Sparse Approximation by Linear Programming: Measuring the Error with the ℓ_1 Norm

Lorenzo Granai and Pierre Vandergheynst
Signal Processing Institute (ITS)
Ecole Polytechnique Fédérale de Lausanne (EPFL)
LTS2-ITS-STI-EPFL, 1015 Lausanne, Switzerland
Technical Report TR-ITS-2005.015
June 13, 2005

Abstract

In this report we study the problem of sparse signal approximation over redundant dictionaries. We focus our attention on the minimization of a cost function where the error is measured using a ℓ_1 norm. We show a constructive equivalence between this minimization and Linear Programming. A recovery condition is then proved and finally we provide an example of the use of such a technique for denoising.

Index Terms

Sparse Approximation, Basis Pursuit Denoising, Relaxation Algorithms, Redundant Dictionaries, Denoising

I. Introduction

We want to approximate a signal $f \in \mathbb{R}^n$ over a redundant set of unit norm functions $\mathcal{D} = \{g_i\}_{i\in\Omega}$, which from now on will be called dictionary. Let us name d the cardinality of the dictionary, with $|\mathcal{D}| = d > n$. Given the overcompleteness of $\mathcal{D}$, the solution to this problem is non-unique and among all the possible approximations we are interested in the one which contains the smallest number of non-zero components, i.e. the sparsest one.

In [1] Chen, Donoho and Saunders introduce the Basis Pursuit Denoising (BPDN) paradigm that consists in the following minimization problem, which can be solved by Quadratic Programming techniques :

$$(P2-1) \quad \min_{\mathbf{b}} \|f - D\mathbf{b}\|_2^2 + \gamma\|\mathbf{b}\|_1. \tag{1}$$

Here D is a $n \times d$ matrix, whose columns are the elements of the the dictionary, f is the column vector corresponding to the signal we want to approximate and $\mathbf{b}$ is the coefficient vector. BPDN can be seen as a convex relaxation of the non-convex, NP-complex Subset Selection problem, where the sparsity constraint is given by the ℓ_0 semi-norm of the coefficient vector:

$$(P2-0) \quad \min_{\mathbf{b}} \|f - D\mathbf{b}\|_2^2 + \tau^2\|\mathbf{b}\|_0. \tag{2}$$

Recently, many interesting contributions showed how, under certain conditions on the dictionary, solving the convex problem of (1) can provide the sparsest approximation of the signal f over $\mathcal{D}$, i.e. the solution of $(P2-0)$ [2], [3], [4], [5], [6], [7], [8].

The problem we propose to solve here substitutes the classical ℓ_2 measure of the error with the ℓ_1 norm:

$$(P1-1) \quad \min_{\mathbf{b}} \|f - D\mathbf{b}\|_1 + \gamma\|\mathbf{b}\|_1. \tag{3}$$

In this way the algorithm gives less importance to "wild" signal samples. The minimization of Eq. (3) can be written as a Linear Programming problem of the following form:

$$\min_{\mathbf{x}} \mathbf{v}^T\mathbf{x} \;\text{ s.t. }\; A\mathbf{x} = s \;\text{ and }\; \mathbf{x} \geq 0. \tag{4}$$

In order to show this equivalence [9] one should create a vector $\mathbf{u} = (\mathbf{u}_+, \mathbf{u}_-)$ with $\mathbf{u}_+, \mathbf{u}_- \geq 0$ such that $\mathbf{b} = \mathbf{u}_+ - \mathbf{u}_-$. The vector $\mathbf{u}_+$ contains only the positive components of $\mathbf{b}$, while the negative ones are in $\mathbf{u}_-$, but with a positive sign. In this way one can see that $\|\mathbf{b}\|_1 = \mathbf{1}^T\mathbf{u}$. The same can be done defining a vector $\mathbf{r} = (\mathbf{r}_+, \mathbf{r}_-)$, with $\mathbf{r}_+, \mathbf{r}_- \geq 0$ and

$$\mathbf{r}_+ - \mathbf{r}_- = f - (D, -D) \cdot \mathbf{u}.$$

It is now clear that Eq. (3) can be written as

$$\min_{\mathbf{u},\mathbf{r}} \mathbf{1}^T\mathbf{r} + \gamma\mathbf{1}^T\mathbf{u} \;\text{ s.t. }\; A \cdot (\mathbf{r}, \mathbf{u}) = f \;\text{ and }\; \mathbf{u}, \mathbf{r} \geq 0,$$

with $A = (I, -I, D, -D)$, where I is a $n \times n$ identity matrix.

