SPARSE APPROXIMATION BY LINEAR PROGRAMMING USING AN L1 DATA-FIDELITY TERM

Lorenzo Granai and Pierre Vandergheynst

Ecole Polytechnique Fédérale de Lausanne, Signal Processing Institute
EPFL-STI-ITS-LTS2, Station 11, CH-1015, Lausanne, Switzerland
Email: {lorenzo.granai,pierre.vandergheynst}@epfl.ch
Tel. +41 +21 6934807, Fax: +41 +21 6937600
Web page: http://lts2www.epfl.ch

ABSTRACT

This paper studies the problem of sparse signal approximation over redundant dictionaries. Our attention is focused on the minimization of a cost function where the error is measured by using the L_1 norm, giving thus less importance to outliers. We show a constructive equivalence between the proposed minimization problem and Linear Programming. A recovery condition is then provided and an example illustrates the use of such a technique for denoising.

1. 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 introduced the Basis Pursuit Denoising (BPDN) paradigm that consists in the following minimization problem, which can be solved by Quadratic Programming (QP) techniques :

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

Here D is the $n \times d$ dictionary synthesis matrix, whose columns are the elements of $\mathcal{D}$, f is the column vector corresponding to the signal 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 quasi-norm of the coefficient vector:

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

The work of Lorenzo Granai has been partly supported by the SNF grant 2100-066912.01/1

In the very last years, many interesting contributions have shown how, under certain conditions on the dictionary, solving the convex problem in (1) can provide the sparsest approximation of the signal f over $\mathcal{D}$, i.e. the solution of (P_{2-0}) [2, 3, 4]. In short, such sufficient conditions pose a limit to the coherence of the redundant dictionary.

In the problem we propose in this paper, an L_1 data-fidelity term substitutes for the classical ℓ_2 measure of the error:

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

In this way the algorithm gives less importance to outliers, or "wild" signal samples.

Recently, the total variation based image denoising model of Rudin, Osher, and Fatemi [5] has been modified by using the L_1 norm to calculate the fidelity term in the cost function [6, 7]. This change brings new interesting implications, as can be seen for example in the pioneering works of Nikolova [8]. In [9] the problem of image restoration is considered, where an original scene f has to be recovered given its observation $\hat{f}$. The problem can be written as:

$$\hat{f} = Hf + w, \tag{4}$$

where H is a blurring matrix and w the additive noise. The authors of [9] propose to solve such a problem by minimizing the following cost function:

$$\min_{f} \|\hat{f} - Hf\|_1 + \gamma\|Rf\|_1, \tag{5}$$

where R is a regularization operator, usually a difference operator. Note the similarity between Eq. (5) and (P_{1-1}). Our choice to introduce (P_{1-1}) from (P_{2-1}), follows a similar but independent idea, even if the background of the two problems is different.

The measure of the approximation error by means of the ℓ_1 norm has been also used in [10]. Moreover, in

the Discussion of [3] Tropp imagines a situation where the Euclidean norm is not the most appropriate way to measure the error in approximating the input signal, but without giving further details. We also addressed a similar problem in [11], while working with multi-component dictionaries, but without providing any theoretical analysis.

It is important to observe that the minimization of Eq. (3) can be written as a Linear Programming (LP) problem of the following form:

$$\min_{\mathbf{x}} \mathbf{v}^T\mathbf{x} \text{ s.t. } A\mathbf{x} = s \text{ and } \mathbf{x} \geq 0, \tag{6}$$

where $\mathbf{v}$ is a vector of known coefficients. In order to show this equivalence (see also [12]), 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 $\|\mathbf{b}\|_1 = \mathbf{1}^T\mathbf{u}$, where $\mathbf{1}$ is a vector of ones. In the same way we define 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. Here we find the form of Eq. (6), with $\mathbf{v} = (\mathbf{1}, \gamma\mathbf{1})$, $\mathbf{x} = (\mathbf{r}, \mathbf{u})$ and $s = f$. In practice, solving a LP is generally faster than solving a QP involved by BPDN.

