Generalized Conditional Gradient with Augmented Lagrangian for Composite Optimization

Antonio Silveti-Falls
Laboratoire GREYC
ENSICAEN
Caen, Normandie 14000
Email: Tonys.falls@gmail.com

Cesare Molinari
Laboratoire GREYC
ENSICAEN
Caen, Normandie 14000
Email: Cecio.molinari@gmail.com

Jalal Fadili
Laboratoire GREYC
ENSICAEN
Caen, Normandie 14000
Email: Jalal.fadili@ensicaen.fr

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 closed, convex, and proper functions over a bounded subset of $\mathcal{H}_p$. The minimization is subject to an affine constraint, which we address by the augmented Lagrangian approach, that allows in particular to deal with composite problems of sum of three or more functions by the usual product space technique. We allow for possibly nonsmooth functions which are simple, i.e., the associated proximal mapping is easily computable. 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, weak convergence of the dual variable to a solution of the dual problem, 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. Experimental results in signal processing are finally reported.

I. Introduction

A. 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 from the context), $A : \mathcal{H}_p \to \mathcal{H}_d$ and $T : \mathcal{H}_p \to \mathcal{H}_v$ are bounded linear operators, $b \in \operatorname{range}(A)$ and f, g, h are proper, convex, and lower semi-continuous functions with $\mathcal{C} \stackrel{\text{def}}{=} \operatorname{dom}(h)$ being a bounded subset of $\mathcal{H}_p$ (or equivalently, since $\mathcal{C}$ is also closed and convex and using [2, Lemma 3.29 and Theorem 3.32], that $\mathcal{C}$ is weakly compact). 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 (I.2)). On the other hand, f is assumed to be differentiable and satisfies a condition that generalizes Lipschitz-continuity of the gradient (see Definition II.1).

Problem ($\mathscr{P}$) can be seen as a generalization of the classical Frank-Wolfe problem in [10] of minimizing a Lipschitz-smooth function f on a convex closed bounded subset $\mathcal{C} \subset \mathcal{H}_p$,

$$\min_{x\in\mathcal{H}_p} \{f(x) : \ x \in \mathcal{C}\} \tag{I.1}$$

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 (I.1) from ($\mathscr{P}$).

B. 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 (I.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 subsequential 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 [21] and [9]. 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 (I.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

$$\min_{s\in\mathcal{H}} h(s) + \langle x, s\rangle \tag{I.2}$$

can be easily solved. Observe that (I.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 [4], [5], [3], [1]. 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 III. These generalizations allow one to apply the algorithm to a plethora of problems arising in signal processing with structure, e.g. sparsity, low-rank, etc.

C. Relation to prior work

In the 1950's Frank and Wolfe developed the so-called Frank-Wolfe algorithm in [10], also commonly referred to as the conditional gradient algorithm [16], [8], [9], for solving problems of the form (I.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., [13], [15], [12], [27], [18], [7]). In the past, composite, constrained problems like ($\mathscr{P}$) have been approached using proximal splitting methods, e.g. generalized forward-backward as developed in [22] or forward-douglas-rachford [17]. Such approaches require one to compute the proximal mapping associated to the function h. 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 ([14], [26]). Unfortunately, the regularity assumptions required by classical Frank-Wolfe style algorithms are too restrictive to apply to general problems like ($\mathscr{P}$).

The recent work of [25], 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,

$$\min_{x\in\mathcal{C}} \{f(x) + g(Tx)\}. \tag{I.3}$$

Like our algorithm, that of [25] is able to handle problems involving both smooth and nonsmooth terms, intersection of multiple constraint sets and affine constraints, however the 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. It is possible to treat the affine constraint as a nonsmooth term g in CGALP and forego the augmented Lagrangian, such that CGALP fully encompasses the conditional gradient framework.

Another recent and parallel work to ours is that of [11], 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,

$$\min_{x\in\mathcal{C}} \{f(x) : \ Ax = 0\}. \tag{I.4}$$

The problem they consider is a particular case of ($\mathscr{P}$), and the algorithm they develop is such that CGALP encompasses this algorithm as well.

