Spread representations

Jean-Jacques Fuchs

SPREAD REPRESENTATIONS.

Jean-Jacques FUCHS

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

ABSTRACT

Sparse representations, where one seeks to represent a vector on a redundant basis using the smallest number of basis vectors, appears to have numerous applications. The other extreme, where one seeks a representation that uses all the basis vectors, might be of interest if one manages to spread the information nearly equally over all of them. Minimizing the ℓ_∞-norm of the vector of weights is one way the find such a representation. Properties of this solution and dedicated fast algorithms allowing to find it are developed. Applications are to be found in robust data coding and improving achievable data rates over amplitude constrained channels.

Index Terms— anti-sparse representations, fast algorithms

1. INTRODUCTION

A (full rank) under-determined system of linear equations has infinitely many solutions and recently much interest has been given to finding the sparsest among them. Because the sparsest solution has many applications, dedicated, more or less fast algorithms have been developed to approximate this goal which indeed can only be attained by an exhaustive and hence generally unfeasible search.

All the remaining solutions are indeed non sparse and generally use all the basis vectors. It is known that expansions on such redundant bases withstand noise and quantization on the coefficients better [1] than orthogonal expansions. A category of such anti-sparse solutions that presents an interest a priori consists in the solutions for which the information content is somehow evenly spread on all vectors. Ideally one could search for the solution for which the absolute values of the components occupy the smallest range, say $(x_{\max} - x_{\min})$. Since this is a difficult non convex problem, one often simply seeks a solution for which $x_{\max}$ is small.

These representation where all the coefficients are of the same order of magnitude or more precisely where the range of the coefficients is small, are of high interest in coding and compression. They are known to withstand errors in their coefficients in a strong way [2]. One can show that the representation error due to quantification, transmission errors or losses, gets bounded by the average, rather than the sum, of the errors in the coefficients.

Representations in which the range of the coefficients is small, have already been considered and are known as Kashin's representation [2]. Minimizing the ℓ_∞-norm of the solution pushes the same idea even further since one explicitly minimizes the range of the coefficients while in the Kashin's this is only done in a loose way. For this optimality to be worthwhile the additional computational cost has to be small, hence the need for fast and dedicated algorithms that are developed below.

2. THE MODEL

Let A be a (n,m) full rank matrix with $m > n$ and columns normalized to one in Euclidean norm. For any $b \in R^n$, the linear system $Ax = b$ has then infinitely many solutions x where generically all components are non zero. Those having the smallest dynamical range are remarkable and have interesting properties in terms of coding or compression. If the range is measured by $\|x\|_\infty = \max_i |x_i|$, the ℓ_∞-norm of x, one should indeed consider solving

$$\min \|x\|_\infty \quad \text{under} \quad Ax = b. \tag{1}$$

It appears to be advantageous to replace the arbitrary matrix A by a tight frame matrix U whose rows are orthonormal and for which one therefore has $UU^T = I_n$. The basic frame representation is then associated with $x = U^T b$ and is such that $\|x\|_2 = \|b\|_2$.

Quite generally, with $\|b\|_2$ the energy in the vector b to be represented by a vector x with $\|x\|_2 = \|b\|_2$ on a m-dimensional redundant basis formed by the columns of a tight frame, the best one can attain, when the energy is equally spread is that each component x_i has absolute value $\|b\|_2/\sqrt{m}$.

Hence the following definition:

Kashin's representation with level K:

An expansion of b in terms of the m columns u_i of U, is called a Kashin's representation with level K of b if

$$b = \sum_1^m x_i \, u_i, \qquad \text{with} \quad \max_i |x_i| \leq \frac{K}{\sqrt{m}} \|b\|_2. \qquad \square$$

From our previous observation, it follows that one can expect that $K \geq 1$ be close to one. In the next section, we briefly sketch part of the work presented in [2] related to this area, we then develop an algorithm that allows to obtain the solution to (1) in an iterative way before we conclude.

3. OBTENTION OF A KASHIN'S REPRESENTATION

Provided the frame matrix U satisfies some additional conditions such as the uncertainty principle (UP) for matrices (see below), it is possible to convert the basic frame representation into a Kashin's representation with all m coefficients guaranteed to be smaller than a fixed constant.

Uncertainty principle (UP) for matrices:

The (n,m) matrix U satisfies the uncertainty principle with parameters η, δ both in $(0,1)$ if

$$|\text{supp}(x)| \leq \delta m \quad \Rightarrow \quad \|Ux\|_2 \leq \eta \|x\|_2. \qquad \square$$

With u_i the columns of U, one can rewrite the principle as

$$\|\sum_{i\in\Omega} u_i x_i\|_2 \leq \eta (\sum_{i\in\Omega} \|x_i\|^2)^{1/2}$$

for any subset Ω of cardinal smaller than δm and the basic idea in [2] is to improve iteratively upon the basic frame solution by projection of the residual sequence on a ℓ_∞-cube in R^m of decreasing radius so that the solution converges to a Kashin's representation.

