Gradient conditionnel généralisé et lagrangien augmenté pour la minimisation composite

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

Generalized Conditional Gradient with Augmented Lagrangian for Composite Minimization

Antonio Silveti-Falls[1], Cesare Molinari[1], Jalal Fadili[1]

[1]Normandie Université, ENSICAEN, UNICAEN, CNRS, GREYC, France.
Tonys.Falls@gmail.com, Cecio.Molinari@gmail.com
Jalal.Fadili@ensicaen.fr.

Résumé – Dans ce travail, nous proposons un schéma d'éclatement en optimisation non lisse, hybridant le gradient conditionnel avec une étape proximale que nous appelons CGALP , pour minimiser la somme de fonctions propres fermées et convexes sur un compact de $\mathbb{R}^n$. La minimisation est de plus sujette à une contrainte affine, que nous prenons en compte par un Lagrangien augmenté, en qui permet en particulier de traiter des problèmes composites à plusieurs fonctions par une technique d'espace produit. Certaines fonctions sont autorisées à être non lisses mais dont l'opérateur proximal est simple à calculer. Notre analyse et garanties de convergence sont assurées pour un large choix de paramètres "en boucle ouverte". Comme résultats principaux, nous montrons la faisabilité asymptotique de la variable primale, la convergence de toute sous-suite vers une solution du problème primal, la convergence de la variable duale à une solution du problème dual, et la convergence du Lagrangien. Des taux de convergence sont aussi fournis. Les implications et illustrations de l'algorithme en traitement des données sont discutées.

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 compact set of $\mathbb{R}^n$. 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, 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.

1 Introduction

1.1 Problem Statement

In this work, we consider the composite optimization problem,

$$\min_{x \in C \subset \mathbb{R}^n} \{f(x) + g(Tx) : \ Ax = b\}, \qquad (\mathscr{P})$$

where $A : \mathbb{R}^n \to \mathbb{R}^m$ and $T : \mathbb{R}^n \to \mathbb{R}^l$ are linear operators, $b \in \operatorname{Im}(A)$, f and g are closed, convex, and proper functions, and $\mathcal{C}$ is a compact subset of $\mathbb{R}^n$. While g is assumed to be prox-friendly it is not necessarily differentiable, however f is assumed to be differentiable with ∇f Lipschitz-continuous.

Problem ($\mathscr{P}$) can be seen as a generalization of the classical Frank-Wolfe (or conditional gradient) problem in [1] of minimizing a differentiable function f with Lipschitz-continuous gradient ∇f on a convex closed bounded subset $\mathcal{C} \subset \mathbb{R}^n$, which is recovered by setting $A \equiv 0$, $b \equiv 0$, and $g \equiv 0$.

1.2 Contribution

The structure of ($\mathscr{P}$) generalizes Frank-Wolfe in two important ways. We consider a possibly nonsmooth term g for which the prox operator is easily computable and problems with an affine constraint which 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 sums of finitely many nonsmooth functions g_i, and, in particular, the intersection of finitely many nonempty compact convex sets $\mathcal{C}_i$ which will be accessed separately; see Section 3.2.1.

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 $\mathcal{C}$ of the form $\min\limits_{s \in \mathcal{C}} \langle v, s\rangle$. This combination of methods provides significant flexibility for the algorithm to be efficiently applied to a wide range of structured problems in both signal processing and machine learning, e.g. problems involving sparsity, low-rank, etc. The linear oracle can be significantly cheaper than proximal alternatives to compute, e.g. projecting on the nuclear ball, and in practice can often be exploited for memory efficient storage, c.f. [2], [3].

Our analysis shows asymptotic feasibility for the affine constraint, convergence of the dual variable to a solution of the dual problem, convergence of the classical Lagrangian 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, i.e. 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, when ($\mathscr{P}$) admits a unique minimizer, convergence of the primal variable to a solution of ($\mathscr{P}$).