2. A BAYESIAN APPROACH

Let us write the model of data approximation from a Bayesian point of view: $f = \hat{f} + r = D\mathbf{b} + r$, where $\hat{f}$ is the approximant and r the residual. Assuming r to be an iid Laplacian set of variables, the probability that $\hat{f}$ corresponds to f, given D and $\mathbf{b}$ is related to:

$$p(f|D, \mathbf{b}) = \frac{1}{\sqrt{2\pi\sigma_r^2}} \cdot \exp\left(-\frac{\|f - D\mathbf{b}\|_1}{2\sigma_r^2}\right), \tag{7}$$

where σ_r^2 is the variance of the residual. In the approximation problem, one aims at maximizing the likelihood $p(\mathbf{b}|f, D)$. Formally, by the Bayes rule, we have

$$p(\mathbf{b}|f, D) = \frac{p(f|D, \mathbf{b}) \cdot p(\mathbf{b})}{p(f, D)},$$

and thus, being $p(f, D)$ uniform for a given signal and dictionary, it follows that the most probable signal representation is:

$$\mathbf{b}_P = \arg\max_{\mathbf{b}} p(f|D, \mathbf{b}) \cdot p(\mathbf{b}). \tag{8}$$

Let us now assume that the coefficients b_i are independent and have a Laplacian distribution with standard deviation σ_i. From (8), by computing the logarithm, it follows that

$$\mathbf{b}_P = \arg\min_{\mathbf{b}} \left(\frac{\|f - D\mathbf{b}\|_1}{2\sigma_r^2} + \sum_i \frac{\sqrt{2}|b_i|}{\sigma_i}\right).$$

Making the hypothesis that σ_i is constant for every index i, the previous equation means that the most probable $\mathbf{b}$ is the one found by solving the problem (P_{1-1}). On the other hand, if r is Gaussian, the most probable coefficient vector is provided by BPDN [13].

3. RECOVERY CONDITION

Let us now study the relationship between (P_{1-1}) and the following non-relaxed minimization problem, where the error is still measured with the ℓ_1 norm:

$$(P_{1-0}) \qquad \min_{\mathbf{c}} \|f - D\mathbf{c}\|_1 + \tau^2\|\mathbf{c}\|_0. \tag{9}$$

The cost function of this problem is a trade-off between the ℓ_0 measure of the sparseness of the approximation and its distance from the input signal. Again (P_{1-0}) is non-convex and here we wonder when and how solving (P_{1-1}) can help us in finding the solution of (9).

Theorem 1 *Let $\mathbf{b}_*$ be the coefficient vector that minimizes (P_{1-1}) and let $\Gamma \subset \Omega$ be the optimal function subset found by solving the non-convex problem (P_{1-0}). D_Γ will be the sub-dictionary 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*

$$\gamma > \frac{\sqrt{n}}{1 - \sup_{i\notin\Gamma} \|D_\Gamma^+ g_i\|_1} \tag{10}$$

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

Proof: This proof is inspired by the proof of the Correlation Condition Lemma that appears in [3]. 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 (P_{1-1}) for both $\mathbf{b}_*$ and its projection onto D_Γ, that is $D_\Gamma^+ D\mathbf{b}_*$. Since $\mathbf{b}_*$ minimizes (P_{1-1}), we have:

$$\gamma\left(\|\mathbf{b}_*\|_1 - \|D_\Gamma^+ D\mathbf{b}_*\|_1\right) \leq \|f - DD_\Gamma^+ D\mathbf{b}_*\|_1 - \|f - D\mathbf{b}_*\|_1. \tag{11}$$

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 $\overline{\Gamma} = \Omega \setminus \Gamma$. The left-hand term of (11) can be bounded as in [3] obtaining:

$$\gamma \left((1 - \sup_{i \notin \Gamma} \|D_\Gamma^+ g_i\|_1) \cdot \|\mathbf{b}_{\overline{\Gamma}}\|_1 \right) \leq \gamma \left(\|\mathbf{b}_*\|_1 - \|D_\Gamma^+ D\mathbf{b}_*\|_1 \right). \tag{12}$$

We now work with the right-hand side of (11):

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

where $P_\Gamma = DD_\Gamma^+ = D_\Gamma D_\Gamma^+$ is an orthogonal projector. Using this result together with (12) we obtain:

$$\gamma(1 - \sup_{i \notin \Gamma} \|D_\Gamma^+ g_i\|_1) \leq \|(I - P_\Gamma)D\|_{1,1}. \tag{13}$$

