FAST IMPLEMENTATION OF A $\ell_\infty - \ell_1$ PENALIZED SPARSE REPRESENTATIONS ALGORITHM: APPLICATIONS IN IMAGE DENOISING AND CODING.

Jean-Jacques FUCHS *and Christine* GUILLEMOT

IRISA/INRIA/Université de Rennes I
Campus de Beaulieu - 35042 Rennes Cedex - France
fuchs@irisa.fr

ABSTRACT

Sparse representation techniques have become an important tool in image processing in recent years, for coding, de-noising and in-painting purposes, for instance. They generally rely on an ℓ_2- ℓ_1 penalized criterion and fast algorithms have been proposed to speed up the applications. We propose to replace the ℓ_2-part of the criterion, which has been chosen both for its easy implementation and its relation to the PSNR quality measure, by a ℓ_∞-part. We present a new fast way to minimize a ℓ_∞-ℓ_1 penalized criterion and assess its potential benefits for image De-noising and coding.

Index Terms**—** Speech, Image and Video Processing: *Image* and Video *Coding*, Multi-rate and Digital Signal Processing: *Signal Representations* and Spectral Analysis.

1. INTRODUCTION

There is growing interest in sparse representations, which is a technique, that allows to decompose a signal into a small number of components chosen from an overcomplete set of vectors or signals often called a dictionary or a redundant basis. More formally, one seeks a sparse representation of $b \in R^n$, in terms of the columns a_k of a $n \times m$ matrix A. Since A is full rank and $m \gg n$, there are an infinity of representations and to select a sparse one, one may solve the linear program:

$$\min_x \|x\|_1 \quad \text{subject to} \quad Ax = b,$$

where $\|x\|_k$ denotes the ℓ_k norm of a vector x, $\|x\|_k = [\sum_1^m |x_j|^k]^{1/k}$ for $k \geq 1$. Since an approximate reconstruction may be sufficient and even preferable, it makes sense to replace the linear program by

$$\min_x \|x\|_1 \quad \text{subject to} \quad \|Ax - b\|_p \leq \rho_p,$$

with ρ_p a tolerance to be defined. Quite generally, and especially so in image processing applications such as image de-noising, in-painting, image expansion and prediction, image and video compression, p is taken equal to 2, i.e., the Euclidean norm is chosen as the error metric. The above criterion is then, somehow, equivalent to the quadratic program [1, 2]

$$\min_x \frac{1}{2}\|Ax - b\|_2^2 + h\|x\|_1, \quad h > 0. \qquad (1)$$

The algorithms, that allow to build such a sparse approximate representation, such as matching pursuit [3], a suboptimal algorithm that selects sequentially the components, or those that minimize the quadratic program (1) exactly, are quite time consuming and a lot of effort is steadily made to develop "fast" variants [4]-[6] and dedicated software architectures leading to fast implementations are developed [7].

Sparse representations used for image [8] and low bit rate video coding [9] are all based on the ℓ_2 error metric and the associated ubiquitous PSNR quality measure. However, it is well-known that this metric is not really appropriate from a perceptual point of view. In [10], it is argued that elements of the spatial-frequency-amplitude response of the human visual system, summarized by the contrast sensitivity function can help choosing an image quality metric from the class of ℓ_p metrics. In that context, it has been argued that the error incurred in image compression should be measured using the ℓ_1 norm rather than the ℓ_2 norm.

In coding applications, the coefficients of the expansion need to be quantized. Their quantized representations together with the indices of the coefficients of the expansion are then entropy coded. The reconstruction error will be bounded by the sum of the quantization and the approximation errors. The number of the expansion coefficients and their quantizers must be such that the reconstruction error is minimized under the constraint of a given bit rate. The authors in [11] show, with predictive and sub-band coding, that optimizing quantizers under the constraint of an ℓ_∞ norm of the error, instead of the global ℓ_2 criterion, which averages the error over the whole image, leads to a sharper image reconstruction with a decreased error range and a better edges reconstruction.

In the sequel, we therefore consider the optimization problem :

$$\min_x \|x\|_1 \quad \text{subject to} \quad \|Ax - b\|_\infty \leq \rho_\infty, \qquad (2)$$

the so-called ℓ_∞-ℓ_1 penalized criterion.