II. Algorithm and assumptions

A. Algorithm

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

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

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: $x_0 \in \mathcal{C} = \operatorname{dom}(h)$; $\mu_0 \in \operatorname{range}(A)$; $(\gamma_k)_{k\in\mathbb{N}}$, $(\beta_k)_{k\in\mathbb{N}}$, $(\theta_k)_{k\in\mathbb{N}}, (\rho_k)_{k\in\mathbb{N}} \in \ell_+$.

$k = 0$

repeat

$y_k = \operatorname{prox}_{\beta_k g}(Tx_k)$

$z_k = \nabla f(x_k) + T^*\left(Tx_k - y_k\right)/\beta_k + A^*\mu_k + \rho_k A^*(Ax_k - b)$

$s_k \in \operatorname{Argmin}_{s\in\mathcal{H}_p}\left\{h(s) + \langle z_k, s\rangle\right\}$

$x_{k+1} = x_k - \gamma_k(x_k - s_k)$

$\mu_{k+1} = \mu_k + \theta_k(Ax_{k+1} - b)$

$k \leftarrow k+1$

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

$$\begin{aligned}\nabla_x\mathcal{E}_k(x,\mu_k) &= \nabla f(x) + T^*[\nabla g^{\beta_k}](Tx) + A^*\mu_k + \rho_k A^*(Ax-b)\\ &= \nabla f(x) + \frac{1}{\beta_k}T^*\left(Tx - \operatorname{prox}_{\beta_k g}(Tx)\right) + A^*\mu_k\\ &\quad + \rho_k A^*(Ax-b).\end{aligned} \tag{II.2}$$

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

$$s_k \in \operatorname*{Argmin}_{s\in\mathcal{H}_p}\left\{h(s) + \langle \nabla_x\mathcal{E}_k(x_k,\mu_k), s\rangle\right\}. \tag{II.3}$$

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.

B. Assumptions

1) Assumptions on the functions: Let $\mathcal{L}$ denote the classical Lagrangian, i.e. $\mathcal{L}(x,\mu) = f(x) + g(Tx) + h(x) + \langle \mu, Ax-b\rangle$. 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 $(x,\mu)\in\mathcal{H}_p\times\mathcal{H}_d$,

$$\mathcal{L}(x^\star,\mu) \le \mathcal{L}(x^\star,\mu^\star) \le \mathcal{L}(x,\mu^\star). \tag{II.4}$$

It is well-known from standard Lagrange duality, see e.g. [2, Proposition 19.19] or [20, 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,

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

As was mentioned in Section I, it is not necessary to assume Lipschitz-continuity of the gradient ∇f as in traditional conditional gradient methods. Instead, we have the following generalization.

Definition II.1. ((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

$$K_{(G,\zeta,\mathcal{C})} \overset{\text{def}}{=} \sup_{\substack{x,s\in\mathcal{C};\ \gamma\in]0,1]\\ z=x+\gamma(s-x)}} \frac{D_G(z,x)}{\zeta(\gamma)} \quad < \quad +\infty. \tag{II.5}$$

where D_G is the Bregman divergence associated to G.

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

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 $\Gamma_0(\mathcal{H}_p)$ (the space of proper, convex, lower-semicontinuous functions).
(A.2) The pair $(f,\mathcal{C})$ is (F,ζ)-smooth (see Definition II.1), where we recall $\mathcal{C}\overset{\text{def}}{=}\operatorname{dom}(h)$.
(A.3) $\mathcal{C}$ is bounded (and thus contained in a ball of radius $R>0$).
(A.4) $T\mathcal{C}\subset\operatorname{dom}(\partial g)$ and $\sup_{x\in\mathcal{C}}\left\|[\partial g(Tx)]^0\right\|<\infty$ where $[\cdot]^0$ is the minimal norm selection.
(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) $\operatorname{range}(A)$ 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

$$\begin{cases} A^{-1}(b)\cap\operatorname{relint}(\operatorname{dom}(g\circ T))\cap\operatorname{relint}(\mathcal{C})\neq\emptyset \\ \text{and} \\ \operatorname{range}(A^*)\cap\operatorname{par}(\operatorname{dom}(g\circ T)\cap\mathcal{C})^\perp=\{0\}, \end{cases} \tag{II.6}$$

where par denotes the parallel subspace.

At this stage, a few remarks are in order.

