A Note on the Hardness of Sparse Approximation

A. Çivril[a]

[a] *Melikşah University, Computer Engineering Department, Talas, Kayseri 38280 Turkey*

Abstract

Given a redundant dictionary Φ, represented by an $M \times N$ matrix ($\Phi \in \mathbb{R}^{M \times N}$) and a target signal $y \in \mathbb{R}^M$, the *sparse approximation problem* asks to find an approximate representation of y using a linear combination of at most k atoms. This note presents a hardness result for sparse approximation problem under a measure of quality, which is essentially the *squared multiple correlation* in statistical analysis. We show that unless $P = NP$, all polynomial time algorithms which provide a k-sparse vector x should satisfy

$$\|\Phi_x \Phi_x^+ y\|_2^2 \leq \left(1 - \frac{1}{e}\right) \|\Phi_{x^*} \Phi_{x^*}^+ y\|_2^2,$$

where x^* is the optimal k-sparse solution and Φ_x denotes the column sub-matrix of Φ which consists of the column vectors with indices of non-zero elements in x. We relate this result to the recent algorithmic results of Das and Kempe [1, 2] and conclude that Forward Regression and Orthogonal Matching Pursuit are almost the best one can hope for solving the sparse approximation problem under the squared multiple correlation metric, especially when the dictionary is near orthogonal.

Key words: Sparse approximation, Subset selection, Complexity, Inapproximability

1. Introduction

Given a redundant dictionary Φ, represented by an $M \times N$ matrix ($\Phi \in \mathbb{R}^{M \times N}$) and a target signal $y \in \mathbb{R}^M$, the *sparse approximation problem* asks to find an approximate representation of y using a linear combination of at most k atoms, i.e. column vectors of Φ. This amounts to finding a coefficient vector $x \in \mathbb{R}^N$ for which one usually solves

$$\min_{\|x\|_0 = k} \|y - \Phi x\|_2 \tag{1}$$

We name the problem with this standard objective function (1) as Sparse. Stated in linear algebraic terms, it is essentially about picking a k-dimensional subspace defined by k column vectors of Φ such that the orthogonal projection of y onto that subspace is as close as possible to y. Thus, if Φ_x denotes the column sub-matrix of Φ which consists of the column vectors with indices of non-zero elements in x, Sparse can also be written as

Email address: acivril@meliksah.edu.tr (A. Çivril)

$$\min_{\|x\|_0=k} \|y - \Phi_x \Phi_x^+ y\|_2, \tag{2}$$

where Φ_x^+ denotes the pseudo-inverse of Φ_x. The problem is of combinatorial nature and the optimal solution can be found by checking all $\binom{N}{k}$ subspaces. SPARSE is NP-hard even to approximate within any factor [3, 8]. The intrinsic difficulty of the problem under this objective function prevents one from designing algorithms which can provably *approximate* the optimal solution. In signal processing community, efforts have been towards analyzing algorithms working on restricted set of dictionaries for which one requires the column vectors to be an almost orthogonal set, namely an incoherent dictionary. Orthogonal Matching Pursuit (OMP) is a common algorithm yielding approximation guarantees under such restrictions [7, 9, 12]. Similar restrictions for matrices have also been suggested in compressive sensing under the name Restricted Isometry Property (RIP) [4]. The problem SPARSE, defined in Hilbert and Banach spaces with full generality, has also been studied as *highly nonlinear approximation* in functional approximation theory [10, 11]. In parallel to the finite dimensional case, definitions of coherence appear in [5] with algorithms providing approximate solutions via the so-called Lebesgue-type inequalities.

A closely related problem to sparse approximation is subset selection in linear regression. In this problem, the goal is to estimate a *predictor variable* Z using a small subset of the *observation variables* $X_1, \ldots, X_n$. One wants to sample a set S of at most k variables X_i, and compute a linear predictor $Z' = \sum_{i \in S} \alpha_i X_i$ of Z. This problem is equivalent to sparse approximation under approximation preserving reductions. Note that in sparse approximation too, one tries to select a subset of k column vectors whose linear combination is a good "predictor" of the input vector y. Given this, the goal is to minimize the *mean square prediction error* $E[(Z - Z')^2]$, which is equivalent to minimizing the expression (2) in the sparse approximation problem. Another alternative, which is the one we take in this note, is to maximize the *squared multiple correlation*

$$R^2_{Z,S} = \frac{Var(Z) - E[(Z - Z')^2]}{Var(Z)}.$$

Analogically, $Var(Z)$ represents $\|y\|_2^2$, and $E[(Z - Z')^2]$ stands for $\|y - \Phi x\|_2^2$. Stated in terms of the sparse approximation problem, this value measures how much the target vector "falls" onto the optimal linear combination of the selected vectors, in contrast to the standard objective function that measures how far away the target vector to the chosen subspace is. Motivated by this, we define the problem DIFF-SPARSE in which we try to optimize the *differential objective function*. It is the same as SPARSE except that one solves

$$\max_{\|x\|_0=k} \|\Phi_x \Phi_x^+ y\|_2^2$$

One could omit the squares, but we preferred not to do so as results in this form are readily related to those of SPARSE. Note also that the optimal solutions of SPARSE and DIFF-SPARSE are the same. In order to underline the motivation behind this new definition, one sees that $R^2_{Z,S} = \frac{Var(Z) - E[(Z-Z')^2]}{Var(Z)}$ is analogous to $\|\Phi_x \Phi_x^+ y\|_2^2$. Our main theorem in this paper is:

Theorem 1.1. DIFF-SPARSE *cannot be approximated within* $\left(1-\frac{1}{e}+\epsilon\right)$ *for all* $\epsilon > 0$ *in polynomial time unless* $P = NP$.

