Generalized Conditional Gradient with Augmented Lagrangian for Composite Minimization

Antonio Silveti-Falls, Cesare Molinari, Jalal M. Fadili

Generalized Conditional Gradient with Augmented Lagrangian for Composite Minimization

Antonio Silveti-Falls* Cesare Molinari* Jalal Fadili*

Abstract. In this paper we propose a splitting scheme which hybridizes generalized conditional gradient with a proximal step which we call CGALP algorithm, for minimizing the sum of three proper convex and lower-semicontinuous functions in real Hilbert spaces. The minimization is subject to an affine constraint, that allows in particular to deal with composite problems (sum of more than three functions) in a separate way by the usual product space technique. While classical conditional gradient methods require Lipschitz-continuity of the gradient of the differentiable part of the objective, CGALP needs only differentiability (on an appropriate subset), hence circumventing the intricate question of Lipschitz continuity of gradients. For the two remaining functions in the objective, we do not require any additional regularity assumption. The second function, possibly nonsmooth, is assumed simple, i.e., the associated proximal mapping is easily computable. For the third function, again nonsmooth, we just assume that its domain is weakly compact and that a linearly perturbed minimization oracle is accessible. In particular, this last function can be chosen to be the indicator of a nonempty bounded closed convex set, in order to deal with additional constraints. Finally, the affine constraint is addressed by the augmented Lagrangian approach. Our analysis is carried out for a wide choice of algorithm parameters satisfying so called "open loop" rules. As main results, under mild conditions, we show asymptotic feasibility with respect to the affine constraint, boundedness of the dual multipliers, and convergence of the Lagrangian values to the saddle-point optimal value. We also provide (subsequential) rates of convergence for both the feasibility gap and the Lagrangian values.

Key words. Conditional gradient; Augmented Lagrangian; Composite minimization; Proximal mapping; Moreau envelope.

AMS subject classifications. 49J52, 65K05, 65K10.

1 Introduction

1.1 Problem Statement

In this work, we consider the composite optimization problem,

$$\min_{x\in\mathcal{H}_p} \{f(x) + g(Tx) + h(x) : \ Ax = b\}, \tag{$\mathscr{P}$}$$

where $\mathcal{H}_p, \mathcal{H}_d, \mathcal{H}_v$ are real Hilbert spaces (the subindices p, d and v denoting the "primal", the "dual" and an auxiliary space - respectively), endowed with the associated scalar products and norms (to be understood

*Normandie Université, ENSICAEN, UNICAEN, CNRS, GREYC, France. E-mail: tonys.falls@gmail.com, cecio.molinari@gmail.com, Jalal.Fadili@ensicaen.fr.

from the context), $A : \mathcal{H}_p \to \mathcal{H}_d$ and $T : \mathcal{H}_p \to \mathcal{H}_v$ are bounded linear operators, $b \in \mathcal{H}_d$ and f, g, h are proper, convex, and lower semi-continuous functions with $\mathcal{C} \stackrel{\text{def}}{=} \operatorname{dom}(h)$ being a weakly compact subset of $\mathcal{H}_p$. We allow for some *asymmetry* in regularity between the functions involved in the objective. While g is assumed to be prox-friendly, for h we assume that it is easy to compute a linearly-perturbed oracle (see (1.2)). On the other hand, f is assumed to be differentiable and satisfies a condition that generalizes Lipschitz-continuity of the gradient (see Definition 2.6).

Problem ($\mathscr{P}$) can be seen as a generalization of the classical Frank-Wolfe problem in [15] of minimizing a Lipschitz-smooth function f on a convex closed bounded subset $\mathcal{C} \subset \mathcal{H}_p$,

$$\min_{x \in \mathcal{H}_p} \{f(x) : \ x \in \mathcal{C}\} \tag{1.1}$$

In fact, if $A \equiv 0$, $b \equiv 0$, $g \equiv 0$, and $h \equiv \iota_{\mathcal{C}}$ is the indicator function of $\mathcal{C}$ then we recover exactly (1.1) from ($\mathscr{P}$).

1.2 Contribution

We develop and analyze a novel algorithm to solve ($\mathscr{P}$) which combines penalization for the nonsmooth function g with the augmented Lagrangian method for the affine constraint $Ax = b$. In turn, this achieves full splitting of all the parts in the composite problem ($\mathscr{P}$) by using the proximal mapping of g (assumed prox-friendly) and a linear oracle for h of the form (1.2). Our analysis shows that the sequence of iterates is asymptotically feasible for the affine constraint, that the sequence of dual variables converges weakly to a solution of the dual problem, that the associated Lagrangian converges to optimality, and establishes convergence rates for a family of sequences of step sizes and sequences of smoothing/penalization parameters which satisfy so-called "open loop" rules in the sense of [31] and [13]. This means that the allowable sequences of parameters do not depend on the iterates, in contrast to a "closed loop" rule, e.g. line search or other adaptive step sizes. Our analysis also shows, in the case where ($\mathscr{P}$) admits a unique minimizer, weak convergence of the whole sequence of primal iterates to the solution.

The structure of ($\mathscr{P}$) generalizes (1.1) in several ways. First, we allow for a possibly nonsmooth term g. Second, we consider h beyond the case of an indicator function where the linear oracle of the form

$$\min_{s \in \mathcal{H}} h(s) + \langle x, s \rangle \tag{1.2}$$

can be easily solved. Observe that (1.2) has a solution over $\operatorname{dom}(h)$ since the latter is weakly compact. This oracle is reminiscent of that in the generalized conditional gradient method [7, 8, 5, 3]. Third, the regularity assumptions on f are also greatly weakened to go far beyond the standard Lipschitz gradient case. Finally, handling an affine constraint in our problem means that our framework can be applied to the splitting of a wide range of composite optimization problems, through a product space technique, including those involving finitely many functions h_i and g_i, and, in particular, intersection of finitely many nonempty bounded closed convex sets; see Section 5.

1.3 Relation to prior work

In the 1950's Frank and Wolfe developed the so-called Frank-Wolfe algorithm in [15], also commonly referred to as the conditional gradient algorithm [24, 12, 13], for solving problems of the form (1.1). The main idea is to replace the objective function f with a linear model at each iteration and solve the resulting linear optimization problem; the solution to the linear model is used as a step direction and the next iterate is

computed as a convex combination of the current iterate and the step direction. We generalize this setting to include composite optimization problems involving both smooth and nonsmooth terms, intersection of multiple constraint sets, and also affine constraints.

Frank-Wolfe algorithms have received a lot of attention in the modern era due to their effectiveness in fields with high-dimensional problems like machine learning and signal processing (without being exhaustive, see, e.g., [20, 6, 22, 17, 39, 26, 10]). In the past, composite, constrained problems like ($\mathscr{P}$) have been approached using proximal splitting methods, e.g. generalized forward-backward as developed in [32] or forward-douglas-rachford [25]. Such approaches require one to compute the proximal mapping associated to the function h. Alternatively, when the objective function satisfies some regularity conditions and when the constraint set is well behaved, one can forgo computing a proximal mapping, instead computing a linear minimization oracle. The computation of the proximal step can be prohibitively expensive; for example, when h is the indicator function of the nuclear norm ball, computing the proximal operator of h requires a full singular value decomposition while the linear minimization oracle over the nuclear norm ball requires only the leading singular vector to be computed ([21], [38]). Unfortunately, the regularity assumptions required by classical Frank-Wolfe style algorithms are too restrictive to apply to general problems like ($\mathscr{P}$).

While finalizing this work, we became aware of the recent work of [37], who independently developed a conditional gradient-based framework which allows one to solve composite optimization problems involving a Lipschitz-smooth function f and a nonsmooth function g,

$$\min_{x\in\mathcal{C}} \{f(x) + g(Tx)\}. \tag{1.3}$$

The main idea is to replace g with its Moreau envelope of index β_k at each iteration k, with the index parameter β_k going to 0. This is equivalent to partial minimization with a quadratic penalization term, as in our algorithm. Like our algorithm, that of [37] is able to handle problems involving both smooth and nonsmooth terms, intersection of multiple constraint sets and affine constraints, however their algorithms employ different methods for these situations. Our algorithm uses an augmented Lagrangian to handle the affine constraint while the conditional gradient framework treats the affine constraint as a nonsmooth term g and uses penalization to smooth the indicator function corresponding to the affine constraint. In particular circumstances, outlined in more detail in Section 6, our algorithms agree completely.

Another recent and parallel work to ours is that of [16], where the Frank-Wolfe via Augmented Lagrangian (FW-AL) is developed to approach the problem of minimizing a Lipschitz-smooth function over a convex, compact set with a linear constraint,

$$\min_{x\in\mathcal{C}} \{f(x) : \ Ax = 0\}. \tag{1.4}$$

The main idea of FW-AL is to use the augmented Lagrangian to handle the linear constraint and then apply the classical augmented Lagrangian algorithm, except that the marginal minimization on the primal variable that is usually performed is replaced by an inner loop of Frank-Wolfe. It turns out that the problem they consider is a particular case of ($\mathscr{P}$), discussed in Section 6.

1.4 Organization of the paper

In Section 2 we introduce the notation and review some necessary material from convex and real analysis. In Section 3 we present the **C**onditional **G**radient with **A**ugmented **L**agrangian and **P**roximal-step **(CGALP)** algorithm and the underlying assumptions. In Section 4, we first state our main convergence results and then turn to their proof. The latter is divided in three main parts. First we show the asymptotic feasibility, then the

boundedness of the dual multiplier in the augmented Lagrangian and finally the optimality guarantees, i.e. weak convergence of the sequence $(\mu_k)_{k\in\mathbb{N}}$ to a solution of the dual problem, weak subsequential convergence of the sequence $(x_k)_{k\in\mathbb{N}}$ to a solution of the primal problem, and convergence of the Lagrangian values, and with convergence rates. In Section 5 we describe how our framework can be instantiated to solve a variety of composite optimization problems. In Section 6 we provide a more detailed discussion comparing CGALP to prior work. Some numerical results are reported in Section 7.

For readers who are primarily interested in the practical perspective, we suggest skipping directly to Section 3 for the algorithms and its assumptions or Section 4 for the main convergence results.

2 Notation and Preliminaries

We first recall some important definitions and results from convex analysis. For a more comprehensive coverage we refer the interested reader to [4, 30] and [33] in the finite dimensional case. Throughout, we let $\mathcal{H}$ denote an arbitrary real Hilbert space and g an arbitrary function from $\mathcal{H}$ to the real extended line, namely $g : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$. The function g is said to belong to $\Gamma_0(\mathcal{H})$ if it is proper, convex, and lower semi-continuous. The *domain* of g is defined to be $\operatorname{dom}(g) \stackrel{\text{def}}{=} \{x \in \mathcal{H} : g(x) < +\infty\}$. The *Legendre-Fenchel conjugate* of g is the function $g^* : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ such that, for every $u \in \mathcal{H}$,

$$g^*(u) \stackrel{\text{def}}{=} \sup_{x\in\mathcal{H}} \{\langle u, x\rangle - g(x)\}.$$

Notice that

$$g_1 \leq g_2 \quad \Longrightarrow \quad g_2^* \leq g_1^*. \tag{2.1}$$

Moreau proximal mapping and envelope The *proximal operator* for the function g is defined to be

$$\operatorname{prox}_g(x) \stackrel{\text{def}}{=} \operatorname*{argmin}_{y\in\mathcal{H}} \left\{ g(y) + \frac{1}{2}\|x-y\|^2 \right\}$$

and its *Moreau envelope* with parameter β as

$$g^\beta(x) \stackrel{\text{def}}{=} \inf_{y\in\mathcal{H}} \left\{ g(y) + \frac{1}{2\beta}\|x-y\|^2 \right\}. \tag{2.2}$$

Denoting $x^+ = \operatorname{prox}_g(x)$, we have the following classical inequality (see, for instance, [30, Chapter 6.2.1]): for every $y \in \mathcal{H}$,

$$2\left[g(x^+) - g(y)\right] + \|x^+ - y\|^2 - \|x-y\|^2 + \|x^+ - x\|^2 \leq 0. \tag{2.3}$$

We recall that the *subdifferential* of the function g is defined as the set-valued operator $\partial g : \mathcal{H} \to 2^{\mathcal{H}}$ such that, for every x in $\mathcal{H}$,

$$\partial g(x) = \left\{u \in \mathcal{H} :\ g(y) \geq g(x) + \langle u, y-x\rangle \quad \forall y \in \mathcal{H}\right\}. \tag{2.4}$$

We denote $\operatorname{dom}(\partial g) \stackrel{\text{def}}{=} \left\{x \in \mathcal{H} :\ \partial g(x) \neq \emptyset\right\}$. When g belongs to $\Gamma_0(\mathcal{H})$, it is well-known that the subdifferential is a maximal monotone operator. If, moreover, the function is Gâteaux differentiable at $x \in \mathcal{H}$, then $\partial g(x) = \{\nabla g(x)\}$. For $x \in \operatorname{dom}(\partial g)$, the *minimal norm selection* of $\partial g(x)$ is defined to be the unique element $\left\{[\partial g(x)]^0\right\} \stackrel{\text{def}}{=} \operatorname*{Argmin}_{y\in\partial g(x)} \|y\|$. Then we have the following fundamental result about Moreau envelopes.

Proposition 2.1. *Given a function $g \in \Gamma_0(\mathcal{H})$, we have the following:*

(i) The Moreau envelope is convex, real-valued, and continuous.

(ii) Lax-Hopf formula: the Moreau envelope is the viscosity solution to the following Hamilton Jacobi equation:

$$\begin{cases} \frac{\partial}{\partial \beta} g^\beta(x) = -\frac{1}{2} \left\| \nabla_x g^\beta(x) \right\|^2 & (x, \beta) \in \mathcal{H} \times (0, +\infty) \\ g^0(x) = g(x) & x \in \mathcal{H}. \end{cases} \tag{2.5}$$

(iii) The gradient of the Moreau envelope is $\frac{1}{\beta}$-Lipschitz continuous and is given by the expression

$$\nabla_x g^\beta(x) = \frac{x - \operatorname{prox}_{\beta g}(x)}{\beta}.$$

(iv) $\forall x \in \operatorname{dom}(\partial g)$, $\left\| \nabla g^\beta(x) \right\| \nearrow \left\| [\partial g(x)]^0 \right\|$ *as* $\beta \searrow 0$.

(v) $\forall x \in \mathcal{H}$, $g^\beta(x) \nearrow g(x)$ *as* $\beta \searrow 0$. *In addition, given two positive real numbers* $\beta' < \beta$, *for all* $x \in \mathcal{H}$ *we have*

$$0 \leq g^{\beta'}(x) - g^\beta(x) \leq \frac{\beta - \beta'}{2} \left\| \nabla_x g^{\beta'}(x) \right\|^2;$$
$$0 \leq g(x) - g^\beta(x) \leq \frac{\beta}{2} \left\| [\partial g(x)]^0 \right\|^2.$$

Proof. (i): see [4, Proposition 12.15]. The proof for (ii) can be found in [2, Lemma 3.27 and Remark 3.32] (see also [19] or [1, Section 3.1]). The proof for claim (iii) can be found in [4, Proposition 12.29] and the proof for claim (iv) can be found in [4, Corollary 23.46]. For the first part in (v), see [4, Proposition 12.32(i)]. To show the first inequality in (v), combine (ii) and convexity of the function $\beta \mapsto g^\beta(x)$ for every $x \in \mathcal{H}$. The second inequality follows from the first one and (iv), taking the limit as $\beta' \to 0$. □

Remark 2.2.

(i) While the regularity claim in Proposition 2.1(iii) of the Moreau envelope $g^\beta(x)$ w.r.t. x is well-known, a less known result is the C^1-regularity w.r.t. β for any $x \in \mathcal{H}$ (Proposition 2.1(ii)). To our knowledge, the proof goes back, at least, to the book of [2]. Though it has been rediscovered in the recent literature in less general settings.

(ii) For given functions $H : \mathcal{H} \to \mathbb{R}$ and $g_0 : \mathcal{H} \to \mathbb{R}$, a natural generalization of the Hamilton-Jacobi equation in (2.5) is

$$\begin{cases} \frac{\partial}{\partial \beta} g(x, \beta) + H\left(\nabla_x g(x, \beta)\right) = 0 & (x, \beta) \in \mathcal{H} \times (0, +\infty) \\ g(x, 0) = g(x) & x \in \mathcal{H}. \end{cases}$$

Supposing that H is convex and that $\lim_{\|p\| \to +\infty} H(p) / \|p\| = +\infty$, the solution of the above system is given by the Lax-Hopf formula (see [14, Theorem 5, Section 3.3.2][1]):

$$g(x, t) \stackrel{\text{def}}{=} \inf_{y \in \mathcal{H}} \left\{ g_0(y) + t H^* \left(\frac{y - x}{t} \right) \right\}.$$

If $H(p) = \frac{1}{2} \|p\|^2$, then $H^*(p) = \frac{1}{2} \|p\|^2$ and we recover the result in Proposition 2.1.

[1] The proof in [14] is given in the finite-dimensional case but it extends readily to any real Hilbert space.

Regularity of differentiable functions In what follows, we introduce some definitions related with regularity of differentiable functions. They will provide useful upper-bounds and descent properties. Notice that the the notions and results of this part are independent from convexity.

Definition 2.3. (ω-smoothness) Consider a function $\omega : \mathbb{R}_+ \to \mathbb{R}_+$ such that $\omega(0) = 0$ and

$$\xi(s) \overset{\text{def}}{=} \int_0^1 \omega(st)\, dt \tag{2.6}$$

is non-decreasing. A differentiable function $g : \mathcal{H} \to \mathbb{R}$ is said to belong to $C^{1,\omega}(\mathcal{H})$ or to be ω-smooth if the following inequality is satisfied for every $x, y \in \mathcal{H}$:

$$\|\nabla g(x) - \nabla g(y)\| \leq \omega(\|x - y\|).$$

Lemma 2.4. *(ω-smooth Descent Lemma) Given a function $g \in C^{1,\omega}(\mathcal{H})$ we have the following inequality: for every x and y in $\mathcal{H}$,*

$$g(y) - g(x) \leq \langle \nabla g(x), y - x \rangle + \|y - x\| \xi(\|y - x\|),$$

where ξ is defined in (2.6).

Proof. We recall here the simple proof for completeness:

$$\begin{aligned} g(y) - g(x) &= \int_0^1 \frac{d}{dt} g(x + t(y - x))\, dt \\ &= \int_0^1 \langle \nabla g(x), y - x \rangle\, dt + \int_0^1 \langle \nabla g(x + t(y - x)) - \nabla g(x), y - x \rangle\, dt \\ &\leq \langle \nabla g(x), y - x \rangle + \|y - x\| \int_0^1 \|\nabla g(x + t(y - x)) - \nabla g(x)\|\, dt \\ &\leq \langle \nabla g(x), y - x \rangle + \|y - x\| \int_0^1 \omega(t\|y - x\|)\, dt, \end{aligned}$$

where in the first inequality we used Cauchy-Schwartz and in the second Definition 2.3. We conclude using the definition of ξ. □

For $L > 0$ and $\omega(t) = Lt^\nu$, $\nu \in]0, 1]$, $C^{1,\omega}(\mathcal{H})$ is the space of differentiable functions with Hölder continuous gradients, in which case $\xi(s) = Ls^\nu/(1 + \nu)$ and the Descent Lemma reads

$$g(y) - g(x) \leq \langle \nabla g(x), y - x \rangle + \frac{L}{1 + \nu} \|y - x\|^{1+\nu}, \tag{2.7}$$

see e.g., [27, 28]. When $\nu = 1$, we have that $C^{1,\omega}(\mathcal{H})$ is the class of differentiable functions with L-Lipschitz continuous gradient, and one recovers the classical Descent Lemma.

Now, following [18], we introduce some notions that allow one to further generalize (2.7). Given a function $G : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$, differentiable on the open set $\mathcal{C}_0 \subset \operatorname{int}(\operatorname{dom}(G))$, define the *Bregman divergence* of G as the function $D_G : \operatorname{dom}(G) \times \mathcal{C}_0 \to \mathbb{R}$,

$$D_G(x, y) = G(x) - G(y) - \langle \nabla G(y), x - y \rangle. \tag{2.8}$$

Then we have the following result.

Lemma 2.5. *(Generalized Descent Lemma, [18, Lemma 1]) Let G and g be differentiable on $\mathcal{C}_0$, where $\mathcal{C}_0$ is an open subset of* $\operatorname{int}(\operatorname{dom}(G))$. *Assume that $G - g$ is convex on $\mathcal{C}_0$. Then, for every x and y in $\mathcal{C}_0$,*

$$g(y) \leq g(x) + \langle \nabla g(x), y - x \rangle + D_G(y, x).$$

Proof. For our purpose, we intentionally weakened the hypothesis needed in the original result of [18, Lemma 1]. We repeat their argument but show the result is still valid under our weaker assumption. Let x and y be in $\mathcal{C}_0$, where, by hypothesis, $\mathcal{C}_0$ is open and contained in $\operatorname{int}(\operatorname{dom}(G))$. As $G - g$ is convex and differentiable on $\mathcal{C}_0$, from the gradient inequality (2.4) we have, for all $y \in \mathcal{C}_0$,

$$(G - g)(y) \geq (G - g)(x) + \langle \nabla (G - g)(x), y - x \rangle.$$

Rearranging the terms and using the definition of D_G in (2.8), we obtain the claim. □

The previous lemma suggests the introduction of the following definition, which extends Definition 2.3.

Definition 2.6. ((G, ζ)-smoothness) Let $G : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ and $\zeta :]0, 1] \to \mathbb{R}_+$. The pair $(g, \mathcal{C})$, where $g : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ and $\mathcal{C} \subset \operatorname{dom}(g)$, is said to be (G, ζ)-smooth if there exists an open set $\mathcal{C}_0$ such that $\mathcal{C} \subset \mathcal{C}_0 \subset \operatorname{int}(\operatorname{dom}(G))$ and

(i) G and g are differentiable on $\mathcal{C}_0$;
(ii) $G - g$ is convex on $\mathcal{C}_0$;
(iii) it holds

$$K_{(G,\zeta,\mathcal{C})} \stackrel{\text{def}}{=} \sup_{\substack{x,s \in \mathcal{C};\ \gamma \in]0,1] \\ z = x + \gamma(s-x)}} \frac{D_G(z, x)}{\zeta(\gamma)} \quad < \quad +\infty. \tag{2.9}$$

$K_{(G,\zeta,\mathcal{C})}$ is a far-reaching generalization of the standard curvature constant widely used in the literature of conditional gradient.

Remark 2.7. Assume that $(g, \mathcal{C})$ is (G, ζ)-smooth. Using first Lemma 2.5 and then the definition in (2.9), we have the following descent property: for every $x, s \in \mathcal{C}$ and for every $\gamma \in]0, 1]$,

$$\begin{aligned} g(x + \gamma(s - x)) &\leq g(x) + \gamma \langle \nabla g(x), s - x \rangle + D_G(x + \gamma(s - x), x) \\ &\leq g(x) + \gamma \langle \nabla g(x), s - x \rangle + K_{(G,\zeta,\mathcal{C})} \zeta(\gamma). \end{aligned}$$

Notice that, as in the previous definition, we do not require $\mathcal{C}$ to be convex. So, in general, the point $z = x + \gamma(s - x)$ may not lie in $\mathcal{C}$.