Remark II.2.
(i) By Assumption (A.1), $\mathcal{C}$ is also closed and convex. This together with Assumption (A.3) entail, upon using [2, 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)

$$\sup_k\left\|[\partial g(Tx_k)]^0\right\|\leq M. \tag{II.7}$$

where M is a positive constant.
(iii) Assumption (A.5) will only be needed in the proof of convergence to optimality (Theorem II.7). It is not needed to show asymptotic feasibility (Theorem II.6).
(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 [2, Proposition 15.24 and Fact 15.25], that ensure the existence of a saddle-point for the Lagrangian $\mathcal{L}$ (see [2, Theorem 19.1 and Proposition 9.19(v)]).
(v) Observe that under the inclusion assumption of Lemma II.3, (A.8)(a) is equivalent to $A^{-1}(b)\cap\operatorname{int}(\mathcal{C})\neq\emptyset$.

The uniform boundedness of the minimal norm selection on $\mathcal{C}$, as required in Assumption (A.4), is important 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 II.3. *Let $\mathcal{C}$ be a nonempty bounded subset of $\mathcal{H}_p$, $g\in\Gamma_0(\mathcal{H}_v)$ and $T:\mathcal{H}_p\to\mathcal{H}_v$ is a bounded linear operator. Suppose that $T\mathcal{C}\subset\operatorname{int}(\operatorname{dom}(g))$. Then the assumption (A.4) holds.*

Proof: Since $g\in\Gamma_0(\mathcal{H}_v)$, it follows from [2, Proposition 16.21] that

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

Moreover, by [2, 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. ∎

2) Assumptions on the parameters: We also use the following assumptions on the parameters of Algorithm 1 (recall the function ζ in Definition II.1):

(P.1) $(\gamma_k)_{k\in\mathbb{N}}\subset]0,1]$ and the sequences $(\zeta(\gamma_k))_{k\in\mathbb{N}}$, $\left(\gamma_k^2/\beta_k\right)_{k\in\mathbb{N}}$ 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 $0<\underline{\rho}\leq\inf_k\rho_k\leq\sup_k\rho_k\leq\overline{\rho}<+\infty$.
(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 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 c in (P.6).

Remark II.4.
(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).

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

Example II.5. Take, 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 $\overline{M} = 2^{1-b}$.

C. Main results

Theorem II.6 (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, i.e., the sequence $(x_k)_{k\in\mathbb{N}}$ is asymptotically feasible for ($\mathscr{P}$) in the strong topology.

(ii) Pointwise rate:

$$\inf_{0\leq i\leq k} \|Ax_i - b\| = O\left(\frac{1}{\sqrt{\Gamma_k}}\right). \tag{II.8}$$

Furthermore, $\exists$ a subsequence $\left(x_{k_j}\right)_{j\in\mathbb{N}}$ such that

$$\|Ax_{k_j} - b\| \leq \frac{1}{\sqrt{\Gamma_{k_j}}},$$

where $\Gamma_k \overset{\text{def}}{=} \sum_{i=0}^k \gamma_i$.

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

$$\|A\bar{x}_k - b\| = O\left(\frac{1}{\sqrt{\Gamma_k}}\right). \tag{II.9}$$

Proof: In lieu of the actual proof, which can be found in the full paper [23], we present a sketch of the proof for the sake of brevity. We define two gap-like quantities, one for the primal and one for the dual,

$$\Delta_k^p = \mathcal{L}_k(x_{k+1}, \mu_k) - \min_x \mathcal{L}_k(x, \mu_k),$$
$$\Delta_k^d = \mathcal{L}(x^\star, \mu^\star) - \min_x \mathcal{L}_k(x, \mu_k),$$

and we estimate the quantity $\Delta_{k+1}^p + \Delta_{k+1}^d - \Delta_k^p - \Delta_k^d$ using standard convex analysis arguments and deduce the convergence and pointwise subsequential rate of convergence of $\|Ax_k - b\|^2$ from the resulting inequalities. The ergodic rate follows from the pointwise rates and Jensen's inequality. ∎

Theorem II.7 (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 II.6, the following holds*

(i) Convergence of the Lagrangian:

$$\lim_{k\to\infty} \mathcal{L}(x_k, \mu^\star) = \mathcal{L}(x^\star, \mu^\star). \tag{II.10}$$

(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 $\mathcal{L}$.

