SOME FURTHER RESULTS ON THE RECOVERY ALGORITHMS.

Jean-Jacques FUCHS

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

ABSTRACT

When seeking a representation of a signal on a redundant basis one generally replaces the quest for the sparsest model by an ℓ_1 minimization and solves thus a linear program. In the presence of noise one has in addition to replace the exact reconstruction constraint by an approximate one. We consider simultaneously several ways to allow for reconstruction errors and detail the optimality conditions of each of the criterion. We then analyze if these conditions are helpful in the implementation of optimization algorithms.

1. INTRODUCTION

We consider the case where a signal can be exactly represented as a linear combination of a small number of elements from an over-complete set of vectors To recover this representation in the presence of noise, very specific criterion that allow for reconstruction errors have to be minimized. We investigate a whole family of them and indicate how the optimization can be made.

Let us first introduce the standard setting and notations used in this context. Let A be a (n,m)-matrix with $m > n$ and columns a_j, let b denote the observed signal, i.e., a vector that admits an exact sparse representation, say $b = Ax_o$. We denote $\|x\|_0$ the number of non-zero entries in x and $\bar{x}_o$ the reduced dimensional vector built upon the nonzero components of x_o. Similarly $\bar{A}_o$ denotes the associated columns in A. We will assume $\bar{A}_o$ to be full rank. One then has, e.g., $Ax_o = \bar{A}_o\bar{x}_o$. We will also use the notation $\bar{\bar{A}}_o$ for the remaining columns in A and thus decompose A as $A = [\bar{A}_o\ \bar{\bar{A}}_o]$. We further assume without loss of generality that the columns a_j of A are normalized to one in Euclidean norm.

It has been shown in [1, 2, 3, 4] that x_o can be recovered from the observation of $b = Ax_o$ by solving the linear program:

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

where $\|x\|_1 = \sum_1^m |x_j|$, if

$$\|x_o\|_0 < \frac{1}{2}(1 + \frac{1}{M}) \qquad (1)$$

where M, is the so-called the mutual coherence [1]

$$M = \max_{1\le i\neq j\le m} |a_i^T\, a_j|, \qquad (2)$$

It is worth noting that (1) is independent of the magnitudes of the nonzero entries of x_o. Being able to recover x_o appears to be only a matter of structure, of angles between vectors.

Now if $b = Ax_o + e$ with e a vector of additive noise, the optimum of (LP), say x^* will be generically unique and attained at a basic feasible solution i.e at a point having n nonzero components. One can then reasonably expect that if e is “small” and smaller than the “smallest” non-zero component in x_o the optimum x^* of (LP) will be quite close to x_o with n-$\|x_o\|_0$ additional small components directly induced by the noise.

A more systematic way to get rid of the noise induced components is however to change the optimization problem and to allow for reconstruction errors. Instead of asking for an x that satisfies $Ax = b$ one asks for an x such that $\|Ax - b\| \le \rho$ where both the norm and the bound ρ remain to be defined. We therefore consider optimization problems of the form

$$\min_x \|x\|_1 \quad \text{subject to} \ \ \|Ax - b\|_p \le \rho_p \qquad (\text{Opt}_p)$$

with $\|x\|_p = (\sum_1^m |x_p|^p)^{\frac{1}{p}}$, the ℓ_p-norm of x. Remember that in this context dual norms are such that $\frac{1}{p} + \frac{1}{q} = 1$, thus ℓ_1 and ℓ_∞ are dual norms while ℓ_2 is its own dual. In the applications we will mainly consider the case $p = 1,\ 2$ and ∞.

