Frank-Wolfe Algorithm for the m-EXACT-SPARSE Problem

Farah Cherfaoui[1]*, Valentin Emiya[1], Liva Ralaivola[1] and Sandrine Anthoine[2]

[1] Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France

[2] Aix Marseille Univ, CNRS, Centrale Marseille, I2M, Marseille, France

***Abstract*— In this paper, we study the properties of the Frank-Wolfe algorithm to solve the m-EXACT-SPARSE reconstruction problem. We prove that when the dictionary is quasi-incoherent, at each iteration, the Frank-Wolfe algorithm picks up an atom indexed by the support. We also prove that when the dictionary is quasi-incoherent, there exists an iteration beyond which the algorithm converges exponentially fast.**

1 Introduction

Given a dictionary of a large number of atoms, the sparse signal approximation problem consists of constructing the best linear combination with a small number of atoms to approximate a given signal. Sparse signal reconstruction is a sub-problem of the sparse signal approximation problem. In the latter case, we suppose that the given signal has an exact representation with m or less atoms from this dictionary. We say that the signal is m-sparse. This subset of atoms is indexed by a set called the support. In this paper, we only consider the sparse signal reconstruction problem, which is called the m-EXACT-SPARSE problem.

Several algorithms have been developed to solve or approximate the m-EXACT-SPARSE problem. The Matching Pursuit algorithm (MP) [6] and Orthogonal Matching Pursuit algorithm (OMP) [7] are two fundamental greedy algorithms used for solving this problem. Tropp [8] and Gribonval and Vandergheynst [3] proved that, if the dictionary is quasi-incoherent, then at each iteration the MP and OMP algorithms pick up an atom indexed by the support. They also proved that these two algorithms converge exponentially fast. In fact, Tropp in [8] demonstrates that OMP converges after exactly m iterations, where m is the size of the support. We study in this paper the properties of the Frank-Wolfe algorithm [2] to solve the m-EXACT-SPARSE problem. The Frank-Wolfe algorithm [2] is an iterative optimization algorithm designed for constrained convex optimization. It has been proven to converge exponentially if the objective function is strongly convex [4] and linearly in the other cases [2]. The atom selection steps in Matching Pursuit and Frank-Wolfe are very similar. This inspired for example Jaggi and al. [5] to use the Frank-Wolfe algorithm to prove the convergence of the MP algorithm when no conditions are made on the dictionary.

In this paper, we use the MP algorithm to prove that the Frank-Wolfe algorithm can have the same recovery and convergence properties as MP. We prove that when the dictionary is quasi-incoherent, the Frank-Wolfe algorithm picks up only atoms indexed by the support. Also, we prove that when the dictionary is quasi-incoherent, the Frank-Wolfe algorithm converges exponentially from a certain iteration even though the function we consider is not strongly convex.

*This work was supported by the Agence Nationale de la Recherche under grant JCJC MAD (ANR-14-CE27-0002).

2 The problem and the algorithm

2.1 The m-EXACT-SPARSE problem

For any vector $x \in \mathbb{R}^n$, we denote by $x(i)$ its i^{th} coordinate. The support of x is the set of indices of nonzero coefficients:

$$\text{support}(x) = \{i | x(i) \neq 0\}.$$

Fix a dictionary $\Phi = [\varphi_1, \ldots, \varphi_n] \in \mathbb{R}^{d \times n}$ of n unit-norm vectors. Assume that y is m-sparse, then the m-EXACT-SPARSE problem is to find:

$$\arg\min_{x \in \mathbb{R}^n} \tfrac{1}{2}\|y - \Phi x\|_2^2 \text{ s.t. } \|x\|_0 \leq m$$

where the l_0 pseudo-norm $\|.\|_0$ counts the number of nonzero components in its argument. This problem has been proven to be NP-hard [1] and has been tackled essentially with two kind of approaches. The first one is the local approach, using a greedy algorithm like MP or OMP. The second approach is a global one where one relaxes the problem. A most popular choice is the l_1 relaxation:

$$\arg\min_{x \in \mathbb{R}^n} \tfrac{1}{2}\|y - \Phi x\|_2^2 \text{ s.t. } \|x\|_1 \leq \beta \qquad (1)$$

where $\|.\|_1$ is the l_1 norm.

