining a straightforward one-dimensional case, for which we can determine the precise minimizer. Subsequently, we extend our analysis to a higher dimensional problem with independently and identically distributed (i.i.d.) Gaussian matrices Q and A. Finally, we delve into a more complex scenario where we generate Q and A with non-trivial covariance matrices.

❖ **One-dimensional** problem

$$\begin{aligned} \min_{x \in \mathbb{R}} \ & \frac{1}{2} \left(\frac{1}{9} x - 2 \right)^2 \\ & 9x = 7 \end{aligned} \tag{1D}$$

The *primal-feasibility* condition within the KKT conditions implies that $x^\star = \frac{7}{9}$. Consequently, in this particular problem, we can gain a more precise assessment of the quality of our approximations for the optimality gap.

❖ **I.I.D. Gaussian matrices:** We examine the LC-LS problem with dimensions set to $n = 20$ and $m = 10$. In this scenario, we generate independently and identically distributed (i.i.d.) Gaussian matrices Q and A. That is:

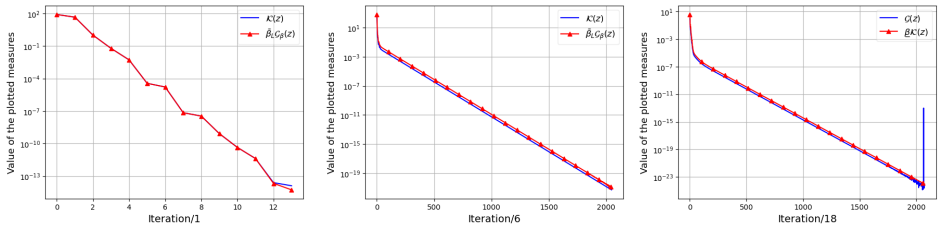
$$\begin{aligned} &Q \in \mathbb{R}^{m \times n} \text{ s.t. } \forall (i,j) \in [\![1,m]\!] \times [\![1,n]\!], \ Q_{ij} \sim \mathcal{N}(0,1) \ \textit{i.i.d.} \\ &A \in \mathbb{R}^{m \times n} \text{ s.t. } \forall (i,j) \in [\![1,m]\!] \times [\![1,n]\!], \ A_{ij} \sim \mathcal{N}(0,1) \ \textit{i.i.d.} \end{aligned} \tag{IIDG}$$

❖ **Non-trivial covariance:** We investigate the LC-LS problem with dimensions set to $n = 20$ and $m = 10$. In this instance, we generate matrices Q and A with non-trivial covariance matrices, defined as follows:

$$A = \Sigma_a X_a \qquad\qquad Q = \Sigma_q X_q \tag{NTC}$$

where

– The matrices $X_a, X_q \in \mathbb{R}^{m\times n}$ are Gaussian matrices, following the pattern established in the previous case.
– The matrices $\Sigma_a, \Sigma_q \in \mathbb{R}^{m\times m}$ serve as covariance matrices, generated using the Python built-in function `toeplitz`.



(a) One-dimensional, $L \approx 0.012$ (b) I.I.D Gaussian $L \approx 49$ (c) Non-trivial covariance

Fig. 3 Numerical illustration of Theorems (5 & 6) on several LC-LS instances. L denotes the Lipschitz constant of the gradient

Figure 3 validates our comparability bound between the smoothed duality gap and the Karush–Kuhn–Tucker error, presented in subsection 6.1. It illustrates their efficiency and tightness across various instances. In addition, we note that the smoother the function (i.e. with smaller L) the tighter the bound as in (3a and 3b).

7.2 Distributed Optimization

We examine the unconstrained optimization problem presented as follows [5]:

$$\min_{x\in\mathbb{R}^n} \ \frac{1}{2}\sum_{i=1}^{M} \|Q_i x - c_i\|^2 \qquad \text{(UC-DO)}$$

In this formulation, for each $i \in [\![1, M]\!]$, the matrix $Q_i \in \mathbb{R}^{m\times n}$ and the vector $c_i \in \mathbb{R}^m$ are real data sourced from the `bodyfat` dataset. This data comprises 252 data-points, each characterized by 14 features. In our case, we divide the dataset into $M = 3, n = 14$, and $m = 84$.

To address security concerns, we assume that the data Q_i and c_i are distributed across M distinct computers. These computers collaborate to solve the problem in a manner where each computer i utilizes the data Q_i and c_i to derive a partial solution, which is then transmitted to the subsequent computer $i + 1$. Repeatedly until finding the optimum. To facilitate this collaborative approach, we need to reformulate the unconstrained problem (UC-DO) into a constrained optimization problem. This involves introducing additional variables and constraints to model the communication process among the various computers.

$$\min_{x_1,\dots,x_M\in\mathbb{R}^n} \frac{1}{2}\sum_{i=1}^{M}\|Q_i x_i - c_i\|^2$$
$$\text{s.t.} \quad x_i = x_{i+1}, \qquad \forall i \in [\![1, M-1]\!]$$

Furthermore,

❖ The constraints can be expressed in matrix form as $AX = 0$, where the matrix A is constructed as follows:

$$A_{(M-1)n\times Mn} := \begin{bmatrix} \mathrm{Id}_n & -\mathrm{Id}_n & \mathbf{0}_n & \dots & \mathbf{0}_n \\ \mathbf{0}_n & \mathrm{Id}_n & -\mathrm{Id}_n & \dots & \mathbf{0}_n \\ \vdots & \dots & \ddots & \ddots & \dots \\ \mathbf{0}_n & \dots & \mathbf{0}_n & \mathrm{Id}_n & -\mathrm{Id}_n \end{bmatrix} \quad \text{and} \quad X_{Mn} := \begin{bmatrix} x_1 \\ \vdots \\ \vdots \\ x_M \end{bmatrix}$$

❖ The objective function $\left(f(X) = \frac{1}{2}\sum_{i=1}^{M}\|Q_i x_i - c_i\|^2\right)$ can be rewritten as:

$$f(X) = \frac{1}{2}\|\mathbf{Q}X - \mathbf{c}\|^2$$

Here, $\mathbf{Q}$ is formed by arranging the matrices $Q_1, Q_2, \dots, Q_M$ along the diagonal, and $\mathbf{c}$ is a stacked vector of $c_1, c_2, \dots, c_M$. That is:

$$\mathbf{Q}_{Mm\times Mn} := \begin{bmatrix} Q_1 & \mathbf{0}_{m\times n} & \dots & \mathbf{0}_{m\times n} \\ \mathbf{0}_{m\times n} & Q_2 & \dots & \mathbf{0}_{m\times n} \\ \vdots & \dots & \ddots & \dots \\ \mathbf{0}_{m\times n} & \dots & \mathbf{0}_{m\times n} & Q_M \end{bmatrix} \qquad \mathbf{c}_{Mm} := \begin{bmatrix} c_1 \\ \vdots \\ c_M \end{bmatrix}$$

Therefore, the unconstrained problem can be reformulated as an instance of the LC-LS problem as follows:

$$\min_{X\in\mathbb{R}^{Mn}} \frac{1}{2}\|\mathbf{Q}X - \mathbf{c}\|^2 \qquad \text{subject to} \qquad AX = 0 \qquad \text{(DO)}$$

It is important to note that all the formulations of the problem are equivalent, specifically (UC-DO $\equiv$ DO). Consequently, since (UC-DO) represents an unconstrained smooth problem, the minimizer $x^\star$ can be easily found:

$$\begin{aligned} x^\star &= \arg\min_{x\in\mathbb{R}^n} \frac{1}{2}\sum_{i=1}^{M}\|Q_i x - c_i\|^2 \\ &\iff 0 = \nabla\left(\frac{1}{2}\sum_{i=1}^{M}\|Q_i x - c_i\|^2\right)(x^\star) \\ &\iff \sum_{i=1}^{M} Q_i^T\left(Q_i x^\star - c_i\right) = 0 \end{aligned}$$

$$\iff \left(\sum_{i=1}^{M} Q_i^T Q_i\right) x^\star = \sum_{i=1}^{M} Q_i^T c_i \tag{47}$$

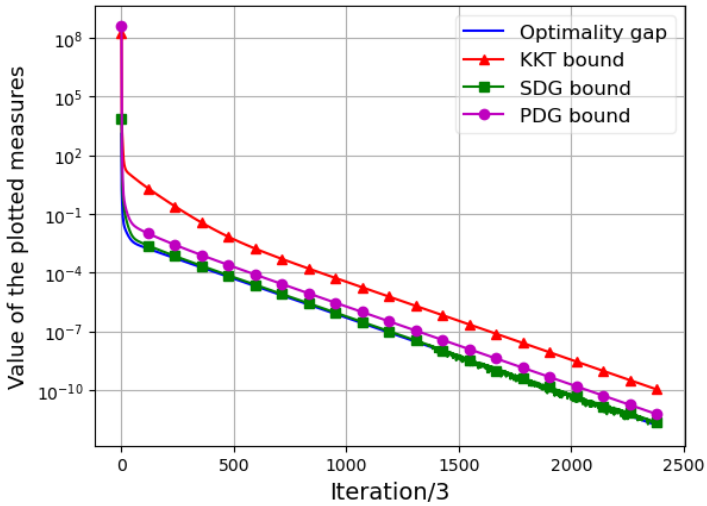


Fig. 4 Numerical illustration of Theorems (2, 3, and 4), optimality gap vs. its bounds. Target: $\varepsilon = 10^{-10}$

Figure 4 validates our computable approximations for the optimality gap, as presented in subsection 5.2. Notably, in this experiment, we were able to precisely plot the optimality gap due to our analytical knowledge of the minimizer (eqn:47). Furthermore, the figure highlights the superior efficiency and tightness of the SDG bound compared to the others, as it closely aligns with the optimality gap curve.

7.3 Quadratic Programming

As we have seen, our theoretical findings have primarily focused on optimization problems under equality affine constraints. However, in this experiment, we aim to extend the applicability of our theoretical insights to address optimization problems that incorporate inequality constraints. Therefore, in this subsection, we delve into the same LC-LS problem as discussed earlier, but now incorporating the additional requirement of non-negativity constraints on the weights [19].