(iii) Pointwise rate:

$$\inf_{0\leq i\leq k} \mathcal{L}(x_{i+1}, \mu^\star) - \mathcal{L}(x^\star, \mu^\star) = O\left(\frac{1}{\Gamma_k}\right)$$

Furthermore, $\exists$ a subsequence $\left(x_{k_j}\right)_{j\in\mathbb{N}}$ such that

$$\mathcal{L}(x_{k_j+1}, \mu^\star) - \mathcal{L}(x^\star, \mu^\star) \leq \frac{1}{\Gamma_{k_j}}.$$

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

$$\mathcal{L}(\bar{x}_k, \mu^\star) - \mathcal{L}(x^\star, \mu^\star) = O\left(\frac{1}{\Gamma_k}\right). \tag{II.11}$$

Proof: As in the previous theorem, we present only a sketch of the proof here; the full proof can be found in [23]. We first show that the dual variable μ_k is bounded by a coercivity argument which strongly depends on (A.8). Let $w_k = \mathcal{L}(x_{k+1}, \mu^\star) - \mathcal{L}(x^\star, \mu^\star) + \frac{\rho_k}{2}\|Ax_k - b\|^2$. Then, with the bounded μ_k result we can have an inequality of the form,

$$r_{k+1} - r_k + \gamma_k w_k \leq \varepsilon_k,$$

with $\varepsilon_k \in \ell_+^1$ and r_k uniformly bounded from below. We are also able to show that $\exists \alpha > 0$ such that

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

The fact that $\gamma_k \notin \ell^1$ combined with the above inequalities directly gives $\lim_{k\to\infty} w_k = 0$ as well as the susequential rate of convergence as a consequence of the Abel-Dini theorem on divergent series (see [24]); the ergodic rate following by Jensen's inequality. Finally, we use a technical argument involving Mosco convergence of functionals (see [6]) and Opial's lemma (see [19]) to show that the dual variable μ_k weakly converges to a solution $\mu^\star$ of the dual problem. ∎

It is important to note that Theorem II.7 actually shows that

$$\lim_{k\to\infty}\left[\mathcal{L}(x_{k+1}, \mu^\star) - \mathcal{L}(x^\star, \mu^\star) + \frac{\rho_k}{2}\|Ax_{k+1} - b\|^2\right] = 0,$$

and subsequentially,

$$\mathcal{L}(x_{k_j+1}, \mu^\star) - \mathcal{L}(x^\star, \mu^\star) + \frac{\rho_{k_j}}{2}\|Ax_{k_j+1} - b\|^2 \leq \frac{1}{\Gamma_{k_j}}. \tag{II.12}$$

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

The following corollary is immediate.

Corollary II.8. *Under the assumptions of* Theorem II.7, *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 II.7(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 [2, Lemma 2.38]. ∎

Example II.9. Suppose that the sequences of parameters are chosen according to Example II.5. 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 $\exists k' \geq 0$ (depending on a and b), such that σ is decreasing for $t \geq k'$. Thus, $\forall k \geq k'$, we have

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

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

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

III. Numerical Experiments

In this section we present some numerical experiments exemplifying the applicability of Algorithm 1 to some composite problems in signal processing. First, a simple problem to demonstrate the effect of the parameters on convergence. After, an inverse problem which demonstrates the flexibility of CGALP by employing the linear oracle for a constraint which would otherwise be computationally intense, e.g. when using proximal methods.

A. Projection problem

First, we consider a simple projection problem,

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

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 (III.1) is nothing but Problem ($\mathscr{P}$) with $f(x) = \frac{1}{2}\|x - y\|_2^2$, $g = 0$, and $\mathcal{C} = \mathbb{B}_1^1$.

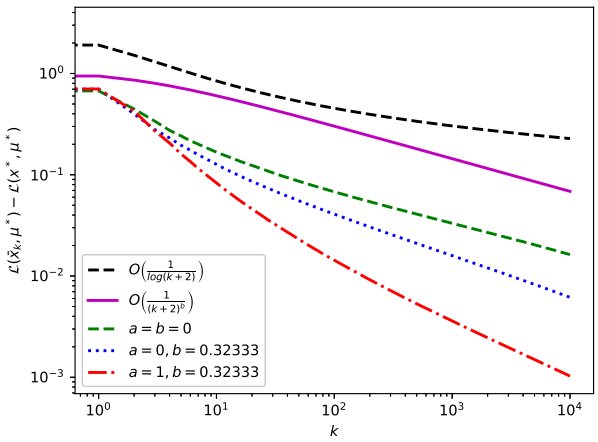

Fig. 1: Ergodic convergence profiles for CGALP applied to the simple projection problem.