This result is a simple extension of Feige's Set-Cover reduction [6]. Yet, it is important in the sense that there are some algorithmic results for DIFF-SPARSE which are comparable to this hardness result. Das and Kempe [1], Gilbert et al. [7] and Tropp [12] obtained algorithms with performance guarantees $1-\Theta(\mu k)$ in the special case where the coherence μ of the dictionary is $O(1/k)$. In [1], the authors also obtain a $\left(1-\frac{1}{e}\right)$ approximation under the special case where there are no "suppressor variables". In a more recent paper, Das and Kempe [2] showed that the algorithm Forward Regression attains the approximation ratio $\left(1-e^{-\lambda_{min}(C,k)}\right)\cdot\Theta\left((\frac{1}{2})^{1/\lambda_{min}(C,k)}\right)$, and the algorithm Orthogonal Matching Pursuit attains the approximation ratio $\left(1-e^{-\lambda_{min}(C,k)^2}\right)\cdot\Theta\left((\frac{1}{2})^{1/\lambda_{min}(C,k)}\right)$, where C is the the matrix of covariances between variables X_i and X_j (the matrix of pair-wise dot products in the sparse approximation problem), and $\lambda_{min}(C,k)$ refers to the smallest eigenvalue of any $k\times k$ sub-matrix of C. Hence, if $\lambda_{min}(C,k)$ is close to 1, which means that the variables are uncorrelated (which corresponds to the case where we have a near orthogonal dictionary in the sparse approximation problem), these approximation ratios are comparable to $(1-\frac{1}{e})$. The recent analyses of the aforementioned algorithms in [2] explain why they work well in practice. It turns out that, according to our result, there are no significantly better algorithms.

2. Inapproximability of DIFF-SPARSE

We first state a simple lemma connecting hardness results for this problem to those of SPARSE. Given Φ, y and k as input to the problem, we have the following:

Proposition 2.1. *If, for some* $c > 0$, DIFF-SPARSE *cannot be approximated within* $(1-c)$ *in polynomial time under some complexity theoretic assumption, then all polynomial time algorithms which find a solution* x *for* SPARSE *should satisfy*

$$\|y-\Phi x\|_2^2 > (1-c)\|y-\Phi x^*\|_2^2 + c\|y\|_2^2$$

under the same complexity theoretic assumption.

Proof. For all polynomial time algorithms, we have

$$\frac{\|\Phi_x\Phi_x^+ y\|_2^2}{\|\Phi_{x^*}\Phi_{x^*}^+ y\|_2^2} = \frac{\|y\|_2^2-\|y-\Phi x\|_2^2}{\|y\|_2^2-\|y-\Phi x^*\|_2^2} < 1-c,$$

where x^* is an optimal solution. Rearranging the last inequality yields the desired result. □

Combining this fact with Theorem 1.1, we have the following:

Corollary 2.2. *Assuming the notation introduced thus far, all polynomial time algorithms solving* SPARSE *satisfy*

$$\|y - \Phi x\|_2^2 \geq (1-c)\|y - \Phi x^*\|_2^2 + c\|y\|_2^2$$

for $\frac{1}{e} > c \geq 0$*, unless* $P = NP$*.*

The corollary above is most significant when the optimal solution has 0 value and c is close to $\frac{1}{e}$. In that case, the corollary states that any polynomial time algorithm devised for SPARSE is stuck with a solution far away from the optimal, unless $P = NP$. As an example, set $y - \Phi x^* = 0$, $c = \frac{1}{e} - \epsilon$ for some small $\epsilon > 0$. Then, we get that such an algorithm should satisfy $\|y - \Phi x\|_2^2 \geq \left(\frac{1}{e} - \epsilon\right)\|y\|_2^2$. A numerical approximation yields $\|y - \Phi x\|_2 > 0.6\|y\|_2$ for some suitably chosen ϵ. It is remarkable how difficult SPARSE is. It is not only inapproximable, but one cannot even hope to get close to 0 when it is the optimal value. Note that, at the other end of the spectrum is the case $c = 0$, which results in the trivial inequality $\|y - \Phi x\|_2 \geq \|y - \Phi x^*\|_2$. We now give the proof of Theorem 1.1:

Proof of Theorem 1.1 We use a result from the classic paper of Feige [6] in which he shows an optimal hardness result for Set Cover and Max k-Cover. In Max k-Cover, we are given a collection of sets $C = \{S_1, S_2, \ldots, S_n\}$ which are subsets of a ground set $E = \{e_1, e_2, \ldots, e_m\}$. We want to choose k sets from C so as to maximize the number of covered elements in E. The following is implicit in Theorem 5.3 of [6]:

Theorem 2.3. *[6] There exists a polynomial time reduction from 3-SAT to Max k-Cover such that if 3-SAT is satisfiable, then there exist k disjoint sets that cover all the elements in E; if 3-SAT is not satisfiable, then no k sets can cover more than* $\left(1 - \frac{1}{e}\right)$ *fraction of the elements.*

Since this is the crux of our result, we would like to briefly mention the reduction which yields this result. In his reduction, Feige uses a multi-prover protocol where each prover receives $n/2$ clauses and $n/2$ *distinguished* variables. These are selected by the verifier from a 3-SAT instance ϕ using a random string, where n is the total number of clauses. The indices of clauses and variables sent to provers are determined by the Hadamard code. Each prover provides an assignment to all the variables it receives and it is said that the answers of two provers are *consistent* if the induced assignments to the distinguished variables are identical. If ϕ is satisfiable, then there is always a strategy such that all the provers are consistent. Otherwise, all the provers are inconsistent with high probability. As for the reduction, a partition system (a ground set together with subsets) is defined for each random string the verifier selects in such a way that if the answers of all the provers are consistent one can find k disjoint sets which perfectly cover the ground set. Otherwise, a certain fraction of the ground set cannot be covered which is quantified by the theorem cited above.

We combine the reduction in the cited theorem with the following reduction from Max k-Cover to DIFF-SPARSE: We construct a matrix $\Phi \in \mathbb{R}^{m \times n}$ where the columns correspond to the sets in C and the rows correspond to the elements in E. Let $\Phi_{ij} = 1$ if e_i is contained in S_j, and 0 otherwise. We also let $y \in \mathbb{R}^m$ be the vector with all entries 1. It is clear that if Max k-Cover has k disjoint sets that perfectly cover E, then there exist k columns of Φ whose some linear combination (in fact with all coefficients 1) is exactly y.

In other words, there exists a vector $x \in \mathbb{R}^n$ such that $\Phi x = y$, i.e. y is in the span of Φ_x. Hence, this gives $\|\Phi_x \Phi_x^+ y\|_2^2 = \|y\|_2^2$. If there are no k sets that cover more than $\left(1 - \frac{1}{e}\right)$ fraction of the elements in E, then no matter how we choose k columns of Φ, no linear combination of them will be able to contribute non-zero elements to the coordinates of y corresponding to the uncovered elements. Since at least $\frac{1}{e}$ fraction of the coordinates of y is uncovered, we have that for any x, $\|y - \Phi x\|_2^2 \geq \frac{1}{e}\|y\|_2^2$. Thus, we get that the orthogonal projection of y, i.e. $\|\Phi_x \Phi_x^+ y\|_2^2 \leq \left(1 - \frac{1}{e}\right)\|y\|_2^2$ by the Pythagoras Theorem. This completes the proof.

3. Final Remark

We have presented an improvement in understanding the hardness of sparse approximation which we can relate to the existing algorithmic results. However, most algorithmic results usually assume further structure on the dictionary or their approximation factors involve parameters of the input. It is a curious open problem whether there exists an algorithm matching the exact lower bound in the general case.

References

[1] A. Das and D. Kempe. Algorithms for subset selection in linear regression. In *Proceedings of the 40th annual ACM symposium on Theory of computing*, STOC '08, pages 45–54, 2008.

[2] A. Das and D. Kempe. Submodular meets spectral: Greedy algorithms for subset selection, sparse approximation and dictionary selection. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, ICML '11, pages 1057–1064, 2011.

[3] G. Davis, S. Mallat, and M. Avellaneda. Adaptive greedy approximations. *Constructive Approximation*, 13:57–98, 1997.

[4] D. L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52:1289–1306, 2006.

[5] D. L. Donoho, M. Elad, and V. N. Temlyakov. On lebesgue-type inequalities for greedy approximation. *Journal of Approximation Theory*, 147:185–795, 2007.

[6] U. Feige. A threshold of $\ln n$ for approximating set cover. *Journal of the ACM*, 45(4):634–652, 1998.

[7] A. C. Gilbert, S. Muthukrishnan, and M. J. Strauss. Approximation of functions over redundant dictionaries using coherence. In *Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms*, SODA '03, pages 243–252, 2003.

[8] B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24(2):227–234, 1995.

[9] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Proceedings of the 27 th Annual Asilomar Conference on Signals, Systems, and Computers*, pages 40–44, 1993.

[10] V. N. Temlyakov. Greedy algorithms and m-term approximation with regard to redundant dictionaries. *Journal of Approximation Theory*, 98:117–145, 1999.

[11] V. N. Temlyakov. Weak greedy algorithms. *Advances in Computational Mathematics*, pages 213–227, 2000.

[12] J. A. Tropp. Greed is good: algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242, 2004.