Gradient conditionnel généralisé et lagrangien augmenté pour la minimisation composite
Antonio Silveti-Falls, Cesare Molinari, Jalal M. Fadili

To cite this version:
Antonio Silveti-Falls, Cesare Molinari, Jalal M. Fadili. Gradient conditionnel généralisé et lagrangien augmenté pour la minimisation composite. GRETSI, Aug 2019, Lille, France. hal-02268271
1 Introduction

1.1 Problem Statement

In this work, we consider the composite optimization problem,

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

where $A : \mathbb{R}^n \to \mathbb{R}^m$ and $T : \mathbb{R}^n \to \mathbb{R}^l$ are linear operators, $b \in \text{Im} (A)$, $f$ and $g$ are closed, convex, and proper functions, and $C$ 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 $\nabla f$ Lipschitz-continuous.

Problem $(\mathcal{P})$ 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 $\nabla f$ on a convex closed bounded subset $C \subseteq \mathbb{R}^n$, which is recovered by setting $A \equiv 0, b \equiv 0$, and $g \equiv 0$.

1.2 Contribution

The structure of $(\mathcal{P})$ 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 $C_i$ which will be accessed separately; see Section 3.2.1.

We develop and analyze a novel algorithm to solve $(\mathcal{P})$ 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 $(\mathcal{P})$ by using the proximal mapping of $g$ (assumed prox-friendly) and a linear oracle for $C$ of the form $\min_{x \in C} \langle v, x \rangle$.

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 \(\mathcal{P}\) admits a unique minimizer, convergence of the primal variable to a solution of \(\mathcal{P}\).

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

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

where \(\lambda_{\mathcal{C}}\) is the indicator function for the set \(\mathcal{C}\), \(\mu\) is the dual variable, and \(\rho\) and \(\beta_k\) 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:** \(x_0 \in \mathcal{C}; \mu_0 \in \text{Im}(A); (\gamma_k)_{k \in \mathbb{N}}, (\beta_k)_{k \in \mathbb{N}} \in \ell_+\); \(\rho > 0\); \(k = 0\).

**repeat**

1. \(y_k = \text{prox}_{\beta_k g}(Tx_k)\)
2. \(z_k = \nabla f(x_k) + T^* (Tx_k - y_k) / \beta_k + A^* (\mu_k + \rho (Ax_k - b))\)
3. \(s_k \in \text{Argmin} \{\langle z_k, s \rangle\} \cap \mathcal{C}\)
4. \(x_{k+1} = x_k - \gamma_k (x_k - s_k)\)
5. \(\mu_{k+1} = \mu_k + \gamma_k (Ax_{k+1} - b)\)

**until** convergence;

**Output:** \(x_{k+1}\).

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

\[
\{y_k\} = \text{Argmin}_{y \in \mathbb{R}^n} \mathcal{J}_k (x_k, y, \mu_k). \tag{2.2}
\]

Now define the functional \(\mathcal{E}_k(x, \mu) \equiv f(x) + g^\beta_k (Tx) + \langle \mu, Ax - b \rangle + \frac{\rho}{2} \|Ax - b\|^2\). It is an augmented Lagrangian where we do not consider the non-differentiable function \(\lambda_{\mathcal{C}}\) and we replace \(g\) by its Moreau envelope \(g^\beta_k\). One can immediately verify that \(z_k\) is just \(\nabla \mathcal{E}_k (x_k, \mu_k)\) and the first three steps of the algorithm can be condensed in

\[
s_k \in \text{Argmin}_{s \in \mathcal{C}} \{\langle \nabla \mathcal{E}_k (x_k, \mu_k), s \rangle\}. \tag{2.3}
\]

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 \(\mu_k\) follows.

2.2 Assumptions

2.2.1 Assumptions on the functions

We define the classical Lagrangian,

\[
\mathcal{L}(x, \mu) \equiv f(x) + g(Tx) + \lambda_{\mathcal{C}}(x) + \langle \mu, Ax - b \rangle \tag{2.4}
\]

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

\[
\mathcal{L}(x^*, \mu) \leq \mathcal{L}(x, \mu^*) \leq \mathcal{L}(x, \mu). \tag{2.5}
\]

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

\[
\min_{\mu \in \mathbb{R}^m} (f + g \circ T + \lambda_{\mathcal{C}})^* (-A^* \mu + \langle \mu, b \rangle). \tag{D'}
\]

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 \(\nabla f\) is Lipschitz continuous on the set \(\mathcal{C}\).