2. THE CRITERION

The previous criterion (2) can indeed be transformed into a linear program and some information about its optimum and its dual can be gained in doing so. Its dual can also be obtained in a more straightforward way and shown to be:

$$\max_d b^T d - \rho_\infty \|d\|_1 \text{ under } \|A^T d\|_\infty \leq 1, \qquad (3)$$

We use this dual expression in the unknowns d to express the joint optimality conditions of (2) and (3) [13]:

Theorem:

The vector x is the optimum of the primal (2) and d the optimum of the dual (3), if and only if

$$Ax - b = -\rho_\infty v \quad \text{and} \quad A^T d = u \qquad (4)$$

$$\text{for some} \quad u \in \partial\|x\|_1 \text{ and } v \in \partial\|d\|_1,$$

where $\partial\|x\|_1$ denotes a sub-gradient of $\|x\|_1$ at x. □

To prove this result, one verifies that both x and d are feasible and lead to the same costs. In addition, let us indicate that, using basic linear programming theory one can establish that the optimal x and d have the same number, say p, of non-zero components, with $p \leq n$.

3. THE ITERATIVE ALGORITHM

3.1. Introduction

It is now possible to use these conditions to build an algorithm that solves (2) exactly in a finite number of steps.

The idea is similar to the one used in [5, 6], one starts with ρ_∞, or to simplify the notations ρ, large, for which the optimum is at zero and follows the optimum $x(\rho)$ for diminishing ρ. The number of nonzero components in $x(\rho)$ essentially increases as ρ diminishes and never exceeds n.

More specifically as ρ decreases from infinity, there is, generically, a first interval in ρ say $]\rho_1,\ \rho_0]$ with $\rho_1 < \rho_0 = \|b\|_\infty$, for which the optimum has just one nonzero component then a second interval $]\rho_2,\ \rho_1]$, for which there are two nonzero components, and so on. In this scheme, while ρ decreases, the boundary values of ρ are given by the values of ρ for which the number of nonzero components in the optimum increases (mainly) or decreases (seldom). From the optimality conditions, one can deduce that, within each *interval*, $x(\rho)$ is continuous and varies linearly while $d(\rho)$ remains constant.

In summary to solve (2), though one is only interested in the optimum for a given value of ρ, it happens to be cheaper to solve the problem for decreasing ρ, i.e., to decompose the positive real axis into intervals $]\rho_k,\ \rho_{k-1}]$ within which the number of nonzero components of the optimum remains constant and to stop the procedure when the ρ of interest is within the current interval.

3.2. Development

We will not give the full details of the algorithm but indicate how it can be developed. Notations will play a important role. If one knows the solution $x,\ d$ that together with $u,\ v$ satisfies the necessary and sufficient condition (4) for a given ρ, the idea is to extend it within an interval, to evaluate the values of the boundaries of the interval and eventually to define how to cross such a boundary. To achieve these goals, we partition the optimal x and d into their zero and non-zero components. We denote $\bar{x}$ the non-zero components of the optimal x and $\bar{\bar{x}}$ its zero components, and partition accordingly (the columns in) A into $\bar{A}$ and $\bar{\bar{A}}$. We partition (the optimal) d into its non-zero components $\underline{d}$ and its zero components $\underline{\underline{d}}$ and partition accordingly the rows of A

into $\underline{A}$ and $\underline{\underline{A}}$. Since the optimal u is related to the optimal x, we partition u into $\bar{u}$ and $\bar{\bar{u}}$ and similarly for the optimal v that we partition into $\underline{v}$ and $\underline{\underline{v}}$. These last four sub-vectors have specific properties as follows from the definition of the sub-gradient. One has, for instance, $\bar{u}$ =sign($\bar{x}$) and $\|\bar{\bar{u}}\|_\infty \leq 1$.

The boundaries of the interval we are seeking, are the values of ρ for which these partitions cease to be valid.

Using these notations, it is now possible to rewrite the relations in (4). The second condition $A^T d = u$ implies $\bar{A}^T d = \bar{u}$ and further $\underline{\bar{A}}^T \underline{d} = \bar{u}$, with $\underline{\bar{A}}$ a square order-p matrix, we assume invertible. This relation tells us that $\underline{d}$ and thus d is invariant within the current interval, since this is the case for $\bar{u}$. It follows from $A^T d = u$ that u also is invariant.

We thus already known that within an interval among the eight sub-vectors that define the optimal quadruple, only two - $\bar{x}$ and $\underline{\underline{v}}$- can and will depend upon ρ.