Web page: http://lts2www.epfl.ch. This work is supported by the SNF grant 2100-066912.01/1

II. Recovery Condition

In this section we study the relationship between (3) and the following non relaxed minimization problem where the error is still measured with the ℓ_1 norm:

$$(P1-0) \quad \min_{\mathbf{c}} \|f - D\mathbf{c}\|_1 + \tau^2 \|\mathbf{c}\|_0. \tag{5}$$

The cost function of this problem is a trade-off between the sparseness of the approximation and its distance from the input signal. Again $(P1-0)$ is not convex and here we wonder when and how solving $(P1-1)$ can help us in finding the solution of (5).

***Theorem** 1:* Let $\mathbf{b}_*$ be the coefficient vector that minimizes $(P1-1)$ and let $\Gamma \subset \Omega$ be the optimal function subset found by solving the non-convex problem $(P1-0)$. D_Γ will be the subdictionary containing only the functions indexed in Γ. Suppose that $\sup_{i \notin \Gamma} \|D_\Gamma^+ g_i\|_1 < 1$, then we can state that if

$$\gamma > \frac{\sqrt{n}}{1 - \sup_{i \notin \Gamma} \|D_\Gamma^+ g_i\|_1} \tag{6}$$

then $support(\mathbf{b}_*) \subset \Gamma$.

Proof: This proof is inspired by the proof of the Correlation Condition Lemma that appears in [7]. Let us call $D_{\overline{\Gamma}}$ the complementary of D_Γ on D, such that $D = D_\Gamma \cup D_{\overline{\Gamma}}$. Suppose that $\mathbf{b}_*$ contains (at least) one element out of Γ, so we can write the cost function of $(P1-1)$ for both $\mathbf{b}_*$ and its projection onto D_Γ, that is $D_\Gamma^+ D\mathbf{b}_*$. Since $\mathbf{b}_*$ minimizes $(P1-1)$, we have:

$$\gamma \left(\|\mathbf{b}_*\|_1 - \|D_\Gamma^+ D\mathbf{b}_*\|_1 \right) \leq \|f - DD_\Gamma^+ D\mathbf{b}_*\|_1 - \|f - D\mathbf{b}_*\|_1. \tag{7}$$

Let us now split the coefficient vector into two parts: $\mathbf{b}_* = \mathbf{b}_\Gamma + \mathbf{b}_{\overline{\Gamma}}$, where the former vector contains the components with indexes in Γ, while the latter the remaining components from $\overline{\Gamma} = \Omega \backslash \Gamma$. The left-hand term of (7) can be bounded as in [7] obtaining:

$$\gamma \left((1 - \sup_{i \notin \Gamma} \|D_\Gamma^+ g_i\|_1) \cdot \|\mathbf{b}_{\overline{\Gamma}}\|_1 \right) \leq \gamma \left(\|\mathbf{b}_*\|_1 - \|D_\Gamma^+ D\mathbf{b}_*\|_1 \right). \tag{8}$$

We now work with the right-hand side of (7):

$$\begin{aligned} \|f - DD_\Gamma^+ D\mathbf{b}_*\|_1 - \|f - D\mathbf{b}_*\|_1 \leq \|D\mathbf{b}_* - P_\Gamma D\mathbf{b}_*\|_1 = \\ \|(I - P_\Gamma) D\mathbf{b}_{\overline{\Gamma}}\|_1 \leq \|(I - P_\Gamma) D\|_{1,1} \cdot \|\mathbf{b}_{\overline{\Gamma}}\|_1, \end{aligned} \tag{9}$$

where $P_\Gamma = DD_\Gamma^+ = D_\Gamma D_\Gamma^+$ is an orthogonal projector. Using this result together with (8) we obtain:

$$\gamma(1 - \sup_{i \notin \Gamma} \|D_\Gamma^+ g_i\|_1) \leq \|(I - P_\Gamma) D\|_{1,1}. \tag{10}$$

The right-hand side of the previous equation is the maximum ℓ_1 norm of the columns of $(I - P_\Gamma D)$, i.e.

$$\|(I - P_\Gamma) D\|_{1,1} = \max_{g \in D_{\overline{\Gamma}}} \|g - P_\Gamma g\|_1 \leq \max_{g \in D_{\overline{\Gamma}}} \|g - P_\Gamma g\|_2 \cdot \sqrt{n} \leq \max_{g \in D_{\overline{\Gamma}}} \|g\|_2 \cdot \sqrt{n} = \sqrt{n}. \tag{11}$$

Finally, we have

$$\gamma(1 - \sup_{i \notin \Gamma} \|D_\Gamma^+ g_i\|_1) \leq \sqrt{n}. \tag{12}$$

If this inequality fails, then $\mathbf{b}_*$ is supported in Γ. ■

Unfortunately, since the optimal set of functions is not known, this condition can not be tested before decomposing a signal. An additional condition based on the cumulative coherence $\mu_1(m)$ of the dictionary can be easily found form (6) using proposition 3.7 in [7]. It turns out that if $|\Gamma| \leq m$ and $\mu_1(m-1) + \mu_1(m) < 1$ then $support(\mathbf{b}_*) \subset \Gamma$ if

$$\gamma = \frac{\sqrt{n}(1 - \mu_1(m-1))}{1 - \mu_1(m-1) - \mu_1(m)}. \tag{13}$$

In this way it is possible to check the new sufficient condition.