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \ & \frac{1}{2}\|Qx - c\|^2 \\ \text{s.t.} \ & Ax = b \\ & x \geq 0 \end{aligned} \tag{QP}$$

Applying our theoretical results to address this problem necessitates a reformulation to align it with our framework. A feasible approach involves encapsulating the non-negativity constraint by introducing an indicator function within the objective function. Thus, the reformulation is as follows:

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \frac{1}{2}\|Qx-c\|^2 + \imath_{\mathbb{R}^n_+}(x) \\ \text{s.t.} \quad Ax = b \end{aligned}$$

Despite the current appropriate formulation, determining the proximal operator of the objective function, essential for computing the smoothed duality gap, has become non-trivial. However, introducing an additional variable, $\tilde{x}$, and incorporating an extra constraint, $x = \tilde{x}$, would render the computation more manageable. That is:

$$\begin{aligned} \min_{x,\tilde{x}\in\mathbb{R}^n} \frac{1}{2}\|Qx-c\|^2 + \imath_{\mathbb{R}^n_+}(\tilde{x}) \\ \text{s.t.} \quad Ax = b \\ x = \tilde{x} \end{aligned}$$

Thanks to this trick, our objective function is now separable, facilitating the computation of its proximal operator. By combining the two constraints into a single matrix form, our problem can be expressed as follows:

$$\begin{aligned} \min_{X\in\mathbb{R}^{2n}} F(X) = \frac{1}{2}\|Qx-c\|^2 + \imath_{\mathbb{R}^n_+}(\tilde{x}) \\ \text{s.t.} \quad \tilde{A}X = B \end{aligned} \tag{PQP}$$

where

$$X := \begin{bmatrix} x \\ \tilde{x} \end{bmatrix} \qquad \tilde{A} := \begin{bmatrix} A & 0 \\ \mathrm{Id} & -\mathrm{Id} \end{bmatrix} \qquad B := \begin{bmatrix} b \\ 0 \end{bmatrix} \tag{48}$$

7.3.1 Problem Analysis

1. **The sub-differential of the objective function**

$$\begin{aligned} \partial F(x,\tilde{x}) &= \partial\left(\frac{1}{2}\|Qx-c\|^2 + \imath_{\mathbb{R}^n_+}(\tilde{x})\right) \\ &\overset{(4)}{=} \nabla\left(\frac{1}{2}\|Qx-c\|^2\right) \times \partial\left(\imath_{\mathbb{R}^n_+}(\tilde{x})\right) \\ &= Q^T(Qx-c) \times \partial\left(\sum_{i=1}^{n} \imath_{\mathbb{R}_+}(\tilde{x}_i)\right) \\ &\overset{(4)}{=} Q^T(Qx-c) \times \prod_{i=1}^{n} \partial\imath_{\mathbb{R}_+}(\tilde{x}_i) \end{aligned}$$

Thus,

$$\partial F(X) = \overbrace{\underbrace{Q^T\left(Qx-c\right)}_{n-\text{tuple}} \times \underbrace{\prod_{i=1}^{n} \mathcal{N}_{\mathbb{R}_+}(\tilde{x}_i)}_{n-\text{tuple}}}^{2n-\text{tuple}} \tag{49}$$

$$\text{where} \quad \mathcal{N}_{\mathbb{R}_+}(\nu) = \begin{cases} \emptyset, & \nu < 0 \\ \mathbb{R}_-, & \nu = 0 \\ \{0\}, & \nu > 0 \end{cases}$$

2. **The stationarity part of the KKT error:**
Let $J = \tilde{A}^T Y \in \mathbb{R}^{2n}$ and define:

$$\underline{J} := \{J_1, \dots, J_n\} \qquad\qquad \bar{J} := \{J_{n+1}, \dots, J_{2n}\}$$

Then,

$$\begin{aligned}
\left\|\partial F(x,\tilde{x}) + \tilde{A}^T Y\right\|_0^2 &= \|\partial F(X) + J\|_0^2 \\
&= \min\left(\{\|q\| \mid q \in \partial F(X) + J\}\right)^2 \\
&= \min\left(\left\{\left\|\left(\left(Q^T\left(Qx-c\right)\right) \times \prod_{i=1}^{n} \mathcal{N}_{\mathbb{R}_+}(\tilde{x}_i)\right) + J\right\|\right\}\right)^2 \\
&= \min\left\{\left\|\left(\left(Q^T\left(Qx-c\right)\right) \times \prod_{i=1}^{n} \mathcal{N}_{\mathbb{R}_+}(\tilde{x}_i)\right) + J\right\|^2\right\} \\
&= \min\left\{\left\|Q^T\left(Qx-c\right) + \underline{J}\right\|^2 + \left\|\prod_{i=1}^{n} \mathcal{N}_{\mathbb{R}_+}(\tilde{x}_i) + \bar{J}\right\|^2\right\} \\
&= \left\|Q^T\left(Qx-c\right) + \underline{J}\right\|^2 + \min\left\{\left\|\prod_{i=1}^{n} \mathcal{N}_{\mathbb{R}_+}(\tilde{x}_i) + \bar{J}\right\|^2\right\}
\end{aligned}$$

Now, let's analyze the second term:

$$\begin{aligned}
\min\left\{\left\|\prod_{i=1}^{n} \mathcal{N}_{\mathbb{R}_+}(\tilde{x}_i) + \bar{J}\right\|^2\right\} &= \min\left\{\sum_{i=1}^{n}\left(\mathcal{N}_{\mathbb{R}_+}(\tilde{x}_i) + \bar{J}_i\right)^2\right\} \\
&= \sum_{i=1}^{n} \min\left\{\left(\mathcal{N}_{\mathbb{R}_+}(\tilde{x}_i) + \bar{J}_i\right)^2\right\} \\
&= \sum_{i=1}^{n} \min\begin{cases} +\infty, & \tilde{x}_i < 0 \\ \left(\mathbb{R}_- + \bar{J}_i\right)^2, & \tilde{x}_i = 0 \\ \bar{J}_i^2, & \tilde{x}_i > 0 \end{cases}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{n} \begin{cases} +\infty, & \tilde{x}_i < 0 \\ \min\limits_{u \in \mathbb{R}_-} \left(u + \bar{J}_i\right)^2, & \tilde{x}_i = 0 \\ \bar{J}_i^2, & \tilde{x}_i > 0 \end{cases} \\
&= \sum_{i=1}^{n} \begin{cases} +\infty, & \tilde{x}_i < 0 \\ \left(\mathrm{Proj}_{\mathbb{R}_-}\left(-\bar{J}_i\right) + \bar{J}_i\right)^2, & \tilde{x}_i = 0 \\ \bar{J}_i^2, & \tilde{x}_i > 0 \end{cases} \\
&= \sum_{i=1}^{n} \begin{cases} +\infty, & \tilde{x}_i < 0 \\ \min\left(0, \bar{J}_i\right)^2, & \tilde{x}_i = 0 \\ \bar{J}_i^2, & \tilde{x}_i > 0 \end{cases}
\end{aligned}
$$

3. **The proximal operator is:**

$$
\begin{aligned}
\mathrm{Prox}_{sf}(x, \tilde{x}) &= \arg\min_{u, \tilde{u} \in \mathbb{R}^n} \underbrace{\frac{1}{2}\|Qu - c\|^2 + \frac{1}{2s}\|u - x\|^2}_{h(u)} + \imath_{\mathbb{R}^n_+}(\tilde{u}) + \frac{1}{2s}\|\tilde{u} - \tilde{x}\|^2 \\
&\overset{(5)}{=} \left(\arg\min_{u \in \mathbb{R}^n} h(u), \arg\min_{\tilde{u} \in \mathbb{R}^n_+} \frac{1}{2s}\|\tilde{u} - \tilde{x}\|^2\right) \\
&\overset{(41)}{=} \left(\left(Q^T Q + \frac{1}{s}\mathrm{Id}\right)^{-1}\left(Q^T c + \frac{1}{s}x\right), \mathrm{Proj}_{\mathbb{R}^n_+}(\tilde{x})\right) \\
&= \left(\left(Q^T Q + \frac{1}{s}\mathrm{Id}\right)^{-1}\left(Q^T c + \frac{1}{s}x\right), \max(\tilde{x}, 0)\right)
\end{aligned}
$$

4. **The Fenchel-Conjugate is:**

$$
\begin{aligned}
f^*(\mu, \tilde{\mu}) &= \sup_{X \in \mathbb{R}^{2n}} \langle \mu, x\rangle + \langle \tilde{\mu}, \tilde{x}\rangle - \frac{1}{2}\|Qx - c\|^2 - \imath_{\mathbb{R}^n_+}(\tilde{x}) \\
&\overset{(6)}{=} \sup_{x \in \mathbb{R}^n}\left(\langle \mu, x\rangle - \frac{1}{2}\|Qx - c\|^2\right) + \sup_{\tilde{x} \in \mathbb{R}^n}\left(\langle \tilde{\mu}, \tilde{x}\rangle - \imath_{\mathbb{R}^n_+}(\tilde{x})\right) \\
&\overset{(42)}{=} \frac{1}{2}\left\langle \mu, \left(Q^T Q\right)^\dagger \mu\right\rangle + \left\langle \mu, \left(Q^T Q\right)^\dagger Q^T c\right\rangle - \frac{1}{2}\left\|Q\left(Q^T Q\right)^\dagger Q^T c - c\right\|^2 \\
&\qquad + \imath_{\mathrm{Ran}(Q^T)}(\mu) + \imath_{\mathbb{R}^n_-}(\tilde{\mu})
\end{aligned}
$$

5. Rewriting $f^*(\mu) = f_1^*(\tilde{\mu}) + f_2^*(\mu)$ such that:

$$
\begin{aligned}
f_1^*(\tilde{\mu}) &= \imath_{\mathbb{R}^n_-}(\tilde{\mu}) \\
f_2^*(\mu) &= \frac{1}{2}\left\langle \mu, \left(Q^T Q\right)^\dagger \mu\right\rangle + \left\langle \mu, \left(Q^T Q\right)^\dagger Q^T c\right\rangle - \frac{1}{2}\left\|Q\left(Q^T Q\right)^\dagger Q^T c - c\right\|^2 \\
&\qquad + \imath_{\mathrm{Ran}(Q^T)}(\mu)
\end{aligned}
$$

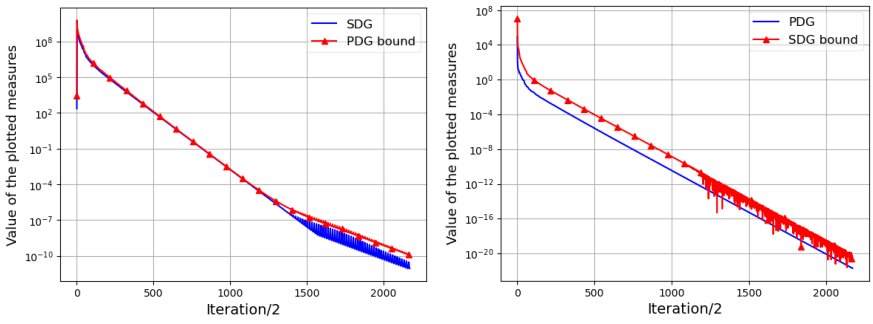
satisfies the set of assumptions, $\mathcal{E}$, of Theorem 8. Thus,

- ❖ f_1^* is $L_{f_1^*} = 0-$Lipschitz on its domain, $\mathbb{R}^n_-$.
- ❖ ∇f_2^* is $L_g = \lambda_{\max}\left(\left(Q^TQ\right)^\dagger\right)-$Lipschitz using the earlier analysis we did for LC-LS.