The first condition in (4): $Ax - b = -\rho v$, becomes $\bar{A}\bar{x} = b - \rho v$ and introducing the d-induced partitions yields: $\underline{\bar{A}}\bar{x} = \underline{b} - \rho\underline{v}$ and $\underline{\underline{\bar{A}}}\bar{x} = \underline{\underline{b}} - \rho\underline{\underline{v}}$. The first of these two relations yields the evolution of the optimal $\bar{x}$ and thus x as a funstion of ρ, within the current interval

$$\bar{x}(\rho) = \underline{\bar{A}}^{-1}\underline{b} - \rho\underline{\bar{A}}^{-1}\underline{v} \qquad (5)$$

since $\underline{v}$ =sign$\underline{d}$ remains constant. Replacing $\bar{x}$ by this expression in the second relation, yields

$$\underline{\underline{v}}(\rho) = \frac{1}{\rho}(\underline{\underline{b}} - \underline{\underline{\bar{A}}}\,\underline{\bar{A}}^{-1}\underline{b}) + \underline{\underline{\bar{A}}}\,\underline{\bar{A}}^{-1}\underline{v}. \qquad (6)$$

These two relations thus fully define how the optimal quadruple $x,\ u,\ d,\ v$ varies within an interval.

It remains then to locate the values of ρ for which these formulas cease to be valid, i.e., the values of ρ for which a component in $\bar{x}(\rho)$ becomes zero and/or for which a component in $\underline{\underline{v}}(\rho)$ becomes equal to ± 1. This is an easy task. A more difficult question is to define the way these partitions have to be modified when crossing the so-obtained boundaries in ρ.

4. APPLICATION TO IMAGE DE-NOISING AND CODING

The sparse representation technique under the l_∞ constraint has first been used for image De-noising using the algorithm described in [12]. The De-noising algorithm uses sparsity over overlapping patches as a global image prior in a MAP estimator. Let X and Y be the image to be estimated and its noisy version (assuming AWG noise) respectively. The initial algorithm proceeds in two steps:

- **Step 1**: Solving $\hat{\alpha_{ij}} = argmin\|\alpha_{ij}\|_1$ subject to $\|A\alpha_{ij} - Z\|_2^2 \leq \rho_2$, e.g. using the OMP and GMF [2] algorithms, as a sliding window sparse coding stage on each overlapping patch (ij);
- **Step 2:** Given all sparse vectors $\hat{\alpha_{ij}}$, update X by solving $\hat{X} = argmin\ \lambda\|X - Y\|_2^2 + \sum_{ij}\|A\hat{\alpha}_{ij} - R_{ij}X\|_2^2$, where R_{ij} is the window extracting every patch from the image.

Here, the first step is replaced by the $\ell_\infty - \ell_1$ penalized sparse representation algorithm. Fig. 4 shows the results of the De-noising algorithm with three different sparse representation techniques: OMP (Orthogonal Matching Pursuit), GMF (Global Matched Filter) with the $\ell_2 - \ell_1$ criterion, and GMF with the $\ell_\infty - \ell_1$ criterion for an additive white Gaussian noise of standard deviation $\sigma = 25$. The GMF algorithm with the $l_2 - l_1$ norm outperforms the OMP algorithm, with a PSNR of 28.94 dB against 28.58 dB. As expected, the PSNR obtained with the $\ell_\infty - \ell_1$ norm is lower (26.97 dB), however with a comparable visual quality.

It has been noted in [10], [11] that minimizing the ℓ_∞−norm instead of the ℓ_2-norm of the reconstruction error in image coding, may lead to a sharper image with a decreased error range and better edges reconstruction. The sparse representation technique with the $\ell_\infty - \ell_1$criterion has thus been considered for image compression. However, it turned out that the representation is far less sparse than the one obtained with the $l_2 - l_1$ criterion, for the same approximation error. This is illustrated in Fig. 2 which shows the average number of atoms per block of size 8×8 with a DCT dictionary (image Goldhill).

5. CONCLUDING REMARKS

We have indicated how it is possible to develop a fast algorithm that minimizes a ℓ_∞- ℓ_1 penalized criterion. While the criterion can be transformed into a linear program and thus solved using standard subroutines, our approach is indeed faster if the number of non-zero component in the optimum is small, corresponding to so-called degenerate linear programs. We applied this criterion to image denoising and coding and in both cases, it appears that, at least for the preliminary investigations we performed, the results are disappointing. In the denoising case, the retained sparse representations are less-sparse than for the competing schemes. The computation time is thus higher and the results at best similar. And similarly in the coding case, the representation are far less sparse then for competing schemes and the resulting disadvantage too important to be compensated for in the later coding steps.