The assumptions mentioned previously, i.e. (A.1)-(A.8), all hold in this setting as f is a closed, convex, and proper function, ∇f is Lipschitz-continuous on $\mathcal{C}$, 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 II.5 with $(a, b) \in \{(0, 0), (0, 1/3 - 0.01), (1, 1/3 - 0.01)\}$ 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 II.7 and Example II.9). 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).

B. Matrix completion problem

We now consider the following more complicated matrix completion problem,

$$\min_{X \in \mathbb{R}^{N \times N}} \left\{ \|\Omega X - y\|_1 : \ \|X\|_* \leq \delta_1, \|X\|_1 \leq \delta_2 \right\}, \tag{III.2}$$

where δ_1 and δ_2 are positive constants, $\Omega : \mathbb{R}^{N \times N} \to \mathcal{H}_v$ is a masking operator, $y \in \mathcal{H}_v$ 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, i.e. 80% of entries were kept. 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 with Ω. The radii of the contraints in (III.2) are chosen according to the nuclear norm and ℓ^1 norm of X_0, $\delta_1 = \frac{\|X_0\|_*}{2}$ and $\delta_2 = \frac{\|X_0\|_1}{2}$.

1) CGALP : Problem (III.2) can be posed in a product space in the following way. Denote $\boldsymbol{X} \overset{\text{def}}{=} \begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix} \in \mathbb{R}^{(N \times N)^2}$, $f = 0$, $g(\boldsymbol{\Omega}\boldsymbol{X}) = \frac{1}{2} \sum_{i=1}^{2} \left\| \Omega X^{(i)} - y \right\|_1$, $\mathcal{C} = \mathbb{B}_*^{\delta_1} \times \mathbb{B}_1^{\delta_2}$ 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. Then problem (III.2) is equivalent to

$$\min_{\boldsymbol{X} \in \mathcal{C} \subset \mathbb{R}^{(N \times N)^2}} \left\{ g(\boldsymbol{\Omega}\boldsymbol{X}) : \ \Pi_{\mathcal{V}^\perp} \boldsymbol{X} = 0 \right\}, \tag{III.3}$$

where $\Pi_{\mathcal{V}^\perp}$ is the orthogonal projection onto $\mathcal{V}^\perp$, the orthogonal complement of $\mathcal{V} \overset{\text{def}}{=} \left\{ \boldsymbol{X} \in \mathbb{R}^{(N \times N)^2} : X^{(1)} = X^{(2)} \right\}$. It is trivial to show that our assumptions (A.1)-(A.8) hold. Indeed, g is closed, convex, and proper and thus (A.1) and (A.2) are verified. The set $\mathcal{C} = \mathbb{B}_*^{\delta_1} \times \mathbb{B}_1^{\delta_2}$ is a non-mepty convex compact set. We also have $\boldsymbol{\Omega}\mathcal{C} \subset \operatorname{dom}(\partial g) = \mathcal{H}_v \times \mathcal{H}_v$, and for any $\boldsymbol{z} \in \mathbb{R}^l \times \mathbb{R}^l$, $\partial g(\boldsymbol{z}) \subset \mathbb{B}_\infty^{1/2} \times \mathbb{B}_\infty^{1/2}$ and thus (A.4) is verified. We also have, since $\operatorname{dom}(g \circ \boldsymbol{\Omega}) = \mathbb{R}^{(N \times N)^2}$,

$$\mathbf{0} \in \mathcal{V} \cap \operatorname{int}\left(\operatorname{dom}(g \circ \boldsymbol{\Omega})\right) \cap \operatorname{int}(\mathcal{C}) = \mathcal{V} \cap \operatorname{int}(\mathbb{B}_*^{\delta_1}) \times \operatorname{int}(\mathbb{B}_1^{\delta_2}), \tag{III.4}$$

which shows that (A.8) is verified. The latter is nothing but the condition in [2, Fact 15.25(i)] which, when combined with (A.8), ensures (A.6).