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

$$\begin{array}{l} \|(I - P_\Gamma)D\|_{1,1} = \max_{g \in D_{\overline{\Gamma}}} \|g - P_\Gamma g\|_1 \leq \\ \max_{g \in D_{\overline{\Gamma}}} \|g - P_\Gamma g\|_2 \cdot \sqrt{n} \leq \\ \max_{g \in D_{\overline{\Gamma}}} \|g\|_2 \cdot \sqrt{n} = \sqrt{n}. \end{array} \tag{14}$$

Finally, we have

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

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

Unfortunately, since the optimal set of functions is not known, the sufficient condition provided by the previous theorem cannot be tested before decomposing a signal. Form (10), one can easily find an additional condition based on the cumulative coherence $\mu_1(m)$ defined as:

$$\mu_1(m, \mathcal{D}) \triangleq \sup_{|\Lambda| = m} \sup_{i \in \Omega \setminus \Lambda} \sum_{\lambda \in \Lambda} |\langle g_\lambda, g_i \rangle|.$$

We know that $1 - \sup_{i \notin \Gamma} \|D_\Gamma^+ g\|_1 > 0$ whenever $\mu_1(m-1) + \mu_1(m) < 1$ (e.g. see Proposition 3.7 in [3]). Applying this to the results of Theorem 1, it turns out that if $|\Gamma| \leq m$ and $\mu_1(m-1) + \mu_1(m) < 1$ then $support(\mathbf{b}_*) \subset \Gamma$ if

$$\gamma = \frac{\sqrt{n}(1 - \mu_1(m-1))}{1 - \mu_1(m-1) - \mu_1(m)}. \tag{15}$$

This new sufficient condition, even if more pessimistic, can be numerically checked. However, computing $\mu_1(m)$ for m and $\mathcal{D}$ not too small can be very computationally expensive.

4. AN EXAMPLE

We offer now an example of the use of the proposed minimization problem. Let us call $\mathbf{b}_*$ the approximation found by solving (P_{1-1}). 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, called $\mathcal{D}_*$. There are no guarantees that the amplitudes of the coefficients are optimal, thus these are recomputed projecting the signal onto the subspace spanned by the elements of $\mathcal{D}_*$ and a new approximation $\mathbf{b}_{**}$ is found. Of course, $support(\mathbf{b}_*) = support(\mathbf{b}_{**})$. Formally, the approximant found after the projection step is:

$$f_{**} = D_*(D_*)^+ f = D\mathbf{b}_{**}. \tag{16}$$

Thus, the minimization of Eq. (3) is used only to select the dictionary subset. The same method can, of course, be adopted for BPDN.

We now decompose a piecewise smooth signal affected by impulse noise. The dictionary in use has redundancy factor 2 and is composed by the union of a wavelet *Symmlet*-4 orthonormal basis and the respective family of footprints for all the possible translations of the Heaviside function (see [14]). The latter is meant to model the discontinuities, while the former should represent the smooth parts of the signal. Figure 1 shows the original noisy signal, and two reconstructions obtained by solving (P_{1-1}) at the top and (P_{2-1}) at the bottom, and then recomputing the coefficients by orthogonal projection as in (16). The Mean Square Error (MSE) is 0.37 and 0.61 respectively, and remark that the MSE is not an error measure favorable to (P_{1-1}), since it is based on the Euclidean norm. It is clear that (P_{1-1}) is less sensible to the wild samples given by the impulse noise, thanks to the ℓ_1 penalization that allows the algorithm to select a better subset of functions. This reflects the fact of assuming r to be Laplacian in Eq. (7).

This example shows a case where the proposed problem can be useful, but it does not satisfy the sufficient condition of equation (15), that turns out to be quite pessimistic.