Lemma 2.8. *Suppose that the set $\mathcal{C}$ is bounded and denote by $d_{\mathcal{C}} \stackrel{\text{def}}{=} \sup_{x,y \in \mathcal{C}} \|x - y\|$ its* diameter. *Moreover, assume that the function g is ω-smooth on some open and convex subset $\mathcal{C}_0$ containing $\mathcal{C}$. Set $\zeta(\gamma) \stackrel{\text{def}}{=} \xi(d_{\mathcal{C}}\gamma)$, where ξ is given in (2.6). Then the pair $(g, \mathcal{C})$ is (g, ζ)-smooth with $K_{(g,\zeta,\mathcal{C})} \leq d_{\mathcal{C}}$.*

Proof. With $G = g$ and g being ω-smooth on $\mathcal{C}_0$, both G and g are differentiable on $\mathcal{C}_0$ and $G - g \equiv 0$ is convex on $\mathcal{C}_0$. Thus, all conditions required in Definition 2.6 hold true. It then remains to show (2.9) with the bound $K_{(g,\zeta,\mathcal{C})} \leq d_{\mathcal{C}}$. First notice that, for every $x, s \in \mathcal{C}$ and for every $\gamma \in]0, 1]$, the point $z = x + \gamma(s - x)$ belongs to $\mathcal{C}_0$. Indeed, $\mathcal{C} \subset \mathcal{C}_0$ and $\mathcal{C}_0$ is convex by hypothesis. In particular, as g is ω-smooth on $\mathcal{C}_0$, the

Descent Lemma 2.4 holds between the points x and z. Then

$$\begin{aligned}
K_{(g,\zeta,\mathcal{C})} &= \sup_{\substack{x,s\in\mathcal{C};\ \gamma\in]0,1]\\ z=x+\gamma(s-x)}} \frac{D_g(z,x)}{\zeta(\gamma)} \\
&= \sup_{\substack{x,s\in\mathcal{C};\ \gamma\in]0,1]\\ z=x+\gamma(s-x)}} \frac{g(z)-g(x)-\langle\nabla g(x), z-x\rangle}{\xi(d_{\mathcal{C}}\gamma)} \\
&\leq \sup_{\substack{x,s\in\mathcal{C};\ \gamma\in]0,1]\\ z=x+\gamma(s-x)}} \frac{\|z-x\|\,\xi(\|z-x\|)}{\xi(d_{\mathcal{C}}\gamma)} \\
&= \sup_{x,s\in\mathcal{C};\ \gamma\in]0,1]} \frac{\gamma\|s-x\|\,\xi(\gamma\|s-x\|)}{\xi(d_{\mathcal{C}}\gamma)} \\
&\leq \sup_{\gamma\in]0,1]} \frac{\gamma d_{\mathcal{C}}\xi(d_{\mathcal{C}}\gamma)}{\xi(d_{\mathcal{C}}\gamma)} = d_{\mathcal{C}}.
\end{aligned}$$

In the first inequality we used Lemma 2.4, while in the second we used that $\|s-x\| \leq d_{\mathcal{C}}$ (both x and s belong to $\mathcal{C}$, that is bounded by hypothesis) and the monotonicity of the function ξ (see Definition 2.3). □

Indicator and support functions Given a subset $\mathcal{C} \subset \mathcal{H}$, we define its *indicator function* as $\iota_{\mathcal{C}}(x) = 0$ if x belongs to $\mathcal{C}$ and $\iota_{\mathcal{C}}(x) = +\infty$ otherwise. Recall that, if $\mathcal{C}$ is nonempty, closed, and convex, then $\iota_{\mathcal{C}}$ belongs to $\Gamma_0(\mathcal{H})$. Remember also the definition of the *support function* of $\mathcal{C}$, $\sigma_{\mathcal{C}} \stackrel{\text{def}}{=} \iota^*_{\mathcal{C}}$. Equivalently, $\sigma_{\mathcal{C}}(x) \stackrel{\text{def}}{=} \sup\{\langle z, x\rangle : z \in \mathcal{C}\}$. We denote by $\mathrm{ri}(\mathcal{C})$ the *relative interior* of the set $\mathcal{C}$ (in finite dimension, it is the interior for the topology relative to its affine full). We denote $\mathrm{par}(C)$ as the subspace parallel to $\mathcal{C}$ which, in finite dimension, takes the form $\mathbb{R}(C-C)$.

We have the following characterization of the support function from the relative interior in finite dimension.

Proposition 2.9. *([35, Lemma 1]) Let $\mathcal{H}$ be finite-dimensional and $\mathcal{C} \subset \mathcal{H}$ a nonempty, closed bounded and convex subset. If $0 \in \mathrm{ri}(\mathcal{C})$, then $\sigma_{\mathcal{C}} \in \Gamma_0(\mathbb{R}^n)$ is sublinear, non-negative and finite-valued, and*

$$\sigma_{\mathcal{C}}(x) = 0 \iff x \in (\mathrm{par}(\mathcal{C}))^{\perp}.$$

Coercivity We recall that a function g is *coercive* if $\lim_{\|x\|\to+\infty} g(x) = +\infty$ and that coercivity is equivalent to the boundedness of the sublevel-sets [4, Proposition 11.11]. We have the following result, that relates coercivity to properties of the Fenchel conjugate.

Proposition 2.10. *([4, Theorem 14.17]) Given g in $\Gamma_0(\mathcal{H})$, g^* is coercive if and only if $0 \in \mathrm{int}(\mathrm{dom}(g))$.*

The *recession function* (sometimes referred to as the horizon function) of g at a given point $d \in \mathbb{R}^n$ is defined to be $g^{d,\infty} : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ such that, for every $x \in \mathbb{R}^n$,

$$g^{d,\infty}(x) \stackrel{\text{def}}{=} \lim_{\alpha\to\infty} \frac{g(d+\alpha x) - g(d)}{\alpha}.$$

Recall that, if g is convex, the recession function is independent from the selection of the point $d \in \mathbb{R}^n$ and can be then simply denoted as g^{∞}. In finite dimension, the following result relates coercivity to properties of the recession function.

Proposition 2.11. *Let $g \in \Gamma_0(\mathbb{R}^n)$ and $A : \mathbb{R}^m \to \mathbb{R}^n$ be a linear operator. Then,*

(i) *g coercive* $\iff$ $g^\infty(x) > 0 \quad \forall x \neq 0$.

(ii) $g^\infty \equiv \sigma_{\mathrm{dom}(g^*)}$.

(iii) $(g \circ A)^\infty \equiv g^\infty \circ A$.

In particular, we deduce that $g \circ A$ is coercive if and only if $\sigma_{\mathrm{dom}(g^*)}(Ax) > 0$ for every $x \neq 0$.

Proof. The proofs can be found in [34, Theorem 3.26], [34, Theorem 11.5] and [23, Corollary 3.2] respectively. □

Real sequences We close this section with some definitions and lemmas for real sequences that will be used to prove the convergence properties of the algorithm. We denote ℓ_+ as the set of all sequences in $[0, +\infty[$. Given $p \in [1, +\infty[$, ℓ^p is the space of real sequences $(r_k)_{k\in\mathbb{N}}$ such that $\left(\sum_{k=1}^{\infty} |r_k|^p\right)^{1/p} < +\infty$. For $p = +\infty$, we denote by ℓ^∞ the space of bounded sequences. Furthermore, we will use the notation $\ell^p_+ \stackrel{\text{def}}{=} \ell^p \cap \ell_+$. In the next, we recall some key results about real sequences.

Lemma 2.12. *([11, Lemma 3.1]) Consider three sequences $(r_k)_{k\in\mathbb{N}} \in \ell_+$, $(a_k)_{k\in\mathbb{N}} \in \ell_+$, and $(z_k)_{k\in\mathbb{N}} \in \ell^1_+$, such that*

$$r_{k+1} \leq r_k - a_k + z_k, \quad \forall k \in \mathbb{N}.$$

Then $(r_k)_{k\in\mathbb{N}}$ is convergent and $(a_k)_{k\in\mathbb{N}} \in \ell^1_+$.

Lemma 2.13. *([36, Theorem 2] and [36, Proposition 2(ii)]) Consider two sequences $(p_k)_{k\in\mathbb{N}} \in \ell_+$ and $(w_k)_{k\in\mathbb{N}} \in \ell_+$ such that $(p_k w_k)_{k\in\mathbb{N}} \in \ell^1_+$ and $(p_k)_{k\in\mathbb{N}} \notin \ell^1$. Then the following holds:*

(i) there exists a subsequence $\left(w_{k_j}\right)_{j\in\mathbb{N}}$ such that

$$w_{k_j} \leq P_{k_j}^{-1},$$

where $P_n = \sum_{k=1}^n p_k$. In particular, $\liminf_k w_k = 0$.

(ii) If moreover there exists a constant $\alpha > 0$ such that $w_k - w_{k+1} \leq \alpha p_k$ for every $k \in \mathbb{N}$, then

$$\lim_k w_k = 0.$$

Lemma 2.14. *Consider the sequences $(r_k)_{k\in\mathbb{N}} \in \ell_+$, $(p_k)_{k\in\mathbb{N}} \in \ell_+$, $(w_k)_{k\in\mathbb{N}} \in \ell_+$, and $(z_k)_{k\in\mathbb{N}} \in \ell_+$. Suppose that $(z_k)_{k\in\mathbb{N}} \in \ell^1_+$, $(p_k)_{k\in\mathbb{N}} \notin \ell^1$, and that, for some $\alpha > 0$, the following inequalities are satisfied for every $k \in \mathbb{N}$:*

$$\begin{aligned} r_{k+1} &\leq r_k - p_k w_k + z_k; \\ w_k - w_{k+1} &\leq \alpha p_k. \end{aligned} \tag{2.10}$$

Then,

(i) *$(r_k)_{k\in\mathbb{N}}$ is convergent and $(p_k w_k)_{k\in\mathbb{N}} \in \ell^1_+$.*

(ii) $\lim_k w_k = 0$.

(iii) *For every $k \in \mathbb{N}$, $\inf_{1\leq i\leq k} w_i \leq (r_0 + E)/P_k$, where, again, $P_n = \sum_{k=1}^n p_k$ and $E = \sum_{k=1}^{+\infty} z_k$.*

(iv) *There exists a subsequence $\left(w_{k_j}\right)_{j\in\mathbb{N}}$ such that, for all $j \in \mathbb{N}$, $w_{k_j} \leq P_{k_j}^{-1}$.*

Proof. (i) See Lemma 2.12.

(ii) Claim (ii) follows by combining (i) and Lemma 2.13(ii).
(iii) Sum (2.10) using a telescoping property and summability of $(z_k)_{k\in\mathbb{N}}$.
(iv) Claim (iv) follows by combining (i) and Lemma 2.13(i).

□

Notice that the conclusions of Lemma 2.14 remain true if non-negativity of the sequence $(r_k)_{k\in\mathbb{N}}$ is replaced with the assumption that it is bounded from below by a trivial translation argument. Observe also that Lemma 2.14 guarantees the convergence of the whole sequence to zero, but it gives a convergence rate only on a subsequence.

3 Algorithm and assumptions

3.1 Algorithm

As described in the introduction, we combine penalization with the augmented Lagrangian approach to form the following functional

$$\mathcal{J}_k(x,y,\mu) = f(x) + g(y) + h(x) + \langle \mu, Ax-b\rangle + \frac{\rho_k}{2}\|Ax-b\|^2 + \frac{1}{2\beta_k}\|y-Tx\|^2, \tag{3.1}$$

where μ is the dual multiplier, and ρ_k and β_k are non-negative parameters. The steps of our scheme, then, are summarized in Algorithm 1.

Algorithm 1: Conditional Gradient with Augmented Lagrangian and Proximal-step (CGALP)

Input: $x_0 \in \mathcal{C} = \operatorname{dom}(h)$; $\mu_0 \in \operatorname{ran}(A)$; $(\gamma_k)_{k\in\mathbb{N}}$, $(\beta_k)_{k\in\mathbb{N}}$, $(\theta_k)_{k\in\mathbb{N}}$, $(\rho_k)_{k\in\mathbb{N}} \in \ell_+$.
$k = 0$
repeat
$y_k = \operatorname{prox}_{\beta_k g}(Tx_k)$
$z_k = \nabla f(x_k) + T^*(Tx_k - y_k)/\beta_k + A^*\mu_k + \rho_k A^*(Ax_k - b)$
$s_k \in \operatorname{Argmin}_{s\in\mathcal{H}_p}\{h(s) + \langle z_k, s\rangle\}$
$x_{k+1} = x_k - \gamma_k(x_k - s_k)$
$\mu_{k+1} = \mu_k + \theta_k(Ax_{k+1} - b)$
$k \leftarrow k+1$
until *convergence*;
Output: x_{k+1}.

For the interpretation of the algorithm, notice that the first step is equivalent to

$$\{y_k\} = \operatorname*{Argmin}_{y\in\mathcal{H}_v} \mathcal{J}_k(x_k, y, \mu_k).$$

Now define the functional $\mathcal{E}_k(x,\mu) \overset{\text{def}}{=} f(x) + g^{\beta_k}(Tx) + \langle \mu, Ax-b\rangle + \frac{\rho_k}{2}\|Ax-b\|^2$. By convexity of the set $\mathcal{C}$ and the definition of x_{k+1} as a convex combination of x_k and s_k, the sequence $(x_k)_{k\in\mathbb{N}}$ remains in $\mathcal{C}$ for all k, although the affine constraint $Ax_k = b$ might only be satisfied asymptotically. It is an augmented

Lagrangian, where we do not consider the non-differentiable function h and we replace g by its Moreau envelope. Notice that

$$\begin{aligned}\nabla_x \mathcal{E}_k(x,\mu_k) &= \nabla f(x) + T^*[\nabla g^{\beta_k}](Tx) + A^*\mu_k + \rho_k A^*(Ax-b)\\ &= \nabla f(x) + \frac{1}{\beta_k}T^*\left(Tx - \operatorname{prox}_{\beta_k g}(Tx)\right) + A^*\mu_k + \rho_k A^*(Ax-b).\end{aligned} \tag{3.2}$$

where in the second equality we used 2.1(iii). Then z_k is just $\nabla_x \mathcal{E}_k(x_k,\mu_k)$ and the first three steps of the algorithm can be condensed in

$$s_k \in \operatorname*{Argmin}_{s\in\mathcal{H}_p}\left\{h(s) + \langle \nabla_x \mathcal{E}_k(x_k,\mu_k), s\rangle\right\}. \tag{3.3}$$

Thus the primal variable update of each step of our algorithm boils down to conditional gradient applied to the function $\mathcal{E}_k(\cdot,\mu_k)$, where the next iterate is a convex combination between the previous one and the new direction s_k. A standard update of the Lagrange multiplier μ_k follows.

3.2 Assumptions

3.2.1 Assumptions on the functions

In order to help the reading, we recall in a compact form the following notation that we will use to refer to various functionals throughout the paper:

$$\begin{aligned}\Phi(x) &\overset{\text{def}}{=} f(x) + g(Tx) + h(x);\\ \Phi_k(x) &\overset{\text{def}}{=} f(x) + g^{\beta_k}(Tx) + h(x) + \frac{\rho_k}{2}\|Ax-b\|^2;\\ \bar{\Phi}(x) &\overset{\text{def}}{=} \Phi(x) + (\overline{\rho}/2)\|Ax-b\|^2;\\ \bar{\varphi}(\mu) &\overset{\text{def}}{=} \bar{\Phi}^*(-A^*\mu) + \langle b,\ \mu\rangle;\\ \mathcal{L}(x,\mu) &\overset{\text{def}}{=} f(x) + g(Tx) + h(x) + \langle \mu, Ax-b\rangle;\\ \mathcal{L}_k(x,\mu) &\overset{\text{def}}{=} f(x) + g^{\beta_k}(Tx) + h(x) + \langle \mu, Ax-b\rangle + \frac{\rho_k}{2}\|Ax-b\|^2;\\ \mathcal{E}_k(x,\mu) &\overset{\text{def}}{=} f(x) + g^{\beta_k}(Tx) + \langle \mu, Ax-b\rangle + \frac{\rho_k}{2}\|Ax-b\|^2,\end{aligned} \tag{3.4}$$

where $\overline{\rho}$ is defined in Assumption (P.4) to be $\overline{\rho} = \sup_k \rho_k$.

In the list (3.4), we can recognize Φ as the objective, Φ_k as the smoothed objective augmented with a quadratic penalization of the constraint, and $\mathcal{L}_k$ as a smoothed augmented Lagrangian. $\mathcal{L}$ denotes the classical Lagrangian. Recall that $(x^\star,\mu^\star) \in \mathcal{H}_p \times \mathcal{H}_d$ is a saddle-point for the Lagrangian $\mathcal{L}$ if for every $(x,\mu) \in \mathcal{H}_p \times \mathcal{H}_d$,

$$\mathcal{L}(x^\star,\mu) \le \mathcal{L}(x^\star,\mu^\star) \le \mathcal{L}(x,\mu^\star). \tag{3.5}$$

It is well-known from standard Lagrange duality, see e.g. [4, Proposition 19.19] or [30, Theorem 3.68], that the existence of a saddle point $(x^\star,\mu^\star)$ ensures strong duality, that $x^\star$ solves ($\mathscr{P}$) and $\mu^\star$ solves the dual problem,

$$\min_{\mu\in\mathcal{H}_d} (f + g\circ T + h)^*(-A^*\mu) + \langle \mu, b\rangle. \tag{$\mathscr{D}$}$$

The following assumptions on the problem will be used throughout the convergence analysis (for some results only a subset of these assumptions will be needed):

(A.1) $f, g \circ T$, and h belong to $\Gamma_0(\mathcal{H}_p)$.
(A.2) The pair $(f, \mathcal{C})$ is (F, ζ)-smooth (see Definition 2.6), where we recall $\mathcal{C} \overset{\text{def}}{=} \operatorname{dom}(h)$.
(A.3) $\mathcal{C}$ is weakly compact (and thus contained in a ball of radius $R > 0$).
(A.4) $T\mathcal{C} \subset \operatorname{dom}(\partial g)$ and $\sup_{x \in \mathcal{C}} \left\| [\partial g(Tx)]^0 \right\| < \infty$.
(A.5) h is Lipschitz continuous relative to its domain $\mathcal{C}$ with constant $L_h \geq 0$, i.e., $\forall (x, z) \in \mathcal{C}^2$, $|h(x) - h(z)| \leq L_h \|x - z\|$.
(A.6) There exists a saddle-point $(x^\star, \mu^\star) \in \mathcal{H}_p \times \mathcal{H}_d$ for the Lagrangian $\mathcal{L}$.
(A.7) $\operatorname{ran}(A)$ is closed.
(A.8) One of the following holds:
  (a) $A^{-1}(b) \cap \operatorname{int}(\operatorname{dom}(g \circ T)) \cap \operatorname{int}(\mathcal{C}) \neq \emptyset$, where $A^{-1}(b)$ is the pre-image of b under A.
  (b) $\mathcal{H}_p$ and $\mathcal{H}_d$ are finite dimensional and

$$\begin{cases} A^{-1}(b) \cap \operatorname{ri}(\operatorname{dom}(g \circ T)) \cap \operatorname{ri}(\mathcal{C}) \neq \emptyset \\ \quad \text{and} \\ \operatorname{ran}(A^*) \cap \operatorname{par}(\operatorname{dom}(g \circ T) \cap \mathcal{C})^\perp = \{0\}. \end{cases} \tag{3.6}$$

At this stage, a few remarks are in order.

Remark 3.1.
(i) By Assumption (A.1), $\mathcal{C}$ is also closed and convex. This together with Assumption (A.3) entail, upon using [4, Lemma 3.29 and Theorem 3.32], that $\mathcal{C}$ is weakly compact.
(ii) Since the sequence of iterates $(x_k)_{k \in \mathbb{N}}$ generated by Algorithm 1 is guaranteed to belong to $\mathcal{C}$ under (P.1), we have from (A.4)

$$\sup_k \left\| [\partial g(Tx_k)]^0 \right\| \leq M. \tag{3.7}$$

where M is a positive constant.
(iii) Assumption (A.5) will only be needed in the proof of convergence to optimality (Theorem 4.2). It is not needed to show asymptotic feasibility (Theorem 4.1).
(iv) Assume that $A^{-1}(b) \cap \operatorname{dom}(g \circ T) \cap \mathcal{C} \neq \emptyset$, which entails that the set of minimizers of ($\mathscr{P}$) is a non-empty convex closed bounded set under (A.1)-(A.3). Then there are various domain qualification conditions, e.g., one of the conditions in [4, Proposition 15.24 and Fact 15.25], that ensure the existence of a saddle-point for the Lagrangian $\mathcal{L}$ (see [4, Theorem 19.1 and Proposition 9.19(v)]).
(v) Observe that under the inclusion assumption of Lemma 3.2, (A.8)(a) is equivalent to $A^{-1}(b) \cap \operatorname{int}(\mathcal{C}) \neq \emptyset$.
(vi) Assumption (A.8) will be crucial to show that $\bar{\varphi}$ is coercive on $\ker(A^*)^\perp = \operatorname{ran}(A)$ (the last equality follows from (A.7)), and hence boundedness of the dual multiplier sequence $(\mu_k)_{k \in \mathbb{N}}$ provided by Algorithm 1 (see Lemma 4.10 and Lemma 4.11).

The uniform boundedness of the minimal norm selection on $\mathcal{C}$, as required in Assumption (A.4), is important when we will invoke Proposition 2.1(v) in our proofs to get meaningful estimates. The following result gives some sufficient conditions under which (A.4) holds (in fact an even stronger claim).

Lemma 3.2. *Let $\mathcal{C}$ be a nonempty bounded subset of $\mathcal{H}_p$, $g \in \Gamma_0(\mathcal{H}_v)$ and $T : \mathcal{H}_p \to \mathcal{H}_v$ be a bounded linear operator. Suppose that $T\mathcal{C} \subset \operatorname{int}(\operatorname{dom}(g))$. Then the assumption (A.4) holds.*

Proof. Since $g \in \Gamma_0(\mathcal{H}_p)$, it follows from [4, Proposition 16.21] that

$$T\mathcal{C} \subset \operatorname{int}(\operatorname{dom}(g)) \subset \operatorname{dom}(\partial g).$$

Moreover, by [4, Corollary 8.30(ii) and Proposition 16.14], we have that ∂g is locally weakly compact on $\operatorname{int}(\operatorname{dom}(g))$. In particular, as we assume that $\mathcal{C}$ is bounded, so is $T\mathcal{C}$, and since $T\mathcal{C} \subset \operatorname{int}(\operatorname{dom}(g))$, it means that for each $z \in T\mathcal{C}$ there exists an open neighborhood of z, denoted by U_z, such that $\partial g(U_z)$ is bounded. Since $(U_z)_{z\in\mathcal{C}}$ is an open cover of $T\mathcal{C}$ and $T\mathcal{C}$ is bounded, there exists a finite subcover $(U_{z_k})_{k=1}^n$. Then,

$$\bigcup_{x\in\mathcal{C}} \partial g(Tx) \subset \bigcup_{k=1}^{n} \partial g(U_{z_k}).$$

Since the right-hand-side is bounded (as it is a finite union of bounded sets),

$$\sup_{x\in\mathcal{C},\ u\in\partial g(Tx)} \|u\| < +\infty,$$

whence the desired conclusion trivially follows. □

3.2.2 Assumptions on the parameters

We also use the following assumptions on the parameters of Algorithm 1 (recall the function ζ in Definition 2.6):

(P.1) $(\gamma_k)_{k\in\mathbb{N}} \subset]0,1]$ and the sequences $(\zeta(\gamma_k))_{k\in\mathbb{N}}$, $(\gamma_k^2/\beta_k)_{k\in\mathbb{N}}$ and $(\gamma_k\beta_k)_{k\in\mathbb{N}}$ belong to ℓ_+^1.