2. OPTIMALITY CONDITIONS

2.1. Previous results

The criterion (Opt$_p$) for $p = 2$ has often been studied recently and the problem then admits several different equivalent forms as for instance

$$\min_x \frac{1}{2}\|Ax - b\|_2^2 + h\|x\|_1, \quad h > 0 \qquad (3)$$

for an adequately chosen parameter h. We will see below that the relation between h and ρ_2 can be made more or less explicit. The Lagrangian dual of (3) has a nice physical interpretation [5]

$$\min_x \ \|Ax\|_2^2 \quad \text{subject to} \quad \|A^T(Ax-b)\|_\infty \leq h$$

Since (3) is a convex program, that can actually be transformed into a quadratic program [5, 4] the necessary and sufficient conditions satisfied at an optimum are well known [5, 4], we do not specify them here since we will recover them below. It is also known that for (Opt$_2$) very specific and dedicated optimization algorithms can be developed [9, 10, 11]. Again we will come back to this point below.

2.2. Optimality conditions

In order to be able to characterize easily the conditions satisfied by the optimum of (Opt$_p$), we introduce $\partial f(x)$ the subdifferential of a convex function f at a point x, it is a set of vectors called the sub-gradients of f at x. For $f(x) = \|x\|_p$ one has [14]

$$\partial\, \|x\|_p = \{u | u^T x = \|x\|_p, \ \|u\|_q \leq 1\} \qquad (4)$$

From the above relation, it follows that

$$\begin{aligned}
&\partial\, \|x\|_1 = \{u | u_i = \text{sign}(x_i) \text{if} x_i \neq 0 \text{ and } |u_i| \leq 1 \text{ else}\} \\
&\partial\, \|x\|_2 = x / \|x\|_2 \\
&\partial\, \|x\|_\infty = \{u | \ |x_i| = \|x\|_\infty \Rightarrow x_i u_i \geq 0, \ |x_i| < \|x\|_\infty \\
&\qquad \Rightarrow u_i = 0, \ \|u\|_1 = 1 \text{ if } x \neq 0, \|u\|_1 \leq 1 \text{ else}\}
\end{aligned}$$

where x_i is the i-th component of x. Note that if f is differentiable at x then $\partial f(x)$ reduces to the gradient.

Before we proceed let us note that (Opt$_p$) is a convex program for $p \geq 1$ and that it admits thus a dual problem (DOpt$_p$) that is convex also. To characterize the optimality conditions of (Opt$_p$) we introduce the dual programs.

Lemma 1. The dual of the convex program (Opt$_p$) is

$$\max_d b^T d - \rho_p \|d\|_q \quad s.t. \ \|A^T d\|_\infty \leq 1 \qquad \text{(DOpt}_p\text{)} \ \square$$

Proof: We first rewrite (Opt$_p$) as

$$\min_{x,\ c} \ \|x\|_1 \quad \text{subject to} \quad \|c\|_p \leq \rho_p \ \text{ and } \ Ax - b = c$$

the Lagrangian of this problem is then

$$\ell(x, c, \lambda, d) = \ \|x\|_1 + \lambda(\|c\|_p - \rho_p) - d^T(Ax - b - c), \quad \lambda \geq 0$$

and defining $\phi(\lambda,\ d) = \min_{x,\ c} \ \ell(x, c, \lambda, d)$, the dual problem is $\max_{\lambda \geq 0,\ d} \phi(\lambda,\ d)$.

In order to evaluate $\phi(\lambda,\ d)$, we first take the minimum of $\ell(\ .\ )$ with respect to x

$$\min_x \|x\|_1 - d^T Ax + ... = \min_x x^T u - x^T A^T d + ...$$

This minimum may not be finite for all d but since we later take the maximum in d theses cases can be ignored. The minimum is finite if and only if $A^T d = u$ for some $u \in \partial\|x\|_1$. From (4), it follows that such a point exists only if $\|A^T d\|_\infty \leq 1$ and the contribution of the terms in x to ℓ or more precisely $\phi(\ .\ ]$ is then zero.