6. **Projection onto the** $\mathrm{dom} f^* = \mathrm{Ran}\left(Q^T\right) \times \mathbb{R}^n_-$

$$\mathrm{Proj}_{\mathrm{dom} f^*}(\mu) = \left(Q^T\left(QQ^T\right)^\dagger Q\mu,\ \min\left(\tilde{\mu}, 0\right)\right)$$

7. While it is theoretically possible to determine the constants for the metric sub-regularity of the sub-differential of the Lagrangian and the quadratic error bound of the smoothed gap, the process is highly complex. Due to the complexity involved, we opt for a practical approach by assigning arbitrary small values to these constants. In this experiment, we set $\gamma = \eta = 10^{-8}$ as convenient and satisfactory choices, allowing us to proceed without extensive efforts in identifying precise values.



(a) PDG bound, (eqn:36) (b) SDG bound, (eqn:83)

Fig. 5 Numerical illustration of Theorems (7 and 8) with $n = 20$ and $m = 10$

7.4 Basis Pursuit

Our final experiment involves a non-smooth convex minimization problem known as the *Basis Pursuit* problem [8,13]. It aims to minimize the ℓ_1 norm while satisfying a system of linear equations and is mathematically formulated as follows:

$$\min_{x\in\mathbb{R}^n} \|x\|_1 \qquad \text{subject to} \qquad Ax = b \tag{BP}$$

Here, we set $n = 20$ and $m = 10$, and generate an i.i.d. Gaussian matrix $A \in \mathbb{R}^{m\times n}$ and vector $b \in \mathbb{R}^m$.

The primary objective of this experiment is to highlight the superior stability of the smoothed duality gap compared to the Karush–Kuhn–Tucker error, as will be demonstrated subsequently.

7.4.1 Problem Analysis

1. **The sub-differential of the objective function**

$$\partial f(x) = \partial(\|x\|_1) = \partial\left(\sum_{i=1}^{n}|x_i|\right) \overset{(4)}{=} \prod_{i=1}^{n}\partial|x_i|$$

where

$$\partial|\nu| = \begin{cases}\{-1\}, & \nu < 0\\ [-1,1], & \nu = 0\\ \{1\}, & \nu > 0\end{cases}$$

2. **The stationarity part of the KKT error:** Applying the same approach as we have previously done for (PQP), we obtain:

$$\left\|\partial f(x) + A^T y\right\|_0^2 = \sum_{i=1}^{n}\begin{cases}\left(\operatorname{sgn}(x_i) + (A^T y)_i\right)^2, & x_i \neq 0\\ \left(\min\left(\max\left(-\left(A^T y\right)_i, -1\right), 1\right) + (A^T y)_i\right)^2, & x_i = 0\end{cases}$$

3. **The proximal operator is:** (*Example 2.16, [9]*)

$$\operatorname{Prox}_{sf}(x) = \left(\left[|x_i| - s\right]_+ \operatorname{sgn}(x_i)\right)_{1\leq i\leq n}$$

4. **The Fenchel-Conjugate is:** (*Example 3.26, [23]*)

$$f^*(\mu) = \imath_{\mathcal{B}_\infty(0,1)}(\mu)$$

5. **The conjugate function is** $L_{f^*} = 0$**−Lipschitz on its domain**, $\mathcal{B}_\infty(0,1)$. So, we can take advantage of using Proposition 9 in Appendix B that has a slightly tighter bound than Theorem 8.
6. **Projection onto** $\operatorname{dom} f^* = \mathcal{B}_\infty(0,1)$.

$$\operatorname{Proj}_{\mathcal{B}_\infty(0,1)}(\mu) = \left(\begin{cases}\operatorname{sgn}(\mu_i), & |\mu_i| > 1\\ \mu_i, & |\mu_i| \leq 1\end{cases}\right)_{1\leq i\leq n}$$

7.4.2 Two versions of PDHG

In this sub-subsection, we consider two versions of the PDHG algorithm designed to solve two formulations of the saddle point problem of (BP):

$$\begin{aligned}&\max_y \min_x \|x\|_1 + \langle Ax - b, y\rangle &(50)\\ \equiv &\max_x \min_y \|y\|_1 + \langle Ay - b, x\rangle\\ \equiv &\max_x \min_y -\left(-\|y\|_1 - \langle Ay - b, x\rangle\right)\\ \equiv &-\min_x \max_y \left(-\|y\|_1 - \langle Ay - b, x\rangle\right) &(51)\end{aligned}$$

Version 1	Interpretation
$\bar{x}_{k+1} = \text{Prox}_{\tau f}\left(x_k - \tau A^T y_k\right)$	Primal Forward-Backward step
$\bar{y}_{k+1} = y_k + \sigma\left(A\bar{x}_{k+1} - b\right)$	Dual Forward-Backward step
$x_{k+1} = \bar{x}_{k+1} - \tau A^T\left(\bar{y}_{k+1} - y_k\right)$	Primal Extrapolation step
$y_{k+1} = \bar{y}_{k+1}$	Dual Extrapolation step

Table 1 Interpretation of each step of PDHG for solving (50)

Version 2	Interpretation
$\bar{y}_{k+1} = y_k + \sigma\left(Ax_k - b\right)$	Dual Forward-Backward step
$\bar{x}_{k+1} = \text{Prox}_{\tau f}\left(x_k - \tau A^T \bar{y}_{k+1}\right)$	Primal Forward-Backward step
$y_{k+1} = \bar{y}_{k+1} + \sigma A\left(\bar{x}_{k+1} - x_k\right)$	Dual Extrapolation step
$x_{k+1} = \bar{x}_{k+1}$	Primal Extrapolation step

Table 2 Interpretation of each step of PDHG for solving (51)

Here, (50) and (51) can be interpreted as swapping the roles of the primal and dual variables (x and y, respectively). Applying the PDHG algorithm to solve (50) and (51) leads us to the following two versions:

The key advantage of implementing the two versions of PDHG for solving the saddle point problems (50) and (51) is to demonstrate the superior stability of the smoothed duality gap over the Karush–Kuhn–Tucker error. To understand this, let's revisit the stationarity part of the KKT error:

$$\left\|\partial f(x) + A^T y\right\|^2 = \sum_{i=1}^{n} \begin{cases} \left(\text{sgn}(x_i) + (A^T y)_i\right)^2, & x_i \neq 0 \\ \left(\min\left(\max\left(-\left(A^T y\right)_i, -1\right), 1\right) + (A^T y)_i\right)^2, & x_i = 0 \end{cases}$$

The instability in the KKT error can be interpreted through the following two remarks:

1. The stationarity expression is highly sensitive to whether the component x_i equals zero for each i. If $|x_i| \leq \varepsilon$ for some i and for some very small $\varepsilon > 0$, the algorithm may face convergence issues.
2. In the first version of the algorithm, $\bar{x}$ could be zero since it represents the proximal of the ℓ_1-norm (as defined in (3)). However, the subsequent update $x = \bar{x} - \tau A^T(\bar{y} - y)$ may result in x being very close to zero but never exactly zero. Consequently, this characteristic could prevent the algorithm from converging. In contrast, the second version of the algorithm performs the last update of x as the proximal of the ℓ_1-norm, which could be exactly zero.

7.5 Benchmark comparison

In this last subsection, we conduct a benchmark analysis by applying our findings to the six experiments previously discussed. Initially, in Table 3, we report the number of iterations required by each problem to identify an $\varepsilon = 10^{-8}$ solution, employing various stopping criteria. Note that, the same data is used to

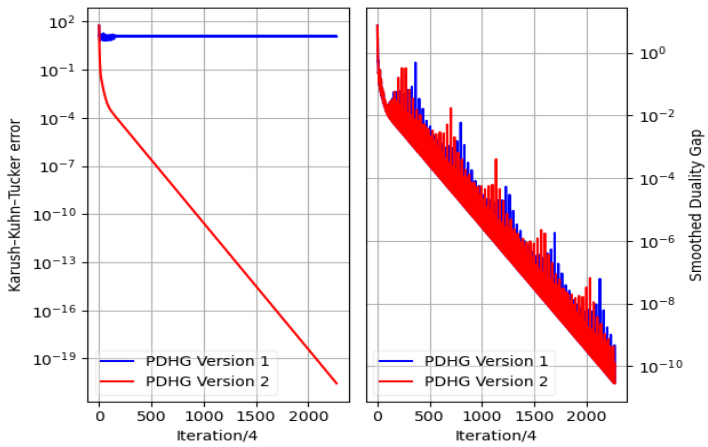


Fig. 6 Version 1 vs. version 2 of PDHG for both: the KKT error and SDG

compute the 3 measures for each experiment. It provides additional evidence supporting our observations from Figure 4, showing that the smoothed duality gap consistently achieves ε-solutions with fewer iterations across nearly all of our experiments, thus demonstrating its superior efficiency.