(P.2) $(\gamma_k)_{k\in\mathbb{N}} \notin \ell^1$.

(P.3) $(\beta_k)_{k\in\mathbb{N}} \in \ell_+$ is non-increasing and converges to 0.

(P.4) $(\rho_k)_{k\in\mathbb{N}} \in \ell_+$ is non-decreasing with $0 < \underline{\rho} = \inf_k \rho_k \leq \sup_k \rho_k = \overline{\rho} < +\infty$.

(P.5) For some positive constants $\underline{M}$ and $\overline{M}$, $\underline{M} \leq \inf_k(\gamma_k/\gamma_{k+1}) \leq \sup_k(\gamma_k/\gamma_{k+1}) \leq \overline{M}$.

(P.6) $(\theta_k)_{k\in\mathbb{N}}$ satisfies $\theta_k = \frac{\gamma_k}{c}$ for all $k \in \mathbb{N}$ for some $c > 0$ such that $\frac{\overline{M}}{c} - \frac{\rho}{2} < 0$.

(P.7) $(\gamma_k)_{k\in\mathbb{N}}$ and $(\rho_k)_{k\in\mathbb{N}}$ satisfy $\rho_{k+1} - \rho_k - \gamma_{k+1}\rho_{k+1} + \frac{2}{c}\gamma_k - \frac{\gamma_k^2}{c} \leq \gamma_{k+1}$ for all $k \in \mathbb{N}$ and for c in (P.6).

Remark 3.3.

(i) One can recognize that the update of the dual multiplier μ_k in Algorithm 1 has a flavour of gradient ascent applied to the augmented dual with step-size θ_k. However, unlike the standard method of multipliers with the augmented Lagrangian, Assumption (P.6) requires θ_k to vanish in our setting. The underlying reason is that our update can be seen as an inexact dual ascent (i.e., exactness stems from the conditional gradient-based update on x_k which is not a minimization of over x of the augmented Lagrangian $\mathcal{L}_k$). Thus θ_k must annihilate this error asymptotically.

(ii) A sufficient condition for (P.7) to hold consists of taking $\rho_k \equiv \rho > 0$ and $\gamma_{k+1} \geq \frac{2}{c(1+\rho)}\gamma_k$. In particular, if $(\gamma_k)_{k\in\mathbb{N}}$ satisfies (P.5), then, for (P.7) to hold, it is sufficient to take $\rho_k \equiv \rho > 2\overline{M}/c$ as supposed in (P.6).

(iii) The relevance of having ρ_k vary is that it allows for more general and less stringent choice of the step-size γ_k. It is, however, possible (and easier in practice), to simply pick $\rho_k \equiv \rho$ for all $k \in \mathbb{N}$ as described above.

There is a large class of sequences that fulfill the requirements (P.1)-(P.7). A typical one is as follows.

Example 3.4. Take[2], for $k \in \mathbb{N}$,

$$\rho_k \equiv \rho > 0, \gamma_k = \frac{(\log(k+2))^a}{(k+1)^{1-b}}, \beta_k = \frac{1}{(k+1)^{1-\delta}}, \quad \text{with}$$
$$a \geq 0, 0 \leq 2b < \delta < 1, \delta < 1-b, \rho > 2^{2-b}/c, c > 0.$$

In this case, one can take the crude bounds $\underline{M} = (\log(2)/\log(3))^a$ and $\overline{M} = 2^{1-b}$, and choose $\rho > 2\overline{M}/c$ as devised in Remark 3.3(ii). In turn, (P.4)-(P.7) hold. In addition, suppose that f has a ν-Hölder continuous gradient (see (2.7)). Thus for (P.1)-(P.2) to hold, simple algebra shows that the allowable choice of b is in $\left[0, \min\left(1/3, \frac{\nu}{1+\nu}\right)\right[$.

4 Convergence analysis

4.1 Main results

We state here our main results.

Theorem 4.1 (Asymptotic feasibility). *Suppose that Assumptions (A.1)-(A.4) and (A.6) hold. Consider the sequence of iterates $(x_k)_{k\in\mathbb{N}}$ from Algorithm 1 with parameters satisfying Assumptions (P.1)-(P.6). Then,*

(i) Ax_k converges strongly to b as $k \to \infty$, i.e., the sequence $(x_k)_{k\in\mathbb{N}}$ is asymptotically feasible for ($\mathscr{P}$) in the strong topology.

(ii) Pointwise rate:

$$\inf_{0\leq i\leq k} \|Ax_i - b\| = O\left(\frac{1}{\sqrt{\Gamma_k}}\right) \text{ and } \exists \text{ a subsequence } \left(x_{k_j}\right)_{j\in\mathbb{N}} \text{ s.t. for all } j \in \mathbb{N}, \|Ax_{k_j} - b\| \leq \frac{1}{\sqrt{\Gamma_{k_j}}}, \tag{4.1}$$

where, for all $k \in \mathbb{N}$, $\Gamma_k \overset{\text{def}}{=} \sum_{i=0}^k \gamma_i$.

(iii) Ergodic rate: for each $k \in \mathbb{N}$, let $\bar{x}_k \overset{\text{def}}{=} \sum_{i=0}^k \gamma_i x_i/\Gamma_k$. Then

$$\|A\bar{x}_k - b\| = O\left(\frac{1}{\sqrt{\Gamma_k}}\right). \tag{4.2}$$

Theorem 4.1 will be proved in Section 4.3.

Theorem 4.2 (Convergence to optimality). *Suppose that assumptions (A.1)-(A.8) and (P.1)-(P.7) hold, with $\underline{M} \geq 1$. Let $(x_k)_{k\in\mathbb{N}}$ be the sequence of primal iterates generated by* Algorithm 1 *and $(x^\star, \mu^\star)$ a saddle-point pair for the Lagrangian. Then, in addition to the results of Theorem 4.1, the following holds*

(i) Convergence of the Lagrangian:

$$\lim_{k\to\infty} \mathcal{L}(x_k, \mu^\star) = \mathcal{L}(x^\star, \mu^\star). \tag{4.3}$$

(ii) Every weak cluster point $\bar{x}$ of $(x_k)_{k\in\mathbb{N}}$ is a solution of the primal problem ($\mathscr{P}$), and $(\mu_k)_{k\in\mathbb{N}}$ converges weakly to $\bar{\mu}$ a solution of the dual problem ($\mathscr{D}$), i.e., $(\bar{x}, \bar{\mu})$ is a saddle point of $\mathcal{L}$.

[2]Of course, one can add a scaling factor in the choice of the parameters which would allow for more practical flexibility. But this does not change anything to our discussion nor to the bahviour of the CGALP algorithm for k large enough.

(iii) Pointwise rate:

$$\begin{aligned} &\inf_{0\le i\le k} \mathcal{L}\left(x_i,\mu^\star\right)-\mathcal{L}\left(x^\star,\mu^\star\right)=O\left(\frac{1}{\Gamma_k}\right) \quad \textit{and} \\ &\exists \textit{ a subsequence } \left(x_{k_j}\right)_{j\in\mathbb{N}} \textit{ s.t. for each } j\in\mathbb{N}, \mathcal{L}\left(x_{k_j+1},\mu^\star\right)-\mathcal{L}\left(x^\star,\mu^\star\right)\le\frac{1}{\Gamma_{k_j}}. \end{aligned} \tag{4.4}$$

(iv) Ergodic rate: for each $k\in\mathbb{N}$, let $\bar{x}_k \overset{\text{def}}{=} \sum_{i=0}^k \gamma_i x_{i+1}/\Gamma_k$. Then

$$\mathcal{L}\left(\bar{x}_k,\mu^\star\right)-\mathcal{L}\left(x^\star,\mu^\star\right)=O\left(\frac{1}{\Gamma_k}\right). \tag{4.5}$$

An important observation is that Theorem 4.2, which will be proved in Section 4.5, actually shows that

$$\lim_{k\to\infty}\left[\mathcal{L}\left(x_k,\mu^\star\right)-\mathcal{L}\left(x^\star,\mu^\star\right)+\frac{\rho_k}{2}\|Ax_k-b\|^2\right]=0,$$

and subsequentially, for each $j\in\mathbb{N}$,

$$\mathcal{L}\left(x_{k_j},\mu^\star\right)-\mathcal{L}\left(x^\star,\mu^\star\right)+\frac{\rho_{k_j}}{2}\|Ax_{k_j}-b\|^2\le\frac{1}{\Gamma_{k_j}}. \tag{4.6}$$

This means, in particular, that the pointwise rate for feasibility and optimality hold simulatenously for the same subsequence.

The following corollary is immediate.

Corollary 4.3. *Under the assumptions of* Theorem 4.2*, if the problem ($\mathscr{P}$) admits a unique solution $x^\star$, then the primal-dual pair sequence $(x_k,\mu_k)_{k\in\mathbb{N}}$ converges weakly to a saddle point $(x^\star,\mu^\star)$.*

Proof. By uniqueness, it follows from Theorem 4.2(ii) that $(x_k)_{k\in\mathbb{N}}$ has exactly one weak sequential cluster point which is the solution to ($\mathscr{P}$). Weak convergence of the sequence $(x_k)_{k\in\mathbb{N}}$ then follows from [4, Lemma 2.38]. □

Example 4.4. Suppose that the sequences of parameters are chosen according to Example 3.4. Let the function $\sigma: t\in\mathbb{R}^+\mapsto(\log(t+2))^a/(t+1)^{1-b}$. We obviously have $\sigma(k)=\gamma_k$ for $k\in\mathbb{N}$. Moreover, it is easy to see that $\exists k'\ge 0$ (depending on a and b), such that σ is decreasing for $t\ge k'$. Thus, $\forall k\ge k'$, we have

$$\Gamma_k\ge\sum_{i=k'}^{k}\gamma_i\ge\int_{k'}^{k+1}\sigma(t)dt\ge\int_{k'+1}^{k+2}(\log(t))^a t^{b-1}dt=\int_{\log(k'+1)}^{\log(k+2)}t^a e^{bt}dt.$$

It is easy to show, using integration by parts for the first case, that

$$\Gamma_k^{-1}=\begin{cases} o\left(\frac{1}{(k+2)^b}\right) & a=1,b>0,\\ O\left(\frac{1}{(k+2)^b}\right) & a=0,b>0,\\ O\left(\frac{1}{\log(k+2)}\right) & a=0,b=0. \end{cases}$$

This result reveals that picking a and b as large as possible results in a faster convergence rate, with the proviso that b satisfy some conditions for (P.1)-(P.7) to hold, see the discussion in Example 3.4 for the largest possible choice of b.

4.2 Preparatory results

The next result is a direct application of the Descent Lemma 2.7 and the generalized one in Lemma 2.5 to the specific case of Algorithm 1. It allows to obtain a descent property for the function $\mathcal{E}_k(\cdot, \mu_k)$ between the previous iterate x_k and next one x_{k+1}.

Lemma 4.5. *Suppose Assumptions (A.1), (A.2) and (P.1) hold. For each $k \in \mathbb{N}$, define the quantity*

$$L_k \stackrel{\text{def}}{=} \frac{\|T\|^2}{\beta_k} + \|A\|^2 \rho_k. \tag{4.7}$$

Then, for each $k \in \mathbb{N}$, we have the following inequality:

$$\begin{aligned}\mathcal{E}_k(x_{k+1}, \mu_k) \leq \mathcal{E}_k(x_k, \mu_k) + \langle \nabla_x \mathcal{E}_k(x_k, \mu_k),\ x_{k+1} - x_k \rangle + K_{(F,\zeta,\mathcal{C})}\zeta(\gamma_k) \\ + \frac{L_k}{2}\|x_{k+1} - x_k\|^2.\end{aligned}$$

Proof. Define for each $k \in \mathbb{N}$,

$$\tilde{\mathcal{E}}_k(x, \mu) \stackrel{\text{def}}{=} g^{\beta_k}(Tx) + \langle \mu, Ax - b \rangle + \frac{\rho_k}{2}\|Ax - b\|^2,$$

so that $\mathcal{E}_k(x, \mu) = f(x) + \tilde{\mathcal{E}}_k(x, \mu)$. Compute

$$\nabla_x \tilde{\mathcal{E}}_k(x, \mu) = T^* \nabla g^{\beta_k}(Tx) + A^* \mu + \rho_k A^*(Ax - b),$$

which is Lipschitz-continuous with constant $L_k = \frac{\|T\|^2}{\beta_k} + \|A\|^2 \rho_k$ by virtue of (A.1) and Proposition 2.1(iii). Then we can use the Descent Lemma (2.7) with $\nu = 1$ on $\tilde{\mathcal{E}}_k(\cdot, \mu_k)$ between the points x_k and x_{k+1}, to obtain, for each $k \in \mathbb{N}$,

$$\tilde{\mathcal{E}}_k(x_{k+1}, \mu_k) \leq \tilde{\mathcal{E}}_k(x_k, \mu_k) + \langle \nabla \tilde{\mathcal{E}}_k(x_k, \mu_k), x_{k+1} - x_k \rangle + \frac{L_k}{2}\|x_{k+1} - x_k\|^2. \tag{4.8}$$

From Assumption (A.2), Lemma 2.5 and Remark 2.7, we have, for each $k \in \mathbb{N}$,

$$\begin{aligned}f(x_{k+1}) &\leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + D_F(x_{k+1}, x_k) \\ &\leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + K_{(F,\zeta,\mathcal{C})}\zeta(\gamma_k),\end{aligned}$$

where we used that both x_k and s_k lie in $\mathcal{C}$, that γ_k belongs to $]0, 1]$ by (P.1) and thus $x_{k+1} = x_k + \gamma_k(s_k - x_k) \in \mathcal{C}$. Summing (4.8) with the latter and recalling that $\mathcal{E}_k(x, \mu_k) = f(x) + \tilde{\mathcal{E}}_k(x, \mu_k)$, we obtain the claim. □

Again for the function $\mathcal{E}_k(\cdot, \mu_k)$, we also have a lower-bound, presented in the next lemma.

Lemma 4.6. *Suppose Assumptions (A.1) and (A.2) hold. Then, for all $k \in \mathbb{N}$, for all $x, x' \in \mathcal{H}_p$ and for all $\mu \in \mathcal{H}_d$,*

$$\mathcal{E}_k(x, \mu) \geq \mathcal{E}_k(x', \mu) + \langle \nabla_x \mathcal{E}_k(x', \mu),\ x - x' \rangle + \frac{\rho_k}{2}\|A(x - x')\|^2.$$

Proof. First, split the function $\mathcal{E}_k(\cdot,\mu)$ as $\mathcal{E}_k(x,\mu) = \mathcal{E}_k^0(x,\mu) + \frac{\rho_k}{2}\|Ax-b\|^2$ for an opportune definition of $\mathcal{E}_k^0(\cdot,\mu)$. For the first term, simply by convexity, we have

$$\mathcal{E}_k^0(x,\mu) \geq \mathcal{E}_k^0(x',\mu) + \langle \nabla_x \mathcal{E}_k^0(x',\mu),\ x-x'\rangle. \tag{4.9}$$

Now use the strong convexity of the term $(\rho_k/2)\|\cdot - b\|^2$ between points Ax and Ax', to affirm that

$$\frac{\rho_k}{2}\|Ax-b\|^2 \geq \frac{\rho_k}{2}\|Ax'-b\|^2 + \langle \nabla\left(\frac{\rho_k}{2}\|\cdot - b\|^2\right)(Ax'), Ax-Ax'\rangle + \frac{\rho_k}{2}\|A(x-x')\|^2. \tag{4.10}$$

Compute

$$\begin{aligned}\langle \nabla\left(\frac{\rho_k}{2}\|\cdot - b\|^2\right)(Ax'), Ax-Ax'\rangle &= \rho_k\langle A^*(Ax'-b), x-x'\rangle \\ &= \langle \nabla\left(\frac{\rho_k}{2}\|A\cdot - b\|^2\right)(x'), x-x'\rangle.\end{aligned}$$

Summing (4.9) and (4.10) and invoking the gradient computation above, we obtain the claim. □

Lemma 4.7. *Suppose that assumptions (A.1)-(A.8) and (P.1)-(P.7) hold, with $\underline{M} \geq 1$. Let $(x_k)_{k\in\mathbb{N}}$ be the sequence of primal iterates generated by* Algorithm 1 *and $\mu^\star$ a solution of the dual problem ($\mathscr{D}$). Then we have the following estimate,*

$$\mathcal{L}(x_k,\mu^\star) - \mathcal{L}(x_{k+1},\mu^\star) \leq \gamma_k d_{\mathcal{C}}\left(M\|T\| + D + L_h + \|A\|\,\|\mu^\star\|\right)$$

Proof. First define $u_k \stackrel{\text{def}}{=} [\partial g(Tx_k)]^0$ and recall that, by (A.4) and its consequence in (3.7), $\|u_k\| \leq M$ for every $k\in\mathbb{N}$. Then,

$$\begin{aligned}\mathcal{L}(x_k,\mu^\star) - \mathcal{L}(x_{k+1},\mu^\star) &= \Phi(x_k) - \Phi(x_{k+1}) + \langle \mu^\star,\ A(x_k - x_{k+1})\rangle \\ &\leq \langle u_k,\ T(x_k - x_{k+1})\rangle + \langle \nabla f(x_k),\ x_k - x_{k+1}\rangle \\ &\quad + L_h\|x_k - x_{k+1}\| + \|\mu^\star\|\,\|A\|\,\|x_k - x_{k+1}\|,\end{aligned}$$

where we used the subdifferential inequality (2.4) on g, the gradient inequality on f, the L_h-Lipschitz continuity of h relative to $\mathcal{C}$ (see (A.5)), and the Cauchy-Schwartz inequality on the scalar product. Since $x_{k+1} = x_k + \gamma_k(x_k - s_k)$, we obtain

$$\begin{aligned}\mathcal{L}(x_k,\mu^\star) - \mathcal{L}(x_{k+1},\mu^\star) &\leq \gamma_k\Big(\langle u_k,\ T(x_k - s_k)\rangle + \langle \nabla f(x_k),\ x_k - s_k\rangle + L_h\|x_k - s_k\| \\ &\quad + \|\mu^\star\|\,\|A\|\,\|x_k - s_k\|\Big) \\ &\leq \gamma_k d_{\mathcal{C}}\left(M\|T\| + D + L_h + \|\mu^\star\|\,\|A\|\right),\end{aligned}$$

where we have denoted by D the constant $D \stackrel{\text{def}}{=} \sup_{x\in\mathcal{C}}\|\nabla f(x)\| < +\infty$ (see (4.32)). □

Lemma 4.8. *Suppose that assumptions (A.3) and (P.4) hold. Let $(x_k)_{k\in\mathbb{N}}$ be the sequence of primal iterates generated by* Algorithm 1. *Then we have the following estimate,*

$$\frac{\rho_k}{2}\|Ax_k - b\|^2 - \frac{\rho_{k+1}}{2}\|Ax_{k+1} - b\|^2 \leq \overline{\rho} d_{\mathcal{C}}\|A\|\left(\|A\|R + \|b\|\right)\gamma_k,$$

where R is the radius of the ball containing $\mathcal{C}$ and $\overline{\rho} = \sup_k \rho_k$.

Proof. By (P.4) and convexity of the function $\frac{\rho_{k+1}}{2}\|A\cdot -b\|^2$, we have

$$\begin{aligned}\frac{\rho_k}{2}\|Ax_k-b\|^2-\frac{\rho_{k+1}}{2}\|Ax_{k+1}-b\|^2 &\leq \frac{\rho_{k+1}}{2}\|Ax_k-b\|^2-\frac{\rho_{k+1}}{2}\|Ax_{k+1}-b\|^2\\ &\leq \langle\nabla\left(\frac{\rho_{k+1}}{2}\|A\cdot -b\|^2\right)(x_k),\ x_k-x_{k+1}\rangle.\end{aligned}$$

Now compute the gradient and use the definition of x_{k+1}, to obtain

$$\begin{aligned}\frac{\rho_k}{2}\|Ax_k-b\|^2-\frac{\rho_{k+1}}{2}\|Ax_{k+1}-b\|^2 &\leq \rho_{k+1}\gamma_k\langle Ax_k-b,\ A\left(x_k-s_k\right)\rangle\\ &\leq \overline{\rho}d_{\mathcal{C}}\|A\|\left(\|A\|R+\|b\|\right)\gamma_k.\end{aligned}$$

In the last inequality, we used Cauchy-Schwartz inequality, triangle inequality, the fact that $\|x_k-s_k\|\leq d_{\mathcal{C}}$, and assumptions (A.3) and (P.4) (respectively, $\sup_{x\in\mathcal{C}}\|x\|\leq R$ and $\rho_{k+1}\leq\overline{\rho}$). □

4.3 Asymptotic feasibility

We begin with an intermediary lemma establishing the main feasibility estimation and some summability results that will also be used in the proof of optimality.