Similarly, the minimum in c may not be finite for all d. It is finite if and only if $\lambda v + d = 0$ for some $v \in \partial\|c\|_p$. Such a point exists only if $\|d\|_q \leq \lambda$ and the contribution of the terms in c to ϕ is then zero. The dual problem is thus

$$\max_{\lambda \geq 0,\ d} d^T b - \lambda\rho_p \quad \text{subject to} \quad \|A^T d\|_\infty \leq 1, \ \ \|d\|_q \leq \lambda$$

and taking the maximum with respect to $\lambda \geq 0$ leads to the announced result. □

The necessary and sufficient conditions for optimality of convex programs admit simple forms when one considers both the primal and the dual and one has

Theorem 1. The optima of (Opt$_p$) and (DOpt$_p$) are respectively x and d if and only

$$\begin{gathered} Ax - b = -\rho_p v \quad \text{and} \quad A^T d = u \\ \text{for some } \ u \in \partial\|x\|_1 \text{ and } v \in \partial\|d\|_q \quad \square \end{gathered} \qquad (5)$$

Proof: The proof is immediate. Both points x and d are feasible and lead to identical costs in both problems. □

These conditions are of course equivalent to the optimality conditions of the primal (Opt$_p$), we introduced them because they are in a form that is more adequate for later use. It is instructive to check it. Since the primal is convex, the first order necessary optimality conditions are also sufficient. The Lagrangian of the primal is

$$\ell(x, \mu) = \|x\|_1 + \mu(\|Ax - b\|_p - \rho_p), \quad \mu \geq 0$$

and the optimality conditions are thus

$$u' + \mu A^T w = 0, \text{ with } \ u' \in \partial\|x\|_1, w \in \partial\|Ax - b\|_p, \ \mu \geq 0$$

To make the link between these conditions and (5) note that from (4) it follows that $\|w\|_q \leq 1$ and $w^T(Ax - b) = \rho_p$. Then take $u' = u$, $w = -d/\|d\|_q$ and $\mu = \|d\|_q$ to transform the solution $\{x,\ d,\ u,\ v\}$ of (5) into a solution of the optimality conditions above.

One can also use this opportunity to establish, for $p = 2$, the link between ρ_2 in (Opt$_2$) and h in (3). The optimality conditions for (3) are

$$A^T(Ax - b) = hu' = 0, \text{for some } \ u' \in \partial\|x\|_1$$

comparing with (5) for $p = 2$, and since $\partial\|d\|_2 = d/\|d\|_2$, it follows that

$$h = \frac{\rho_2}{\|d\|_2} \ \text{ where } \ A^T d = u, \ \text{ i.e. } d = \bar{A}^{T+}\text{sign}(\bar{x})$$

3. THE ITERATIVE ALGORITHMS

We will use the two relations in (5) to try to construct some kind of iterative algorithms that yield the solution to (Opt$_p$) in a very economical way.

Note that due to the presence of u and v the two relations in (5) are far from defining the optimal x and d that can only be obtained by an iterative procedure. They nevertheless carry a lot of information that is helpful if one is interested in the way the optima x and d or more precisely $x(\rho_p)$ and $d(\rho_p)$ vary with ρ_p. We will mainly investigate the case $p = 2, \infty$ and 1.

3.1. The case p=q=2

This case is far easier to analyze than the others. There is no need to use duality in that case. We have seen that (Opt$_2$) can be rewritten (3) whose optimality conditions are simply

$$A^T(Ax - b) + hu = 0 \qquad \text{for some } \ u \in \partial\|x\|_1$$

Assume we have the optimum $\{\ x,\ u\}$ for a given h we show how to extend this optimum in an interval around the current h. The problem consists in extending the current solution $\{\ x,\ u\}$ which is valid for a specific h to possibly all the values in $h > 0$. We will see that the current solution is easily extendible to a whole interval in h around its current value, it thus remains to define the boundaries of this interval and to indicate the changes that need to be done in the expressions of x and u to cross such a boundary and make them valid in the contiguous interval. At last one needs to specify how to initialize for h large the procedure to entirely characterize the *algorithm*. To solve (3) for a given h, one then i initializes the procedure for h large, proceeds from interval to interval until the h of interest belongs to the current interval.