One starts with $b_0 = b = Ux_0$ and truncates its coefficients $x_0 = U^T b_0$ at level $M = \|b\|_2/[\sqrt{\delta m}$ to build $t(b_0)$, one then defines $b_1 = b_0 - t(b_0)$ the first residual vector. With this specific M, it has a sparse representation with support of cardinal smaller than δm and applying (UP) it follows that

$$\|b_1\|_2 = \|b_0 - t(b_0)\|_2 \leq \eta \|b\|_2 = \eta \|b_0\|_2.$$

One now represents b_1 and truncates its coefficients at level ηM and the same reasoning now applies to $b_2 = b_1 - t(b_1)$ and one iterates the procedure. The process converges since $\|b_k\|_2 = \eta^k \|b_0\|_2$ and the residual thus tends toward zero. From $b = \sum_k t(b_k)$ and the slightly decreasing truncation level, it follows that all the coefficients in this representation of b have absolute value smaller than $M/(1-\eta)$ and thus a Kashin's representation of level $K = \delta^{(-1/2)}/(1-\eta)$.

The problem with this approach is that one needs to dispose of a frame matrix with known (and certified) constants $(\delta,\ \eta)$ from which one then deduces a conservative Kashin constant K that is used to tune the algorithm that turns the basic frame solution to a Kashin's solution. These are quite restrictive prerequisites that limit the feasibility and the performance of the approach. It is fair to say that for reasonable redundancy factors, say $\lambda = m/n < 2$, the potential gains are only attainable by the optimal strategy (1) described below and are only a theoretical perspective in Kashin's approach.

4. MINIMIZING THE INFINITE NORM

4.1. Generalities

The optimization problem (1) can be transformed into a linear program and solved using the simplex or interior point methods. One can also consider the optimization problem, parametrized by $h \in R^+$:

$$\min \frac{1}{2}\|Ax - b\|_2^2 + h\|x\|_\infty, \tag{2}$$

which can be transformed into a quadratic program. Its optimum say $x^*(h)$ converges to the optimum x^* of (1) when h decreases to zero. From this last observation, an optimization algorithm that converges to the solution of (1) in a number of steps much smaller than the number of steps required by, say, the simplex algorithm, will be developed. It is a path-following method similar to the one presented in [2] and is also related to the continuation techniques, which have been studied in the optimization literature [3].

Though it will not be used in the sequel, it is interesting to note that using basic Linear Programming theory, one can establish the following result.

Proposition: Generically, the optimum x^* of (1) has $m-n+1$ components equal to $\pm\|x^*\|_\infty$ and the $n-1$ remaining ones in between these two extreme values. $\square$

This result makes sense since the n equalities in $Ax = b$ allow to fix the n degrees of freedom consisting in $\|x^*\|_\infty$ and the $n-1$ components in between these two extreme values. Indeed if one wants to further diminish the spread, one might think about solving

$$\min_{x,v,u} v - u, \qquad \text{under} \quad Ax = b, \quad 0 \leq u \leq |x| \leq v$$

but this problem is not convex and difficult to solve one thus replaces it by

$$\min_{x,v,u} v - u, \qquad \text{under} \quad Ax = b, \quad 0 \leq u \leq x \leq v$$

that is convex and from LP theory, it follows that generically there are $n-2$ components strictly in between the two extreme values u and v. A result that again makes sense and can be deduced from the same reasoning as above.

4.2. Optimality conditions

The problem (2) is a convex program that can be transformed into a quadratic program. One can rewrite it as

$$\min \frac{1}{2}\|Ax - b\|_2^2 + ht$$

$$\text{s.t. } x = x^+ - x^-, \quad 0 \leq x^+, \ x^- \leq t\mathbf{1}$$

whose dual can be shown to be

$$\min \|Ax\|_2^2 \quad \text{s.t.} \quad \|A^T(Ax - b)\|_1 \leq h. \tag{3}$$

Note the dual of (1) is $\max_d b^T d$ under $\|A^T d\|_1 \leq 1$.