Lemma 4.9. *Suppose that Assumptions (A.1)-(A.4) and (A.6) hold. Consider the sequence of iterates $(x_k)_{k\in\mathbb{N}}$ from Algorithm 1 with parameters satisfying Assumptions (P.1)-(P.6). Define the two quantities Δ_k^p and Δ_k^d in the following way,*

$$\Delta_k^p \overset{\text{def}}{=} \mathcal{L}_k\left(x_{k+1},\mu_k\right)-\tilde{\mathcal{L}}_k\left(\mu_k\right),\quad \Delta_k^d \overset{\text{def}}{=} \tilde{\mathcal{L}}-\tilde{\mathcal{L}}_k\left(\mu_k\right),$$

where we have denoted $\tilde{\mathcal{L}}_k\left(\mu_k\right)\overset{\text{def}}{=}\min_x\mathcal{L}_k\left(x,\mu_k\right)$ and $\tilde{\mathcal{L}}\overset{\text{def}}{=}\mathcal{L}\left(x^\star,\mu^\star\right)$. Denote the sum $\Delta_k\overset{\text{def}}{=}\Delta_k^p+\Delta_k^d$. Then we have the following estimation,

$$\begin{aligned}\Delta_{k+1}\leq\Delta_k-\gamma_{k+1}\left(\frac{M}{c}\left\|A\tilde{x}_{k+1}-b\right\|^2+\delta\left\|A\left(x_{k+1}-\tilde{x}_{k+1}\right)\right\|^2\right)+\frac{L_{k+1}}{2}\gamma_{k+1}^2d_{\mathcal{C}}^2\\ +K_{(F,\zeta,\mathcal{C})}\zeta\left(\gamma_{k+1}\right)+\frac{\beta_k-\beta_{k+1}}{2}M+\left(\frac{\rho_{k+1}-\rho_k}{2}\right)\left\|Ax_{k+1}-b\right\|^2,\end{aligned}$$

and, moreover,

$$\left(\gamma_k\left\|A\tilde{x}_k-b\right\|^2\right)_{k\in\mathbb{N}}\in\ell_+^1,\quad \left(\gamma_k\left\|A\left(x_k-\tilde{x}_k\right)\right\|^2\right)_{k\in\mathbb{N}}\in\ell_+^1,\ \text{ and }\ \left(\gamma_k\left\|Ax_k-b\right\|^2\right)_{k\in\mathbb{N}}\in\ell_+^1.$$

Proof. First notice that the quantity $\Delta_k^p\geq 0$ and can be seen as a primal gap at iteration k while Δ_k^d may be negative but is bounded from below by our assumptions. Indeed, in view of (A.1), (A.6) and Remark 3.1(iv), $\tilde{\mathcal{L}}_k\left(\mu_k\right)$ is bounded from above since

$$\begin{aligned}\tilde{\mathcal{L}}_k\left(\mu_k\right)&\leq\mathcal{L}_k\left(x^\star,\mu_k\right)\\ &=f\left(x^\star\right)+g^{\beta_k}\left(Tx^\star\right)+h\left(x^\star\right)+\langle\mu_k,Ax^\star-b\rangle+\frac{\rho_k}{2}\left\|Ax^\star-b\right\|^2\\ &=f\left(x^\star\right)+g^{\beta_k}\left(Tx^\star\right)+h\left(x^\star\right)\\ &\leq f\left(x^\star\right)+g\left(Tx^\star\right)+h\left(x^\star\right)<+\infty,\end{aligned}$$

where we used Proposition 2.1(v) in the last inequality.

We denote a minimizer of $\mathcal{L}_k(x,\mu_k)$ by $\tilde{x}_k \in \operatorname*{Argmin}_{x\in\mathcal{H}_p} \mathcal{L}_k(x,\mu_k)$, which exists and belongs to $\mathcal{C}$ by (A.1)-(A.3). Then, we have

$$\Delta_{k+1}^d - \Delta_k^d = \mathcal{L}_k(\tilde{x}_k,\mu_k) - \mathcal{L}_{k+1}(\tilde{x}_{k+1},\mu_{k+1}). \tag{4.11}$$

Since $\tilde{x}_k$ is a minimizer of $\mathcal{L}_k(x,\mu_k)$ we have that $\mathcal{L}_k(\tilde{x}_k,\mu_k) \le \mathcal{L}_k(\tilde{x}_{k+1},\mu_k)$ which leads to,

$$\begin{aligned}\mathcal{L}_k(\tilde{x}_{k+1},\mu_k) &= \mathcal{L}_{k+1}(\tilde{x}_{k+1},\mu_k) + g^{\beta_k}(T\tilde{x}_{k+1}) - g^{\beta_{k+1}}(T\tilde{x}_{k+1}) + \tfrac{\rho_k-\rho_{k+1}}{2}\|A\tilde{x}_{k+1}-b\|^2 \\ &\le \mathcal{L}_{k+1}(\tilde{x}_{k+1},\mu_k),\end{aligned}$$

where the last inequality comes from Proposition 2.1(v) and the assumptions (P.3) and (P.4). Combining this with (4.11),

$$\begin{aligned}\Delta_{k+1}^d - \Delta_k^d &\le \mathcal{L}_{k+1}(\tilde{x}_{k+1},\mu_k) - \mathcal{L}_{k+1}(\tilde{x}_{k+1},\mu_{k+1}) \\ &= \langle \mu_k - \mu_{k+1}, A\tilde{x}_{k+1} - b\rangle \\ &= -\theta_k \langle Ax_{k+1} - b, A\tilde{x}_{k+1} - b\rangle,\end{aligned} \tag{4.12}$$

where in the last equality we used the definition of μ_{k+1}. Meanwhile, for the primal gap we have

$$\Delta_{k+1}^p - \Delta_k^p = (\mathcal{L}_{k+1}(x_{k+2},\mu_{k+1}) - \mathcal{L}_k(x_{k+1},\mu_k)) + (\mathcal{L}_k(\tilde{x}_k,\mu_k) - \mathcal{L}_{k+1}(\tilde{x}_{k+1},\mu_{k+1})).$$

Note that

$$\mathcal{L}_k(x_{k+1},\mu_k) = \mathcal{L}_k(x_{k+1},\mu_{k+1}) - \theta_k\|Ax_{k+1}-b\|^2$$

and estimate $\mathcal{L}_k(\tilde{x}_k,\mu_k) - \mathcal{L}_{k+1}(\tilde{x}_{k+1},\mu_{k+1})$ as in (4.12), to get

$$\begin{aligned}\Delta_{k+1}^p - \Delta_k^p \le \mathcal{L}_{k+1}(x_{k+2},\mu_{k+1}) - \mathcal{L}_k(x_{k+1},\mu_{k+1}) + \theta_k\|Ax_{k+1}-b\|^2 \\ - \theta_k\langle Ax_{k+1}-b, A\tilde{x}_{k+1}-b\rangle.\end{aligned} \tag{4.13}$$

Using (4.12) and (4.13), we then have

$$\begin{aligned}\Delta_{k+1} - \Delta_k \le \mathcal{L}_{k+1}(x_{k+2},\mu_{k+1}) - \mathcal{L}_k(x_{k+1},\mu_{k+1}) + \theta_k\|Ax_{k+1}-b\|^2 \\ - 2\theta_k\langle Ax_{k+1}-b, A\tilde{x}_{k+1}-b\rangle.\end{aligned}$$

Note that

$$\mathcal{L}_k(x_{k+1},\mu_{k+1}) = \mathcal{L}_{k+1}(x_{k+1},\mu_{k+1}) - \left[g^{\beta_{k+1}} - g^{\beta_k}\right](Tx_{k+1}) - \left(\frac{\rho_{k+1}-\rho_k}{2}\right)\|Ax_{k+1}-b\|^2.$$

Then

$$\begin{aligned}\Delta_{k+1} - \Delta_k \le \mathcal{L}_{k+1}(x_{k+2},\mu_{k+1}) - \mathcal{L}_{k+1}(x_{k+1},\mu_{k+1}) + g^{\beta_{k+1}}(Tx_{k+1}) - g^{\beta_k}(Tx_{k+1}) \\ + \left(\frac{\rho_{k+1}-\rho_k}{2}\right)\|Ax_{k+1}-b\|^2 + \theta_k\|Ax_{k+1}-b\|^2 - 2\theta_k\langle Ax_{k+1}-b, A\tilde{x}_{k+1}-b\rangle.\end{aligned}$$

We denote by $\mathbf{T1} = \mathcal{L}_{k+1}(x_{k+2}, \mu_{k+1}) - \mathcal{L}_{k+1}(x_{k+1}, \mu_{k+1})$ and the remaining part of the right-hand side by $\mathbf{T2}$. For the moment, we focus our attention on $\mathbf{T1}$. Recall that $\mathcal{L}_k(x, \mu_k) = \mathcal{E}_k(x, \mu_k) + h(x)$ and apply Lemma 4.5 between points x_{k+2} and x_{k+1}, to get

$$\mathbf{T1} \leq h(x_{k+2}) - h(x_{k+1}) + \langle \nabla_x \mathcal{E}_{k+1}(x_{k+1}, \mu_{k+1}), x_{k+2} - x_{k+1} \rangle + K_{(F,\zeta,\mathcal{C})}\zeta(\gamma_{k+1}) + \frac{L_{k+1}}{2}\|x_{k+2} - x_{k+1}\|^2.$$

By (A.1) we have that h is convex and thus, since x_{k+2} is a convex combination of x_{k+1} and s_{k+1}, we get

$$\mathbf{T1} \leq -\gamma_{k+1}(h(x_{k+1}) - h(s_{k+1}) + \langle \nabla_x \mathcal{E}_{k+1}(x_{k+1}, \mu_{k+1}), x_{k+1} - s_{k+1} \rangle) + \frac{L_{k+1}}{2}\|x_{k+2} - x_{k+1}\|^2 + K_{(F,\zeta,\mathcal{C})}\zeta(\gamma_{k+1}).$$

Applying the definition of s_k as the minimizer of the linear minimization oracle and Lemma 4.6 at the points $\tilde{x}_{k+1}$, x_{k+1}, and μ_{k+1} gives,

$$\begin{aligned}
\mathbf{T1} &\leq -\gamma_{k+1}(h(x_{k+1}) - h(\tilde{x}_{k+1}) + \langle \nabla_x \mathcal{E}_{k+1}(x_{k+1}, \mu_{k+1}), x_{k+1} - \tilde{x}_{k+1} \rangle) \\
&\quad + \frac{L_{k+1}}{2}\|x_{k+2} - x_{k+1}\|^2 + K_{(F,\zeta,\mathcal{C})}\zeta(\gamma_{k+1}) \\
&\leq -\gamma_{k+1}\Big(h(x_{k+1}) - h(\tilde{x}_{k+1}) + \mathcal{E}_{k+1}(x_{k+1}, \mu_{k+1}) - \mathcal{E}_{k+1}(\tilde{x}_{k+1}, \mu_{k+1}) \\
&\quad + \frac{\rho_{k+1}}{2}\|A(x_{k+1} - \tilde{x}_{k+1})\|^2\Big) + \frac{L_{k+1}}{2}\|x_{k+2} - x_{k+1}\|^2 + K_{(F,\zeta,\mathcal{C})}\zeta(\gamma_{k+1}) \\
&= -\gamma_{k+1}\left(\mathcal{L}_{k+1}(x_{k+1}, \mu_{k+1}) - \mathcal{L}_{k+1}(\tilde{x}_{k+1}, \mu_{k+1}) + \frac{\rho_{k+1}}{2}\|A(x_{k+1} - \tilde{x}_{k+1})\|^2\right) \\
&\quad + \frac{L_{k+1}}{2}\|x_{k+2} - x_{k+1}\|^2 + K_{(F,\zeta,\mathcal{C})}\zeta(\gamma_{k+1}) \\
&\leq -\frac{\gamma_{k+1}\rho_{k+1}}{2}\|A(x_{k+1} - \tilde{x}_{k+1})\|^2 + \frac{L_{k+1}}{2}\|x_{k+2} - x_{k+1}\|^2 + K_{(F,\zeta,\mathcal{C})}\zeta(\gamma_{k+1}),
\end{aligned}$$

where we used that $\tilde{x}_{k+1}$ is a minimizer of $\mathcal{L}_{k+1}(\cdot, \mu_{k+1})$ in the last inequality. Now, combining $\mathbf{T1}$ and $\mathbf{T2}$ and using the Pythagoras identity we have

$$\begin{aligned}
\Delta_{k+1} - \Delta_k \leq -\theta_k\|A\tilde{x}_{k+1} - b\|^2 + \left(\theta_k - \gamma_{k+1}\frac{\rho_{k+1}}{2}\right)\|A(x_{k+1} - \tilde{x}_{k+1})\|^2 \\
+ \frac{L_{k+1}}{2}\|x_{k+2} - x_{k+1}\|^2 + K_{(F,\zeta,\mathcal{C})}\zeta(\gamma_{k+1}) + \left[g^{\beta_{k+1}} - g^{\beta_k}\right](Tx_{k+1}) \\
+ \frac{\rho_{k+1} - \rho_k}{2}\|Ax_{k+1} - b\|^2. \quad (4.14)
\end{aligned}$$

Under (P.6) we have $\theta_k = \frac{\gamma_k}{c}$ for some $c > 0$ such that

$$\exists \delta > 0, \quad \frac{\overline{M}}{c} - \frac{\underline{\rho}}{2} = -\delta < 0,$$

where $\overline{M}$ is the constant such that $\gamma_k \leq \overline{M}\gamma_{k+1}$ (see Assumption (P.5)). Then, using (P.5) and the above inequality,

$$\theta_k - \gamma_{k+1}\frac{\rho_{k+1}}{2} \leq \left(\frac{\overline{M}}{c} - \frac{\rho_{k+1}}{2}\right)\gamma_{k+1} \leq \left(\frac{\overline{M}}{c} - \frac{\underline{\rho}}{2}\right)\gamma_{k+1} = -\delta\gamma_{k+1} \text{ and } \theta_k \geq \tfrac{\underline{M}\gamma_{k+1}}{c}. \quad (4.15)$$

Now use the fact that $x_{k+2} = x_{k+1} + \gamma_{k+1}(s_{k+1} - x_{k+1})$ to estimate

$$\|x_{k+2} - x_{k+1}\|^2 \leq \gamma_{k+1}^2 d_{\mathcal{C}}^2. \tag{4.16}$$

Moreover, by the two assumptions (P.3), (A.4) and Proposition 2.1(v), (3.7) holds with a constant $M > 0$, and thus with Proposition 2.1(iv) we obtain

$$\left[g^{\beta_{k+1}} - g^{\beta_k}\right](Tx_{k+1}) \leq \frac{\beta_k - \beta_{k+1}}{2} \left\| \left[\partial g(Tx_{k+1})\right]^0 \right\|^2 \leq \frac{\beta_k - \beta_{k+1}}{2} M. \tag{4.17}$$

Plugging (4.15), (4.16) and (4.17) into (4.14), we get

$$\begin{aligned} \Delta_{k+1} - \Delta_k \leq & -\frac{M}{c}\gamma_{k+1}\|A\tilde{x}_{k+1} - b\|^2 - \delta\gamma_{k+1}\|A(x_{k+1} - \tilde{x}_{k+1})\|^2 + \frac{L_{k+1}}{2}\gamma_{k+1}^2 d_{\mathcal{C}}^2 \\ & + K_{(F,\zeta,\mathcal{C})}\zeta(\gamma_{k+1}) + \frac{\beta_k - \beta_{k+1}}{2} M + \left(\frac{\rho_{k+1} - \rho_k}{2}\right)\|Ax_{k+1} - b\|^2. \end{aligned} \tag{4.18}$$

Because of the assumptions (P.1) and (P.4), and in view of the definition of L_k in (4.7), we have the following,

$$\frac{L_k}{2}\gamma_k^2 d_{\mathcal{C}}^2 = \frac{1}{2}\left(\frac{\|T\|^2}{\beta_k} + \|A\|^2 \rho_k\right)\gamma_k^2 d_{\mathcal{C}}^2 \in \ell_+^1.$$

For the telescopic terms from the right hand side of (4.18) we have

$$\frac{\beta_k - \beta_{k+1}}{2} \in \ell_+^1 \text{ and } \left(\frac{\rho_{k+1} - \rho_k}{2}\right)\|Ax_{k+1} - b\|^2 \leq (\rho_{k+1} - \rho_k)\left(\|A\|^2 R^2 + \|b\|^2\right) \in \ell_+^1,$$

where R is the constant arising from (A.3). Under (P.1) we also have that

$$K_{(F,\zeta,\mathcal{C})}\zeta(\gamma_{k+1}) \in \ell_+^1.$$

Using the notation of Lemma 2.14, we set

$$\begin{aligned} & r_k = \Delta_k, \quad p_k = \gamma_{k+1}, \quad w_k = \left(\frac{M}{c}\|A\tilde{x}_{k+1} - b\|^2 + \delta\|A(x_{k+1} - \tilde{x}_{k+1})\|^2\right), \\ & z_k = \frac{L_{k+1}}{2}\gamma_{k+1}^2 d_{\mathcal{C}}^2 + K_{(F,\zeta,\mathcal{C})}\zeta(\gamma_{k+1}) + \frac{\beta_k - \beta_{k+1}}{2} M + \left(\frac{\rho_{k+1} - \rho_k}{2}\right)\|Ax_{k+1} - b\|^2. \end{aligned}$$

We have shown above that

$$r_{k+1} \leq r_k - p_k w_k + z_k,$$

where $(z_k)_{k\in\mathbb{N}} \in \ell_+^1$, and r_k is bounded from below. We then deduce using Lemma 2.14(i) that $(r_k)_{k\in\mathbb{N}}$ is convergent and

$$\left(\gamma_k\|A\tilde{x}_k - b\|^2\right)_{k\in\mathbb{N}} \in \ell_+^1, \quad \left(\gamma_k\|A(x_k - \tilde{x}_k)\|^2\right)_{k\in\mathbb{N}} \in \ell_+^1. \tag{4.19}$$

Consequently,

$$\left(\gamma_k\|Ax_k - b\|^2\right)_{k\in\mathbb{N}} \in \ell_+^1, \tag{4.20}$$

since, by Jensen's inequality,

$$\sum_{k=1}^{\infty}\gamma_k\|Ax_k - b\|^2 \leq 2\sum_{k=1}^{\infty}\gamma_k\left(\|A(x_k - \tilde{x}_k)\|^2 + \|A\tilde{x}_k - b\|^2\right) < +\infty.$$

□

We are now ready to prove Theorem 4.1, i.e., to show that the sequence of iterates $(x_k)_{k\in\mathbb{N}}$ is asymptotically feasible.

Proof. (i) By Lemma 4.8 with $\rho_k \equiv \rho_{k+1} \equiv 2$, we have

$$\|Ax_k - b\|^2 - \|Ax_{k+1} - b\|^2 \leq 2\gamma_k d_{\mathcal{C}} \|A\| \left(\|A\| R + \|b\|\right).$$

Using this together with Lemma 4.9 and Assumption (P.2), we are in position to apply Lemma 2.14(ii) to conclude that $\lim_{k\to\infty} \|Ax_k - b\| = 0$.

(ii) The rates in (4.1) follow respectively from Lemma 2.14(iii) and Lemma 2.14(iv).

(iii) We have, by Jensen's inequality and Lemma 4.9, that

$$\|A\bar{x}_k - b\|^2 \leq \frac{1}{\Gamma_k} \sum_{i=0}^{k} \gamma_i \|Ax_i - b\|^2 \leq \frac{1}{\Gamma_k} \sum_{i=0}^{+\infty} \gamma_i \|Ax_i - b\|^2 = O\left(\frac{1}{\Gamma_k}\right).$$

□

4.4 Dual multiplier boundedness

In this part we provide a lemma that shows the sequence of dual variables $(\mu_k)_{k\in\mathbb{N}}$ generated by Algorithm 1 is bounded.

We start by studying coercivity of $\bar{\varphi}$.

Lemma 4.10. *Suppose that Assumptions (A.1)-(A.3) and (A.6)-(A.8) hold. Then $\bar{\varphi}$ is coercive on* $\operatorname{ran}(A)$.

Proof. From (3.4), we have, for any $c \in A^{-1}(b)$, that

$$\bar{\varphi}(\mu) = \left(\bar{\Phi}^* + \langle -c, \cdot \rangle\right)(-A^*\mu).$$

Moreover, Assumptions (A.1) and (A.7) entail that $\bar{\Phi} \in \Gamma_0(\mathcal{H}_p)$. We now consider separately the two assumptions.

(a) Case of (A.8)(a): If follows from the Fenchel-Moreau theorem ([4, Theorem 13.32]) that

$$\left(\bar{\Phi}^* - \langle c, \cdot \rangle\right)^* = \bar{\Phi}^{**}(\cdot + c) = \bar{\Phi}(\cdot + c).$$

Using this, together with Proposition 2.10 and (A.2), we can assert that $\bar{\Phi}^* - \langle c, \cdot \rangle$ is coercive if and only if

$$\begin{aligned} 0 \in \operatorname{int}\left(\operatorname{dom}\left(\bar{\Phi}(\cdot + c)\right)\right) = \operatorname{int}\left(\operatorname{dom}(\Phi)\right) - c &= \operatorname{int}\left(\operatorname{dom}(g \circ T) \cap \mathcal{C}\right) - c \\ &= \operatorname{int}\left(\operatorname{dom}(g \circ T)\right) \cap \operatorname{int}(\mathcal{C}) - c. \end{aligned}$$

But this is precisely what (A.8)(a) guarantees. In turn, using [4, Proposition 14.15], (A.8)(a) is equivalent to

$$\exists (a > 0, \beta \in \mathbb{R}), \quad \bar{\Phi}^* - \langle c, \cdot \rangle \geq a \|\cdot\| + \beta.$$

Using standard results on linear operators in Hilbert spaces [4, Facts 2.18 and 2.19], we have

$$\text{(A.7)} \iff (\exists \alpha > 0), (\forall \mu \in \operatorname{ran}(A)), \quad \|A^*\mu\| \geq \alpha \|\mu\|.$$

Combining the last two inequalities, we deduce that under (A.8)(a),

$$\exists (a > 0, \alpha > 0, \beta \in \mathbb{R}), (\forall \mu \in \operatorname{ran}(A)), \quad \bar{\varphi}(\mu) \geq a \|A^*\mu\| + \beta \geq a\alpha \|\mu\| + \beta,$$

which in turn is equivalent to coercivity of $\bar{\varphi}$ on $\operatorname{ran}(A)$ by [4, Proposition 14.15].

(b) Case of (A.8)(b): Since $\mathcal{H}_d$ is finite dimensional, We have, $\forall u \in \mathcal{H}_d$,

$$\begin{aligned}
\bar{\varphi}^{\infty}(u) &= \left(\left(\bar{\Phi}^* + \langle -c, \cdot \rangle\right) \circ (-A^*)\right)^{\infty}(u) \\
\text{(Proposition 2.11(iii))} \quad &= \left(\bar{\Phi}^* + \langle -c, \cdot \rangle\right)^{\infty}(-A^* u) \\
\text{(Proposition 2.11(ii))} \quad &= \sigma_{\mathrm{dom}\left(\bar{\Phi}^* + \langle -c, \cdot \rangle\right)^*}(-A^* u) \\
&= \sigma_{\mathrm{dom}\left(\bar{\Phi}(\cdot + c)\right)}(-A^* u) \\
&= \sigma_{\mathrm{dom}\left(\bar{\Phi}\right) - c}(-A^* u) \\
\text{(by (A.2))} \quad &= \sigma_{\mathrm{dom}(g \circ T) \cap \mathcal{C} - c}(-A^* u)\,.
\end{aligned}$$

Notice that, by Assumption (A.4), we have $\mathrm{dom}\,(g \circ T) \cap \mathcal{C} = \mathcal{C}$. Thus, using Proposition 2.11(i), we have the following chain of equivalences

$$\begin{aligned}
\bar{\varphi} \text{ is coercive on } \mathrm{ran}\,(A) &\iff \bar{\varphi}^{\infty}(u) > 0, \quad \forall u \in \mathrm{ran}\,(A) \setminus \{0\} \\
&\iff \sigma_{\mathcal{C}-c}(-A^* u) > 0, \quad \forall u \in \mathrm{ran}\,(A) \setminus \{0\}\,.
\end{aligned}$$

For this to hold, and since $\mathrm{ran}\,(A) = \ker\,(A^*)^{\perp}$, a sufficient condition is that

$$\sigma_{\mathcal{C}-c}(x) > 0, \quad \forall x \in \mathrm{ran}\,(A^*) \setminus \{0\}\,. \tag{4.21}$$

It remains to check that the latter condition holds under (A.8)(b). First, observe that $\mathcal{C}$ is a nonempty bounded convex set thanks to (A.1) and (A.3). The first condition in (A.8)(b) is equivalent to $0 \in \mathrm{ri}(\mathcal{C} - c)$ for some $c \in A^{-1}(b)$. It then follows from Proposition 2.9 that

$$\sigma_{\mathcal{C}-c}(x) > 0, \forall x \notin \mathrm{par}(\mathcal{C} - c)^{\perp} = \mathrm{par}(\mathcal{C})^{\perp},$$

which then implies (4.21) thanks to the second condition in (A.8)(b).