2 Algorithm and Theoretical Guarantees

2.1 Algorithm

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

$$\mathcal{J}_k(x,y,\mu) = f(x) + g(y) + \iota_{\mathcal{C}}(x) + \langle \mu, Ax-b\rangle + \frac{\rho}{2}\|Ax-b\|^2 + \frac{1}{2\beta_k}\|y-Tx\|^2, \tag{2.1}$$

where $\iota_{\mathcal{C}}$ is the indicator function for the set $\mathcal{C}$, μ is the dual variable, and ρ 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: $x_0 \in \mathcal{C}$; $\mu_0 \in \mathrm{Im}(A)$; $(\gamma_k)_{k\in\mathbb{N}}$, $(\beta_k)_{k\in\mathbb{N}} \in \ell_+$;
       $\rho > 0$; $k = 0$.
repeat
    $y_k = \mathrm{prox}_{\beta_k g}(Tx_k)$
    $z_k = \nabla f(x_k) + T^*(Tx_k - y_k)/\beta_k + A^*(\mu_k + \rho(Ax_k - b))$
    $s_k \in \mathrm{Argmin}_{s\in\mathcal{C}} \{\langle z_k, s\rangle\}$
    $x_{k+1} = x_k - \gamma_k(x_k - s_k)$
    $\mu_{k+1} = \mu_k + \gamma_k(Ax_{k+1} - b)$
    $k \leftarrow k+1$
until convergence;
Output: $x_{k+1}$.
```

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

$$\{y_k\} = \operatorname*{Argmin}_{y\in\mathbb{R}^l} \mathcal{J}_k(x_k, y, \mu_k). \tag{2.2}$$

Now define the functional $\mathcal{E}_k(x,\mu) \overset{\text{def}}{=} f(x) + g^{\beta_k}(Tx) + \langle \mu, Ax-b\rangle + \frac{\rho}{2}\|Ax-b\|^2$. It is an augmented Lagrangian where we do not consider the non-differentiable function $\iota_{\mathcal{C}}$ and we replace g by its Moreau envelope g^{β_k}. One can immediately verify that 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

$$s_k \in \operatorname*{Argmin}_{s\in\mathcal{C}} \{\langle \nabla_x \mathcal{E}_k(x_k,\mu_k), s\rangle\}. \tag{2.3}$$

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. 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 $k \in \mathbb{N}$. Meanwhile, the affine constraint $Ax_k = b$ might only be satisfied asymptotically. A standard update of the Lagrange multiplier μ_k follows.

2.2 Assumptions

2.2.1 Assumptions on the functions

We define the classical Lagrangian,

$$\mathcal{L}(x,\mu) \overset{\text{def}}{=} f(x) + g(Tx) + \iota_{\mathcal{C}}(x) + \langle \mu, Ax-b\rangle \tag{2.4}$$

Recall that $(x^\star,\mu^\star) \in \mathbb{R}^n \times \mathbb{R}^m$ is a saddle-point for the Lagrangian $\mathcal{L}$ if, for every $(x,\mu) \in \mathbb{R}^n \times \mathbb{R}^m$,

$$\mathcal{L}(x^\star,\mu) \le \mathcal{L}(x^\star,\mu^\star) \le \mathcal{L}(x,\mu^\star). \tag{2.5}$$

It is well-known from standard Lagrange duality, see e.g. [4, Proposition 19.19], 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,

$$\min_{\mu\in\mathbb{R}^m} (f + g\circ T + \iota_{\mathcal{C}})^*(-A^*\mu) + \langle \mu, b\rangle. \tag{$\mathscr{D}$}$$

The following assumptions on the problem will be necessary for the theoretical guarantees:

(A.1) The functions f and $g\circ T$ are closed, convex, and proper.

(A.2) The gradient ∇f is Lipschitz continuous on the set $\mathcal{C}$.

(A.3) The set $\mathcal{C} \subset \mathbb{R}^n$ is compact.