5. CONCLUSIONS

The first term in the expression "sparse approximation" refers to the number of non-zero elements used to approximate a target. Such a measure is often "relaxed" switching to a ℓ_1 norm in order to obtain a convex problem. The term "approximation" says that an error is tolerated, but how should such an error be measured? Usually an Euclidean norm is used, however this is far from being optimal for every circumstance. This paper presents a minimization of a cost function where the data-fidelity is

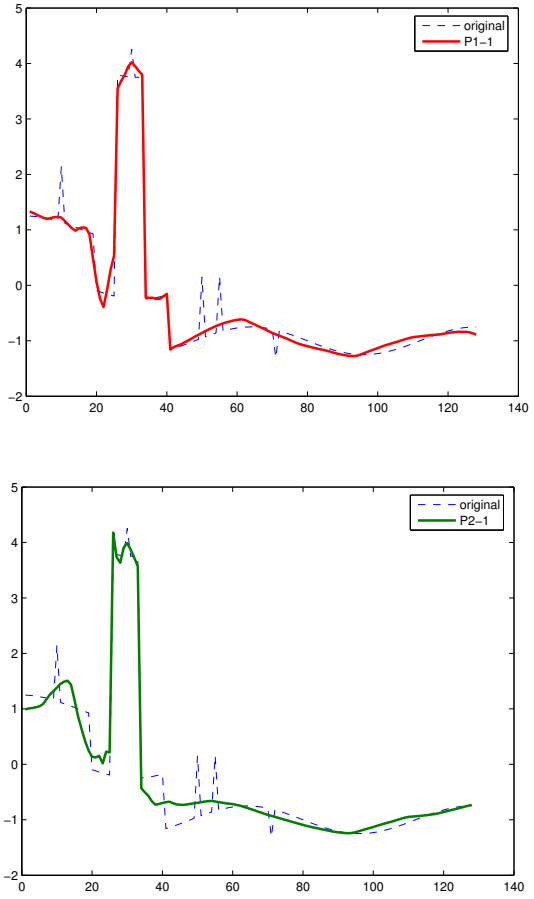

Fig. 1. The original, noisy signal and the approximants obtained with 9 coefficients by solving (P_{1-1}) (top) and (P_{2-1}) (bottom).

measured in L_1. An interesting application is the elimination of impulse noise, as shown in the example. Remark that the goodness of the achieved results would not be possible without the projection step that follows the dictionary subset selection (see Eq. (16)).

The proposed approach can be also applied to image denoising, provided that the LP problem involved in its solution does not become too complex.

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] J. A. Tropp, "Just relax: Convex programming methods for subset selection and sparse approximation," Texas Institute for Computational Engineering and Sciences, Tech. Rep., 2004.

[4] J. J. Fuchs, "On sparse representations in arbitrary redundant bases," *IEEE Trans. Inform. Theory*, vol. 50, no. 6, 2004.

[5] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," in *Proc. of the Eleventh Annual Int. Conf. of the Center for Nonlinear Studies on Experimental mathematics: Computational issues in nonlinear science.* Amsterdam, The Netherlands: Elsevier North-Holland, Inc., 1992, pp. 259–268.

[6] M. Nikolova, "Minimizers of cost-functions involving nonsmooth data-fidelity terms. Application to the processing of outliers," *SIAM Journal on Num. Analysis*, vol. 20, pp. 965–994, 2002.

[7] 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.

[8] M. Nikolova, "A variational approach to remove outliers and impulse noise," *J. Math. Imaging Vis.*, vol. 20, no. 1-2, pp. 99–120, 2004.

[9] H. Fu, M. Ng, M. Nikolova, and J. Barlow, "Efficient minimization methods of mixed L1-L1 and L2-L1 norms for image restoration," *SIAM Journal on Scientific Computing*, 2005, to Appear.

[10] E. J. Candes, M. Rudelson, T. Tao, and R. Vershynin, "Error correction via linear programming," in *In Proc. of 46th Annual IEEE Symposium on Foundations of Computer Science (FOCS05).* IEEE, 2005.

[11] L. Granai and P. Vandergheynst, "Sparse decomposition over multi-component redundant dictionaries," in *Proc. of Multimedia Signal Processing, Workshop on (MMSP04).* IEEE, September 2004, pp. 494–497.

[12] S. Sardy, "Regularization techniques for linear regression with a large set of carriers," Ph.D. dissertation, Univ. Washington, Seattle, 1998.

[13] L. Granai, "Nonlinear approximation with redundant multi-component dictionaries," Ph.D. dissertation, EPFL, 1015 Lausanne, Switzerland, 2005.

[14] P. L. Dragotti and M. Vetterli, "Wavelet footprints: Theory, algorithms and applications," *IEEE Trans. Signal Processing*, vol. 51, no. 5, pp. 1306–1323, May 2003.