□

Lemma 4.11. *Suppose that assumptions (A.1)-(A.3) and (A.6)-(A.8) and (P.1)-(P.6) hold. Then the sequence of dual iterates $(\mu_k)_{k\in\mathbb{N}}$ generated by* Algorithm 1 *is bounded.*

Proof. Using the notation in (3.4), the primal problem:

$$\min_{x \in \mathcal{H}_p} \{\Phi(x) : \ Ax = b\} = \min_{x \in \mathcal{H}_p} \sup_{\mu \in \mathcal{H}_d} \mathcal{L}(x, \mu)\,,$$

is obviously equivalent to

$$\min_{x \in \mathcal{H}_p} \left\{\Phi(x) + \frac{\rho_k}{2}\|Ax - b\|^2 : \ Ax = b\right\} = \min_{x \in \mathcal{H}_p} \sup_{\mu \in \mathcal{H}_d} \left\{\mathcal{L}(x, \mu) + \frac{\rho_k}{2}\|Ax - b\|^2\right\}.$$

We associate to the previous the following regularized primal problem:

$$\min_{x \in \mathcal{H}_p} \{\Phi_k(x) : \ Ax = b\} = \min_{x \in \mathcal{H}_p} \sup_{\mu \in \mathcal{H}_d} \mathcal{L}_k(x, \mu)$$

and its Lagrangian dual, namely:

$$\sup_{\mu \in \mathcal{H}_d} \inf_{x \in \mathcal{H}_p} \mathcal{L}_k(x, \mu) = -\inf_{\mu \in \mathcal{H}_d} \sup_{x \in \mathcal{H}_p} -\mathcal{L}_k(x, \mu)\,.$$

Now consider the dual function in the latter, namely $\varphi_k(\mu) \overset{\text{def}}{=} -\inf_{x\in\mathcal{H}_p} \mathcal{L}_k(x,\mu)$. Observe that the minimum is actually attained owing to (A.1) and (A.3). Now we claim that φ_k is continuously differentiable with $L_{\nabla\varphi_k}$-Lipschitz gradient, and $1/\underline{\rho}$ (see (P.4)) is an upper-bound for $(L_{\nabla\varphi_k})_{k\in\mathbb{N}}$. In order to show it, introduce the notation

$$\begin{aligned} F_k(x) &\overset{\text{def}}{=} f(x) + g^{\beta_k}(Tx) + h(x); \\ G_k(v) &\overset{\text{def}}{=} \frac{\rho_k}{2}\|v-b\|^2. \end{aligned}$$

By definition, we have

$$\begin{aligned} \varphi_k(\mu) &= -\min_{x\in\mathcal{H}_p}\left\{f(x) + g^{\beta_k}(Tx) + h(x) + \langle \mu,\ Ax-b\rangle + \frac{\rho_k}{2}\|Ax-b\|^2\right\} \\ &= -\min_{x\in\mathcal{H}_p}\left\{F_k(x) + \langle A^*\mu,\ x\rangle + G_k(Ax)\right\} + \langle \mu,\ b\rangle. \end{aligned} \tag{4.22}$$

Using Fenchel-Rockafellar duality and strong duality, which holds by (P.4) and continuity of G_k (see, for instance, [30, Theorem 3.51]), we have the following equality,

$$\begin{aligned} \min_{x\in\mathcal{H}_p}\left\{F_k(x) + \langle A^*\mu, x\rangle + G_k(Ax)\right\} &= -\min_{v\in\mathcal{H}_d}\left\{\left(F_k(\cdot) + \langle A^*\mu,\cdot\rangle\right)^*(-A^*v) + G_k^*(v)\right\} \\ &= -\min_{v\in\mathcal{H}_d}\left\{F_k^*(-A^*v - A^*\mu) + G_k^*(v)\right\} \end{aligned}$$

where we have used the fact that the conjugate of a linear perturbation is the translation of the conjugate in the last line. Substituting the above into (4.22) we find

$$\begin{aligned} \varphi_k(\mu) &= \min_{v\in\mathcal{H}_d}\left\{F_k^*(-A^*(v+\mu)) + \frac{1}{2\rho_k}\|v\|^2 + \langle v,\ b\rangle\right\} + \langle \mu,\ b\rangle \\ &= \min_{v\in\mathcal{H}_d}\left\{F_k^*(-A^*(v+\mu)) + \frac{1}{2\rho_k}\|v+\rho_k b\|^2\right\} + \langle \mu,\ b\rangle - \frac{\rho_k}{2}\|b\|^2 \end{aligned}$$

Moreover, from the primal-dual extremality relationships [30, Theorem 3.51(i)], we have

$$-\tilde{v} = \nabla G_k(A\tilde{x}) = \rho_k\left(A\tilde{x} - b\right), \tag{4.23}$$

where $\tilde{x}$ is a minimizer (which exists and belongs to $\mathcal{C}$) of the primal objective $\mathcal{L}_k(\cdot,\mu)$ and $\tilde{v}$ is the unique minimizer to the associated dual objective. Now, using the change of variable $u = v+\mu$, we get

$$\begin{aligned} \varphi_k(\mu) &= \inf_{u\in\mathcal{H}_d}\left\{F_k^*(-A^*u) + \frac{1}{2\rho_k}\|u-\mu+\rho_k b\|^2\right\} + \langle \mu,\ b\rangle - \frac{\rho_k}{2}\|b\|^2 \\ &= \left[F_k^*\circ(-A^*)\right]^{\rho_k}(\mu-\rho_k b) + \langle \mu,\ b\rangle - \frac{\rho_k}{2}\|b\|^2, \end{aligned}$$

where the notation $[\cdot]^{\rho_k}$ denotes the Moreau envelope with parameter ρ_k as defined in (2.2). It follows from Proposition 2.1(i) and (iii), that φ_k is convex, real-valued and its gradient, given by

$$\nabla\varphi_k(\mu) = \rho_k^{-1}\left(\mu - \rho_k b - \tilde{u}\right) + b = \rho_k^{-1}\left(\mu - \tilde{u}\right), \quad \text{where} \quad \tilde{u} = \operatorname{prox}_{\rho_k F_k^*\circ(-A^*)}(\mu - \rho_k b), \tag{4.24}$$

is $1/\rho_k$-Lipschitz continuous since the gradient of a Moreau envelope with parameter ρ_k is $1/\rho_k$-Lipschitz continuous (see Proposition 2.1(iii)). As ρ_k is non-decreasing, $1/\rho_k \leq 1/\underline{\rho}$ and the sequence of functions

$(\nabla\varphi_k)_{k\in\mathbb{N}}$ is uniformly Lipschitz-continuous with constant $1/\underline{\rho}$. In addition, combining (4.23) and (4.24), and recalling the change of variable $\tilde{u} = \tilde{v} + \mu$, we get that

$$\nabla\varphi_k(\mu) = \rho_k^{-1}(\mu - \tilde{u}) = -\rho_k^{-1}\tilde{v} = A\tilde{x} - b. \tag{4.25}$$

As in Lemma 4.9, we are going to denote $\tilde{x}_k$ a minimizer of $\mathcal{L}_k(x, \mu_k)$. Then, from the Descent Lemma (see Proposition 2.4 and inequality (2.7)), we have

$$\varphi_k(\mu_{k+1}) \leq \varphi_k(\mu_k) + \langle \nabla\varphi_k(\mu_k),\ \mu_{k+1} - \mu_k\rangle + \frac{1}{2\underline{\rho}}\|\mu_{k+1} - \mu_k\|^2.$$

Now substitute in the right-hand-side the expression $\nabla\varphi_k(\mu_k) = A\tilde{x}_k - b$ in (4.25) and the update $\mu_{k+1} = \mu_k + \theta_k(Ax_{k+1} - b)$ from the algorithm, to obtain

$$\begin{aligned}\varphi_k(\mu_{k+1}) &\leq \varphi_k(\mu_k) + \theta_k\langle A\tilde{x}_k - b,\ Ax_{k+1} - b\rangle + \frac{\theta_k^2}{2\underline{\rho}}\|Ax_{k+1} - b\|^2 \\ &\leq \varphi_k(\mu_k) + \frac{\theta_k}{2}\|A\tilde{x}_k - b\|^2 + \frac{\theta_k}{2}\left(\frac{\theta_k}{\underline{\rho}} + 1\right)\|Ax_{k+1} - b\|^2,\end{aligned} \tag{4.26}$$

where we estimated the scalar product by Cauchy-Schwartz and Young inequality. Moreover, by definition,

$$\begin{aligned}\varphi_{k+1}(\mu_{k+1}) &= -\inf_{x\in\mathcal{H}_p}\left\{f(x) + g^{\beta_{k+1}}(Tx) + h(x) + \langle\mu_{k+1},\ Ax - b\rangle + \frac{\rho_{k+1}}{2}\|Ax - b\|^2\right\} \\ &= \sup_{x\in\mathcal{H}_p}\left\{-\mathcal{L}_k(x, \mu_{k+1}) + \left[g^{\beta_k} - g^{\beta_{k+1}}\right](Tx) + \frac{1}{2}(\rho_k - \rho_{k+1})\|Ax - b\|^2\right\}.\end{aligned} \tag{4.27}$$

Now recall assumptions (P.3) and (P.4): for β_k non-increasing, $\left[g^{\beta_k} - g^{\beta_{k+1}}\right](Tx) \leq 0$ for every $x \in \mathcal{H}_p$ by Proposition 2.1(v) and, for ρ_k non-decreasing, $\rho_k - \rho_{k+1} \leq 0$. Then we can estimate the right-hand-side of (4.27) to obtain

$$\varphi_{k+1}(\mu_{k+1}) \leq \sup_{x\in\mathcal{H}_p} -\mathcal{L}_k(x, \mu_{k+1}) = \varphi_k(\mu_{k+1}).$$

Sum (4.26) with the latter, to obtain

$$\varphi_{k+1}(\mu_{k+1}) - \varphi_k(\mu_k) \leq \frac{\theta_k}{2}\|A\tilde{x}_k - b\|^2 + \frac{\theta_k}{2}\left(\frac{\theta_k}{\underline{\rho}} + 1\right)\|Ax_{k+1} - b\|^2.$$

By Assumption (P.6), $\theta_k = \gamma_k/c$ where $\gamma_k \leq 1$. Moreover, by assumption (P.5), $\gamma_k \leq \overline{M}\gamma_{k+1}$. Then,

$$\varphi_{k+1}(\mu_{k+1}) - \varphi_k(\mu_k) \leq \frac{\gamma_k}{2c}\|A\tilde{x}_k - b\|^2 + \frac{\overline{M}}{2c}\left(\frac{1}{\underline{\rho}c} + 1\right)\gamma_{k+1}\|Ax_{k+1} - b\|^2. \tag{4.28}$$

Notice that the right-hand-side is in ℓ^1_+, because both $\left(\gamma_k\|Ax_k - b\|^2\right)_{k\in\mathbb{N}}$ and $\left(\gamma_k\|A\tilde{x}_k - b\|^2\right)_{k\in\mathbb{N}}$ are in ℓ^1_+ by Lemma 4.9. Additionally, $(\varphi_k(\mu_k))_{k\in\mathbb{N}}$ is bounded from below. Indeed, by virtue of (A.6) and Remark 3.1(iv), we have

$$\begin{aligned}\varphi_k(\mu_k) &\geq -\mathcal{L}_k(x^\star, \mu_k) \\ &\geq -\left[f(x^\star) + g(Tx^\star) + h(x^\star)\right] > -\infty.\end{aligned}$$

Then we can use Lemma 2.14(i) on inequality (4.28) to conclude that $(\varphi_k(\mu_k))_{k\in\mathbb{N}}$ is convergent and, in particular, bounded. Now recall Φ_k, $\bar{\Phi}$ and $\bar{\varphi}$ from (3.4). Notice that

$$\begin{aligned}\varphi_k(\mu) &= \sup_{x\in\mathcal{H}_p}\{\langle\mu,\, b-Ax\rangle - \Phi_k(x)\}\\ &= \sup_{x\in\mathcal{H}_p}\{\langle -A^*\mu,\, x\rangle - \Phi_k(x)\} + \langle b,\, \mu\rangle\\ &= \Phi_k^*(-A^*\mu) + \langle b,\, \mu\rangle.\end{aligned}$$

It then follows that

$$g^{\beta_k} \leq g \quad\Longrightarrow\quad \Phi_k \leq \bar{\Phi} \quad\Longleftrightarrow\quad \bar{\Phi}^* \leq \Phi_k^* \quad\Longrightarrow\quad \bar{\varphi} \leq \varphi_k, \tag{4.29}$$

where we used Proposition 2.1(v) and the fact in (2.1). We are now in position to invoke Lemma 4.10 which shows that $\bar{\varphi}$ is coercive on $\mathrm{ran}(A)$, and thus, by (4.29), $(\varphi_k)_{k\in\mathbb{N}}$ is coercive uniformly in k on $\mathrm{ran}(A)$. In turn, since $\mathrm{ran}(A)$ is closed and $(\mu_k)_{k\in\mathbb{N}} \subset \mathrm{ran}(A) = \ker(A^*)^\perp$, we have from (4.29) and the proof of Lemma 4.10 that

$$\exists(a>0, \alpha>0, \beta\in\mathbb{R}), (\forall k\in\mathbb{N}), \quad \varphi_k(\mu_k) \geq \bar{\varphi}(\mu_k) \geq a\left\|A^*\mu_k\right\| + \beta \geq a\alpha\left\|\mu_k\right\| + \beta,$$

which shows that $(\mu_k)_{k\in\mathbb{N}}$ is indeed bounded by boundedness of $(\varphi_k(\mu_k))_{k\in\mathbb{N}}$. □

4.5 Optimality

In this section we prove Theorem 4.2 by establishing convergence of the Lagrangian values to the optimum (i.e., the value at the saddle-point).

We start by showing some boundedness claims that will be important in our proof.

Lemma 4.12. *Under assumptions (A.1)-(A.8) and (P.1)-(P.6), the objective Φ is bounded on $\mathcal{C}$, and thus*

$$\tilde{M} \stackrel{\text{def}}{=} \sup_{x\in\mathcal{C}}|\Phi(x)| + \sup_{k\in\mathbb{N}}\|\mu_k\|\left(\|A\|\, R + \|b\|\right) < +\infty, \tag{4.30}$$

where we recall the radius R from assumption (A.3).

Proof. By assumption (A.4), g is subdifferentiable at Tx for any $x\in\mathcal{C}$. Thus convexity of g implies that for any $x\in\mathcal{C}$

$$\begin{aligned}g(Tx) &\leq g(Tx^\star) + \left\langle [\partial g(Tx)]^0, Tx - Tx^\star\right\rangle \leq g(Tx^\star) + \left\|[\partial g(Tx)]^0\right\|\|T\| d_{\mathcal{C}}\\ g(Tx) &\geq g(Tx^\star) + \left\langle [\partial g(Tx^\star)]^0, Tx - Tx^\star\right\rangle \geq g(Tx^\star) - \left\|[\partial g(Tx^\star)]^0\right\|\|T\| d_{\mathcal{C}}.\end{aligned} \tag{4.31}$$

From assumptions (A.1) and (A.2), f belongs to $\Gamma_0(\mathcal{H}_p)$ and is differentiable on an open set $\mathcal{C}_0$ that contains $\mathcal{C} \subset \mathrm{dom}(f)$ (see Definition 2.6). Thus the continuity set of f contains $\mathcal{C}$, and it follows from [4, Corollary 8.30(ii)] that $\mathcal{C} \subset \mathrm{int}(\mathrm{dom}(f))$. Consequently, arguing as in the proof of Lemma 3.2, we deduce that

$$\sup_{x\in\mathcal{C}}\|\nabla f(x)\| < +\infty. \tag{4.32}$$

In turn, convexity entails that for any $x \in \mathcal{C}$

$$\begin{aligned} f(x) &\leq f(x^\star) + \langle \nabla f(x), x - x^\star \rangle \leq f(x^\star) + \|\nabla f(x)\| \, d_{\mathcal{C}}, \\ f(x) &\geq f(x^\star) + \langle \nabla f(x^\star), x - x^\star \rangle \geq f(x^\star) - \|\nabla f(x^\star)\| \, d_{\mathcal{C}}. \end{aligned} \tag{4.33}$$

From assumption (A.5), we also have for any $x \in \mathcal{C}$

$$h(x^\star) - L_h d_{\mathcal{C}} \leq h(x) \leq h(x^\star) + L_h d_{\mathcal{C}}. \tag{4.34}$$

Summing (4.31), (4.33) and (4.34), using (4.32) and assumption (A.4), we get

$$|\Phi(x)| \leq |\Phi(x^\star)| + \left(L_h + \|T\| \sup_{x \in \mathcal{C}} \left\| [\partial g(Tx)]^0 \right\| + \sup_{x \in \mathcal{C}} \|\nabla f(x)\| \right).$$

From Lemma 4.11, we know that the sequence of dual variables $(\mu_k)_{k \in \mathbb{N}}$ is bounded which concludes the proof. □

Define $C_k \overset{\text{def}}{=} \frac{L_k}{2} d_{\mathcal{C}}^2 + d_{\mathcal{C}} \left(D + M\|T\| + L_h + \|A\| \, \|\mu^\star\| \right)$, where L_k is given in (4.7) and the constants D, M, and L_h are as in Lemma 4.7. We then have the following lemma, in which we state the main energy estimation.

Lemma 4.13. *Suppose that assumptions (A.1)-(A.8) and (P.1)-(P.6) hold, with* $\underline{M} \geq 1$. *Consider the sequence of primal-dual iterates* $((x_k, \mu_k))_{k \in \mathbb{N}}$ *generated by* Algorithm 1 *and* $(x^\star, \mu^\star)$ *a saddle-point point of the Lagrangian as in* (3.5). *Let*

$$r_k \overset{\text{def}}{=} (1 - \gamma_k) \mathcal{L}_k(x_k, \mu_k) + \frac{c}{2} \|\mu_k - \mu^\star\|^2 + \frac{\beta_k}{2} M^2 + \gamma_k \tilde{M}. \tag{4.35}$$

Then, we have the following energy estimate

$$\begin{aligned} &r_{k+1} - r_k + \gamma_k \left[\mathcal{L}(x_k, \mu^\star) - \mathcal{L}(x^\star, \mu^\star) + \frac{\rho_k}{2} \|Ax_k - b\|^2 \right] \leq \\ &\frac{1}{2} \left[\rho_{k+1} - \rho_k - \gamma_{k+1} \rho_{k+1} + \frac{2}{c} \gamma_k - \frac{\gamma_k^2}{c} \right] \|Ax_{k+1} - b\|^2 + \frac{\gamma_k \beta_k}{2} M^2 + K_{(F,\zeta,\mathcal{C})} \zeta(\gamma_k) + C_k \gamma_k^2. \end{aligned} \tag{4.36}$$

Proof. Notice that the dual update $\mu_{k+1} = \mu_k + \theta_k (Ax_{k+1} - b)$ can be re-written as

$$\{\mu_{k+1}\} = \operatorname*{Argmin}_{\mu \in \mathcal{H}_d} \left\{ -\mathcal{L}_k(x_{k+1}, \mu) + \frac{1}{2\theta_k} \|\mu - \mu_k\|^2 \right\}.$$

Then, from firm nonexpansiveness of the proximal mapping (see (2.3)),

$$\begin{aligned} 0 &\geq \theta_k \left[\mathcal{L}_k(x_{k+1}, \mu^\star) - \mathcal{L}_k(x_{k+1}, \mu_{k+1}) \right] + \frac{1}{2} \left[\|\mu_{k+1} - \mu^\star\|^2 - \|\mu_k - \mu^\star\|^2 \right. \\ &\qquad \left. + \|\mu_{k+1} - \mu_k\|^2 \right] \\ &= \theta_k \left[\mathcal{L}_k(x_{k+1}, \mu^\star) - \mathcal{L}_k(x_{k+1}, \mu_{k+1}) \right] + \frac{1}{2} \left[\|\mu_{k+1} - \mu^\star\|^2 - \|\mu_k - \mu^\star\|^2 \right] \\ &\qquad + \frac{\theta_k^2}{2} \|Ax_{k+1} - b\|^2. \end{aligned} \tag{4.37}$$

Notice that

$$\mathcal{L}_k\left(x_{k+1},\mu_k\right)-\mathcal{L}_k\left(x_k,\mu_k\right)=\left[\mathcal{E}_k\left(x_{k+1},\mu_k\right)+h(x_{k+1})\right]-\left[\mathcal{E}_k\left(x_k,\mu_k\right)+h(x_k)\right]$$

and that, by the definition of x_{k+1} in the algorithm and by convexity of function h,

$$\begin{aligned} h(x_{k+1})-h(x_k) &= h((1-\gamma_k)\,x_k+\gamma_k s_k)-h(x_k)\\ &\leq \gamma_k\left(h(s_k)-h(x_k)\right). \end{aligned}$$

Then,

$$\mathcal{L}_k\left(x_{k+1},\mu_k\right)-\mathcal{L}_k\left(x_k,\mu_k\right)\leq\mathcal{E}_k\left(x_{k+1},\mu_k\right)-\mathcal{E}_k\left(x_k,\mu_k\right)+\gamma_k\left(h(s_k)-h(x_k)\right). \tag{4.38}$$

Now apply Lemma 4.6 at the points $x^\star$, x_k, and μ_k to affirm that

$$\mathcal{E}_k\left(x^\star,\mu_k\right)\geq\mathcal{E}_k\left(x_k,\mu_k\right)+\left\langle\nabla_x\mathcal{E}_k\left(x_k,\mu_k\right),\ x^\star-x_k\right\rangle+\frac{\rho_k}{2}\|A(x^\star-x_k)\|^2.$$

From the latter, by the alternative definition of s_k in the algorithm (see (3.3)), we obtain

$$\mathcal{E}_k\left(x^\star,\mu_k\right)\geq\mathcal{E}_k\left(x_k,\mu_k\right)-h(x^\star)+h(s_k)+\left\langle\nabla_x\mathcal{E}_k\left(x_k,\mu_k\right),\ s_k-x_k\right\rangle+\frac{\rho_k}{2}\|Ax_k-b\|^2. \tag{4.39}$$