To be able to characterize easily the conditions satisfied by the optimum of (2), we introduce $\partial f(x)$ the sub-differential of a convex function [5] f at a point x, it is a set of vectors called the sub-gradients of f at x. For $f(x) = \|x\|_\infty$ one has

$$\partial\|x\|_\infty = \{v|\ |x_i| = \|x\|_\infty \Rightarrow x_i v_i \geq 0, |x_i| < \|x\|_\infty \quad (4)$$
$$\Rightarrow v_i = 0;\ \|v\|_1 = 1 \text{ if } x \neq 0, \|v\|_1 \leq 1 \text{ else}\}$$

Note that if f is differentiable at x then $\partial f(x)$ reduces to the gradient.

Since (2) is a convex program the first order optimality conditions (zeroing the sub-differential) are necessary and sufficient conditions for optimality and one thus gets

Lemma 1. The optimum of (2) is x iff the vector 0 is a sub-gradient of the criterion at x, i.e., iff :

$$A^T(Ax - b) + hv = 0 \quad \text{for some} \quad v \in \partial\|x\|_\infty \quad \diamond \quad (5)$$

4.3. Some specific notations

To exploit these conditions in which some parts of v are not uniquely defined, we need to introduce some notations. Let us denote x_∞ the the ℓ_∞ norm of x. To take care of $v \in \partial\|x\|_\infty$, we partition the optimal x into $\bar{x}$ its q $middle$-components associated with $\bar{v} = 0$ and $\bar{\bar{x}}$ the remaining components that are equal to $\pm\ x_\infty$ associated with $\bar{\bar{v}}$. For non-zero x one then has $\|\bar{\bar{v}}\|_1 = 1$, $\bar{\bar{x}}_i^T \bar{\bar{v}}_i \geq 0$, $\bar{\bar{v}}^T \bar{\bar{x}} = \|x\|_\infty$ and thus generically sign($\bar{\bar{v}}$)=sign ($\bar{\bar{x}}$). The above defined partition of x induces similarly the partition of v, we already introduced but also the partition of the (columns of) matrix A into $\bar{A}$ and $\bar{\bar{A}}$. One then has, for instance $Ax = \bar{A}\bar{x} + \bar{\bar{A}}\text{sign}(\bar{\bar{v}})x_\infty$

One can observe, that, provided its partition is known, x has only $q+1$ degrees of freedom the q components in $\bar{x}$ and $\|x\|_\infty$ we denote x_∞ for short.

One can now rewrite (5) in a more usable way as

$$A^T(\bar{A}\bar{x} + \bar{\bar{A}}\text{sign}(\bar{\bar{v}})x_\infty - b) + hv = 0,$$

which can be divided into two parts

$$\bar{A}^T(\bar{A}\bar{x} + \bar{\bar{A}}\text{sign}(\bar{\bar{v}})x_\infty) = \bar{A}^T b \quad (6)$$
$$\bar{\bar{A}}^T(\bar{A}\bar{x} + \bar{\bar{A}}\text{sign}(\bar{\bar{v}})x_\infty - b) = -h\bar{\bar{v}}. \quad (7)$$

4.4. Development

Provided $\bar{A}$ is a full (column) rank matrix, the relation (6) yields an expression of $\bar{x}$ in terms of x_∞ of the form, say,

$$\bar{x} = X_1 + X_2 x_\infty. \quad (8)$$

Pre-multiplying then (7) by sign$(\bar{\bar{v}})^T$ and replacing $\bar{x}$ by (8), yields an expression of h in terms of x_∞ of the form, say,

$$h = H_1 + H_2 x_\infty, \quad (9)$$

with

$$H_2 = -sign(\bar{\bar{v}})^T \bar{\bar{A}}^T\ (I - \bar{A}(\bar{A}^T\bar{A})^{-1}\bar{A}^T)\ \bar{\bar{A}}\ sign(\bar{\bar{v}}),$$

a negative real scalar. There is thus a one-to-one relation between h and x_∞ and as h decreases, x_∞ increases which is what one would expect. Replacing similarly $\bar{x}$ in (7) yields a relation of the form,

$$h\bar{\bar{v}} = V_1 + V_2 x_\infty. \quad (10)$$

The three expressions (8, 9, 10) are all one needs to extend an optimal x valid for a fixed h to its neighborhood. Indeed to extend the optimal x as h -or equivalently x_∞- varies, we need to guarantee that the quantities $\bar{x}(x_\infty)$, $\bar{\bar{v}}(x_\infty)$ and $h(x_\infty)$ we propose, satisfy (5) or equivalently (6) and (7).

And the three expressions we have obtained do exactly that as long as they are valid, i.e., as long as the components in $\bar{x}$ are smaller than x_∞ and as long as in $\bar{\bar{v}}$ no components becomes zero.

As x_∞ increases the first value of x_∞ for which one of these two events happens defines the upper bound of the interval in x_∞, and similarly lower bound of the interval in h, in which one can extend the current optimum.