We present, in the next parts, the Frank-Wolfe algorithm [2] for the m-EXACT-SPARSE problem, and then the recovery properties and convergence rate of this algorithm.

2.2 The Frank-Wolfe algorithm

The Frank-Wolfe algorithm solves the optimization problem

$$\min_{x \in \mathcal{C}} f(x) \text{ s.t. } x \in \mathcal{C}$$

where f is a convex and continuously differentiable function and $\mathcal{C}$ is a compact and convex set. In the original version of the Frank-Wolfe algorithm, each iterate x_{k+1} is defined as a convex combination between x_k and s_k with $s_k = \arg\min_{s \in \mathcal{C}} \langle s, \nabla f(x_k) \rangle$.

In the case of the relaxation of the m-EXACT-SPARSE problem (Equation (1)), $f(x) = \frac{1}{2}\|y - \Phi x\|_2^2$ and $\mathcal{C} = \{x : \|x\|_1 \leq \beta\} = \mathcal{B}_1(\beta)$ is the l_1 ball of radius β. Noting that $\mathcal{B}_1(\beta) = \text{conv}\{\pm\beta e_i | i \in \{1, \ldots, n\}\}$ and that $\nabla f(x) = \Phi^t(\Phi x - y)$, we obtain that s_k can be calculated as in line 4 and 5 of Algorithm 1. Note also that we initialize x_0 by zero (line 1) and that we select the convex combination parameter γ_k as in line 6.

In the analysis of Algorithm 1, we use the residual $r_k = y - \Phi x_k$ whose norm is also the minimized objective function $f(x_k) = \frac{1}{2}\|r_k\|_2^2$.

Algorithm 1: Frank-Wolfe algorithm

Data: signal y, dictionary $\Phi = [\varphi_1, \dots, \varphi_n]$, scalar β.

1 $x_0 = 0$
2 $k = 0$
3 **while** *stopping criterion not verified* **do**
4 $\quad i_k = \arg\max_{i \in \{1,\dots,n\}} |\langle \varphi_i, \Phi x_k - y \rangle|$
5 $\quad s_k = -\operatorname{sign}(\langle \varphi_{i_k}, \Phi x_k - y \rangle) \beta e_{i_k}$
6 $\quad \gamma_k = \arg\min_{\gamma \in [0,1]} \|y - \Phi(x_k + \gamma(s_k - x_k))\|_2^2$
7 $\quad x_{k+1} = x_k + \gamma_k(s_k - x_k)$
8 $\quad k = k + 1$
9 **end**

3 Recovery property and convergence rate

For a dictionary Φ, we denote by $\mu = \max_{j \neq k} |\langle \varphi_j, \varphi_k \rangle|$ the coherence of Φ and by $\mu_1(m) = \max_{|\Lambda|=m} \max_{i \notin \Lambda} \sum_{j \in \Lambda} |\langle \varphi_i, \varphi_j \rangle|$ the Babel function. These two quantities measure how much the elements of the dictionary look alike. More details can be found in [8].

In this section we present our major results. Theorem 1 gives the recovery property for the Frank-Wolfe algorithm. We prove that when the dictionary is quasi-incoherent (i.e. $m < \frac{1}{2}(\mu^{-1} + 1)$), the Frank-Wolfe algorithm reconstructs every m-sparse signal. Theorem 2 shows that when the dictionary is quasi-incoherent, the Frank-Wolfe algorithm converges exponentially. We recall that a sequence $(a_k)_{k=0}^{\infty}$ converges exponentially if: $\forall\, k \in \{1, \dots, +\infty\}$, $a_{k+1} \leq q a_k$ with $0 < q < 1$.