From Lemma 4.5, we have also that

$$\mathcal{E}_k\left(x_{k+1},\mu_k\right)\leq\mathcal{E}_k\left(x_k,\mu_k\right)+\left\langle\nabla_x\mathcal{E}_k\left(x_k,\mu_k\right),\ x_{k+1}-x_k\right\rangle+K_{(F,\zeta,\mathcal{C})}\zeta\left(\gamma_k\right)+\frac{L_k}{2}\|x_{k+1}-x_k\|^2.$$

Recall that, from the algorithm, $x_{k+1}=x_k+\gamma_k\left(s_k-x_k\right)$. Then,

$$\begin{aligned} \mathcal{E}_k\left(x_{k+1},\mu_k\right) &\leq \mathcal{E}_k\left(x_k,\mu_k\right)+\gamma_k\left\langle\nabla_x\mathcal{E}_k\left(x_k,\mu_k\right),\ s_k-x_k\right\rangle+K_{(F,\zeta,\mathcal{C})}\zeta\left(\gamma_k\right)+\frac{L_k\gamma_k^2}{2}\|s_k-x_k\|^2\\ &\leq \mathcal{E}_k\left(x_k,\mu_k\right)+\gamma_k\left[\mathcal{E}_k\left(x^\star,\mu_k\right)+h(x^\star)-\mathcal{E}_k\left(x_k,\mu_k\right)-h(s_k)-\frac{\rho_k}{2}\|Ax_k-b\|^2\right]\\ &\quad+K_{(F,\zeta,\mathcal{C})}\zeta\left(\gamma_k\right)+\frac{L_k}{2}d_{\mathcal{C}}^2\gamma_k^2, \end{aligned}$$

where in the last inequality we used (4.39). Using the latter in (4.38), we obtain

$$\begin{aligned} \mathcal{L}_k\left(x_{k+1},\mu_k\right)-\mathcal{L}_k\left(x_k,\mu_k\right)\leq&\gamma_k\left[\mathcal{L}_k\left(x^\star,\mu_k\right)-\mathcal{L}_k\left(x_k,\mu_k\right)-\frac{\rho_k}{2}\|Ax_k-b\|^2\right]\\ &+K_{(F,\zeta,\mathcal{C})}\zeta\left(\gamma_k\right)+\frac{L_k}{2}d_{\mathcal{C}}^2\gamma_k^2. \end{aligned} \tag{4.40}$$

Notice also that, from the definitions of $\mathcal{L}_k\left(x_{k+1},\cdot\right)$ and μ_{k+1} as $\mu_{k+1}=\mu_k+\theta_k\left(Ax_{k+1}-b\right)$,

$$\mathcal{L}_k\left(x_{k+1},\mu_{k+1}\right)-\mathcal{L}_k\left(x_{k+1},\mu_k\right)=\left\langle\mu_{k+1}-\mu_k,\ Ax_{k+1}-b\right\rangle=\theta_k\|Ax_{k+1}-b\|^2.$$

So, from the latter and (4.40),

$$\begin{aligned} \mathcal{L}_k\left(x_{k+1},\mu_{k+1}\right)-\mathcal{L}_k\left(x_k,\mu_k\right)\leq\theta_k\|Ax_{k+1}-b\|^2+\gamma_k\left[\mathcal{L}_k\left(x^\star,\mu_k\right)-\mathcal{L}_k\left(x_k,\mu_k\right)\right]\\ -\frac{\rho_k\gamma_k}{2}\|Ax_k-b\|^2+K_{(F,\zeta,\mathcal{C})}\zeta\left(\gamma_k\right)+\frac{L_k}{2}d_{\mathcal{C}}^2\gamma_k^2. \end{aligned}$$

Now recall that, by assumption (P.6), $\theta_k = \gamma_k/c$. Multiply (4.37) by c and sum with the latter, to obtain

$$\begin{aligned}
&(1-c\theta_k)\mathcal{L}_k\left(x_{k+1},\mu_{k+1}\right) - (1-c\theta_k)\mathcal{L}_k\left(x_k,\mu_k\right) + \tfrac{c}{2}\left[\|\mu_{k+1}-\mu^\star\|^2 - \|\mu_k-\mu^\star\|^2\right] \\
\leq\ & \left(\theta_k - \tfrac{c\theta_k^2}{2}\right)\|Ax_{k+1}-b\|^2 + \gamma_k\left[\mathcal{L}_k\left(x^\star,\mu_k\right) - \mathcal{L}_k\left(x_k,\mu_k\right)\right] - c\theta_k\left[\mathcal{L}_k\left(x_{k+1},\mu^\star\right) - \mathcal{L}_k\left(x_k,\mu_k\right)\right] \\
& -\tfrac{\rho_k\gamma_k}{2}\|Ax_k-b\|^2 + K_{(F,\zeta,\mathcal{C})}\zeta\left(\gamma_k\right) + \tfrac{L_k}{2}d_{\mathcal{C}}^2\gamma_k^2.
\end{aligned}$$

The previous inequality can be re-written, by trivial manipulations, as

$$\begin{aligned}
&(1-c\theta_{k+1})\mathcal{L}_{k+1}\left(x_{k+1},\mu_{k+1}\right) - (1-c\theta_k)\mathcal{L}_k\left(x_k,\mu_k\right) + \frac{c}{2}\left[\|\mu_{k+1}-\mu^\star\|^2 - \|\mu_k-\mu^\star\|^2\right] \\
\leq\ & (1-c\theta_{k+1})\mathcal{L}_{k+1}\left(x_{k+1},\mu_{k+1}\right) - (1-c\theta_k)\mathcal{L}_k\left(x_{k+1},\mu_{k+1}\right) + \left(\theta_k - \frac{c\theta_k^2}{2}\right)\|Ax_{k+1}-b\|^2 \\
& + \gamma_k\left[\mathcal{L}_k\left(x^\star,\mu_k\right) - \mathcal{L}_k\left(x_k,\mu_k\right)\right] - c\theta_k\left[\mathcal{L}_k\left(x_{k+1},\mu^\star\right) - \mathcal{L}_k\left(x_k,\mu_k\right)\right] - \frac{\rho_k\gamma_k}{2}\|Ax_k-b\|^2 \\
& + K_{(F,\zeta,\mathcal{C})}\zeta\left(\gamma_k\right) + \frac{L_k}{2}d_{\mathcal{C}}^2\gamma_k^2 \\
=\ & c\left(\theta_k-\theta_{k+1}\right)\left[f+h+\langle\mu_{k+1},\ A\cdot-b\rangle\right]\left(x_{k+1}\right) + \left[\left(1-c\theta_{k+1}\right)g^{\beta_{k+1}} - \left(1-c\theta_k\right)g^{\beta_k}\right]\left(Tx_{k+1}\right) \\
& + \frac{1}{2}\left[\left(1-c\theta_{k+1}\right)\rho_{k+1} - \left(1-c\theta_k\right)\rho_k + 2\theta_k - c\theta_k^2\right]\|Ax_{k+1}-b\|^2 \\
& + \gamma_k\left[\mathcal{L}_k\left(x^\star,\mu_k\right) - \mathcal{L}_k\left(x_k,\mu_k\right)\right] - c\theta_k\left[\mathcal{L}_k\left(x_{k+1},\mu^\star\right) - \mathcal{L}_k\left(x_k,\mu_k\right)\right] - \frac{\rho_k\gamma_k}{2}\|Ax_k-b\|^2 \\
& + K_{(F,\zeta,\mathcal{C})}\zeta\left(\gamma_k\right) + \frac{L_k}{2}d_{\mathcal{C}}^2\gamma_k^2.
\end{aligned} \tag{4.41}$$

By (P.5) and (P.6), and the assumption that $\underline{M} \geq 1$, we have $\theta_{k+1} \leq \underline{M}^{-1}\theta_k \leq \theta_k$. In view of (P.3), we also have $\beta_{k+1} \leq \beta_k$ by (P.3). In particular, $g^{\beta_k} \leq g^{\beta_{k+1}} \leq g$. Now, by Proposition 2.1(iv) and the definition of the constant M in (3.7), we are able to estimate the quantity

$$\begin{aligned}
&\left[\left(1-c\theta_{k+1}\right)g^{\beta_{k+1}} - \left(1-c\theta_k\right)g^{\beta_k}\right]\left(Tx_{k+1}\right) \\
&= \left[g^{\beta_{k+1}} - g^{\beta_k}\right]\left(Tx_{k+1}\right) + c\left[\theta_k g^{\beta_k} - \theta_{k+1}g^{\beta_{k+1}}\right]\left(Tx_{k+1}\right) \\
&\leq \frac{1}{2}\left(\beta_k-\beta_{k+1}\right)\|\left[\partial g(Tx_{k+1})\right]^0\|^2 + c\left[\theta_k g^{\beta_k} - \theta_{k+1}g^{\beta_k}\right]\left(Tx_{k+1}\right) \\
&\leq \frac{1}{2}\left(\beta_k-\beta_{k+1}\right)M^2 + c\left(\theta_k-\theta_{k+1}\right)g(Tx_{k+1}).
\end{aligned}$$

Then,

$$\begin{aligned}
& c\left(\theta_k-\theta_{k+1}\right)\left[f+h+\langle\mu_{k+1},\ A\cdot-b\rangle\right]\left(x_{k+1}\right) + \left[\left(1-c\theta_{k+1}\right)g^{\beta_{k+1}} - \left(1-c\theta_k\right)g^{\beta_k}\right]\left(Tx_{k+1}\right) \\
\leq\ & c\left(\theta_k-\theta_{k+1}\right)\mathcal{L}\left(x_{k+1},\mu_{k+1}\right) + \frac{1}{2}\left(\beta_k-\beta_{k+1}\right)M^2.
\end{aligned} \tag{4.42}$$

Recall that, by assumption (A.3), $\mathcal{C}$ is convex and bounded and that, by the update $x_{k+1} = x_k + \gamma_k\left(s_k - x_k\right)$ with $s_k \in \mathcal{C}$ and $\gamma_k \in]0,1]$ by (P.1), x_k always belongs to $\mathcal{C}$. From the assumptions, the functions f, h and

$g \circ T$ are bounded on $\mathcal{C}$ and, from the algorithm and convexity, $(x_k)_{k\in\mathbb{N}} \subset \mathcal{C}$. By Lemma 4.11, also the sequence $(\mu_k)_{k\in\mathbb{N}}$ is bounded. Then, recalling $\tilde{M}$ from Lemma 4.12, we can use the Cauchy-Schwartz and the triangular inequality to affirm that

$$\mathcal{L}(x_k, \mu_k) = \Phi(x_k) + \langle \mu_k,\ Ax_k - b\rangle \leq \tilde{M}. \tag{4.43}$$

Recall the definition of r_k in (4.35). Coming back to (4.41) and using both (4.42) and (4.43), we obtain

$$\begin{aligned} r_{k+1} - r_k \ \leq\ & \frac{1}{2}\left[(1-\gamma_{k+1})\,\rho_{k+1} - (1-\gamma_k)\,\rho_k + \frac{2}{c}\gamma_k - \frac{\gamma_k^2}{c}\right]\|Ax_{k+1}-b\|^2 \\ & + \gamma_k\left[\mathcal{L}_k(x^\star,\mu_k) - \mathcal{L}_k(x_{k+1},\mu^\star)\right] - \frac{\rho_k\gamma_k}{2}\|Ax_k-b\|^2 + K_{(F,\zeta,\mathcal{C})}\zeta(\gamma_k) + \frac{L_k}{2}d_{\mathcal{C}}^2\gamma_k^2. \end{aligned} \tag{4.44}$$

Recall that, by feasibility of $x^\star$, $\mathcal{L}(x^\star,\mu_k) = \mathcal{L}(x^\star,\mu^\star)$. Now compute

$$\begin{aligned} \mathcal{L}_k(x^\star,\mu_k) - \mathcal{L}_k(x_{k+1},\mu^\star) &= \mathcal{L}(x^\star,\mu_k) - \mathcal{L}(x_{k+1},\mu^\star) + \left[g^{\beta_k} - g\right](Tx^\star) + \left[g - g^{\beta_k}\right](Tx_{k+1}) \\ &\quad - \frac{\rho_k}{2}\|Ax_{k+1}-b\|^2 \\ &\leq \mathcal{L}(x^\star,\mu^\star) - \mathcal{L}(x_{k+1},\mu^\star) + \frac{\beta_k}{2}M^2 - \frac{\rho_k}{2}\|Ax_{k+1}-b\|^2, \end{aligned}$$

where in the inequality we used the facts that $g^{\beta_k} \leq g$ and that, by Proposition 2.1(v) and (3.7),

$$\left[g - g^{\beta_k}\right](Tx_{k+1}) \leq \frac{\beta_k}{2}\|\left[\partial g(Tx_{k+1})\right]^0\|^2 \leq \frac{\beta_k}{2}M^2.$$

Then, using the latter in (4.44), we obtain

$$\begin{aligned} r_{k+1} - r_k \leq \frac{1}{2}\left[\rho_{k+1} - \rho_k - \gamma_{k+1}\rho_{k+1} + \frac{2}{c}\gamma_k - \frac{\gamma_k^2}{c}\right]\|Ax_{k+1}-b\|^2 + \gamma_k\left[\mathcal{L}(x^\star,\mu^\star) - \mathcal{L}(x_{k+1},\mu^\star)\right] \\ + \frac{\gamma_k\beta_k}{2}M^2 - \frac{\rho_k\gamma_k}{2}\|Ax_k-b\|^2 + K_{(F,\zeta,\mathcal{C})}\zeta(\gamma_k) + \frac{L_k}{2}d_{\mathcal{C}}^2\gamma_k^2. \end{aligned}$$

We replace the term $[\mathcal{L}(x^\star,\mu^\star) - \mathcal{L}(x_{k+1},\mu^\star)]$ with $[\mathcal{L}(x^\star,\mu^\star) - \mathcal{L}(x_k,\mu^\star)] + [\mathcal{L}(x_k,\mu^\star) - \mathcal{L}(x_{k+1},\mu^\star)]$ and estimate using Lemma 4.7 to get the following,

$$\begin{aligned} r_{k+1} - r_k \leq \frac{1}{2}\left[\rho_{k+1} - \rho_k - \gamma_{k+1}\rho_{k+1} + \frac{2}{c}\gamma_k - \frac{\gamma_k^2}{c}\right]\|Ax_{k+1}-b\|^2 + \gamma_k\left[\mathcal{L}(x^\star,\mu^\star) - \mathcal{L}(x_k,\mu^\star)\right] \\ + \frac{\gamma_k\beta_k}{2}M^2 - \frac{\rho_k\gamma_k}{2}\|Ax_k-b\|^2 + K_{(F,\zeta,\mathcal{C})}\zeta(\gamma_k) + C_k\gamma_k^2. \end{aligned}$$

We conclude by trivial manipulations. □

We are now ready to prove Theorem 4.2.

Proof. Our starting point is the main energy estimate (4.36). Let us focus on its right-hand-side. Under assumption (P.7),

$$\frac{1}{2}\left[\rho_{k+1} - \rho_k - \gamma_{k+1}\rho_{k+1} + \frac{2}{c}\gamma_k - \frac{\gamma_k^2}{c}\right]\|Ax_{k+1}-b\|^2 \leq \gamma_{k+1}\|Ax_{k+1}-b\|^2,$$

where the right hand side is in ℓ^1_+ by Lemma 4.9. Now remember that $C_k = \frac{L_k}{2} d_{\mathcal{C}}^2 + d_{\mathcal{C}} \left(D + M\|T\| + L_h + \|A\| \, \|\mu^\star\|\right)$, where $L_k = \|T\|^2/\beta_k + \|A\|^2 \rho_k$. Then we have

$$\gamma_k \beta_k M^2/2 + K_{(F,\zeta,\mathcal{C})} \zeta\left(\gamma_k\right) + C_k \gamma_k^2 = \gamma_k \beta_k M^2/2 + K_{(F,\zeta,\mathcal{C})} \zeta\left(\gamma_k\right) + \|T\|^2 \gamma_k^2 d_{\mathcal{C}} / \left(2\beta_k\right) + \|A\|^2 \rho_k \gamma_k^2 d_{\mathcal{C}}/2 + d_{\mathcal{C}} \left(D + M\|T\| + L_h + \|A\| \, \|\mu^\star\|\right) \gamma_k^2 \in \ell^1_+.$$

Indeed, under assumption (P.1), the sequences $\left(\gamma_k \beta_k\right)_{k\in\mathbb{N}}, \left(\zeta\left(\gamma_k\right)\right)_{k\in\mathbb{N}}$, and $\left(\gamma_k^2/\beta_k\right)_{k\in\mathbb{N}}$ belong to ℓ^1_+. Moreover, we have by assumptions (P.3) and (P.4) that $\underline{\rho}\gamma_k^2 \leq \rho_k \gamma_k^2 \leq \beta_0 \overline{\rho} \gamma_k^2/\beta_k$, whence we get that $\left(\rho_k \gamma_k^2\right)_{k\in\mathbb{N}} \in \ell^1_+$ and $\left(\gamma_k^2\right)_{k\in\mathbb{N}} \in \ell^1_+$ after invoking assumption (P.1). Thus all terms on the right hand side are summable. Let

$$w_k \overset{\text{def}}{=} \left[\mathcal{L}\left(x_k, \mu^\star\right) - \mathcal{L}\left(x^\star, \mu^\star\right)\right] + \frac{\rho_k}{2} \|Ax_k - b\|^2$$
$$z_k \overset{\text{def}}{=} \gamma_{k+1} \|Ax_{k+1} - b\|^2 + \gamma_k \beta_k M^2/2 + K_{(F,\zeta,\mathcal{C})} \zeta\left(\gamma_k\right) + C_k \gamma_k^2.$$

So far, we have shown that

$$r_{k+1} \leq r_k - \gamma_k w_k + z_k, \tag{4.45}$$

where r_k is bounded from below, and $\left(z_k\right)_{k\in\mathbb{N}} \in \ell^1_+$. The rest of the proof consists of invoking properly Lemma 2.14.

(i) In order to use Lemma 2.14(ii), we need to show that for some positive constant α,

$$w_k - w_{k+1} \leq \alpha \gamma_k.$$

Notice that the term $\mathcal{L}\left(x_k, \mu^\star\right) - \mathcal{L}\left(x^\star, \mu^\star\right)$ is proportional to γ_k by Lemma 4.7. For the second term of w_k, we have by Lemma 4.8 that $\frac{\rho_k}{2} \left\|Ax_k - b\right\|^2 - \frac{\rho_{k+1}}{2} \left\|Ax_{k+1} - b\right\|^2$ is proportional to γ_k. The desired claim then follows from Lemma 2.14(ii).

(ii) By [4, Lemma 2.37], we can assert that $\left(x_k\right)_{k\in\mathbb{N}}$ possesses a weakly convergent subsequence, say $\left(x_{k_j}\right)_{j\in\mathbb{N}}$, with cluster point $\bar{x} \in \mathcal{C}$. Since $\|A \cdot - b\| \in \Gamma_0(\mathcal{H}_p)$ and in view of [4, Theorem 9.1], we have

$$\|A\bar{x} - b\| \leq \liminf_j \left\|Ax_{k_j} - b\right\| = \lim_k \|Ax_k - b\| = 0,$$

where we used lower semicontinuity of the norm and Theorem 4.2. Thus $A\bar{x} = 0$, meaning that $\bar{x}$ is a feasible point of ($\mathscr{P}$). In turn, $\mathcal{L}\left(\bar{x}, \mu^\star\right) = \Phi(\bar{x})$. The function $\mathcal{L}\left(\cdot, \mu^\star\right)$ is lower semicontinuous by (A.1) and (A.6). Thus, using [4, Theorem 9.1] and by virtue of claim (i), we have

$$\Phi(\bar{x}) = \mathcal{L}\left(\bar{x}, \mu^\star\right) \leq \liminf_j \mathcal{L}\left(x_{k_j}, \mu^\star\right) = \lim_k \mathcal{L}\left(x_k, \mu^\star\right) = \mathcal{L}\left(x^\star, \mu^\star\right) \leq \mathcal{L}\left(x, \mu^\star\right)$$

for all $x \in \mathcal{H}_p$, and in particular for all $x \in A^{-1}(b)$. Thus, for every $x \in A^{-1}(b)$, we deduce that

$$\Phi(\bar{x}) \leq \mathcal{L}\left(x, \mu^\star\right) = \Phi(x),$$

meaning that $\bar{x}$ is a solution for problem ($\mathscr{P}$).

Meanwhile, as the sequence $(\mu_k)_{k\in\mathbb{N}}$ is bounded by Lemma 4.11, we can again invoke [4, Lemma 2.37] to extract a weakly convergent subsequence $\left(\mu_{k_j}\right)_{j\in\mathbb{N}}$ with cluster point $\bar{\mu}$. By Fermat's rule ([4, Theorem 16.2]), the weak sequential cluster point $\bar{\mu}$ is a solution to $(\mathscr{D})$ if and only if

$$0 \in \partial\left(\Phi^* \circ (-A^*)\right)(\bar{\mu}) + b.$$

Since the proximal operator is the resolvent of the subdifferential, it follows that (4.24) is equivalent to

$$\nabla\varphi_{k_j}\left(\mu_{k_j}\right) - b \in \partial\left(\Phi_{k_j}^* \circ (-A^*)\right)\left(\mu_{k_j} - \rho_{k_j}\nabla\varphi_{k_j}\left(\mu_{k_j}\right)\right). \tag{4.46}$$

By Lemma 4.9 it follows that $A\tilde{x}_k$ converges strongly to b and, combined with (4.25), thus $\nabla\varphi_{k_j}\left(\mu_{k_j}\right)$ converges strongly to 0. On the other hand, $\mu_{k_j} - \rho_{k_j}\nabla\varphi_{k_j}\left(\mu_{k_j}\right)$ converges weakly to $\bar{\mu}$. We now argue that we can pass to the limit in (4.46) by showing sequential closedness.

When $g \equiv 0$, we have, for all $j \in \mathbb{N}$, $\Phi_{k_j} \equiv f + h$ and the rest of the argument relies on sequential closedness of the graph of the subdifferential of $\Phi^* \circ (-A^*) \in \Gamma_0(\mathcal{H}_d)$ in the weak-strong topology. For the general case, our argument will rely on the fundamental concept of Mosco convergence of functions, which is epigraphical convergence for both the weak and strong topology (see [9] and [2, Definition 3.7]).