III. An Example

In this section we offer an example of the use of the proposed minimization problem. Let us call $\mathbf{b}_*$ the approximation found by solving $(P1-1)$. This vector is thresholded, removing the numerically negligible components, and in this way we are able to individuate a sparse support and thus a subset of the dictionary. Let us label the subdictionary found by $(P1-1)$ with $\mathcal{D}_*$ (composed by the atoms corresponding to the non-zero elements of $\mathbf{b}_*$). Once this is given, there are no guarantees that the coefficients that represent f are optimal. These are, thus, recomputed projecting the signal onto $\mathcal{D}_*$ and a new approximation of f named $\mathbf{b}_{**}$ is found. Of course, $support(\mathbf{b}_*) = support(\mathbf{b}_{**})$. Formally the approximant found by $(P1-1)$ after the projection step is:

$$f_{**} = D_*(D_*)^+ f = D\mathbf{b}_{**}. \tag{14}$$

So the minimization of Eq. (3) is used only to select the dictionary subset. Of course the very same method can be used for the BPDN paradigm.

We now decompose a piecewise smooth signal affected by "pointwise" noise. The dictionary used has redundancy factor 2 and is composed by the union of a wavelet *Symmlet*-4 orthonormal basis [10] and the respective family of footprints for all the possible translations of the Heaviside function (see [11]). The latter is meant to model the discontinuities, while the former should represent the smooth parts of the signal [12]. Figure 1 shows the original noisy signal, and two reconstructions obtained by solving $(P1-1)$ on the left and $(P2-1)$ on the right, and then recomputing the coefficients by orthogonal projection as in (14). The MSE is respectively 0.37 and 0.61. It can be seen how $(P1-1)$ is less sensible to wild samples given by the pointwise noise, thanks to the ℓ_1 penalization that allows the algorithm to select a better subset of functions.

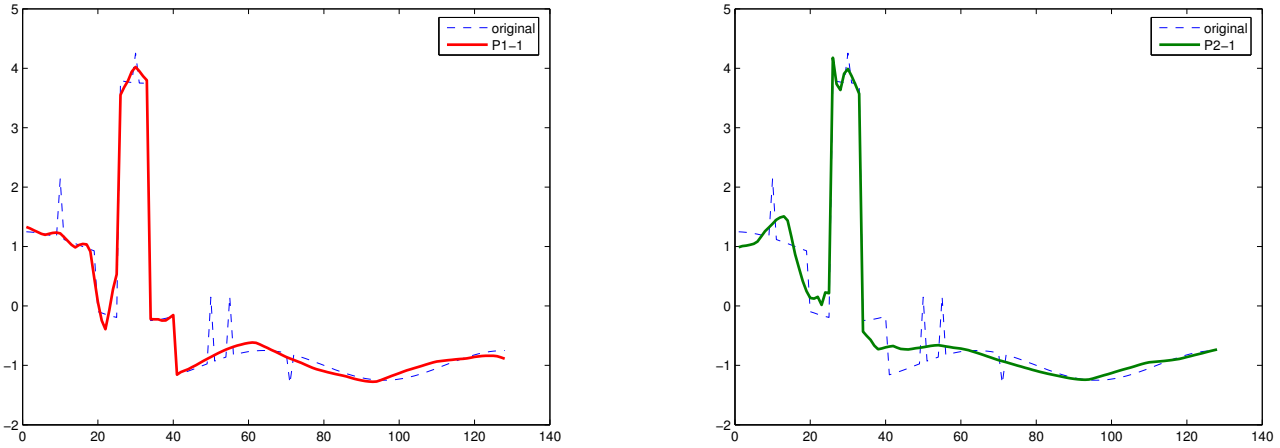

Fig. 1. The original noisy signal and the approximations obtained with 9 coefficients by solving $(P1-1)$ on the left and $(P2-1)$ on the right

Figure 2 shows the error decay for both approximation methods versus the number of selected functions. Although the MSE is a criterion of evaluation that is clearly favorable to BPDN, because it measures the ℓ_2 distance between two signals, in same cases we obtain that the the solution found by solving $(P1-1)$ overcomes BPDN.

This example shows a case where the proposed problem can be useful, but it does not satisfy the sufficient condition of equation (13). That condition turns out to be quite pessimistic. Can we find another (toy) example that satisfies the hypothesis?