We use Algorithm 1 by choosing the sequence of parameters $\gamma_k = \frac{1}{k+1}$, $\beta_k = \frac{1}{\sqrt{k+1}}$, and $\rho = 15$, which verify all our assumptions (P.1)-(P.7) in view of Example II.5.

2) GFB: We will use a similar product space to apply GFB. Denote $\boldsymbol{W} \overset{\text{def}}{=} \begin{pmatrix} W^{(1)} \\ W^{(2)} \\ W^{(3)} \end{pmatrix} \in \mathbb{R}^{(N \times N)^3}$, $Q(\boldsymbol{W}) = \left\| \Omega W^{(1)} - y \right\|_1 + \iota_{\mathbb{B}_{\|\cdot\|_*}^{\delta_1}}\left(W^{(2)}\right) + \iota_{\mathbb{B}_{\|\cdot\|_1}^{\delta_2}}\left(W^{(3)}\right)$. Then we reformulate problem (III.2) as

$$\min_{\boldsymbol{W} \in \boldsymbol{\mathcal{H}}_p} \left\{ Q(\boldsymbol{W}) : \ \boldsymbol{W} \in \mathcal{V} \right\}, \tag{III.5}$$

which fits the framework to apply the GFB algorithm proposed in [22] (in fact Douglas-Rachford since the smooth part vanishes). We choose the step sizes $\lambda_k = \gamma = 1$.

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

[17] 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.

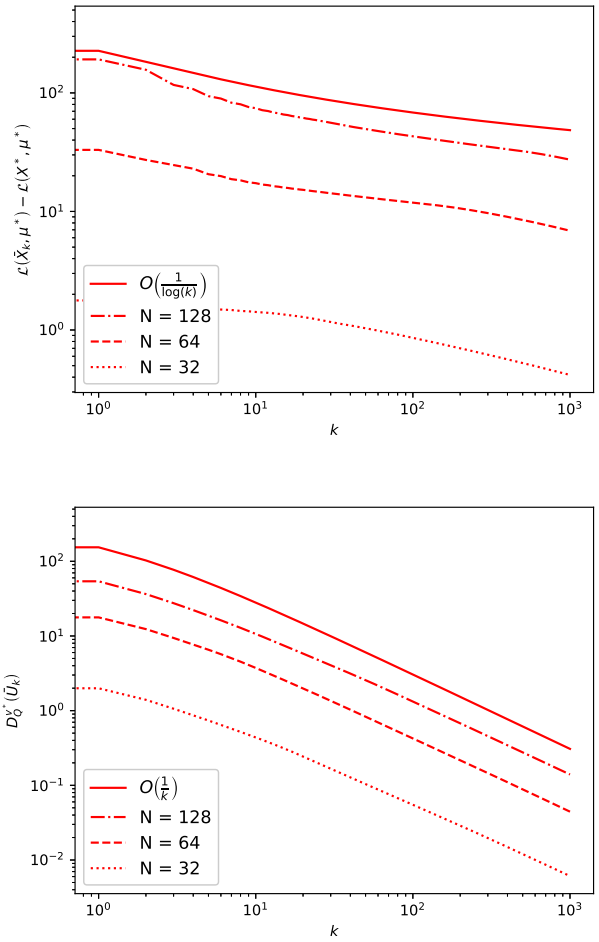

Fig. 2: Ergodic convergence profiles for CGALP (above) and GFB (below) for $N = 32$, $N = 64$, and $N = 128$.

It can be observed that our theoretically predicted rate 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 computing an entire 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 worse than for GFB per iteration, one can expect CGALP to be a more time computationally efficient algorithm for large N as each iteration is cheaper.

Acknowledgment

We would like to warmly thank Gabriel Peyré for his support and very fruitful and inspiring discussions.

References

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

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

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

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

[5] 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.

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

[7] 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.

[8] 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.

[9] 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.

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

[11] 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).

[12] 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.

[13] 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.

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

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

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

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

[18] 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.

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

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

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

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

[23] A. Silveti-Falls, C. Molinari, and J. Fadili. Generalized conditional gradient with augmented lagrangian for composite minimization. *arXiv preprint arXiv:1901.01287*, 2019.

[24] 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.

[25] 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.

[26] 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.

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