Problems \ Measures	KKT error	SDG	PDG
One-dimensional (1D)	12	**11**	13
I.I.D. Gaussian matrices (IIDG)	5334	**5244**	11538
Non-trivial covariance (NTC)	14274	**13968**	31806
Distributed Optimization (DO)	5652	**4014**	4659
Quadratic programming (PQP)	3492	**3172**	3358
Basis Pursuit (BP)	$\gg 10^6$	**6920**	7196

Table 3 Iterations needed to identify $\varepsilon = 10^{-8}$ solution using various stopping criteria across experiments

Subsequently, in Table 4, we demonstrate the tightness of our comparability bounds, as outlined in Section 6, across each experiment. This is achieved by presenting the average and standard deviation of the ratio of each result, more specifically, for the bound $\mathcal{M}_1 \leq \mathcal{W}(\mathcal{M}_2)$. The displayed values represent the average and standard deviation of the term $\frac{\mathcal{W}(\mathcal{M}_2)}{\mathcal{M}_1}$. We observe the following:

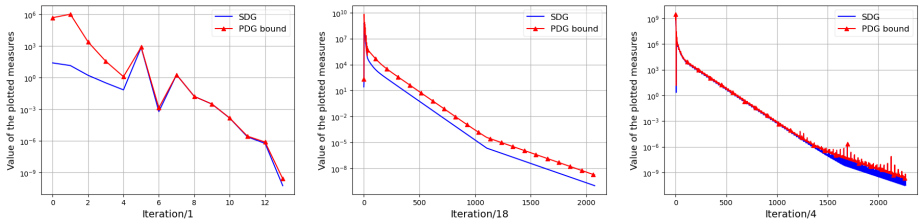
- ❖ We observe that Theorem 5 provides a tight bound overall. Notably, in the Basis Pursuit experiment, the non-convergence of the KKT error explains the huge average and standard deviation observed.
- ❖ Concerning Theorem 6, we note a relatively less tight bound, particularly when the gradient of the objective has a larger Lipschitz constant. The Lipschitz constants for the experiments: the I.I.D. Gaussian matrices, the non-trivial covariance, and the distributed optimization are, approximately: 49,

Problem	Theorem 5 $\mathcal{G}_\beta \leq \underline{\beta}\mathcal{K}$	Theorem 6 $\mathcal{K} \leq \bar{\beta}_L \mathcal{G}_\beta$	Theorem 7 $\mathcal{G}_\beta \leq \mathcal{W}_1(\mathcal{D})$	Theorem 8 $\mathcal{D} \leq \mathcal{W}_2(\mathcal{G}_\beta)$
(1D)	1.76 ± 0.6	0.95 ± 0.15	$(6.82 \pm 19.3)10^4$	5.0 ± 3.2
(IIDG)	2.02 ± 0.06	2.2 ± 0.05	101.65 ± 25.13	5.23 ± 5.43
(NTC)	2.03 ± 0.34	5.44 ± 0.47	15.21 ± 2.47	42.19 ± 35.72
(DO)	3.72 ± 3.9	$(2.12 \pm 25.9)10^8$	11.8 ± 9.15	28.17 ± 10.75
(PQP)	1.25 ± 2.03	$+\infty$	4.21 ± 4.7	32.3 ± 20.98
(BP)	$(7.29 \pm 146)10^9$	$+\infty$	2.49 ± 2.02	20.52 ± 4.25

Table 4 Average $\pm$ (Standard deviation) of the ratio of each result across each experiment. The ratio of a result, $\mathcal{M}_1 \leq \mathcal{W}(\mathcal{M}_2)$, is: $\frac{\mathcal{W}(\mathcal{M}_2)}{\mathcal{M}_1}$

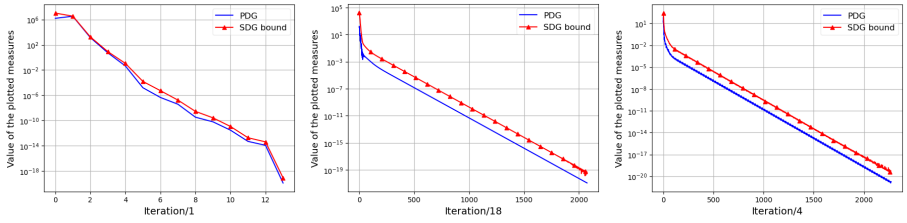
133.5, and 220.6, respectively. Furthermore, the non-smooth nature of the objective function in both the quadratic programming and basis pursuit experiments justifies the occurrence of $+\infty$ in our results.

- ❖ Theorems (7 and 8) provide tight bounds overall. Even in instances where the average ratio may appear elevated, such as observed in Theorem 7 during the one-dimensional experiment, Figure 7 illustrates that this phenomenon occurs primarily within the initial iterations.



(a) One-dimensional (b) Non-trivial covariance (c) Basis pursuit

Fig. 7 Numerical illustration of Theorem 7



(a) One-dimensional (b) Non-trivial covariance (c) Basis pursuit

Fig. 8 Numerical illustration of Theorem 8

8 Conclusion and Perspectives

In this paper, we have studied several stopping criteria: OGFE, the KKT error, PDG, and SDG to determine under which conditions they are accurate to detect $\varepsilon-$solutions. In the realm of convex optimization problems under affine-equality constraints, our findings have led to significant insights:

- ❖ The efficacy of SDG stands on par with both: the KKT error and PDG given specific conditions.
- ❖ By assuming MSRSDL or leveraging QEBSG, we have derived that the KKT error, or SDG and PDG, respectively, serve as practical upper bounds for the optimality gap, providing an effective approximation.
- ❖ Our investigation vividly demonstrates the superior stability of SDG over the KKT error.
- ❖ Although our methodology is rooted in affine-equality constraints, our findings extend their applicability to encompass other problems entailing inequality constraints.

This work opens several perspectives:

- ❖ Given our utilization of the QEBSG assumption in establishing the PDG approximation for the OG, it prompts the question: could enhancing PDG with a distinct regularity assumption potentially improve its performance?
- ❖ Is it feasible to develop a novel algorithm leveraging SDG as its stopping criterion?
- ❖ What about the applicability of our findings in non-convex optimization settings?

A MSR and QEB for LC-LS

▶ **Lemma [10]** *For the LC-LS problem, let* $z^\star = \mathrm{Proj}_{\mathcal{Z}^\star}(z)$. *Then, for any* $z \in \mathcal{Z}$ *the Lagrangian's sub-differential satisfies:*

$$\|\nabla_x \mathcal{L}(x,y)\| + \|\nabla_y \mathcal{L}(x,y)\| \geq |\lambda(\mathcal{M})|_{\min} \operatorname{dist}(z, \mathcal{Z}^\star)$$
$$\mathcal{M} = \begin{bmatrix} Q^T Q & A^T \\ A & 0 \end{bmatrix}$$

where $|\lambda(\mathcal{M})|_{\min}$ *is the smallest positive eigenvalue of* $\mathcal{M}$.

Proof For any $z \in \mathcal{Z}$, we have:

$$\begin{aligned} \|\nabla_x \mathcal{L}(x,y)\| + \|\nabla_y \mathcal{L}(x,y)\| &\geq \|\nabla \mathcal{L}(x,y)\| \\ &= \|\nabla \mathcal{L}(x,y) - \nabla \mathcal{L}(x^\star, y^\star)\| \\ &= \left\| \begin{bmatrix} Q^T Q(x - x^\star) + A^T(y - y^\star) \\ A(x - x^\star) \end{bmatrix} \right\| \\ &= \|\mathcal{M}(z - z^\star)\| \\ &\overset{(4)}{\geq} |\lambda(\mathcal{M})|_{\min} \operatorname{dist}(z, \mathcal{Z}^*) \end{aligned}$$

□

▶ **Lemma [11]** *For the LC-LS problem, the self-centered smoothed gap could be reformulated into a quadratic form. That is, for any $z \in \mathcal{Z}$:*

$$\mathcal{G}_\beta(z) = z^T \mathcal{H} z + \langle z, v \rangle + cst$$

where,

$$v_{(n+m)} = \begin{bmatrix} v_x \\ v_y \end{bmatrix} \qquad \mathcal{H}_{(n+m)\times(n+m)} = \begin{bmatrix} M_{xx} & \frac{1}{2}M_{xy} \\ \frac{1}{2}M_{xy}^T & M_{yy} \end{bmatrix}$$

Vector - Matrix	Size
$v_x = \frac{\beta_x}{\tau} B^{-1} Q^T c - Q^T c$	$n \times 1$
$v_y = -AB^{-1}Q^T c$	$m \times 1$
$B = Q^T Q + \frac{\beta_x}{\tau} \mathrm{Id}_n$	$n \times n$
$M_{xx} = \frac{1}{2}Q^T Q + \frac{\sigma}{2\beta_y} A^T A + \frac{\beta_x^2}{2\tau^2} B^{-1} - \frac{\beta_x}{2\tau}\mathrm{Id}_n$	$n \times n$
$M_{xy} = A^T - \frac{\beta_x}{\tau} B^{-1} A^T$	$n \times m$
$M_{yy} = \frac{1}{2} A B^{-1} A^T$	$m \times m$

Proof We start with the definition of the smoothed duality gap (eqn:13):

$$\begin{aligned}
\mathcal{G}_\beta(z) &= \sup_{z' \in \mathcal{Z}} \mathcal{L}(x, y') - \mathcal{L}(x', y) - \frac{\beta_x}{2\tau}\|x' - x\|^2 - \frac{\beta_y}{2\sigma}\|y' - y\|^2 \\
&= \frac{1}{2}\|Qx - c\|^2 + \langle b, y\rangle + \sup_{y'} \left(\langle Ax - b, y' \rangle - \frac{\beta_y}{2\sigma}\|y' - y\|^2 \right) \\
&\qquad + \sup_{x'} \left(-\frac{1}{2}\|Qx' - c\|^2 - \langle x', A^T y\rangle - \frac{\beta_x}{2\tau}\|x' - x\|^2 \right) \\
&= \frac{1}{2}\|Qx - c\|^2 + \langle b, y\rangle + \left(\langle Ax - b, \tilde{y} \rangle - \frac{\beta_y}{2\sigma}\|\tilde{y} - y\|^2 \right) \\
&\qquad - \left(\frac{1}{2}\|Q\tilde{x} - c\|^2 - \langle \tilde{x}, A^T y\rangle - \frac{\beta_x}{2\tau}\|\tilde{x} - x\|^2 \right)
\end{aligned}$$

$$\text{with} \begin{cases} \tilde{y} & = \arg\max_{y'} \left(\langle Ax - b, y' \rangle - \frac{\beta_y}{2\sigma}\|y' - y\|^2 \right) \\ & = y + \frac{\sigma}{\beta_y}(Ax - b) \\ \tilde{x} & = \arg\max_{x'} \left(-\frac{1}{2}\|Qx' - c\|^2 - \langle x', A^T y\rangle - \frac{\beta_x}{2\tau}\|x' - x\|^2 \right) \\ & = B^{-1}\left(Q^T c - A^T y + \frac{\beta_x}{\tau} x \right) \text{ where } B = Q^T Q + \frac{\beta_x}{\tau}\mathrm{Id} \end{cases}$$