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

(A.4) \(T \subset \text{dom}(\partial g)\) and \(\sup_{x \in C} \left(\inf_{g \in \partial g(Tx)} \|g\|\right) < \infty\).

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

(A.6) The following holds

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

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 \((\gamma_k)_{k \in \mathbb{N}}, (\beta_k)_{k \in \mathbb{N}}\) belong to \(\ell_+ \) with \((\gamma_k)_{k \in \mathbb{N}} \notin \ell_+\).

(P.2) \((\beta_k)_{k \in \mathbb{N}} \in \ell_+\) is non-increasing and converges to 0.
Theorem 2.2. Suppose that assumptions (A.1)-(A.6) and (P.1)-(P.4) hold. Let \((x^*, \mu^*)\) be a saddle-point pair for the Lagrangian. Then,

(i) Asymptotic feasibility:

\[
\lim_{k \to \infty} A x_k = b.
\]  

(ii) Convergence of the Lagrangian:

\[
\lim_{k \to \infty} \mathcal{L}(x_k, \mu_k) = \mathcal{L}(x^*, \mu^*).
\]  

(iii) Every cluster point \(\bar{x}\) of \(\{x_k\}_{k \geq 1}\) is a solution of the primal problem \((\mathcal{P})\), and \(\{\mu_k\}_{k \geq 1}\) converges to \(\bar{\mu}\) a solution of the dual problem \((\mathcal{D})\), i.e., \((\bar{x}, \bar{\mu})\) is a saddle point of \(\mathcal{L}\).

(iv) Pointwise rate: there exists a subsequence \(\{x_{k_j}\}_{j \geq 1}\) such that

\[
\mathcal{L}(x_{k_j}, \mu_{k_j}) - \mathcal{L}(x^*, \mu^*) + \frac{\rho}{2} \|Ax_{k_j} - b\|^2 \leq \frac{1}{\Gamma_{k_j}}.
\]  

(v) Ergodic rate: let \(\bar{x}_{k} = \frac{1}{k} \sum_{i=0}^{k} x_i / \Gamma_k\), then

\[
\mathcal{L}(\bar{x}_k, \mu_k) - \mathcal{L}(x^*, \mu^*) \in O\left(\frac{1}{\Gamma_k}\right).
\]  

Corollary 2.3. Under the assumptions of Theorem 2.2, if the problem \((\mathcal{P})\) admits a unique solution \(x^*\), then the sequence of primal iterates \(\{x_k\}_{k \geq 1}\) converges to \(x^*\). Moreover, \(\forall k \in \mathbb{N}\),

\[
\mathcal{L}(x_k, \mu_k) - \mathcal{L}(x^*, \mu^*) \leq \frac{1}{\Gamma_k} \text{ and } \|Ax_k - b\| \leq \frac{1}{\sqrt{\Gamma_k}}.
\]  

3 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.

3.1 Projection problem

First, we consider a simple projection problem,

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

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 \(y \not\in B^1_1 \cap \ker(A)\), where \(B^1_1\) is the unit \(\ell^1\) ball centered at the origin. Then Problem (3.1) is nothing but Problem \((\mathcal{P})\) with \(f(x) = \frac{1}{2} \|x - y\|^2\), \(g = 0\), and \(C = B^1_1\).

![Fig. 1: Ergodic convergence profiles for CGALP applied to the simple projection problem.](image)

The assumptions mentioned previously, i.e. (A.1)-(A.6), all hold in this setting as \(f\) is a closed, convex, and proper function, \(\nabla f\) is Lipschitz-continuous on \(C\), \(g\) has full domain, and \(0 \in \ker(A) \cap \text{int}(C)\). Regarding the parameters and the associated assumptions, we choose \(\gamma_k\) according to Example 2.1 with \((a, b) \in \{(0, 0), (0, 0.3 - 0.01), (1, 1/3 - 0.01)\}\) and \(\rho = 2^{a-b} + 1\). 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 \(O\left(\frac{1}{\log(k+2)}\right)\), \(O\left(\frac{1}{(k+2)^{1/2}}\right)\) and \(o\left(\frac{1}{(k+2)^{1/2}}\right)\) for the respective choices of \((a, b)\).