We will need the notations presented in the introduction, i.e., we split or partition the optimum x, we now denote $x(h)$ to emphasize its dependency on h, into its non-zero components $\bar{x}(h)$ and its zero components $\bar{\bar{x}}(h)$ and partition accordingly A into $\bar{A}$ and $\bar{\bar{A}}$.

By definition the interval around the current h is such that this $x(h)$-induced partition remains unchanged within the interval and the values of the boundaries are those values of h for which this partition has to be modified.

Using these notations,we rewrite $A^T(Ax-b)+hu = 0$ first as $A^T(\bar{A}\bar{x}(h) - b) + hu(h) = 0$ and splitting these m equations into two parts, the first associated with $\bar{A}^T$ is

$$\bar{A}^T\bar{A}\bar{x}(h) - \bar{A}^T b = -h\ \bar{u}(h)$$

since $\bar{u}(h)$=sign$\bar{x}(h)$ is constant over the interval, one has

$$\bar{x}(h) = \bar{A}^+ b - h(\bar{A}^T\bar{A})^{-1}\bar{u}$$

which says that $\bar{x}(h)$ varies linearly within each interval. This variation of $x(h)$ deduced from the first part of the m equations induces changes in the second part

$$\bar{\bar{A}}^T(\bar{A}\bar{x}(h) - b) = -h\bar{\bar{u}}(h)$$
$$\Rightarrow\ \ \bar{\bar{u}}(h) = \bar{\bar{A}}^T\bar{A}^+\bar{u} + \frac{1}{h}\bar{\bar{A}}^T(I - \bar{A}\bar{A}^+)b$$

Since both $\bar{u}$ and $\bar{\bar{x}}$ remains constant, we have thus completely defined how the solution $\{x(h),\ u(h)\}$ varies as a function of h in the current interval. To define the boundaries in h of the current interval one monitors the components in $\bar{x}(h)$ to check if a component becomes zero and the components in $\bar{\bar{u}}(h)$ to check if a component becomes equal to ± 1 whatever happens first.

If a component in $\bar{x}(h)$ becomes zero, one moves it from $\bar{x}$ to $\bar{\bar{x}}$ and modifies accordingly the partition of A and u. If a component in $\bar{\bar{u}}(h)$ becomes, say, $+1$, one moves the corresponding component from $\bar{\bar{x}}$ to $\bar{x}$, and modifies accordingly the partition of A and u. the initialization is easy of the procedure is quite easy since for $h > \|A^T b\|_\infty$ the optimum is at the origin and the first component to become non-zero is x_j with $j = \arg\max_i |a_i^T b|$ and this happens for $h = \|A^T b\|_\infty$ which is thus the upper-boundary of the first interval. These results are known [9, 10, 11, 8].

3.2. The case p=∞, q=1

The optimality conditions are

$$Ax - b = -\rho_\infty v \quad \text{and} \quad A^T\delta = u$$
$$\text{for some } \ u \in \partial\|x\|_1 \text{ and } v \in \partial\|\delta\|_1$$

We shall try to derive a algorithm from these equations that proceeds as for the case $p = q = 2$. The task is more difficult.

The first step is to observe both programs (Opt$_\infty$) and (DOpt$_\infty$) can be transformed into linear programs and to use this observation to show that if for the current ρ_∞, the optimum $x(\rho_\infty)$ has $p \leq n \leq m = \dim(x)$ non zero components then the same holds generically for $\delta(\rho_\infty)$. Both x and δ are thus sparse and partitioning x as above and δ in a similar way, we will define *intervals* in ρ_∞ which are such that these partitions change only at the boundaries.