Substituting $\tilde{x}$ and $\tilde{y}$, simplifying, and gathering the related terms we get:

$$\begin{aligned}
\mathcal{G}_\beta(z) = \overbrace{\frac{1}{2}\|Qx\|^2 + \frac{\sigma}{2\beta_y}\|Ax\|^2 - \frac{1}{2}\left\|\frac{\beta_x}{\tau} Q B^{-1} x\right\|^2 - \frac{\beta_x}{2\tau}\left\|\left(\frac{\beta_x}{\tau} B^{-1} - \mathrm{Id}\right) x\right\|^2}^{T_{xx}} \\
\underbrace{+\langle x, A^T y\rangle - \frac{\beta_x}{\tau}\left(\left\langle B^{-1} x, A^T y \right\rangle - \left\langle Q B^{-1} x, Q B^{-1} A^T y \right\rangle - \left\langle \left(\frac{\beta_x}{\tau} B^{-1} - \mathrm{Id}\right) x, B^{-1} A^T y \right\rangle \right)}_{T_{xy}} \\
\underbrace{+\left\langle B^{-1} A^T y, A^T y \right\rangle - \frac{1}{2}\left\| Q B^{-1} A^T y \right\|^2 - \frac{\beta_x}{2\tau}\left\| B^{-1} A^T y \right\|^2}_{T_{yy}}
\end{aligned}$$

$$\underbrace{-\langle Qx, c\rangle - \frac{\sigma}{\beta_y}\langle Ax, b\rangle - \left\langle \frac{\beta_x}{\tau} QB^{-1}x, QB^{-1}Q^Tc - c\right\rangle - \frac{\beta_x}{\tau}\left\langle \left(\frac{\beta_x}{\tau}B^{-1} - \mathrm{Id}\right)x, B^{-1}Q^Tc\right\rangle}_{T_x}$$

$$\underbrace{-\left\langle B^{-1}Q^Tc, A^Ty\right\rangle + \left\langle QB^{-1}A^Ty, QB^{-1}Q^Tc - c\right\rangle + \frac{\beta_x}{\tau}\left\langle B^{-1}Q^Tc, B^{-1}A^Ty\right\rangle}_{T_y}$$

$$\underbrace{+\frac{1}{2}\|c\|^2 + \frac{\sigma}{2\beta_y}\|b\|^2 - \frac{1}{2}\left\|QB^{-1}Q^Tc - c\right\|^2 - \frac{\beta_x}{2\tau}\left\|B^{-1}Q^Tc\right\|^2}_{cst}$$

Now, we will further simplify each sub-term, for instance:

$$\begin{aligned}
T_{xx} &= \frac{1}{2}\|Qx\|^2 + \frac{\sigma}{2\beta_y}\|Ax\|^2 - \frac{1}{2}\left\|\frac{\beta_x}{\tau}QB^{-1}x\right\|^2 - \frac{\beta_x}{2\tau}\left\|\left(\frac{\beta_x}{\tau}B^{-1} - \mathrm{Id}\right)x\right\|^2 \\
&= \frac{1}{2}\langle Qx, Qx\rangle + \frac{\sigma}{2\beta_y}\langle Ax, Ax\rangle - \frac{\beta_x^2}{2\tau^2}\langle QB^{-1}x, QB^{-1}x\rangle \\
&\qquad - \frac{\beta_x}{2\tau}\left\langle \left(\frac{\beta_x}{\tau}B^{-1} - \mathrm{Id}\right)x, \left(\frac{\beta_x}{\tau}B^{-1} - \mathrm{Id}\right)x\right\rangle \\
&= \langle x, M_{xx}x\rangle
\end{aligned}$$

where

$$\begin{aligned}
M_{xx} &= \frac{1}{2}Q^TQ + \frac{\sigma}{2\beta_y}A^TA - \frac{\beta_x^2}{2\tau^2}B^{-1}Q^TQB^{-1} - \frac{\beta_x}{2\tau}\left(\frac{\beta_x}{\tau}B^{-1} - \mathrm{Id}\right)^T\left(\frac{\beta_x}{\tau}B^{-1} - \mathrm{Id}\right) \\
&= \frac{1}{2}Q^TQ + \frac{\sigma}{2\beta_y}A^TA - \frac{\beta_x^2}{2\tau^2}B^{-1}Q^TQB^{-1} - \frac{\beta_x^3}{2\tau^3}B^{-1}B^{-1} + \frac{\beta_x^2}{\tau^2}B^{-1} - \frac{\beta_x}{2\tau}\mathrm{Id} \\
&= \frac{1}{2}Q^TQ + \frac{\sigma}{2\beta_y}A^TA - \frac{\beta_x^2}{2\tau^2}B^{-1}\underbrace{\left(Q^TQ + \frac{\beta_x}{\tau}\mathrm{Id}\right)}_{B}B^{-1} + \frac{\beta_x^2}{\tau^2}B^{-1} - \frac{\beta_x}{2\tau}\mathrm{Id} \\
&= \frac{1}{2}Q^TQ + \frac{\sigma}{2\beta_y}A^TA - \frac{\beta_x^2}{2\tau^2}B^{-1}BB^{-1} + \frac{\beta_x^2}{\tau^2}B^{-1} - \frac{\beta_x}{2\tau}\mathrm{Id} \\
&= \frac{1}{2}Q^TQ + \frac{\sigma}{2\beta_y}A^TA + \frac{\beta_x^2}{2\tau^2}B^{-1} - \frac{\beta_x}{2\tau}\mathrm{Id}
\end{aligned}$$

Same kind of simplifications could be done for the other sub-terms. Therefore, we can rewrite the smoothed duality gap as follows:

$$\begin{aligned}
\mathcal{G}_\beta(z) &= \langle x, M_{xx}x\rangle + \langle x, M_{xy}y\rangle + \langle y, M_{yy}y\rangle + \langle x, v_x\rangle + \langle y, v_y\rangle + cst \\
&= z^T\mathcal{H}z + \langle z, v\rangle + cst
\end{aligned}$$

□

B PDG in terms of SDG

Within this appendix, we provide an exhaustive proof of Theorem 8. We start by defining two vectors pivotal to demonstrating the result, along with outlining their key properties.

1. The first one is:

$$p^* := \mathrm{Prox}_{\beta_x f^*}\left(\beta_x x - A^Ty\right) \tag{52}$$

Where $f^* \in \Gamma_0(\mathcal{X})$ is the Fenchel-Conjugate of the function f. Then, by Moreau's identity (6), we get:

$$p + \frac{1}{\beta_x}p^* = x \tag{53}$$

2. The second one comes from Lemma 5: $p = \text{Prox}_{\beta_x^{-1} f}\left(x - \frac{1}{\beta_x} A^T y\right) \iff \beta_x(x-p) \in \partial f(p) + A^T y$. So, we define:

$$\tilde{a} := -A^T y + \beta_x(x-p) \in \partial f(p) \tag{54}$$

Remark that since $\tilde{a} \in \partial f(p)$ this ensures that $\tilde{a} \in \text{dom} f^*$, which is a necessary condition for the following properties to hold.

$$\text{By (2)} \qquad \left\|a + A^T y\right\| \leq \left\|\tilde{a} + A^T y\right\| = \|p^*\| \tag{55}$$

$$\text{By (3)} \qquad \left\langle -A^T y - a, \tilde{a} - a \right\rangle \leq 0 \tag{56}$$

$$\text{By (4)} \qquad f(p) + f^*(\tilde{a}) = \langle p, \tilde{a} \rangle \tag{57}$$

Lemma 12 *For a, p^*, and $\tilde{a}$ defined, respectively, in (12), (52), and (54) we have:*

$$\|a - \tilde{a}\| \leq \|p^*\| \tag{58}$$

Proof

$$\begin{aligned}
\|a-\tilde{a}\|^2 &= \left\|a + A^T y\right\|^2 + \left\|\tilde{a} + A^T y\right\|^2 - 2\left\langle a + A^T y, \tilde{a} + A^T y\right\rangle \\
&= \left\|a + A^T y\right\|^2 + \left\|\tilde{a} + A^T y\right\|^2 + 2\left\langle -a - A^T y, \tilde{a} - a\right\rangle + 2\left\langle -a - A^T y, a + A^T y\right\rangle \\
&\overset{(56)}{\leq} \left\|a + A^T y\right\|^2 + \left\|\tilde{a} + A^T y\right\|^2 + 0 - 2\left\|a + A^T y\right\|^2 \\
&\leq \left\|\tilde{a} + A^T y\right\|^2
\end{aligned}$$

□

Corollary 3 *Given $\beta = (\beta_x, \beta_y) \in [0, +\infty]^2$, then for p^* defined in (52) and for all $z \in \mathcal{Z}$ the self-centered smoothed duality gap satisfies:*

$$\mathcal{G}_\beta(z) \geq \frac{1}{2\beta_x}\|p^*\|^2 \tag{59}$$

Proof From Lemma 6, we know:

$$\mathcal{G}_\beta(z) \geq \frac{\beta_x}{2}\|x-p\|^2 + \frac{1}{2\beta_y}\|Ax-b\|^2 \geq \frac{\beta_x}{2}\|x-p\|^2 \overset{(53)}{=} \frac{\beta_x}{2}\left\|\frac{1}{\beta_x}p^*\right\|^2$$

□

Lemma 13 *Given $\beta = (\beta_x, \beta_y) \in [0, +\infty]^2$, a function $f \in \Gamma_0(\mathcal{X})$, and let $a = \text{Proj}_{\text{dom} f^*}(\mu)$. Then, for any $z \in \mathcal{Z}$ the self-centered smoothed duality gap defined in (14) satisfies:*

$$f(x) + f^*(a) + \langle b, y\rangle \geq -\left(\sqrt{2\beta_x}\|x\| + \sqrt{2\beta_y}\|y\|\right)\sqrt{\mathcal{G}_\beta(z)} \tag{60}$$

Proof

$$\begin{aligned}
f(x) + f^*(a) + \langle b, y\rangle &= f(x) + f^*(a) - \langle x, a\rangle + \langle b - Ax, y\rangle + \left\langle x, A^T y + a\right\rangle \\
&\overset{(4)}{\geq} 0 + \langle y, b - Ax\rangle + \left\langle x, a + A^T y\right\rangle \\
&\geq -\|y\|\|Ax - b\| - \|x\|\left\|a + A^T y\right\| \\
&\overset{(17)}{\geq} -\|y\|\sqrt{2\beta_y \mathcal{G}_\beta(z)} - \|x\|\left\|a + A^T y\right\| \\
&\overset{(55)}{\geq} -\|y\|\sqrt{2\beta_y \mathcal{G}_\beta(z)} - \|x\|\|p^*\| \\
&\overset{(59)}{\geq} -\|y\|\sqrt{2\beta_y \mathcal{G}_\beta(z)} - \|x\|\sqrt{2\beta_x \mathcal{G}_\beta(z)}
\end{aligned}$$

□

Lemma 14 *Given $\beta = (\beta_x, \beta_y) \in [0, +\infty]^2$ and a function $f \in \Gamma_0(\mathcal{X})$. Then, for $\tilde{a}$ defined in (54), and for any $z \in \mathcal{Z}$ the self-centered smoothed duality gap defined in (14) satisfies:*

$$f(x) + f^*(\tilde{a}) - \langle x, \tilde{a} \rangle \leq \mathcal{G}_\beta(z) \tag{61}$$