By Proposition 2.1(v) and assumptions (A.1)-(A.2), $\left(\Phi_{k_j}\right)_{j\in\mathbb{N}}$ is an increasing sequence of functions in $\Gamma_0\left(\mathcal{H}_d\right)$. It follows from [2, Theorem 3.20(i)] that Φ_{k_j} Mosco-converges to $\sup_{j\in\mathbb{N}}\Phi_{k_j} = \sup_{j\in\mathbb{N}} f + g^{\beta_{k_j}} \circ T + h = f + g\circ T + h = \Phi$ since $\beta_{k_j} \to 0$ by (P.3). Bicontinuity of the Legendre-Fenchel conjugation for the Mosco convergence (see [2, Theorem 3.18]) entails that $\Phi_{k_j}^* \circ -(A^*)$ Mosco-converges to $(f + g \circ T + h)^* \circ (-A^*) = \Phi^* \circ (-A^*)$. This implies, via [2, Theorem 3.66], that $\partial\Phi_{k_j}^* \circ (-A^*)$ graph-converges to $\partial\Phi^* \circ (-A^*)$, and [2, Proposition 3.59] shows that $\left(\partial\Phi_{k_j} \circ (-A^*)\right)_{j\in\mathbb{N}}$ is sequentially closed for graph-convergence in the weak-strong topology on $\mathcal{H}_d$, i.e., for any sequence (v_{k_j}, η_{k_j}) in the graph of $\partial\Phi_{k_j}^* \circ (-A^*)$ such that v_{k_j} converges weakly to $\bar{v}$ and η_{k_j} converges strongly to $\bar{\eta}$, we have $\bar{\eta} \in \partial\Phi^* \circ (-A^*)(\bar{v})$. Taking $v_{k_j} = \nabla\varphi_{k_j}\left(\mu_{k_j}\right) - b$ and $\eta_{k_j} = \mu_{k_j} - \rho_{k_j}\nabla\varphi_{k_j}\left(\mu_{k_j}\right)$, we conclude that

$$0 \in \partial\left(\Phi^* \circ (-A^*)\right)(\bar{\mu}) + b,$$

i.e., $\bar{\mu}$ is a solution of the dual problem $(\mathscr{D})$.

Recall r_k from (4.35) which verifies (4.45). From Lemma 2.14(i), $(r_k)_{k\in\mathbb{N}}$ is convergent. By (P.1) and (P.3), γ_k and β_k both converge to 0. We also have that

$$\begin{aligned}
-\mathcal{L}_k\left(x_k, \mu_k\right) &= \left(\mathcal{L}(x_k, \mu^\star) - \mathcal{L}_k\left(x_k, \mu_k\right)\right) - \mathcal{L}(x_k, \mu^\star) \\
&= g(Tx_k) - g^{\beta_k}(Tx_k) + \langle \mu^\star - \mu_k, Ax_k - b\rangle - \frac{\rho_k}{2}\left\|Ax_k - b\right\|^2 \\
&\quad - \mathcal{L}(x_k, \mu^\star).
\end{aligned}$$

We have from Theorem 4.1(i) that $\frac{\rho_k}{2}\left\|Ax_k - b\right\|^2 \to 0$. In turn, $\langle \mu^\star - \mu_k, Ax_k - b\rangle \to 0$ since $(\mu_k)_{k\in\mathbb{N}}$ is bounded (Lemma 4.11). We also have $\mathcal{L}(x_k, \mu^\star) \to \mathcal{L}(x^\star, \mu^\star)$ by claim (i) above. By Proposition 2.1(v) and (3.7), we get that

$$0 \leq \left(g(Tx_k) - g^{\beta_k}(Tx_k)\right) \leq \frac{\beta_k}{2}M^2.$$

Passing to the limit and in view of (P.3), we conclude that $g(Tx_k) - g^{\beta_k}(Tx_k) \to 0$. Altogether, this shows that $\mathcal{L}_k(x_k, \mu_k) \to \mathcal{L}(x^\star, \mu^\star)$. In turn, we conclude that the limit

$$\lim_{k\to\infty} \|\mu_k - \mu^\star\|^2 = 2/c \left(\lim_{k\to\infty} r_k - \mathcal{L}(x^\star, \mu^\star) \right)$$

exists. Since $\mu^\star$ was an arbitray optimal dual point, and we have shown above that each subsequence of $(\mu_k)_{k\in\mathbb{N}}$ converges weakly to an optimal dual point, we are in position to invoke Opial's lemma [29] to conclude that the whole dual multiplier sequence weakly converges to a solution of the dual problem.

(iii) Recalling that $(\gamma_k)_{k\in\mathbb{N}} \notin \ell^1_+$ (see assumption (P.2)), the rates in (4.4) follow by applying Lemma 2.14(iii)-(iv) to (4.45). Notice that both terms in w_k are positive and that $\rho_k \geq \underline{\rho} > 0$ (see again assumption (P.4)). Therefore we have that, for the same subsequence $\left(x_{k_j}\right)_{j\in\mathbb{N}}$, (4.6) holds.

(iv) The ergodic rate (4.2) follows by applying the Jensen's inequality to the convex function $\mathcal{L}(\cdot, \mu^\star)$.

□

5 Applications

5.1 Sum of several nonsmooth functions

In this section we explore the applications of Algorithm 1 to splitting in composite optimization problems, where we allow the presence of more than one nonsmooth function g or h in the objective:

$$\min_{x\in\mathcal{H}_p} \left\{ f(x) + \sum_{i=1}^n g_i(T_i x) + \sum_{i=1}^n h_i(x) \right\}. \tag{5.1}$$

First, we denote the product space by $\boldsymbol{\mathcal{H}}_p \overset{\text{def}}{=} \mathcal{H}_p^n$ endowed with the scalar product $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \frac{1}{n}\sum_{i=1}^n \langle x^{(i)}, y^{(i)} \rangle$, where $\boldsymbol{x}$ and $\boldsymbol{y}$ are vectors in $\boldsymbol{\mathcal{H}}_p$ with $\boldsymbol{x} \overset{\text{def}}{=} \left(x^{(1)}, \ldots, x^{(n)}\right)^\top$. We define also $\mathcal{V}$ as the diagonal subspace of $\boldsymbol{\mathcal{H}}_p$, i.e. $\mathcal{V} \overset{\text{def}}{=} \{\boldsymbol{x} \in \boldsymbol{\mathcal{H}}_p : x^{(1)} = \ldots = x^{(n)}\}$, $\mathcal{V}^\perp$ the orthogonal subspace to $\mathcal{V}$, and $\Pi_{\mathcal{V}}, \Pi_{\mathcal{V}^\perp}$ the orthogonal projections onto $\mathcal{V}$, $\mathcal{V}^\perp$ - respectively. We finally introduce the (diagonal) linear operator $\boldsymbol{T} : \boldsymbol{\mathcal{H}}_p \to \boldsymbol{\mathcal{H}}_p$ defined by

$$[\boldsymbol{T}(\boldsymbol{x})]^{(i)} = T_i x^{(i)}$$

and the functions

$$F(\boldsymbol{x}) \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n f\left(x^{(i)}\right); \quad G(\boldsymbol{T}\boldsymbol{x}) \overset{\text{def}}{=} \sum_{i=1}^n g_i\left(T_i x^{(i)}\right); \quad H(\boldsymbol{x}) \overset{\text{def}}{=} \sum_{i=1}^n h_i\left(x^{(i)}\right).$$

Then problem (5.1) is obviously equivalent to

$$\min_{\boldsymbol{x}\in\boldsymbol{\mathcal{H}}_p} \left\{ F(\boldsymbol{x}) + G(\boldsymbol{T}\boldsymbol{x}) + H(\boldsymbol{x}) : \ \Pi_{\mathcal{V}^\perp}\boldsymbol{x} = 0 \right\}, \tag{5.2}$$

which fits in the setting of our main problem ($\mathscr{P}$). In order to make more clear the presentation, we separate the two cases of multiple g and multiple h, that can be trivially combined. Moreover, we focus on the main case $h_i = \iota_{\mathcal{C}_i}$.

5.2 Sum of several simple functions over a compact set

Consider the following composite minimization problem,

$$\min_{x\in\mathcal{C}} \left\{ f(x) + \sum_{i=1}^{n} g_i(T_i x) \right\}. \tag{5.3}$$

We can reformulate the problem in the product space $\boldsymbol{\mathcal{H}}_p$ using the above notation to get,

$$\min_{\boldsymbol{x}\in\mathcal{C}^n\cap\mathcal{V}} \{F(\boldsymbol{x}) + G(\boldsymbol{T}\boldsymbol{x})\}.$$

Applying Algorithm 1 to this problem gives a completely separable scheme; we first compute the direction,

$$\boldsymbol{s}_k \in \operatorname*{Argmin}_{\boldsymbol{s}\in\mathcal{C}^n\cap\mathcal{V}} \langle \nabla \left(F(\boldsymbol{x}_k) + G^{\beta_k}(\boldsymbol{T}\boldsymbol{x}_k) \right), \boldsymbol{s} \rangle,$$

which reduces to the following computation since $\boldsymbol{s}_k = \begin{pmatrix} s_k \\ \vdots \\ s_k \end{pmatrix}$ has identical components,

$$s_k \in \operatorname*{Argmin}_{s\in\mathcal{C}} \left\langle \sum_{i=1}^{n} \left(\frac{1}{n}\nabla f\left(x_k^{(i)}\right) + \nabla g_i^{\beta_k}\left(T_i x_k^{(i)}\right) \right), s \right\rangle.$$

The term $\nabla g_i^{\beta_k}$ has a closed form given in Proposition 2.1 which can be used to get the following formula for the direction,

$$s_k \in \operatorname*{Argmin}_{s\in\mathcal{C}} \left\langle \sum_{i=1}^{n} \left(\frac{1}{n}\nabla f\left(x_k^{(i)}\right) + \frac{1}{\beta_k} T_i^* \left(T_i x_k^{(i)} - \operatorname{prox}_{\beta g}\left(T_i x_k^{(i)}\right)\right)\right), s \right\rangle.$$

5.3 Minimizing over intersection of compact sets

A classical problem found in machine learning is to minimize a Lipschitz-smooth function f over the intersection of convex, compact sets $\mathcal{C}_i$ in some real Hilbert space $\mathcal{H}$,

$$\min_{x\in\bigcap_{i=1}^{n}\mathcal{C}_i} f(x) = \min_{x\in\mathcal{H}} \left\{ f(x) + \sum_{i=1}^{n} h_i(x) \right\},$$

where $h_i \equiv \iota_{\mathcal{C}_i}$. Reformulating the problem in the product space $\boldsymbol{\mathcal{H}}_p$ gives,

$$\min_{\substack{\boldsymbol{x}\in\boldsymbol{\mathcal{H}}_p \\ \Pi_{\mathcal{V}^\perp}\boldsymbol{x}=0}} \{F(\boldsymbol{x}) + H(\boldsymbol{x})\}.$$

Then, we can apply Algorithm 1 and compute the step direction

$$\boldsymbol{s}_k \in \operatorname*{Argmin}_{\boldsymbol{s}\in\mathcal{C}_1\times\ldots\times\mathcal{C}_n} \langle \boldsymbol{s}, \nabla \left[F(\boldsymbol{x}) + \langle \boldsymbol{\mu}_k, \Pi_{\mathcal{V}^\perp}\boldsymbol{x}_k \rangle_+ \frac{\rho_k}{2} \|\Pi_{\mathcal{V}^\perp}\boldsymbol{x}_k\|^2 \right] \rangle$$

which gives a separable scheme for each component of $\boldsymbol{s}_k = \begin{pmatrix} s_k^{(1)} \\ \vdots \\ s_k^{(n)} \end{pmatrix}$,

$$\begin{aligned} s_k^{(i)} &\in \underset{s \in \mathcal{C}_i}{\operatorname{Argmin}} \left\langle s, \frac{1}{n} \nabla f\left(x_k^{(i)}\right) + \left(\Pi_{\mathcal{V}^\perp} \boldsymbol{\mu}_k\right)^{(i)} + \rho_k \left(\Pi_{\mathcal{V}^\perp} \boldsymbol{x}_k\right)^{(i)} \right\rangle \\ &= \underset{s \in \mathcal{C}_i}{\operatorname{Argmin}} \left\langle s, \frac{1}{n} \nabla f\left(x_k^{(i)}\right) + \mu_k^{(i)} - \frac{1}{n} \sum_{j=1}^{n} \mu_k^{(j)} + \rho_k \left(x_k^{(i)} - \frac{1}{n} \sum_{j=1}^{n} x_k^{(j)} \right) \right\rangle . \end{aligned} \tag{5.4}$$

6 Comparison

6.1 Conditional Gradient Framework

In [37] the following problem was analyzed in the finite-dimensional setting,

$$\min_{x \in \mathcal{C}} \{ f(x) + g(Tx) \} \tag{6.1}$$

where $f \in C^{1,1}(\mathbb{R}^n) \cap \Gamma_0(\mathbb{R}^n)$, $T \in \mathbb{R}^{d \times n}$ is a linear operator, $g \circ T \in \Gamma_0(\mathbb{R}^n)$, and $\mathcal{C}$ is a compact, convex subset of $\mathbb{R}^n$. They develop an algorithm which avoids projecting onto the set $\mathcal{C}$, instead utilizing a linear minimization oracle $\operatorname{lmo}_{\mathcal{C}}(v) = \underset{x \in \mathcal{C}}{\operatorname{Argmin}} \langle x, v \rangle$, and replaces the function $g \circ T$ with the smooth function $g_k^\beta \circ T$. They consider only functions f which are Lipschitz-smooth and finite dimensional spaces, i.e. $\mathbb{R}^n$, compared to CGALP which weakens the assumptions on f to be differentiable and (F, ζ)-smooth (see Definition 2.6) with an arbitrary real Hilbert space $\mathcal{H}_p$ (possibly infinite dimensional). Furthermore, the analysis in [37] is restricted to the parameter choices $\gamma_k = \frac{2}{k+1}$ and $\beta_k = \frac{\beta_0}{\sqrt{k+1}}$ exclusively, although they do include a section in which they consider two variants of an inexact linear minimization oracle: one with additive noise and one with multiplicative noise. In contrast, the results we present in Section 3 show optimality and feasibility for a wider choice for both the sequence of stepsizes $(\gamma_k)_{k \in \mathbb{N}}$ and the sequence of smoothing parameters $(\beta_k)_{k \in \mathbb{N}}$, although we only consider exact linear perturbation oracles of the form $\underset{s \in \mathcal{H}_p}{\operatorname{Argmin}} \{ h(s) + \langle x, s \rangle \}$. Finally, for solving (6.1) with an exact linear minimization oracle, our algorithm encompasses the algorithm in [37] by choosing $h(x) = \iota_{\mathcal{C}}(x)$, $A \equiv 0$, and restricting f to be in $C^{1,1}(\mathcal{H})$ with $\mathcal{H} = \mathbb{R}^n$.

In [37, Section 5] there is a discussion on splitting and affine constraints using the conditonal gradient framework presented. In this setting, i.e. assuming exact oracles, the primary difference between CGALP and the conditional gradient framework is the approach each algorithm takes to handle affine constraints. In CGALP, the augmented Lagrangian formulation is used to account for the affine constraints, introducing a dual variable μ and both a linear and quadratic term for the constraint $Ax - b = 0$. In contrast, in [37] the affine constraint is treated the same as the nonsmooth term $g \circ T$ and thus handled by quadratic penalization/smoothing alone. The consequence of smoothing for the affine constraint $Ax = b$ comes from calculating the gradient of the squared-distance to the constraint. This will involve solving a least squares problem at each iteration which can be computationally expensive. Our algorithm does not need to solve such a linear system.

The difference in the approaches is highlighted when both methods are applied to problem presented in Section 5.3 with $n = 2$ since this problem necessitates an affine constraint $\Pi_{\mathcal{V}^\perp} \boldsymbol{x} = 0$ for splitting. According

to [37, Section 5], we reformulate the problem to be

$$\min_{\substack{x^{(1)}\in\mathcal{C}_1\\ x^{(2)}\in C_2}} \left\{ \frac{1}{2}\left(f\left(x^{(1)}\right) + f\left(x^{(2)}\right)\right) + \iota_{\left\{x^{(1)}\right\}}\left(x^{(2)}\right)\right\}.$$

Note that the inclusion of the function $\iota_{\left\{x^{(1)}\right\}}\left(x^{(2)}\right)$ in the objective is equivalent to the affine constraint $\Pi_{\mathcal{V}^\perp}\boldsymbol{x} = 0$ in the $n = 2$ case. Apply the conditional gradient framework on the variable $\left(x^{(1)}, x^{(2)}\right)$ to get

$$\boldsymbol{s}_k \in \underset{\substack{s^{(1)}\in\mathcal{C}_1\\ s^{(2)}\in\mathcal{C}_2}}{\operatorname{Argmin}} \left\{ \left\langle \begin{pmatrix} s^{(1)}\\ s^{(2)}\end{pmatrix}, \begin{pmatrix} \nabla_{x^{(1)}}\left[\frac{1}{2}f\left(x_k^{(1)}\right) + \iota^{\beta_k}_{x_k^{(2)}}\left(x_k^{(1)}\right)\right]\\ \nabla_{x^{(2)}}\left[\frac{1}{2}f\left(x_k^{(2)}\right) + \iota^{\beta_k}_{x_k^{(1)}}\left(x_k^{(2)}\right)\right]\end{pmatrix}\right\rangle\right\},$$

which leads to a separable scheme that can be computed component-wise,

$$\begin{aligned} s_k^{(1)} &\in \underset{s\in\mathcal{C}_1}{\operatorname{Argmin}} \left\langle s, \frac{1}{2}\nabla f\left(x_k^{(1)}\right) + \frac{x_k^{(1)} - x_k^{(2)}}{\beta_k}\right\rangle\\ s_k^{(2)} &\in \underset{s\in\mathcal{C}_2}{\operatorname{Argmin}} \left\langle s, \frac{1}{2}\nabla f\left(x_k^{(2)}\right) + \frac{x_k^{(2)} - x_k^{(1)}}{\beta_k}\right\rangle. \end{aligned} \tag{6.2}$$

Compare the direction obtained in (6.2) to the one obtained in (5.4), the components of which we rewrite below for $n = 2$,

$$\begin{aligned} s_k^{(1)} &\in \underset{s\in\mathcal{C}_1}{\operatorname{Argmin}} \left\langle s, \frac{1}{2}\nabla f\left(x_k^{(1)}\right) + \frac{1}{2}\left(\mu_k^{(1)} - \mu_k^{(2)}\right) + \frac{\rho_k}{2}\left(x_k^{(1)} - x_k^{(2)}\right)\right\rangle\\ s_k^{(2)} &\in \underset{s\in\mathcal{C}_2}{\operatorname{Argmin}} \left\langle s, \frac{1}{2}\nabla f\left(x_k^{(2)}\right) + \frac{1}{2}\left(\mu_k^{(2)} - \mu_k^{(1)}\right) + \frac{\rho_k}{2}\left(x_k^{(2)} - x_k^{(1)}\right)\right\rangle. \end{aligned} \tag{6.3}$$

Due to affine constraint, the computation of the direction in (6.2) necessitates smoothing and, as a consequence, the parameter β_k, which is necessarily going to 0. In CGALP , the introduction of the dual variable $\boldsymbol{\mu}_k$ in place of smoothing the affine constraint avoids the parameter β_k. Instead, we have the parameter ρ_k but ρ_k can be picked to be constant without issue.

6.2 FW-AL Algorithm

In [16] the following problem was analyzed,

$$\min_{\substack{x\in\bigcap_{i=1}^n \mathcal{C}_i\\ Ax=0}} f(x)$$

using a combination of the Frank-Wolfe algorithm with the augmented Lagrangian to account for the constraint $Ax = 0$. The function f is assumed to be Lipschitz-smooth, in contrast to our approach. The perspective used in their paper is to modify the classic ADMM algorithm, replacing the marginal minimization with respect to the primal variable by a Frank-Wolfe step instead, although their analysis is not restricted only to Frank-Wolfe steps. Indeed, in all the scenarios where one can apply FW-AL using a Frank-Wolfe step our

algorithm encompasses FW-AL as a special case, discussed in Section 5.3. The primary differences between CGALP and FW-AL are in the convergence results and the generality of CGALP. The results in [16] prove convergence of the objective in the case where the sets $\mathcal{C}_i$ are polytopes and convergence of the iterates in the case where the sets $\mathcal{C}_i$ are polytopes and f is strongly convex, but they do not prove convergence of the objective, weak convergence of the dual variable, or asymptotic feasibility of the iterates in the general case where each $\mathcal{C}_i$ is a compact, convex set. Instead, they prove two theorems which imply subsequential convergence of the objective and subsequential asymptotic feasibility in the general case and subsequential convergence of the iterates to the optimum in the strongly convex case in [16, Theorem 2] and [16, Corollary 2] respectively. Unfortunately, each of these results is obtained separately and so the subsequences that produce each result are not guaranteed to coincide with one another.

Interestingly, the results they obtain are not unique to Frank-Wolfe style algorithms as their analysis is from the perspective of a modified ADMM algorithm; they only require that the algorithm used to replace the marginal minimization on the primal variable in ADMM produces sublinear decrease in the objective. Finally, they do not provide conditions for the dual multiplier sequence, μ_k in our notation, to be bounded as they discuss in their analysis of issues with similar proofs, e.g. in GDMM. This is a crucial issue as the constants in their bounds depend on the norm of these dual multipliers.

7 Numerical Experiments

In this section we present some numerical experiments comparing the performance of Algorithm 1 and a proximal algorithm applied to splitting in composite optimization problems.

7.1 Projection problem

First, we consider a simple projection problem,

$$\min_{x\in\mathbb{R}^2}\left\{\frac{1}{2}\|x-y\|_2^2:\ \|x\|_1\leq 1, Ax=0\right\}, \tag{7.1}$$

where $y\in\mathbb{R}^2$ is the vector to be projected and $A:\mathbb{R}^2\to\mathbb{R}^2$ is a rank-one matrix. To exclude trivial projections, we choose randomly $y\notin\mathbb{B}_1^1\cap\ker(A)$, where $\mathbb{B}_1^1$ is the unit ℓ^1 ball centered at the origin. Then Problem (7.1) is nothing but Problem ($\mathscr{P}$) with $f(x)=\frac{1}{2}\|x-y\|_2^2$, $g\equiv 0$, $h\equiv\iota_{\mathbb{B}_1^1}$ and $\mathcal{C}=\mathbb{B}_1^1$.

The assumptions mentioned previously, i.e. (A.1)-(A.8), all hold in this finite-dimensional case as f, g, and h are all in $\Gamma_0\left(\mathbb{R}^2\right)$, f is Lipschitz-smooth, h is the indicator function for a compact convex set, g has full domain and $0\in\ker(A)\cap\operatorname{int}(\mathcal{C})$. Regarding the parameters and the associated assumptions, we choose γ_k according to Example 3.4 with $(a,b)\in\{(0,0),(0,1/3-0.01),(1,1/3-0.01)\}$, $\theta_k=\gamma_k$, and $\rho=2^{2-b}+1$. The ergodic convergence profiles of the Lagrangian are displayed in Figure 1 along with the theoretical rates (see Theorem 4.2 and Example 4.4). The observed rates agree with the predicted ones of $O\left(\frac{1}{\log(k+2)}\right)$, $O\left(\frac{1}{(k+2)^b}\right)$ and $o\left(\frac{1}{(k+2)^b}\right)$ for the respective choices of (a,b).