IV. Brief Discussion

Recently total variation based image denoising model of Rudin, Osher, and Fatemi has been modified by using the L_1 norm to calculate the fidelity term in the cost function to minimize [13]. This modification have interesting new implications. Our choice to introduce $(P1-1)$ from $(P2-1)$, follows a similar idea, even if the background of the two problems is different.

The measure of the approximation error with ℓ_1 norm has been also used by Candes and Tao in [14], [15]. The problem we solve here was also addressed in [16]. Moreover, in the Discussion of [7] Tropp imagines the situation where the ℓ_2 norm is not the most appropriate way to measure the error in approximating the input signal, but without giving further details.

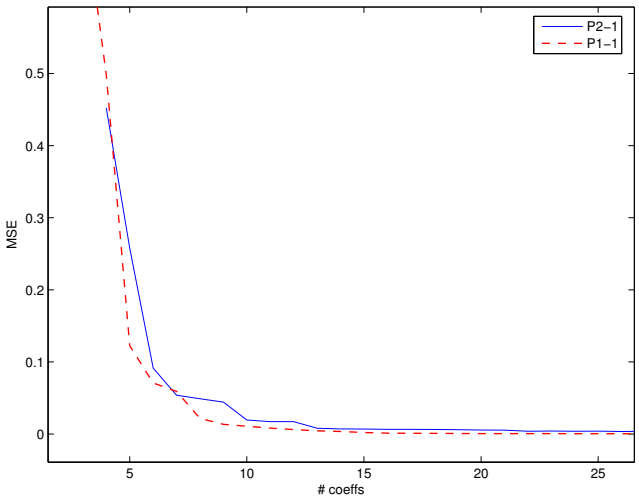

Fig. 2. MSE versus number of selected coefficients

Acknowledgments

We would like to thank Lorenzo Peotta and Jean-Jacques Fuchs for fruitful discussions.

References

[1] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comp.*, vol. 20, no. 1, pp. 33–61, 1999.
[2] D. L. Donoho and X. Huo, "Uncertainty principles and ideal atom decomposition," *IEEE Trans. Inform. Theory*, vol. 47, no. 7, pp. 2845–2862, Nov 2001.
[3] M. Elad and A. M. Bruckstein, "A generalized uncertainty principles and sparse representation in pairs of bases," *IEEE Trans. Inform. Theory*, vol. 48, no. 9, pp. 2558–2567, Sep 2002.
[4] D. L. Donoho and M. Elad, "Maximal sparsity representation in general (non-orthogonal) dictionaries via l1 minimization," Stanford University, Stanford, CA, Tech. Rep., 2002.
[5] J. A. Tropp, "Greed is good : Algorithmic results for sparse approximation," *IEEE Trans. Inform. Theory*, vol. 50, no. 10, pp. 2231–2242, Oct 2004.
[6] R. Gribonval and M. Nielsen, "Approximation with highly redundant dictionaries," in *Proc. of 48th SPIE annual meeting*, San Diego, USA, August 2003.
[7] J. A. Tropp, "Just relax: Convex programming methods for subset selection and sparse approximation," Texas Institute for Computational Engineering and Sciences, Tech. Rep., 2004.
[8] J. J. Fuchs, "On sparse representations in arbitrary redundant bases," *IEEE Trans. Inform. Theory*, vol. 50, no. 6, 2004.
[9] S. Sardy, "Regularization techniques for linear regression with a large set of carriers," Ph.D. dissertation, Univ. Washington, Seattle, 1998.
[10] S. Mallat, *A Wavelet Tour of Signal Processing*. Academic Press, 1998.
[11] P. Dragotti and M. Vetterli, "Wavelet footprints: Theory, algorithms and applications," *IEEE Transactions on Signal Processing*, vol. 51, no. 5, pp. 1306–1323, May 2003.
[12] O. Divorra Escoda, L. Granai, and P. Vandergheynst, "On the use of *a Priori* information for sparse signal approximations," ITS/LTS-2 EPFL, Tech. Rep. 23.2004, November 2004.
[13] T. F. Chan and S. Esedoglu, "Aspects of total variation regularized L^1 function approximation," UCLA, Tech. Rep. CAM Report 04-07, February 2004, to appear in SIAM J. Appl. Math.
[14] E. Candes and T. Tao, "Decoding by linear programming," December 2004, submitted.
[15] ——, "Error correction via linear programming," in *FOCS 2005*, 2005, submitted.
[16] L. Granai and P. Vandergheynst, "Sparse decomposition over multi-component redundant dictionaries," in *Proc. of Multimedia Signal Processing, Workshop on. MMSP04*, September 2004, pp. 494–497.