Proof

$$\begin{aligned}
\mathcal{G}_\beta(z) &= f(x) - f(p) + \langle A(x-p), y \rangle - \frac{\beta_x}{2}\|x-p\|^2 + \frac{1}{2\beta_y}\|Ax-b\|^2 \\
&\overset{(57)}{=} f(x) + f^*(\tilde{a}) - \langle p, \tilde{a} \rangle + \langle A(x-p), y \rangle - \frac{\beta_x}{2}\|x-p\|^2 + \frac{1}{2\beta_y}\|Ax-b\|^2 \\
&= f(x) + f^*(\tilde{a}) - \left\langle p, -A^T y + \beta_x(x-p) \right\rangle + \langle A(x-p), y \rangle - \frac{\beta_x}{2}\|x-p\|^2 + \frac{1}{2\beta_y}\|Ax-b\|^2 \\
&= f(x) + f^*(\tilde{a}) - \langle x, \beta_x(x-p) \rangle + \langle x-p, \beta_x(x-p) \rangle - \frac{\beta_x}{2}\|x-p\|^2 + \langle Ax, y \rangle + \frac{1}{2\beta_y}\|Ax-b\|^2 \\
&= f(x) + f^*(\tilde{a}) - \langle x, \tilde{a} \rangle + \frac{\beta_x}{2}\|x-p\|^2 + \frac{1}{2\beta_y}\|Ax-b\|^2 \\
&\geq f(x) + f^*(\tilde{a}) - \langle x, \tilde{a} \rangle
\end{aligned}$$

□

Lemma 15 *Given $\beta = (\beta_x, \beta_y) \in [0, +\infty]^2$ and a function $f \in \Gamma_0(\mathcal{X})$. Let $a = \mathrm{Proj}_{\mathrm{dom} f^*}(\mu)$, and $\beta_{\max} = \max(\beta_x, \beta_y)$. Then, the primal-dual feasibility terms in (11) could be approximated in terms of the self-centered smoothed gap defined in (14). More precisely, for any $z \in \mathcal{Z}$:*

$$\left\| a + A^T y \right\|^2 + \|Ax - b\|^2 \leq 2\beta_{\max} \mathcal{G}_\beta(z) \tag{62}$$

Proof From Lemma 6, we know:

$$\begin{aligned}
\mathcal{G}_\beta(z) &\geq \frac{\beta_x}{2}\|x-p\|^2 + \frac{1}{2\beta_y}\|Ax-b\|^2 \\
&\overset{(53)}{=} \frac{1}{2\beta_x}\|p^*\|^2 + \frac{1}{2\beta_y}\|Ax-b\|^2 \\
&\overset{(55)}{\geq} \frac{1}{2\beta_x}\left\| a + A^T y \right\|^2 + \frac{1}{2\beta_y}\|Ax-b\|^2 \\
&\geq \frac{1}{2\beta_{\max}}\left(\left\| a + A^T y \right\|^2 + \|Ax-b\|^2 \right)
\end{aligned}$$

□

Lemma 16 *Let $\mathcal{A} \subseteq \mathbb{R}^n$ be a non-empty affine space, and $x_0 \in \mathcal{A}$. Then for the vector space $V := \mathcal{A} - x_0$, which could be seen as a smooth manifold, one can define a unique coordinate chart $(\mathcal{U}, \phi)$. Moreover, for any $x, \tilde{x} \in \mathcal{A}$,*

1. *The vector $x \in \mathcal{A}$ could be rewritten in terms of the diffeomorphism ϕ. Say otherwise:*

$$x = x_0 + \phi^{-1}(\lambda), \qquad \forall \lambda \in \phi(\mathcal{U}) \tag{63}$$

2. *The defined diffeomorphism preserves the norm. Say otherwise,*

$$\|\phi(x - x_0) - \phi(\tilde{x} - x_0)\| = \|x - \tilde{x}\| \tag{64}$$

3. *Denoting $J_{\phi^{-1}}$ the Jacobian of the diffeomorphism ϕ^{-1}, we have:*

$$J_{\phi^{-1}}\left(\phi(x - x_0) - \phi(\tilde{x} - x_0)\right) = x - \tilde{x} \tag{65}$$

Proof 1. Since V is a vector space, one can find an orthonormal basis for it. Say, $\mathcal{B} = \{v_1, \ldots, v_n\}$

Thus, for any $v \in V, \exists!\ \lambda_1, \ldots, \lambda_n \in \mathbb{R}$ such that $v = \sum_{i=1}^{n} \lambda_i v_i$

Therefore, one could define the unique diffeomorphism as follows:

$$\phi\colon \mathcal{U} = V \longrightarrow \hat{\mathcal{U}} \subseteq \mathbb{R}^n \qquad \text{s.t.} \qquad \phi(v) = \phi\left(\sum_i \lambda_i v_i\right) = \lambda \tag{66}$$

Equivalently, for any $x \in \mathcal{A}, \exists! \lambda_1, \ldots, \lambda_n \in \mathbb{R}$ such that:

$$x = x_0 + v = x_0 + \sum_j \lambda_j v_j \overset{(66)}{=} x_0 + \phi^{-1}(\lambda) \tag{67}$$

□

2. Let $x, \tilde{x} \in \mathcal{A}$, then by (63), we have:

$$x = x_0 + \sum_{i=1}^{n} \lambda_i v_i \qquad\qquad \tilde{x} = x_0 + \sum_{i=1}^{n} \tilde{\lambda}_i v_i$$

Since $x - x_0$ & $\tilde{x} - x_0 \in \mathcal{U}$, we define:

$$\phi(x - x_0) := \lambda =\in \mathbb{R}^n \qquad\qquad \phi(\tilde{x} - x_0) := \tilde{\lambda} =\in \mathbb{R}^n$$

Then, by *Pythagorean theorem*, we obtain:

$$\begin{aligned}
\|x - \tilde{x}\|^2 &= \left\| \left(x_0 + \sum_{i=1}^{n} \lambda_i v_i\right) - \left(x_0 + \sum_{i=1}^{n} \tilde{\lambda}_i v_i\right) \right\|^2 \\
&= \left\| \sum_{i=1}^{n} \left(\lambda_i - \tilde{\lambda}_i\right) v_i \right\|^2 \\
&= \sum_{i=1}^{n} (\lambda_i - \tilde{\lambda}_i)^2 \\
&= \|\lambda - \tilde{\lambda}\|^2
\end{aligned}$$

□

3. Since ϕ is a diffeomorphism, then:

$$\phi(x - x_0) = \lambda \iff \phi^{-1}(\lambda) = x - x_0 \overset{(67)}{=} \sum_{i=1}^{n} \lambda_i v_i \Longrightarrow J_{\phi^{-1}} = \begin{bmatrix} v_1 & \ldots & v_n \end{bmatrix}$$

Hence,

$$\begin{aligned}
J_{\phi^{-1}}\left(\phi(x - x_0) - \phi(\tilde{x} - x_0)\right) &= J_{\phi^{-1}}(\lambda - \tilde{\lambda}) \\
&= \begin{bmatrix} v_1 & \ldots & v_n \end{bmatrix} \begin{bmatrix} \lambda_1 - \tilde{\lambda}_1 \\ \vdots \\ \lambda_n - \tilde{\lambda}_n \end{bmatrix} \\
&= \sum_{i=1}^{n} (\lambda_i - \tilde{\lambda}_i) v_i \\
&\overset{(67)}{=} x - \tilde{x}
\end{aligned}$$

□

▶ **Theorem [8]** *Given $\beta = (\beta_x, \beta_y) \in (0, +\infty)^2$, $z \in \mathcal{Z}$, and a function $f \in \Gamma_0(\mathcal{X})$. Then, under the following set of assumptions (we denote it $\mathcal{E}$):*

– *The Fenchel-Conjugate of the objective function, f^*, could be written in a separable way:*

$$f^*(\mu) = f_1^*(\mu_1) + f_2^*(\mu_2), \qquad \mu \in \mathcal{X} \tag{68}$$

– *f_1^* is $L_{f_1^*}-$Lipschitz on its domain,* $\mathrm{dom} f_1^*$.
– *The domain of f_2^* is a non-empty affine space.*
– *Let $\mu_0 \in \mathrm{dom} f_2^*$, then $\forall \mu_2 \in \mathrm{dom} f_2^*$, we define*

$$g(\lambda) = f_2^*\left(\mu_0 + \phi^{-1}(\lambda)\right) = f_2^*(\mu_2) \tag{69}$$

where ϕ is the diffeomorphism defined in Lemma 16 (eqn:66).
– *The function g is differentiable and has an L_g-Lipschitz gradient.*

The projected duality gap and the smoothed duality gap defined, respectively, in (11) and (14) satisfy:

$$\mathcal{D}(z) \leq \left((3+\beta_x L_g)\,\mathcal{G}_\beta(z) + \left(\sqrt{2\beta_x}\left(2\|x\| + L_{f_1^*}\right) + \sqrt{2\beta_y}\|y\|\right)\sqrt{\mathcal{G}_\beta(z)}\right)^2 + 2\beta_{\max}\mathcal{G}_\beta(z)$$

$$a = \mathrm{Proj}_{\mathrm{dom} f^*}\left(-A^T y\right) \qquad p = \mathrm{Prox}_{\beta_x^{-1} f}\left(x - \frac{1}{\beta_x} A^T y\right)$$

Proof We initiate by analyzing and interpreting the assumptions:

❖ $f^*(\mu) = f_1^*(\mu_1) + f_2^*(\mu_2)$ is separable implies that:

$$\partial f^*(\mu) = \partial f_1^*(\mu_1) \times \partial f_2^*(\mu_2) \tag{70}$$
$$\mathrm{dom} f^* = \mathrm{dom} f_1^* \times \mathrm{dom} f_2^* \tag{71}$$

So, $p \in \partial f^*(\tilde{a}) \overset{(70)}{=} \partial f_1^*(\tilde{a}_1) \times \partial f_2^*(\tilde{a}_2)$ if, and only if,

$$p = (p_1, p_2)\ \&\ \tilde{a} = (\tilde{a}_1, \tilde{a}_2) \quad \text{s.t.} \quad p_1 \in \partial f_1^*(\tilde{a}_1)\ \&\ p_2 \in \partial f_2^*(\tilde{a}_2) \tag{72}$$

Also, for

$$\begin{aligned} a = \mathrm{Proj}_{\mathrm{dom} f^*}\left(-A^T y\right) &\overset{(71)}{=} \mathrm{Proj}_{\mathrm{dom} f_1^* \times \mathrm{dom} f_2^*}\left(-A^T y\right) \\ &= \left(\mathrm{Proj}_{\mathrm{dom} f_1^*}\left(-\left(A^T y\right)_1\right), \mathrm{Proj}_{\mathrm{dom} f_2^*}\left(-\left(A^T y\right)_2\right)\right) \end{aligned}$$