Looking at both programs (Opt$_\infty$), (DOpt$_\infty$) and at the optimality conditions, one can deduce that $\bar{x}(\rho_\infty)$ varies linearly in ρ_∞ locally, i.e., each component is piecewise linear and that $\delta(\rho_\infty)$ remains constant within each *interval*.

So far we only considered x-induced partitions of A in terms of columns we will now need in addition δ-induced partitions of A in terms of rows. We will partition both δ (and v) into $\underline{\delta}$ and $\underline{\underline{\delta}} = 0$ and accordingly the rows of A (and $\bar{A}$) into $\underline{A}$ and $\underline{\underline{A}}$.

Assume we have the optimum $x,\ u,\ \delta,\ v$ for a given ρ_∞ we extend them within an interval in ρ_∞. By definition $\bar{u}$ =sign$\bar{x}(\rho_\infty)$ and $\underline{v}$ =sign$\underline{\delta}$ remain constant within an interval, but since δ also remains invariant it follows from $A^T\delta = u$ that the whole vector u remains constant.

Introducing the different partitions, we get successively from $Ax - b = -\rho_\infty v$

$$\bar{A}\bar{x} = b - \rho_\infty v \;\;\Rightarrow\;\; \underline{\bar{A}}\bar{x} = \underline{b} - \rho_\infty \underline{v}$$
$$\Rightarrow\;\; \bar{x}(\rho_\infty) = \underline{\bar{A}}^{-1} b - \rho_\infty \underline{\bar{A}}^{-1} \underline{v}$$

where we assume $\underline{\bar{A}}^{-1}$ to exist. Substituting $\bar{x}$, it follows then from $\underline{\underline{\bar{A}}}\bar{x} = \underline{\underline{b}} - \rho_\infty \underline{\underline{v}}$, that

$$\underline{\underline{v}}(\rho_\infty) = \frac{1}{\rho_\infty}\underline{\underline{b}} - \frac{1}{\rho_\infty}\underline{\underline{\bar{A}}}\,\underline{\bar{A}}^{-1}\underline{b} + \underline{\underline{\bar{A}}}\,\underline{\bar{A}}^{-1}\underline{v}$$

In summary as ρ_∞ varies within the current interval, only $\bar{x}(\rho_\infty)$ and $\underline{\underline{v}}(\rho_\infty)$ are varying, all the remaining vectors in $x,\ u,\ \delta,\ v$ are invariant. When ρ_∞ increases (decreases) two things can happen a component in $\underline{\underline{v}}(\rho_\infty)$ becomes equal to ± 1 or a component in $\bar{x}(\rho_\infty)$ becomes zero. The upper bound (lower bound) of the current interval is the ρ_∞ denoted ρ_u (ρ_l) associated with the event that happens first.

If a component in $\underline{\underline{v}}(\rho_\infty)$ becomes say $+1$ this means that the corresponding component in δ, which was zero becomes positive, the row in $\underline{\underline{\bar{A}}}$ for which this happens is removed from $\underline{\underline{\bar{A}}}$ and added to $\underline{\bar{A}}$ which becomes say $\underline{\bar{A}}_a$ ($_a$ for augmented), it remains to add to this matrix a column drawn from $\underline{\bar{\bar{A}}}_a$ to get the new augmented square matrix. To choose this column one has to solve

$$\min_x \|x\|_1 \;\text{ s.t. }\; \underline{A}_a x = \underline{b}_a - \rho_l^- \underline{v}_a$$

where ρ_l^- is just slightly smaller than the lower bound ρ_l. By continuity, the optimum for ρ_l would have say p non zero components those associated with the columns in $\underline{\bar{\bar{A}}}$ while for ρ_l^- an additional component is just emerging.