7.2 Matrix completion problem

We also consider the following, more complicated matrix completion problem,

$$\min_{X\in\mathbb{R}^{N\times N}}\left\{\|\Omega X-y\|_1:\ \|X\|_*\leq\delta_1,\|X\|_1\leq\delta_2\right\}, \tag{7.2}$$

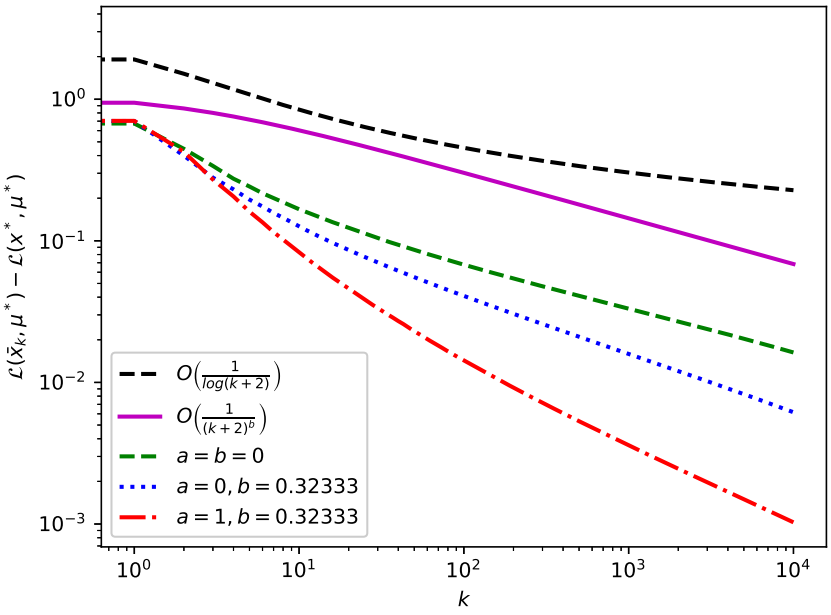

Figure 1: Ergodic convergence profiles for CGALP applied to the simple projection problem.

where δ_1 and δ_2 are positive constants, $\Omega : \mathbb{R}^{N\times N} \to \mathbb{R}^p$ is a masking operator, $y \in \mathbb{R}^p$ is a vector of observations, and $\|\cdot\|_*$ and $\|\cdot\|_1$ are respectively the nuclear and ℓ^1 norms. The mask operator Ω is generated randomly by specifying a sampling density, in our case 0.8. We generate the vector y randomly in the following way. We first generate a sparse vector $\tilde{y} \in \mathbb{R}^N$ with $N/5$ non-zero entries independently uniformly distributed in $[-1, 1]$. We take the exterior product $\tilde{y}\tilde{y}^\top = X_0$ to get a rank-1 sparse matrix which we then mask to get ΩX_0. The radii of the contraints in (7.2) are chosen according to the nuclear norm and ℓ^1 norm of X_0, $\delta_1 = \frac{\|X_0\|_*}{2}$ and $\delta_2 = \frac{\|X_0\|_1}{2}$.

7.2.1 CGALP

Problem (7.2) is a special instance of (5.1) with $n = 2$, $f \equiv 0$, $g_i = \|\cdot - y\|_1/2$, $T_i = \Omega$, $h_1 = \iota_{\mathbb{B}_*^{\delta_1}}$, $h_2 = \iota_{\mathbb{B}_1^{\delta_2}}$, where $\mathbb{B}_*^{\delta_1}$ and $\mathbb{B}_1^{\delta_2}$ are the nuclear and ℓ^1 balls of radii δ_1 and δ_2. We then follow the same steps as in Section 5.1. Let $\mathcal{H}_p = \mathbb{R}^{N\times N}$, $\boldsymbol{\mathcal{H}}_p = \mathcal{H}_p^2$, $\boldsymbol{X} = \begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix} \in \boldsymbol{\mathcal{H}}_p$. We then have $G(\Omega \boldsymbol{X}) = \frac{1}{2}\left(\left\|\Omega X^{(1)} - y\right\|_1 + \left\|\Omega X^{(2)} - y\right\|_1\right)$, and $H(\boldsymbol{X}) = \iota_{\mathbb{B}_*^{\delta_1}}(X^{(1)}) + \iota_{\mathbb{B}_1^{\delta_2}}(X^{(2)})$. Then problem (7.2) is obviously equivalent to

$$\min_{\boldsymbol{X}\in\boldsymbol{\mathcal{H}}_p} \left\{G(\Omega \boldsymbol{X}) + H(\boldsymbol{X}) : \Pi_{\mathcal{V}^\perp} \boldsymbol{X} = 0\right\}, \tag{7.3}$$

which is a special case of (5.2) with $F \equiv 0$. It is immediate to check that our assumptions (A.1)-(A.8) hold. Indeed, all functions are in $\Gamma_0(\mathcal{H}_p)$ and $F \equiv 0$, and thus (A.1) and (A.2) are verified. $\mathcal{C} = \mathbb{B}_*^{\delta_1} \times \mathbb{B}_1^{\delta_2}$ which is a non-mepty convex compact set. We also have $\Omega\mathcal{C} \subset \mathrm{dom}(\partial G) = \mathbb{R}^p \times \mathbb{R}^p$, and for any $\boldsymbol{z} \in \mathbb{R}^p \times \mathbb{R}^p$,

$\partial G(\boldsymbol{z}) \subset \mathbb{B}_\infty^{1/2} \times \mathbb{B}_\infty^{1/2}$ and thus (A.4) is verified. (A.5) also holds with $L_h = 0$. $\mathcal{V}$ is closed as we are in finite dimension, and thus (A.7) is fulfilled. We also have, since $\operatorname{dom}(G \circ \Omega) = \boldsymbol{\mathcal{H}}_p$,

$$\mathbf{0} \in \mathcal{V} \cap \operatorname{int}\left(\operatorname{dom}(G \circ \Omega)\right) \cap \operatorname{int}\left(\mathcal{C}\right) = \mathcal{V} \cap \operatorname{int}(\mathbb{B}_*^{\delta_1}) \times \operatorname{int}(\mathbb{B}_1^{\delta_2}),$$

which shows that (A.8) is verified. The latter is nothing but the condition in [4, Fact 15.25(i)]. It then follows from the discussion in Remark 3.1(iv) that (A.6) holds true.

We use Algorithm 1 by choosing the sequence of parameters $\gamma_k = \frac{1}{k+1}$, $\beta_k = \frac{1}{\sqrt{k+1}}$, $\theta_k = \gamma_k$, and $\rho_k \equiv 15$, which verify all our assumptions (P.1)-(P.7) in view of Example 3.4. Our choice of γ_k is the most common in the literature, and it can be improved according to our discussion in the previous section.

Finding the direction $\boldsymbol{S}_k$ by solving the linear minimization oracle is a separable problem, and thus each component is given by,

$$\begin{aligned} S_k^{(1)} \in \underset{S^{(1)} \in \mathbb{B}_{\|\cdot\|_*}^{\delta_1}}{\operatorname{Argmin}} \Bigg\langle & \frac{\Omega^*\left(\Omega X_k^{(1)} - y - \operatorname{prox}_{\frac{\beta_k}{2}\|\cdot\|_1}\left(\Omega X_k^{(1)} - y\right)\right)}{\beta_k} \\ & + \frac{1}{2}\left(\mu_k^{(1)} - \mu_k^{(2)} + \rho_k\left(X_k^{(1)} - X_k^{(2)}\right)\right), S^{(1)} \Bigg\rangle, \\ S_k^{(2)} \in \underset{S^{(2)} \in \mathbb{B}_{\|\cdot\|_1}^{\delta_2}}{\operatorname{Argmin}} \Bigg\langle & \frac{\Omega^*\left(\Omega X_k^{(2)} - y - \operatorname{prox}_{\frac{\beta_k}{2}\|\cdot\|_1}\left(\Omega X_k^{(2)} - y\right)\right)}{\beta_k} \\ & + \frac{1}{2}\left(\mu_k^{(2)} - \mu_k^{(1)} + \rho_k\left(X_k^{(2)} - X_k^{(1)}\right)\right), S^{(2)} \Bigg\rangle. \end{aligned} \tag{7.4}$$

Because of the structure of the sets $\mathbb{B}_{\|\cdot\|_*}^{\delta_1}$ and $\mathbb{B}_{\|\cdot\|_1}^{\delta_2}$, finding the first component of $\boldsymbol{S}_k$ reduces to computing the leading right and left singular vectors of

$$\frac{\Omega^*\left(\Omega X_k^{(1)} - y - \operatorname{prox}_{\frac{\beta_k}{2}\|\cdot\|_1}\left(\Omega X_k^{(1)} - y\right)\right)}{\beta_k} + \frac{1}{2}\left(\mu_k^{(1)} - \mu_k^{(2)} + \rho_k\left(X_k^{(1)} - X_k^{(2)}\right)\right)$$

while finding the second component reduces to computing the largest entry of

$$\left| \left(\frac{\Omega^*\left(\Omega X_k^{(2)} - y - \operatorname{prox}_{\frac{\beta_k}{2}\|\cdot\|_1}\left(\Omega X_k^{(2)} - y\right)\right)}{\beta_k} + \frac{1}{2}\left(\mu_k^{(2)} - \mu_k^{(1)} + \rho_k\left(X_k^{(2)} - X_k^{(1)}\right)\right) \right)_{(i,j)} \right|$$

over all the entries (i, j). The dual variable update is given by,

$$\boldsymbol{\mu}_{k+1} \overset{\text{def}}{=} \begin{pmatrix} \mu_{k+1}^{(1)} \\ \mu_{k+1}^{(2)} \end{pmatrix} = \begin{pmatrix} \mu_k^{(1)} \\ \mu_k^{(2)} \end{pmatrix} + \frac{\gamma_k}{2} \begin{pmatrix} X_{k+1}^{(1)} - X_{k+1}^{(2)} \\ X_{k+1}^{(2)} - X_{k+1}^{(1)} \end{pmatrix}$$

7.2.2 GFB

Let $\mathcal{H}_p = \mathbb{R}^{N \times N}$, $\boldsymbol{\mathcal{H}}_p = \mathcal{H}_p^3$, $\boldsymbol{W} = \begin{pmatrix} W^{(1)} \\ W^{(2)} \\ W^{(3)} \end{pmatrix} \in \boldsymbol{\mathcal{H}}_p$, $Q\left(\boldsymbol{W}\right) = \left\|\Omega W^{(1)} - y\right\|_1 + \iota_{\mathbb{B}_{\|\cdot\|_*}^{\delta_1}}\left(W^{(2)}\right) + \iota_{\mathbb{B}_{\|\cdot\|_1}^{\delta_2}}\left(W^{(3)}\right)$. Then we reformulate problem (7.2) as

$$\min_{\boldsymbol{W} \in \boldsymbol{\mathcal{H}}_p} \left\{ Q\left(\boldsymbol{W}\right) : \ \boldsymbol{W} \in \mathcal{V} \right\}, \tag{7.5}$$

which fits the framework to apply the GFB algorithm proposed in [32] (in fact Douglas-Rachford since the smooth part vanishes).

The algorithm has three steps, each of which is separable in the components. We choose the step sizes $\lambda_k = \gamma = 1$ in the GFB to get,

$$\begin{cases} \boldsymbol{U}_{k+1} = \begin{pmatrix} 2W_k^{(1)} - Z_k^{(1)} + \Omega^* \left(y - \Omega \left(2W_k^{(1)} - Z_k^{(1)} \right) + \operatorname{prox}_{\|\cdot\|_1} \left(\Omega \left(2W_k^{(1)} - Z_k^{(1)} \right) - y \right) \right) \\ \Pi_{\mathbb{B}^{\delta_1}_{\|\cdot\|_*}} \left(2W_k^{(2)} - Z_k^{(2)} \right) \\ \Pi_{\mathbb{B}^{\delta_2}_{\|\cdot\|_1}} \left(2W_k^{(3)} - Z_k^{(3)} \right) \end{pmatrix} \\ \boldsymbol{Z}_{k+1} = \boldsymbol{Z}_k + \boldsymbol{U}_{k+1} - \boldsymbol{W}_k \\ \boldsymbol{W}_{k+1} = \begin{pmatrix} \sum_{i=1}^3 Z_{k+1}^{(i)}/3 \\ \sum_{i=1}^3 Z_{k+1}^{(i)}/3 \\ \sum_{i=1}^3 Z_{k+1}^{(i)}/3 \end{pmatrix} \end{cases} \tag{7.6}$$

We know from [32] that $\boldsymbol{Z}_k$ converges to $\boldsymbol{Z}^\star$, and $\boldsymbol{W}_k$ and $\boldsymbol{U}_k$ both converge to $\boldsymbol{W}^\star = \Pi_{\mathcal{V}}(\boldsymbol{Z}^\star) = (X^\star, X^\star, X^\star)$, where $X^\star$ is a minimizer of (7.2).

7.2.3 Results

We compare the performance of CGALP with GFB for varying dimension, N, using their respective ergodic convergence criteria. For CGALP this is the quantity $\mathcal{L}\left(\bar{\boldsymbol{X}}_k, \mu^*\right) - \mathcal{L}\left(\boldsymbol{X}^\star, \boldsymbol{\mu}^\star\right)$ where $\bar{\boldsymbol{X}}_k = \sum_{i=0}^{k} \gamma_i \boldsymbol{X}_i / \Gamma_k$. Meanwhile, for GFB, we know from [25] that the Bregman divergence $D_Q^{\boldsymbol{v}^\star}\left(\bar{\boldsymbol{U}}_k\right) = Q(\bar{\boldsymbol{U}}_k) - Q(\boldsymbol{W}^\star) - \left\langle \boldsymbol{v}^\star, \bar{\boldsymbol{U}}_k - \boldsymbol{W}^\star \right\rangle$, with $\bar{\boldsymbol{U}}_k = \sum_{i=0}^{k} \boldsymbol{U}_i/(k+1)$ and $\boldsymbol{v}^\star = (\boldsymbol{W}^\star - \boldsymbol{Z}^\star)/\gamma$, converges at the rate $O(1/(k+1))$. To compute the convergence criteria, we first run each algorithm for 10^5 iterations to approximate the optimal variables ($\boldsymbol{X}^\star$ and $\boldsymbol{\mu}^\star$ for CGALP, and $\boldsymbol{Z}^\star$ and $\boldsymbol{W}^\star$ for GFB). Then, we run each algorithm again for 10^5 iterations, this time recording the convergence criteria at each iteration. The results are displayed in Figure 2.

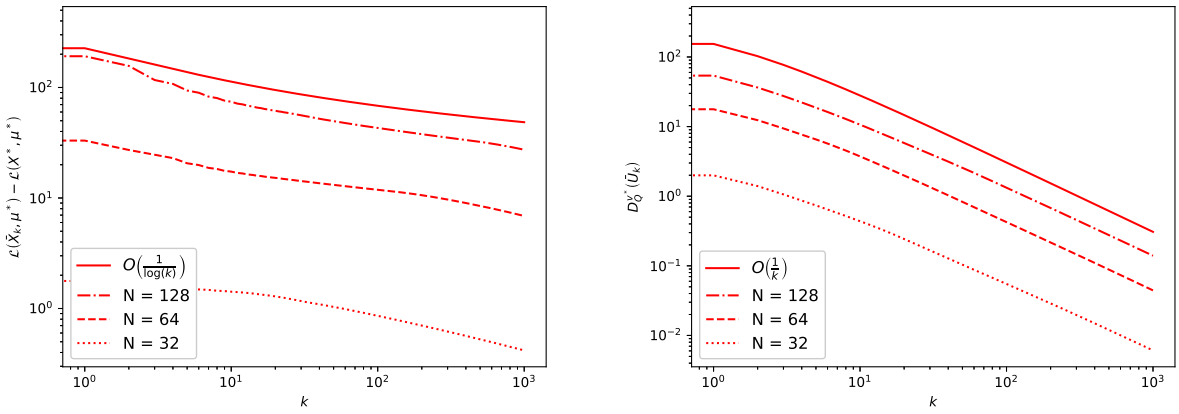

Figure 2: Convergence profiles for CGALP (left) and GFB (right) for $N = 32$, $N = 64$, and $N = 128$.

It can be observed that our theoretically predicted rate (which is $O\left(1/\log(k+2)\right)$ for CGALP according to Theorem 4.2 and Example 4.4) is in close agreement with the observed one. On the other hand, as is very

well-known, employing a proximal step for the nuclear ball constraint will necessitate to compute an SVD which is much more time consuming than computing the linear minimization oracle for large N. For this reason, even though the rates of convergence guaranteed for CGALP are slower than for GFB, one can expect CGALP to be a more time computationally efficient algorithm for large N.

Acknowledgements

ASF was supported by the ERC Consolidated grant NORIA. JF was partly supported by Institut Universitaire de France. CM was supported by Project MONOMADS funded by Conseil Régional de Normandie. We would like to warmly thank Gabriel Peyré for his support and very fruitful and inspiring discussions.

References

[1] L. Ambrosio, N. Gigli, and G. Savare. *Gradient Flows*. Lectures in Mathematics. ETH Zürich. Birkhäuser Basel, 2008.

[2] H. Attouch. *Variational convergence for functions and operators*. Applicable mathematics series. Pitman Advanced Publishing Program, 1984.

[3] F. Bach. Duality between subgradient and conditional gradient methods. *SIAM Journal on Optimization*, 25(1):115–129, 2015.

[4] H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, 2011.

[5] A. Beck, E. Pauwels, and S. Sabach. The cyclic block conditional gradient method for convex optimization problems. *SIAM Journal on Optimization*, 25, 02 2015.

[6] Amir Beck and Marc Teboulle. A conditional gradient method with linear rate of convergence for solving convex linear systems. *Mathematical Methods of Operations Research*, 59(2):235–247, Jun 2004.

[7] K. Bredies and D. Lorenz. Iterated hard shrinkage for minimization problems with sparsity constraints. *SIAM Journal on Scientific Computing*, 30(2):657–683, 2008.

[8] K. Bredies, D. A. Lorenz, and P. Maass. A generalized conditional gradient method and its connection to an iterative shrinkage method. *Computational Optimization and Applications*, 42(2):173–193, Mar 2009.

[9] H. Brezis and A. Pazy. Convergence and approximation of semigroups of nonlinear operators in banach spaces. *J. Functional Analysis*, 9:63–74, 1972.

[10] P. Catala, V. Duval, and G. Peyré. A low-rank approach to off-the-grid sparse deconvolution. In *Journal of Physics: Conference Series*, volume 904, page 012015. IOP Publishing, 2017.

[11] P.L. Combettes. Quasi-fejérian analysis of some optimization algorithms. In *Studies in Computational Mathematics*, volume 8, pages 115–152. Elsevier, 2001.

[12] V. F. Dem'yanov and A. M. Rubinov. The minimization of a smooth convex functional on a convex set. *SIAM J. Control*, 5(2):280–294, 1967.

[13] J.C. Dunn and S. Harshbarger. Conditional gradient algorithms with open loop step size rules. *Journal of Mathematical Analysis and Applications*, 62(2):432 – 444, 1978.

[14] L.C. Evans. *Partial differential equations*. American Mathematical Society, Providence, R.I., 2010.

[15] M. Franke and P. Wolfe. An algorithm for quadratic programming. *Naval research logitistics quarterly*, 3(1-2):95–110, 1956.

[16] F. Pedregosa G. Gidel and S. Lacoste-Julien. Frank-wolfe splitting via augmented lagrangian method. *10th NIPS Workshop on Optimization for Machine Learning*, 2018. (arXiv:1804.03176).

[17] Z. Harchaoui, A. Juditsky, and A. Nemirovski. Conditional gradient algorithms for norm-regularized smooth convex optimization. *Math. Program.*, 152(1-2):75–112, August 2015.

[18] J. Bolte H.H. Bauschke and M. Teboulle. A descent lemma beyond lipschitz gradient continuity: first-order methods revisited and applications. *Math. Oper. Res.*, 42(2):330–348, 2017.

[19] C. Imbert. Convex analysis techniques for hopf-lax formulae in hamilton-jacobi equations. *Journal of Nonlinear and Convex Analysis*, 2(3):333–343, 2001.

[20] M. Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In Sanjoy Dasgupta and David McAllester, editors, *ICML*, volume 28, pages 427–435, Atlanta, Georgia, USA, 17–19 Jun 2013.

[21] M. Jaggi and M. Sulovsk. A simple algorithm for nuclear norm regularized problems. In *ICML*, pages 471–478, 2010.

[22] S. Lacoste-Julien and M. Jaggi. On the global linear convergence of frank-wolfe optimization variants. In *NIPS*, pages 496–504, 2015.

[23] M. Laghdir and M. Volle. A general formula for the horizon function of a convex composite function. *Archiv der Mathematik*, 73(4):291–302, Oct 1999.

[24] E.S. Levitin and B.T. Polyak. Constrained minimization methods. *USSR Computational Mathematics and Mathematical Physics*, 6(5):1 – 50, 1966.

[25] C. Molinari, J. Liang, and J. Fadili. Convergence rates of Forward–Douglas–Rachford splitting method. *ArXiv e-prints*, January 2018.

[26] H. Narasimhan. Learning with complex loss functions and constraints. In Amos Storkey and Fernando Perez-Cruz, editors, *AISTATS*, volume 84, pages 1646–1654, 09–11 Apr 2018.

[27] Yu Nesterov. Universal gradient methods for convex optimization problems. *Mathematical Programming*, 152(1):381–404, Aug 2015.

[28] Yu. Nesterov. Complexity bounds for primal-dual methods minimizing the model of objective function. *Mathematical Programming*, 171(1):311–330, Sep 2018.

[29] Z. Opial. Weak convergence of the sequence of successive approximations for nonexpansive mappings. *Bulletin of the American Mathematical Society*, 73(4):591–597, 1967.

[30] J. Peypouquet. *Convex optimization in normed spaces: theory, methods and examples*. Springer, 2015.

[31] E. Polak. An historical survey of computational methods in optimal control. *SIAM Review*, 15(2):553–584, 1973.

[32] H. Raguet, M. J. Fadili, and G. Peyré. Generalized forward-backward splitting. *SIAM Journal on Imaging Sciences*, 6(3):1199–1226, 2013.

[33] R. T. Rockafellar. *Convex analysis*, volume 28. Princeton university press, 1997.

[34] R. T. Rockafellar and R. Wets. *Variational analysis*, volume 317. Springer Verlag, 1998.

[35] S. Vaiter, M. Golbabaee, J. Fadili, and G. Peyré. Model selection with low complexity priors. *Information and Inference: A Journal of the IMA (IMAIAI)*, 4(3):230–287, 2015.

[36] A. N. Iusem Ya. I. Alber and M. V. Solodov. On the projected subgradient method for nonsmooth convex optimization in a hilbert space. *Mathematical Programming*, 81(1):23–35, 1998.

[37] A. Yurtsever, O. Fercoq, F. Locatello, and V. Cevher. A conditional gradient framework for composite convex minimization with applications to semidefinite programming. *ICML*, 80:5713–5722, 2018.

[38] A. Yurtsever, M. Udell, J. A Tropp, and V. Cevher. Sketchy decisions: Convex low-rank matrix optimization with optimal storage. *arXiv preprint arXiv:1702.06838*, 2017.

[39] X. Zhang, D. Schuurmans, and Y.L. Yu. Accelerated training for matrix-norm regularization: A boosting approach. In *NIPS*, pages 2906–2914, 2012.