So, for $A^T y = \left(\left(A^T y\right)_1, \left(A^T y\right)_2\right)$, we let $a = (a_1, a_2)$ be defined as follows:

$$a_1 = \mathrm{Proj}_{\mathrm{dom} f_1^*}\left(-\left(A^T y\right)_1\right) \qquad a_2 = \mathrm{Proj}_{\mathrm{dom} f_2^*}\left(-\left(A^T y\right)_2\right) \tag{73}$$

❖ f_1^* is $L_{f_1^*}-$Lipschitz on its domain implies:

$$f_1^*(a_1) \leq f_1^*(\tilde{a}_1) + L_{f_1^*}\|a_1 - \tilde{a}_1\| \tag{74}$$

❖ By Lemma 16, we have seen that we can rewrite any vector $\mu_2 \in \mathrm{dom} f_2^*$ as:

$$\mu_2 = \mu_0 + \sum_{i=1}^{n} \lambda_i v_i = \mu_0 + \phi^{-1}(\lambda) \tag{75}$$

where $\mu_0 \in \mathrm{dom} f_2^*, \mathcal{B} = \{v_1, \ldots, v_n\}$ is an orthonormal basis for $V = \mathrm{dom} f_2^* - \mu_0$, and $\phi(\mu_2 - \mu_0 \in V) := \lambda$. Thus, for any $\mu_2 \in \mathrm{dom} f_2^*$:

$$f_2^*(\mu_2) = f_2^*\left(\mu_0 + \sum_{i=1}^{n} \lambda_i v_i\right) = f_2^*\left(\mu_0 + \phi^{-1}(\lambda)\right) \tag{76}$$

Hence, by defining $g(\lambda) = f_2^*(\mu_0 + \phi^{-1}(\lambda))$ and assuming that g is differentiable, we mean that the function f_2^* is differentiable on its domain.

- The function g is differentiable, so by the chain rule: $\{\nabla g(\lambda)\} = J_{\phi^{-1}}^T \partial f_2^*(\mu_0 + \phi^{-1}(\lambda))$. Hence,
$$\nabla g(\lambda) = J_{\phi^{-1}}^T.\omega, \qquad \forall \omega \in \partial f_2^*(\mu_0 + \phi^{-1}(\lambda)) \tag{77}$$
- The function g has an L_g-Lipschitz gradient implies that the Taylor-Lagrange inequality holds. That is: for any $\lambda, \tilde{\lambda} \in \hat{\mathcal{U}} \subseteq \mathbb{R}^n$, we have:
$$g(\lambda) \leq g(\tilde{\lambda}) + \left\langle \nabla g(\tilde{\lambda}), \lambda - \tilde{\lambda} \right\rangle + \frac{L_g}{2}\|\lambda - \tilde{\lambda}\|^2 \tag{78}$$

Now, for the points $a_2, \tilde{a}_2 \in \mathrm{dom} f_2^*$, we define:
$$\lambda := \phi(a_2 - \mu_0) \qquad \tilde{\lambda} := \phi(\tilde{a}_2 - \mu_0) \tag{79}$$

Thus, by (72, 77, and 78), we obtain:
$$g(\lambda) \leq g(\tilde{\lambda}) + \left\langle J_{\phi^{-1}}^T.p_2, \lambda - \tilde{\lambda} \right\rangle + \frac{L_g}{2}\|\lambda - \tilde{\lambda}\|^2 \tag{80}$$

Equivalently:
$$f_2^*(a_2) \leq f_2^*(\tilde{a}_2) + \left\langle p_2, J_{\phi^{-1}}(\lambda - \tilde{\lambda}) \right\rangle + \frac{L_g}{2}\|\lambda - \tilde{\lambda}\|^2 \tag{81}$$

Everything is ready now to be used to upper-bound the term:

$$\begin{aligned}
&f(x) + f^*(a) + \langle b, y\rangle \\
&\overset{(68)}{=} f(x) + f_1^*(a_1) + f_2^*(a_2) + \langle b, y\rangle \\
&\overset{(74,81)}{\leq} f(x) + f_1^*(\tilde{a}_1) + L_{f_1^*}\|a_1 - \tilde{a}_1\| + f_2^*(\tilde{a}_2) + \left\langle p_2, J_{\phi^{-1}}(\lambda - \tilde{\lambda}) \right\rangle + \frac{L_g}{2}\|\lambda - \tilde{\lambda}\|^2 + \langle b, y\rangle \\
&\overset{(64,65)}{=} f(x) + f_1^*(\tilde{a}_1) + L_{f_1^*}\|a_1 - \tilde{a}_1\| + f_2^*(\tilde{a}_2) + \langle p_2, a_2 - \tilde{a}_2\rangle + \frac{L_g}{2}\|a_2 - \tilde{a}_2\|^2 + \langle b, y\rangle \\
&= f(x) + f_1^*(\tilde{a}_1) + f_2^*(\tilde{a}_2) + L_{f_1^*}\|a_1 - \tilde{a}_1\| + \langle p_2, a_2 - \tilde{a}_2\rangle + \frac{L_g}{2}\|a_2 - \tilde{a}_2\|^2 + \langle b, y\rangle \\
&= f(x) + f^*(\tilde{a}) + L_{f_1^*}\|a_1 - \tilde{a}_1\| + \langle p_2, a_2 - \tilde{a}_2\rangle + \frac{L_g}{2}\|a_2 - \tilde{a}_2\|^2 - \langle x, \tilde{a}\rangle - \langle x, a - \tilde{a}\rangle \\
&\qquad + \langle x, a + A^T y\rangle - \langle Ax - b, y\rangle \\
&= f(x) + f^*(\tilde{a}) - \langle x, \tilde{a}\rangle + L_{f_1^*}\|a_1 - \tilde{a}_1\| + \langle p_2, a_2 - \tilde{a}_2\rangle + \frac{L_g}{2}\|a_2 - \tilde{a}_2\|^2 - \langle x, a - \tilde{a}\rangle \\
&\qquad + \langle x, a + A^T y\rangle - \langle Ax - b, y\rangle \\
&\overset{(61)}{\leq} \mathcal{G}_\beta(z) + L_{f_1^*}\|a_1 - \tilde{a}_1\| + \langle p_2 - x_2, a_2 - \tilde{a}_2\rangle + \frac{L_g}{2}\|a_2 - \tilde{a}_2\|^2 - \langle x_1, a_1 - \tilde{a}_1\rangle \\
&\qquad + \langle x, a + A^T y\rangle - \langle Ax - b, y\rangle \\
&\leq \mathcal{G}_\beta(z) + L_{f_1^*}\|a_1 - \tilde{a}_1\| + \|x_1\|\|a_1 - \tilde{a}_1\| + \|p_2 - x_2\|\|a_2 - \tilde{a}_2\| + \frac{L_g}{2}\|a_2 - \tilde{a}_2\|^2 \\
&\qquad + \|x\|\left\|a + A^T y\right\| + \|Ax - b\|\,\|y\| \\
&\overset{(53)}{=} \mathcal{G}_\beta(z) + \left(L_{f_1^*} + \|x_1\|\right)\|a_1 - \tilde{a}_1\| + \frac{1}{\beta_x}\|p_2^*\|\|a_2 - \tilde{a}_2\| + \frac{L_g}{2}\|a_2 - \tilde{a}_2\|^2 \\
&\qquad + \|x\|\left\|a + A^T y\right\| + \|Ax - b\|\,\|y\| \\
&\leq \mathcal{G}_\beta(z) + \left(L_{f_1^*} + \|x\|\right)\|a_1 - \tilde{a}_1\| + \frac{1}{\beta_x}\|p^*\|\|a_2 - \tilde{a}_2\| + \frac{L_g}{2}\|a_2 - \tilde{a}_2\|^2 + \|x\|\left\|a + A^T y\right\| \\
&\qquad + \|Ax - b\|\,\|y\| \\
&\overset{(55)}{\leq} \mathcal{G}_\beta(z) + \left(L_{f_1^*} + \|x\|\right)\|a_1 - \tilde{a}_1\| + \frac{1}{\beta_x}\|p^*\|\|a_2 - \tilde{a}_2\| + \frac{L_g}{2}\|a_2 - \tilde{a}_2\|^2 + \|x\|\|p^*\| \\
&\qquad + \|Ax - b\|\,\|y\|
\end{aligned}$$

$$
\begin{aligned}
&\overset{(58)}{\leq} \mathcal{G}_\beta(z) + \left(L_{f_1^*} + \|x\|\right)\|p^*\| + \frac{1}{\beta_x}\|p^*\|^2 + \frac{L_g}{2}\|p^*\|^2 + \|x\|\|p^*\| + \|Ax-b\|\,\|y\| \\
&\overset{(59)}{\leq} \mathcal{G}_\beta(z) + \left(L_{f_1^*} + \|x\|\right)\sqrt{2\beta_x\mathcal{G}_\beta(z)} + 2\mathcal{G}_\beta(z) + \beta_x L_g \mathcal{G}_\beta(z) + \|x\|\sqrt{2\beta_x\mathcal{G}_\beta(z)} \\
&\qquad\qquad + \|Ax-b\|\|y\| \\
&\overset{(17)}{\leq} (3+\beta_x L_g)\,\mathcal{G}_\beta(z) + \left(L_{f_1^*} + 2\|x\|\right)\sqrt{2\beta_x\mathcal{G}_\beta(z)} + \|y\|\sqrt{2\beta_y\mathcal{G}_\beta(z)}
\end{aligned}
$$

Consequently, we have derived an upper bound for the term $f(x) + f^*(a) + \langle b, y\rangle$. Thanks to Lemma 13 that provides us with a lower bound of that term as well. By combining these two bounds, we obtain:

$$
|f(x) + f^*(a) + \langle b, y\rangle| \leq (3+\beta_x L_g)\,\mathcal{G}_\beta(z) + \left(\sqrt{2\beta_x}\left(2\|x\| + L_{f_1^*}\right) + \sqrt{2\beta_y}\|y\|\right)\sqrt{\mathcal{G}_\beta(z)} \tag{82}
$$

Lastly, Lemma 15 along with this last bound (82) conclude the proof:

$$
\begin{aligned}
\mathcal{D}(z) &= |f(x) + f^*(a) + \langle b, y\rangle|^2 + \|Ax-b\|^2 + \left\|a + A^T y\right\|^2 \\
&\overset{(82)}{\leq} \left(\mathcal{G}_\beta(z) + \left(\sqrt{2\beta_x}\left(\|x\| + L_{f^*}\right) + \sqrt{2\beta_y}\|y\|\right)\sqrt{\mathcal{G}_\beta(z)}\right)^2 + \|Ax-b\|^2 + \left\|a + A^T y\right\|^2 \\
&\overset{(62)}{\leq} \left(\mathcal{G}_\beta(z) + \left(\sqrt{2\beta_x}\left(\|x\| + L_{f^*}\right) + \sqrt{2\beta_y}\|y\|\right)\sqrt{\mathcal{G}_\beta(z)}\right)^2 + 2\beta_{\max}\mathcal{G}_\beta(z) \qquad \square
\end{aligned}
$$

Upon setting $f_2^* = 0$ in our preceding theorem, we managed to derive a slightly tighter bound. The detailed expression is presented below.