(A.4) $T\mathcal{C} \subset \mathrm{dom}(\partial g)$ and $\sup_{x\in\mathcal{C}}\left(\inf_{g'\in\partial g(Tx)}\|g'\|\right) < \infty$.

(A.5) There exists a saddle-point $(x^\star,\mu^\star) \in \mathbb{R}^n \times \mathbb{R}^m$ for the Lagrangian $\mathcal{L}$.

(A.6) The following holds

$$\begin{cases} A^{-1}(b) \cap \mathrm{relint}(\mathrm{dom}(g\circ T)) \cap \mathrm{relint}(\mathcal{C}) \neq \emptyset \\ \quad \text{and} \\ \mathrm{Im}(A^*) \cap \mathrm{par}(\mathrm{dom}(g\circ T) \cap \mathcal{C})^\perp = \{0\}. \end{cases} \tag{2.6}$$

where int denotes the interior, relint the relative interior, $A^{-1}(b)$ the pre-image of b under A, and par denotes the parallel subspace.

2.2.2 Assumptions on the parameters

We also use the following assumptions on the parameters of Algorithm 1:

(P.1) $\forall k \in \mathbb{N}, \gamma_k \in]0,1]$ and the sequences $(\gamma_k^2)_{k\in\mathbb{N}}$, $\left(\frac{\gamma_k^2}{\beta_k}\right)_{k\in\mathbb{N}}$ and $(\gamma_k\beta_k)_{k\in\mathbb{N}}$ belong to ℓ^1_+ with $(\gamma_k)_{k\in\mathbb{N}} \notin \ell^1$.

(P.2) $(\beta_k)_{k\in\mathbb{N}} \in \ell_+$ is non-increasing and converges to 0.

(P.3) There exist positive constants $\underline{M}$ and $\overline{M}$ such that, $1 \leq \underline{M} \leq \inf_k (\gamma_k/\gamma_{k+1}) \leq \sup_k (\gamma_k/\gamma_{k+1}) \leq \overline{M}$.

(P.4) $\rho > 2\overline{M}$ where $\overline{M}$ is defined above.

There is a large class of sequences that fulfill the requirements (P.1)-(P.4). A typical one is as follows.

Example 2.1. Take, $\forall k \in \mathbb{N}$,

$$\gamma_k = \frac{(\log(k+2))^a}{(k+1)^{1-b}}, \beta_k = \frac{1}{(k+1)^{1-\delta}}, \quad \text{with} \quad a \geq 0,\ 0 \leq 2b < \delta < 1,\ \delta < 1-b, \text{ and } \rho > 2^{2-b}. \tag{2.7}$$

One can then take the crude bounds $\underline{M} = (\log(2)/\log(3))^a$ and $\overline{M} = 2^{1-b}$.

2.3 Main results

Theorem 2.2. *Suppose that assumptions (A.1)-(A.6) and (P.1)-(P.4) hold. Let $(x^\star, \mu^\star)$ be a saddle-point pair for the Lagrangian. Then,*

(i) Asymptotic feasibility:

$$\lim_{k\to\infty} Ax_k = b. \tag{2.8}$$

(ii) Convergence of the Lagrangian:

$$\lim_{k\to\infty} \mathcal{L}(x_k, \mu^\star) = \mathcal{L}(x^\star, \mu^\star). \tag{2.9}$$

(iii) Every 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 to $\bar{\mu}$ a solution of the dual problem ($\mathscr{D}$), i.e., $(\bar{x}, \bar{\mu})$ is a saddle point of $\mathcal{L}$.

(iv) Pointwise rate: there exists a subsequence $\left(x_{k_j}\right)_{j\in\mathbb{N}}$ such that

$$\mathcal{L}\left(x_{k_j}, \mu^\star\right) - \mathcal{L}(x^\star, \mu^\star) + \frac{\rho}{2}\|Ax_{k_j} - b\|^2 \leq \frac{1}{\Gamma_{k_j}}. \tag{2.10}$$

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

$$\mathcal{L}(\bar{x}_k, \mu^\star) - \mathcal{L}(x^\star, \mu^\star) \in O\left(\frac{1}{\Gamma_k}\right). \tag{2.11}$$