If a component in $\bar{x}(\rho_\infty)$ becomes zero, for $\rho_\infty = \rho_u$, this means that a component in $\bar{x}$ has to be removed from the selection, one removes the corresponding columns from $\bar{A}$ which becomes say $\bar{A}^r$ (^{r} for reduced) and in order to detect the corresponding component in $\underline{\delta}$ that will become zero we solve

$$\min_\delta \|\delta\|_1 \;\text{ s.t. }\; \bar{A}^{rT}\delta = \bar{u}^r$$

again this amounts to remove a line in $\underline{\bar{\bar{A}}}$ in which one has already removed a column. We have thus indicated how to obtain both the boundaries and the modification that need to be done to cross them, it remains to indicate how to initialize the procedure again this is easy since for $\rho_\infty > \|b\|_\infty$ the optimum is at the origin and the first component of x to become non-zero as ρ_∞ is x_j with $j = \arg\max_i |a_{i_1,i}|$ with $i_1 = \arg\max_i |b_i|$.

3.3. The case p=1 and q=∞

A similar analysis can be performed in that case, as for $p = \infty$ one first draws some information from the fact that both problems are LP's and proceeds in a way that is quite similar to what we did above.

4. CONCLUSIONS AND PERSPECTIVES

We have considered simultaneously different optimization criteria (Opt$_p$) that allow to recover sparse representations in the presence of noise. We have defined the dual problems (Dopt$_p$) in the Lagrangian sense and detailed the optimality conditions satisfied by their optimum (5). Due to the presence of the ℓ_1-norm in all of these criteria they have some very specific properties. This has been observed first for the case $p = 2$ where though (Opt$_2$) or an equivalent criterion can be transformed into a quadratic program it can be solved in a far more computationally efficient way by using its very specific structure. Our objective was to analyze if a similar property holds for $p = \infty$ or $p = 1$. For these cases the criteria can be transformed into linear programs and though the ℓ_1-norm induced specific structure can somehow be exploited it is no clear if a similar benefit can be expected from it. Further investigations are required.

5. REFERENCES

[1] D.L. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Trans. on I.T.*, 47, 11, 2845-2862, Nov. 2001.

[2] M. Elad an A.M. Bruckstein. A generalized uncertainty principle and sparse representation in pairs of bases. *IEEE Trans. on I.T.*, 48, 9, 2558-256, Sept. 2002.

[3] R. Gribonval and M. Nielsen. Sparse representations in unions of bases. *IEEE Trans. on I.T.* 49, 12, 3320-3325, Dec. 2003.

[4] J.J. Fuchs. More on sparse representations in arbitrary bases. *IEEE Trans. on I.T* 50, 6, 1341-1344, June 2004

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

[6] J.A. Tropp, "Greed is good: Algorithmic Results for Sparse Approximations," *IEEE Trans. on I.T.*, 50, 10, 2231-2242, Oct. 2004.

[7] J.J. Fuchs, "Recovery of exact sparse representations in the presence of noise," Proceedings of the *IEEE ICASSP* Conference, vol. II, pp. 533-536, Montreal, May 2004.

[8] J.J. Fuchs. Recovery of exact sparse representations in the presence of bounded noise *IEEE Trans. on I.T.* 51, 10, 3601-3608, Oct. 2005.

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

[10] M. R. Osborne, B. Presnell, and B. A. Turlach, A new approach to variable selection in least squares problems. *IMA Journal of Numerical Analysis*, 20, 3, 389-403, 2000.

[11] D.M. Malioutov, M. Cetin and A.S. Willsky. "Homotopy continuation for sparse signal representation" *Philadelphia, IEEE ICASSP*, 2005.

[12] D.L. Donoho, M. Elad and V. Temlyakov, "Stable Recovery of Sparse Overcomplete Representations in the Presence of Noise," submitted to *IEEE Trans. on I.T.*, Feb. 2004.

[13] J.A. Tropp, "Just relax: Convex Programming Methods for Identifying Sparse Signals in Noise" submitted to *IEEE Trans. on I.T.*,revised Feb. 2005.

[14] R. Fletcher. Practical methods of optimization. *John Wiley and Sons*, 1987.