Theorem 1. *Let $\Phi \in \mathbb{R}^{d \times n}$ be a dictionary, μ its coherence, and $y = \Phi x^*$ a m-sparse signal (i.e. $|\operatorname{support}(x^*)| = m$). If $m < \frac{1}{2}(\mu^{-1} + 1)$, then at each iteration, Algorithm 1 picks up a correct atom, i.e. $\forall\, k$, $i_k \in \operatorname{support}(x^*)$.*

Sketch of proof. The proof of this theorem is very similar to the proof of Theorem 3.1 in [8]. □

Theorem 2. *Let $\Phi \in \mathbb{R}^{d \times n}$ be a dictionary, μ its coherence, and $y = \Phi x^*$ a m-sparse signal (i.e. $|\operatorname{support}(x^*)| = m$). If $m < \frac{1}{2}(\mu^{-1} + 1)$ and $\|x^*\|_1 < \beta$, then there exists a K such that for all iteration $k \geq K$ of Algorithm 1, we have:*

$$\|r_{k+1}\|^2 \leq \|r_k\|^2 \left(1 - \frac{\epsilon^2(1 - \mu_1(m-1))}{4\beta^2}\right)$$

where $\epsilon = \frac{1}{2}(\beta - \|x^*\|_1)$.

Sketch of proof. The general idea of the proof can be summarized as follows. The first step will be to prove that if the dictionary is quasi-incoherent, then the step γ_k chosen in line 6 of Algorithm 1 is in $(0, 1)$. A consequence of this is that:

$$\gamma_k = \arg\min_{\gamma \in \mathbb{R}} \|y - \Phi(x_k + \gamma(s_k - x_k))\|_2^2 \tag{2}$$

$$= \frac{\langle r_k, \Phi(s_k - x_k) \rangle}{\|\Phi(s_k - x_k)\|_2^2} \tag{3}$$

We can then write the expression of $\|r_{k+1}\|_2^2$:

$$\|r_{k+1}\|_2^2 = \|y - \Phi x_{k+1}\|_2^2 = \|r_k + \gamma_k \Phi(s_k - x_k)\|_2^2,$$

which yields using Eq. (3):

$$\|r_{k+1}\|_2^2 = \|r_k\|_2^2 - \frac{\langle r_k, \Phi(s_k - x_k) \rangle^2}{\|\Phi(s_k - x_k)\|_2^2}.$$

The second step is to bound $\langle r_k, \Phi(s_k - x_k) \rangle$. Using Theorem 1, we can show that the sequence of $\|x_k - x^*\|$ is bounded by the sequence $f(x_k) - f(x^*)$. Since the sequence $f(x_k) - f(x^*)$ converges to zero, then the sequence of $\|x_k - x^*\|$ also converges to zero. Therefore, there exists an iteration K such that for all $k \geq K$: $x_k \in \mathcal{B}_2(x^*, \epsilon)$ where $\mathcal{B}_2(x^*, \epsilon)$ is l_2 ball centered in x^* and of radius ϵ. As a result, $x_k - \epsilon \frac{\nabla f(x_k)}{\|\nabla f(x_k)\|} \in \mathcal{B}_2(x^*, 2\epsilon)$. Since $\|x^*\|_1 + 2\epsilon < \beta$, we have $\mathcal{B}_2(x^*, 2\epsilon) \subseteq \mathcal{B}_1(\beta)$.

By definition of s_k:

$$\langle s_k, \nabla f(x_k) \rangle \leq \langle x_k - \epsilon \frac{\nabla f(x_k)}{\|\nabla f(x_k)\|}, \nabla f(x_k) \rangle.$$

Noting that $\nabla f(x_k) = -\Phi^t r_k$, one obtains

$$\langle r_k, \Phi(s_k - x_k) \rangle \geq \epsilon \|\Phi^t r_k\|.$$

By Theorem 1, r_k lies in the linear span of atoms indexed by $\operatorname{support}(x^*)$. Since we assume that these atoms are linearly independent, we have

$$\|\Phi^t r_k\| \geq \lambda_{min}^{\Phi_{\operatorname{support}(x^*)}} \|r_k\|_2,$$

where $\Phi_{\operatorname{support}(x^*)}$ is the matrix whose columns are the atoms indexed by $\operatorname{support}(x^*)$ and $\lambda_{min}^{\Phi_{\operatorname{support}(x^*)}}$ its smallest singular. So, By Lemma 2.3 of [8], $\lambda_{min}^{\Phi_{\operatorname{support}(x^*)}} \geq (1 - \mu_1(m-1))$ and we obtain:

$$\langle r_k, \Phi(s_k - x_k) \rangle \geq \epsilon(1 - \mu_1(m-1)) \|r_k\|_2.$$