Proposition 9 *Given $\beta = (\beta_x, \beta_y) \in (0, +\infty)^2$, $z \in \mathcal{Z}$, and a function $f \in \Gamma_0(\mathcal{X})$ such that its Fenchel-conjugate, f^*, is $L_{f^*}-$Lipschitz on its domain. Let $\beta_{\max} = \max(\beta_x, \beta_y)$. Then, for the projected duality gap and the smoothed duality gap defined, respectively, in (11) and (14) we have:*

$$
\begin{aligned}
&\mathcal{D}(z) \leq \left(\mathcal{G}_\beta(z) + \left(\sqrt{2\beta_x}\left(\|x\| + L_{f^*}\right) + \sqrt{2\beta_y}\|y\|\right)\sqrt{\mathcal{G}_\beta(z)}\right)^2 + 2\beta_{\max}\mathcal{G}_\beta(z) \\
&a = \mathrm{Proj}_{\mathrm{dom} f^*}\left(-A^T y\right) \qquad\qquad p = \mathrm{Prox}_{\beta_x^{-1} f}\left(x - \frac{1}{\beta_x}A^T y\right)
\end{aligned} \tag{83}
$$

Proof The main contrast between the proof of Theorem 8 and this one lies in the upper bound established for the term $f(x) + f^*(a) + \langle b, y\rangle$. However, the rest of the proof remains unchanged. Since f^* is $L_{f^*}-$Lipschitz on its domain, then:

$$
f^*(a) \leq L_{f^*}\|a - \tilde{a}\| + f^*(\tilde{a}) \tag{84}
$$

Therefore,

$$
\begin{aligned}
f(x) + f^*(a) + \langle b, y\rangle &\overset{(84)}{\leq} f(x) + f^*(\tilde{a}) + L_{f^*}\|a-\tilde{a}\| + \langle b, y\rangle \\
&= f(x) + f^*(\tilde{a}) - \langle x, \tilde{a}\rangle - \langle Ax - b, y\rangle + \left\langle x, \tilde{a} + A^T y\right\rangle + L_{f^*}\|a-\tilde{a}\| \\
&\overset{(61)}{\leq} \mathcal{G}_\beta(z) - \langle Ax-b, y\rangle + \left\langle x, \tilde{a} + A^T y\right\rangle + L_{f^*}\|a-\tilde{a}\| \\
&\leq \mathcal{G}_\beta(z) + \|Ax-b\|\,\|y\| + \|x\|\left\|\tilde{a} + A^T y\right\| + L_{f^*}\|a-\tilde{a}\| \\
&\overset{(55)}{\leq} \mathcal{G}_\beta(z) + \|Ax-b\|\,\|y\| + \|x\|\|p^*\| + L_{f^*}\|a-\tilde{a}\| \\
&\leq \mathcal{G}_\beta(z) + \|y\|\sqrt{2\beta_y\mathcal{G}_\beta(z)} + \|x\|\sqrt{2\beta_x\mathcal{G}_\beta(z)} + L_{f^*}\sqrt{2\beta_x\mathcal{G}_\beta(z)}
\end{aligned} \tag{85}
$$

where in the last line we utilized the upper bounds (17, 59, 58), respectively. Now, from this upper bound (85) and the earlier-proven lower-bound (60), we get:

$$|f(x) + f^*(a) + \langle b, y\rangle| \leq \mathcal{G}_\beta(z) + \left(\sqrt{2\beta_x}\left(\|x\| + L_{f^*}\right) + \sqrt{2\beta_y}\|y\|\right)\sqrt{\mathcal{G}_\beta(z)} \tag{86}$$

Therefore, Lemma 15 along with this last bound (86) conclude the proof:

$$\begin{aligned}\mathcal{D}(z) &= |f(x) + f^*(a) + \langle b, y\rangle|^2 + \|Ax - b\|^2 + \left\|a + A^T y\right\|^2 \\ &\leq \left((3 + \beta_x L_g)\,\mathcal{G}_\beta(z) + \left(\sqrt{2\beta_x}\left(2\|x\| + L_{f_1^*}\right) + \sqrt{2\beta_y}\|y\|\right)\sqrt{\mathcal{G}_\beta(z)}\right)^2 + 2\beta_{\max}\mathcal{G}_\beta(z)\end{aligned}$$

□

Acknowledgements This work was supported by the Agence National de la Recherche grant ANR-20-CE40-0027, Optimal Primal-Dual Algorithms (APDO), and Télécom Paris's research and teaching chair Data Science and Artificial Intelligence for Digitalized Industry and Services DSAIDIS.

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Alacaoglu, A., Fercoq, O., Cevher, V.: On the convergence of stochastic primal-dual hybrid gradient. SIAM Journal on Optimization **32**(2), 1288–1318 (2022). DOI 10.1137/19M1296252. URL https://doi.org/10.1137/19M1296252
2. Andersen, E.D., Andersen, K.D.: The Mosek Interior Point Optimizer for Linear Programming: An Implementation of the Homogeneous Algorithm, pp. 197–232. Springer US, Boston, MA (2000). DOI 10.1007/978-1-4757-3216-0_8. URL https://doi.org/10.1007/978-1-4757-3216-0_8
3. Applegate, D., Díaz, M., Hinder, O., Lu, H., Lubin, M., O'Donoghue, B., Schudy, W.: Practical large-scale linear programming using primal-dual hybrid gradient. Advances in Neural Information Processing Systems **34**, 20243–20257 (2021)
4. Arablouei, R., Doğançay, K.: Performance analysis of linear-equality-constrained least-squares estimation. IEEE Transactions on Signal Processing **63**(14), 3762–3769 (2015). DOI 10.1109/TSP.2015.2424199
5. Bianchi, P., Hachem, W., Iutzeler, F.: A stochastic primal-dual algorithm for distributed asynchronous composite optimization. In: 2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP), pp. 732–736 (2014). DOI 10.1109/GlobalSIP.2014.7032215
6. Byrd, R.H., Nocedal, J., Waltz, R.A.: Knitro: An Integrated Package for Nonlinear Optimization, pp. 35–59. Springer US, Boston, MA (2006). DOI 10.1007/0-387-30065-1_4. URL https://doi.org/10.1007/0-387-30065-1_4
7. Chambolle, A., Pock, T.: A first-order primal-dual algorithm for convex problems with applications to imaging. Journal of Mathematical Imaging and Vision **40** (2011). DOI 10.1007/s10851-010-0251-1
8. Chen, S., Donoho, D.L.: Basis pursuit. Proceedings of 1994 28th Asilomar Conference on Signals, Systems and Computers **1**, 41–44 vol.1 (1994). URL https://api.semanticscholar.org/CorpusID:96447294
9. Combettes, P.L., Wajs, V.R.: Signal recovery by proximal forward-backward splitting. Multiscale Modeling & Simulation **4**(4), 1168–1200 (2005). DOI 10.1137/050626090. URL https://doi.org/10.1137/050626090
10. Condat, L.: A primal–dual splitting method for convex optimization involving lipschitzian, proximable and linear composite terms. Journal of Optimization Theory and Applications **158** (2013). DOI 10.1007/s10957-012-0245-9

11. Dantzig, G.B.: Linear programming. Operations Research **50**(1), 42–47 (2002). DOI 10.1287/opre.50.1.42.17798. URL https://doi.org/10.1287/opre.50.1.42.17798
12. Daubechies, I.: Ten Lectures on Wavelets. CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics (1992). URL https://books.google.fr/books?id=B3C5aG4OboIC
13. Donoho, D.: Compressed sensing. Information Theory, IEEE Transactions on **52**, 1289 – 1306 (2006). DOI 10.1109/TIT.2006.871582
14. Dutta, J., Deb, K., Tulshyan, R., Arora, R.: Approximate kkt points and a proximity measure for termination. Journal of Global Optimization **56** (2013). DOI 10.1007/s10898-012-9920-5
15. Fercoq, O.: Quadratic error bound of the smoothed gap and the restarted averaged primal-dual hybrid gradient. Open Journal of Mathematical Optimization **4**, 6 (2023). DOI 10.5802/ojmo.26. URL https://ojmo.centre-mersenne.org/articles/10.5802/ojmo.26/
16. Haeser, G., Melo, V.: Convergence detection for optimization algorithms: Approximate-kkt stopping criterion when lagrange multipliers are not available. Operations Research Letters **43**, 484–488 (2015). DOI 10.1016/j.orl.2015.06.009
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). DOI 10.1109/TAC.2019.2906924
18. Lu, H., Yang, J.: On the infimal sub-differential size of primal-dual hybrid gradient method and beyond (2023)
19. McCarl, B.A., Moskowitz, H., Furtan, H.: Quadratic programming applications. Omega **5**(1), 43–55 (1977). DOI https://doi.org/10.1016/0305-0483(77)90020-2. URL https://www.sciencedirect.com/science/article/pii/0305048377900202
20. Nesterov, Y.: Smooth minimization of non-smooth functions. Mathematical Programming **103**, 127–152 (2005). URL https://api.semanticscholar.org/CorpusID:2391217
21. R. Tyrrell Rockafellar, R.J.B.W.: Variational Analysis. Springer Berlin, Heidelberg (26 June 2009). URL libgen.li/file.php?md5=b8624d493f5332a34a826b9cc74b88b1
22. Rockafellar, R.: Convex Analysis. Princeton Landmarks in Mathematics and Physics. Princeton University Press (1970). URL https://books.google.fr/books?id=1TiOka9bx3sC
23. Stephen Boyd, L.V.: Convex Optimization. Cambridge University Press (2004). URL libgen.li/file.php?md5=b8624d493f5332a34a826b9cc74b88b1
24. Taha, H.: Operations research: An introduction. Journal of Manufacturing Systems **17**(1), 78 (1998)
25. Tibshirani, R.J.: The lasso problem and uniqueness. Electronic Journal of Statistics (2013)
26. 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). DOI 10.1137/16M1093094. URL https://doi.org/10.1137/16M1093094
27. Van Loan, C.: On the method of weighting for equality-constrained least-squares problems. SIAM Journal on Numerical Analysis **22**(5), 851–864 (1985). DOI 10.1137/0722051. URL https://doi.org/10.1137/0722051
28. Zou, H.: The adaptive lasso and its oracle properties. Journal of the American statistical association **101**(476), 1418–1429 (2006)