6. REFERENCES

[1] S. Chen, D. Donoho and M. Saunders, "Atomic Decomposition by Basis Pursuit" *SIAM J. on Scientific Comput.*, 20, 1, 33-61, 1999.

[2] J.J. Fuchs. "On the application of the global matched filter to DOA estimation with uniform circular arrays." *IEEE-T-SP*, 49, p. 702–709, Apr. 2001.

[3] S. Mallat and Z. Zhang, "Matching Pursuit with time-frequency dictionaries," *IEEE Trans. on S.P.*, 41, 3397-3415, 1993.

[4] R. Gribonval and E. Bacry "Harmonic decomposition of audio signals with matching pursuit" *IEEE Trans.on S.P.*, vol.51, 1, 101-111, Jan. 2003.

[5] B. Efron, T. Hastie, I. Johnstone and R.Tibshirani, "Least angle regression," *Annals of Statistics*, 32, pp. 407–499, Apr. 2004.

[6] S. Maria and J.J. Fuchs, "Application of the global matched filter to STAP data, an efficient algorithmic approach" Proceedings of the *IEEE ICASSP* Conference, Toulouse, May 2006.

[7] S. Krstulovic and R. Gribonval MPTK: Matching Pursuit made tractable Proceedings of the *IEEE ICASSP* Conference, Toulouse, May 2006.

[8] P. Frossard, P. Vandergheynst, R. M. Figueras i Ventura, and M. Kunt, "A Posteriori Quantization of Progressive Matching Pursuit Streams" *IEEE Trans. on Signal Processing*, vol. 52, 2, pp. 525-535, Feb. 2004.

[9] R.Neff and A. Zakhor, "Very low bit-rate video coding based on matching pursuit video coder", *IEEE Trans. on Circuits and systems for video technology*, vol. 7, 1, Feb. 1997.

[10] R. A. DeVore, B. Jawerth and B. J. Lucier, "Image Compression Through Wavelet Transform Coding", *IEEE Trans. on Information Theory*, vol. 38, 2, March. 1992.

[11] L. Karray, P. Duhamel and O. Rioul "Image Coding with an ℓ_∞ Norm and Confidence Interval Criteria", *IEEE Trans. on Signal Processing*, vol. 7, 5, May 1998.

[12] M. Elad and M. Aharon "Image Denoising Via Sparse and Redundant Representations Over Learned Dictionaries", *IEEE Trans. on Image Processing*, vol. 15, 12, Dec. 2006.

[13] J.J. Fuchs, "Some further results on the recovery algorithms", *SPARS'05*, Rennes, Nov. 2005.

Fig. 1. Denoising with three sparse representation techniques: OMP (Orthogonal Matching Pursuit), GMF (Global Matched Filter) with the $\ell_2 - \ell_1$ criterion, and GMF with the $\ell_\infty - \ell_1$ criterion, for an AWGN ($\sigma = 25$).

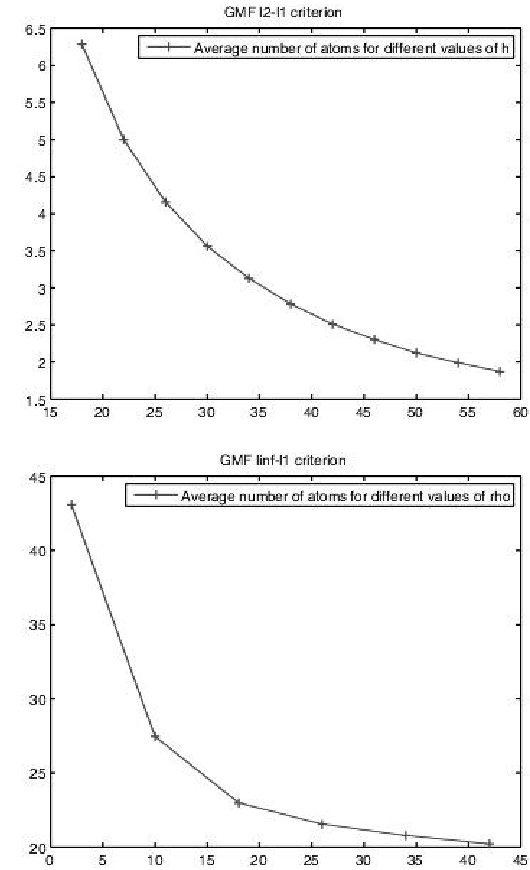

Fig. 2. Average number of atoms per block of size 8×8 with a DCT dictionary (image Goldhill).