Finally, we show that $\|\Phi(s_k - x_k)\|_2 \leq 2\beta$ using the fact that $\|\Phi((s_k - x_k))\|_2 \leq \|s_k - x_k\|_1$ since the φ_i are of unit norm. □

Note that Tropp in [8] has already proved that if the dictionary is incoherent, then $\mu_1(m) + \mu_1(m-1) < 1$. As a result, $1 - \mu_1(m-1)$ is in $(0, 1)$. We also have that $\frac{\epsilon^2}{\beta^2} < 1$ because $\epsilon < \beta$. Finally, since d is greater that 1, we have that $\frac{\epsilon^2(1-\mu_1(m-1))}{4\beta^2 d}$ is in $(0, 1)$. We conclude that Theorem 2 gives the exponential convergence rate of the residual norm. As $f(x_k) = \frac{1}{2}\|r_k\|_2^2$, this implies that this theorem also gives the exponential convergence rate of the objective function beyond a certain iteration.

It is possible to guarantee an exponential convergence from the first iteration if β is big enough. Lemma 1 gives a lower bound of β to obtain this result.

Lemma 1. *Let Φ be a dictionary of coherence μ, $y = \Phi x^*$ a m-sparse signal (i.e. $|\operatorname{support}(x^*)| = m$) and $\epsilon \in (0, 1)$. If*

$$\beta > \frac{m\|y\|_2}{\epsilon \lambda_{min}^{\Phi_{\operatorname{support}(x^*)}}} \left(1 + \frac{\lambda_{max}^{\Phi_{\operatorname{support}(x^*)}}}{\lambda_{min}^{\Phi_{\operatorname{support}(x^*)}}}\right)$$

then Algorithm 1 converges exponentially from the first iteration. Here, $\Phi_{\operatorname{support}(x^*)}$ is the matrix whose columns are the atoms indexed by $\operatorname{support}(x^*)$.

We proved in Theorem 2 that when the iterates x_k enter the ball $\mathcal{B}_1(x^*, \epsilon)$, the Frank-Wolfe algorithm converges exponentially. The intuition of this lemma is to grow the value of β compared to $\|x^*\|$ (then ϵ also grows). This implies that the iterates x_k enter the ball $\mathcal{B}_1(x^*, \epsilon)$ earlier and the exponential convergence starts earlier.

References

[1] Geoff Davis, Stephane Mallat, and Marco Avellaneda. Adaptive greedy approximations. *Constructive approximation*, 13(1):57–98, 1997.

[2] Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval Research Logistics (NRL)*, 3(1-2):95–110, 1956.

[3] Rémi Gribonval and Pierre Vandergheynst. On the exponential convergence of matching pursuits in quasi-incoherent dictionaries. *IEEE Transactions on Information Theory*, 52(1):255–261, 2006.

[4] Jacques Guélat and Patrice Marcotte. Some comments on Wolfe's 'away step'. *Mathematical Programming*, 35(1):110–119, 1986.

[5] Francesco Locatello, Rajiv Khanna, Michael Tschannen, and Martin Jaggi. A unified optimization view on generalized matching pursuit and Frank-Wolfe. *arXiv preprint arXiv:1702.06457*, 2017.

[6] Stéphane G Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on signal processing*, 41(12):3397–3415, 1993.

[7] Yagyensh Chandra Pati, Ramin Rezaiifar, and Perinkulam Sambamurthy Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. *Signals, Systems and Computers, 1993. 1993 Conference Record of The Twenty-Seventh Asilomar Conference on*, pages 40–44, 1993.

[8] Joel A Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information theory*, 50(10):2231–2242, 2004.