Corollary 2.3. *Under the assumptions of* Theorem 2.2, *if the problem ($\mathscr{P}$) admits a unique solution $x^\star$, then the sequence of primal iterates $(x_k)_{k\in\mathbb{N}}$ converges to $x^\star$. Moreover, $\forall k \in \mathbb{N}$,*

$$\mathcal{L}(x_k, \mu^\star) - \mathcal{L}(x^\star, \mu^\star) \leq \frac{1}{\Gamma_k} \quad \textit{and} \quad \|Ax_k - b\| \leq \frac{1}{\sqrt{\Gamma_k}}.$$

For obvious space limitations, the proofs of all these results can be found in the long version [5].

Example 2.4. Suppose that the sequences of parameters are chosen according to Example 2.1. Then one can show that

$$\Gamma_k^{-1} \in \begin{cases} o\left(\frac{1}{(k+2)^b}\right) & a = 1, b > 0, \\ O\left(\frac{1}{(k+2)^b}\right) & a = 0, b > 0, \\ O\left(\frac{1}{\log(k+2)}\right) & a = 0, b = 0. \end{cases} \tag{2.12}$$

3 Numerical Experiments

In this section we present some numerical experiments exemplifying the applicability of Algorithm 1 to some compoosite 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.

3.1 Projection problem

First, we consider a simple projection problem,

$$\min_{x\in\mathbb{R}^2} \left\{ \frac{1}{2}\|x - y\|_2^2 : \ \|x\|_1 \leq 1, Ax = 0 \right\}, \tag{3.1}$$

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

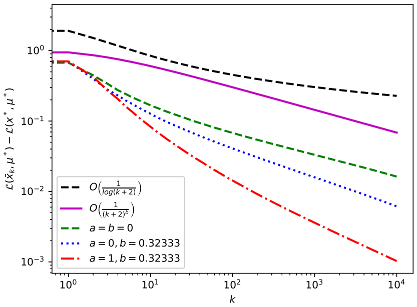

FIG. 1: Ergodic convergence profiles for CGALP applied to the simple projection problem.

The assumptions mentioned previously, i.e. (A.1)-(A.6), 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 $0 \in \ker(A) \cap \mathrm{int}(\mathcal{C})$. Regarding the parameters and the associated assumptions, we choose γ_k according to Example 2.1 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 2.2 and Example 2.4). 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).

3.2 Matrix completion problem

We consider the following more complicated matrix completion problem,

$$\min_{X\in\mathbb{R}^{N\times N}} \left\{ \|\Omega X - y\|_1 : \ \|X\|_* \leq \delta_1, \|X\|_1 \leq \delta_2 \right\}, \tag{3.2}$$

where δ_1 and δ_2 are positive constants, $\Omega : \mathbb{R}^{N\times N} \to \mathbb{R}^l$ is a masking operator, $y \in \mathbb{R}^l$ is a vector of observations, and $\left\|\cdot\right\|_*$ and $\left\|\cdot\right\|_1$ 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 (3.2) are chosen according to the nuclear norm and ℓ^1 norm of X_0, $\delta_1 = \frac{\|X_0\|_*}{2}$ and $\delta_2 = \frac{\|X_0\|_1}{2}$.

3.2.1 CGALP

Problem (3.2) can be posed in a product space in the following way. Denote $\boldsymbol{X} \overset{\text{def}}{=} \begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix} \in \mathbb{R}^{(N\times N)^2}$, $f = 0$, $g\left(\boldsymbol{\Omega X}\right) = \frac{1}{2}\sum_{i=1}^{2}\left\|\Omega X^{(i)} - y\right\|_1$, $\mathcal{C} = \mathbb{B}_*^{\delta_1} \times \mathbb{B}_1^{\delta_2}$ 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 (3.2) is equivalent to

$$\min_{\boldsymbol{X}\in\mathcal{C}\subset\mathbb{R}^{(N\times N)^2}} \left\{g\left(\boldsymbol{\Omega X}\right) : \; \Pi_{\mathcal{V}^\perp}\boldsymbol{X} = 0\right\}, \tag{3.3}$$