3.2 Matrix completion problem

We consider the following more complicated matrix completion problem,

\[
\min_{X \in \mathbb{R}^{2 \times N}} \{ \|\Omega X - y\|_1 : \|X\|_* \leq \delta_1, \|X\|_1 \leq \delta_2 \},
\]  

where \(\Omega\) is an \(m \times n\) matrix with \(m \ll n\), \(\Omega_{ij} = 1\) if \((i, j)\) is a sampled index, and \(\Omega_{ij} = 0\) otherwise.
where $\delta_1$ and $\delta_2$ 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 $\|\cdot\|_1$ and $\|\cdot\|_2$ are respectively the nuclear and $\ell^1$ norms. The mask operator $\Omega$ 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 $\tilde{y} \in \mathbb{R}^l$ 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 $\Omega$. The radii of the contraints in (3.2.1) are chosen according to the nuclear norm and $\ell^1$ norm of $X_0$, $\delta_1 = \frac{\|X_0\|_2}{2}$ 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 $X = \begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix} \in \mathbb{R}^{(N \times N)^2}$, $f = 0$, $g(\Omega X) = \frac{1}{2} \sum_{i=1}^2 \|\Omega X^{(i)} - y\|_1$, $C = \mathbb{B}^{\delta_1}_2 \times \mathbb{B}^{\delta_2}_2$ where $\mathbb{B}^{\delta_1}_2$ and $\mathbb{B}^{\delta_2}_2$ are the nuclear and $\ell^1$ balls of radii $\delta_1$ and $\delta_2$. Then problem (3.2) is equivalent to

$$\min_{X \in C} \left\{ g(\Omega X) : \Pi_{\mathcal{C}} X = 0 \right\}, \quad (3.3)$$

where $\Pi_{\mathcal{C}}$ is the orthogonal projection onto $\mathcal{C}$, the orthogonal complement of $\mathcal{V} = \left\{ X \in \mathbb{R}^{(N \times N)^2} : X^{(1)} = X^{(2)} \right\}$. 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 $\mathcal{C} = \mathbb{B}^{\delta_1}_2 \times \mathbb{B}^{\delta_2}_2$ is a non-empty convex compact set. We also have $\Omega \mathcal{C} \subset \text{dom}(\partial g) = \mathbb{R}^l \times \mathbb{R}^l$, and for any $z \in \mathbb{R}^l \times \mathbb{R}^l$, $\partial g(z) \subset \mathbb{B}^{1/2}_2 \times \mathbb{B}^{1/2}_2$ and thus (A.4) is verified. We also have, since $\text{dom}(g \circ \Omega) = \mathbb{R}^{(N \times N)^2}$, $\mathcal{V} = \mathcal{C} = \mathcal{V} \cap \text{int}(\mathbb{B}^{\delta_1}_2) \times \text{int}(\mathbb{B}^{\delta_2}_2),$

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 $\gamma_k = \frac{1}{k+1}$, $\beta_k = \sqrt{k+1}$, 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 $W = \begin{pmatrix} W^{(1)} \\ W^{(2)} \\ W^{(3)} \end{pmatrix} \in \mathbb{R}^{(N \times N)^3}$, $Q(W) = \|\Omega W^{(1)} - y\|_1 + \sum_{i=1}^2 \|W^{(i)}\|_1$, then reformulate problem (3.2) as

$$\min_{W \in \mathcal{H}_p} \{ Q(W) : W \in \mathcal{V} \}, \quad (3.5)$$

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 $L(\tilde{X}_k, \mu^*) - L(X^*, \mu^*)$ where $\tilde{X}_k = \sum_{i=0}^k \gamma_i X_i / T_k$. Meanwhile, for GFB, we know from [7] that the Bregman divergence $D_{\Omega}^{\mu^*}(\tilde{U}_k) = Q(\tilde{U}_k) - Q(W^*) - \langle W^*, \tilde{U}_k - W^* \rangle$, with $\tilde{U}_k = \sum_{i=0}^k U_i / (k+1)$ and $W^* = (W^* - Z^*) / \gamma$, converges at the rate $O(1/(k+1))$. To compute the convergence criteria, we first run each algorithm for $10^6$ iterations to approximate the optimal variables $(X^* \mu^*)$ for CGALP, and $(Z^* \mu^*)$ 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 $N = 128$.](image)

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