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.6) 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 $\boldsymbol{\Omega}\mathcal{C} \subset \operatorname{dom}(\partial g) = \mathbb{R}^l \times \mathbb{R}^l$, 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 $\operatorname{dom}(g \circ \boldsymbol{\Omega}) = \mathbb{R}^{(N\times N)^2}$,

$$\boldsymbol{0} \in \mathcal{V} \cap \operatorname{int}\left(\operatorname{dom}(g \circ \boldsymbol{\Omega})\right) \cap \operatorname{int}\left(\mathcal{C}\right) = \mathcal{V} \cap \operatorname{int}(\mathbb{B}_*^{\delta_1}) \times \operatorname{int}(\mathbb{B}_1^{\delta_2}), \tag{3.4}$$

which shows that (A.6) is verified. The latter is nothing but the condition in [4, Fact 15.25(i)] which, when combined with (A.6), ensures (A.5).

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.4) in view of Example 2.1.

3.2.2 GFB

We will use a similar product space to apply GFB. Denote $\boldsymbol{W} \overset{\text{def}}{=} \begin{pmatrix} W^{(1)} \\ W^{(2)} \\ W^{(3)} \end{pmatrix} \in \mathbb{R}^{(N\times N)^3}, Q\left(\boldsymbol{W}\right) = \left\|\Omega W^{(1)} - y\right\|_1 + \iota_{\mathbb{B}^{\delta_1}_{\|\cdot\|_*}}\left(W^{(2)}\right) + \iota_{\mathbb{B}^{\delta_2}_{\|\cdot\|_1}}\left(W^{(3)}\right)$. Then we reformulate problem (3.2) as

$$\min_{\boldsymbol{W}\in\boldsymbol{\mathcal{H}}_p} \left\{Q\left(\boldsymbol{W}\right) : \; \boldsymbol{W} \in \mathcal{V}\right\}, \tag{3.5}$$

which fits the framework to apply the GFB algorithm proposed in [6] (in fact Douglas-Rachford since the smooth part vanishes). We choose the step sizes $\lambda_k = \gamma = 1$.

3.2.3 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 [7] 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.

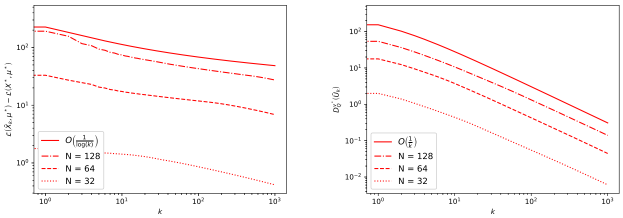

FIG. 2: Convergence profiles for CGALP (left) and GFB (right) for $N = 32$, $N = 64$, and $N = 128$.

It can be observed that our theoretically predicted rate 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 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.

References

[1] M. Franke and P. Wolfe. *An algorithm for quadratic programming*. Naval research logitistics quarterly, 1956.

[2] F. Bach. *Duality Between Subgradient and Conditional Gradient Methods*. SIAM J. Optim., 2015.

[3] A. Yurtsever, M. Udell, J. Tropp, and V. Cevher. *Sketchy Decisions: Convex Low-Rank Matrix Optimizationwith Optimal Storage*. AISTATS, 2017.

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

[5] A. Silveti-Falls, C. Molinari, and J. Fadili. *Generalized Conditional Gradient with Augmented Lagrangian for Composite Minimization*. ArXiv e-prints, 1901.01287, 2019.

[6] H. Raguet, J. Fadili, and G. Peyré. *Generalized forward-backward splitting*. SIAM J. on Imaging Sciences, 2013.

[7] C. Molinari, J. Liang, and J. Fadili. *Convergence rates of Forward–Douglas–Rachford splitting method*. J. Opt. Th. and App., in press, 2018.