It remains then to start the procedure, i.e., to get the optimal triplet in a first interval and to indicate how to cross such a boundary, i.e., how to get these same optimal expressions within the next interval.

4.4.1. The initial step

For h large, the optimal x is at the origin. Indeed $x = 0$ and $v = A^T b/h$ satisfies (5) as long as $h \geq \|A^T b\|_1$ which is thus the first boundary value, we denote h^0. These observations follow trivially from the dual (3) problem of (2). From the expansion of the criterion in (2) around the origin

$$f(x) \simeq -b^T Ax + h\|x\|_\infty,$$

it also appears that the most efficient way to diminish the cost which is equal to $b^T b/2$ for $x = 0$ is to take $x =$sign$(A^T b)\alpha$ and that taking the scalar α positive and small is beneficial only if $h \leq \|A^T b\|_1$.

For h within the first interval $[h^1,\ h^0]$, with h^1 yet to be defined, one has sign(x) =sign(v)=sign$(A^T b)$ and $x = \bar{\bar{x}} =$ sign$(\bar{\bar{v}})x_\infty$. In this very specific first interval the only possible event that can happen is a component in $\bar{\bar{v}}$ becoming zero. From (7) with $\bar{A}$ missing, one gets (10) and x^1_∞ is the smallest component in the vector $-V_1./V_2$ that is greater than $x^0_\infty = 0$. One then deduces h^1 from (9) and if it is $\bar{\bar{v}}_{j_1}$ that became zero, one must change accordingly the partition, i.e., remove column a_{j_1} from $\bar{\bar{A}}$ and introduce it into $\bar{A}$ which was empty so far. The number q of middle components in the optimal x is now equal to one. And we enter the general step unless h^1 is smaller than, say h_d , the h for which one seeks the solution.

4.4.2. The standard step

As x_∞ increases from its current (boundary) value say x_∞^-, we seek the event that happens first among

$\diamond$ a component in $\bar{x}$ in (8) becomes equal to $\pm x_\infty$ or

$\diamond$ a component in $\bar{\bar{v}}$ in (10) becomes zero,

and denote x_∞^+ the value of x_∞ for which it happens. In the first case, one seeks the smallest component in the vector $X_1./(1 - X_2)$ or in $-X_1./(1 + X_2)$ that is greater than x_∞^- in the second case one similarly inspects $-V_1./V_2$. The new boundary value is the smallest of these three values. One then computes the associated value of h^+ using (9).

If h^+ is smaller than h_d, one deduces the associated x_∞ using (9) and replaces it into (8) to build $\bar{x}$ and thus the optimal $x(h_d)$, otherwise one changes the partition, in the first case q decreases by one and one moves one column from $\bar{A}$ to $\bar{\bar{A}}$ and adds the sign of this new component of $\bar{\bar{x}}$ to sign($\bar{\bar{v}}$), in the second case q increases by one and one moves a column the other way. This concludes the standard step and the description of the dedicated algorithm

As a matter of fact, one can ignore the initialization step by simply initializing the standard step with $\bar{\bar{A}} = A$, $\bar{A} = []$, sign($\bar{\bar{v}}$) = sign($A^T b$), $h^- = \|A^T b\|_1$ and $x_\infty^- = 0$.

5. CONCLUSIONS

Besides their robustness against noise, representations with limited dynamical range are well adapted to improve the achievable data rate over amplitude constrained channels. Indeed while power constrained channels are well investigated, it might be more realistic to consider constraints on the amplitude to avoid the damaging effects of non linearities often present at higher amplitudes. Since, on the other hand, redundancy causes a direct loss in the data rate, only low redundancy factors are of interest and this is precisely where the optimal strategy considered here is the only one that allows to achieve the potential gains.

Indeed it appears that even this 'optimal' strategy fails to make this approach viable. So while the application of such models in robust data coding and in improving achievable data rates over amplitude constrained channels seems to be wishful thinking, its use in indexing techniques appears to be quite promising. In this context, one further replaces the optimal vector by its sign vector (potentially associated with a re-evaluated scalar weight) to get a binary vector that is not only cheap to store and (somehow) easy to search for but also allows for an explicit reconstruction unlike all other Hamming embedding functions used to map real vectors into binary vectors [6, 7].

6. REFERENCES

[1] Z. Cvetkovic, "Resilience properties of redundant expansions under additive noise and quantization," IEEE-T-IT, 49, 7, 644-656, 2003.

[2] Y. Lyubarskii and R. Vershynin, "Uncertainty principles and vector quantization," IEEE-T-IT, 56, 7, 3491-3501, July 2010.

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

[4] E. Allgower and K. Georg. "Continuation and path following" Acta Numerica, 2, 31-64, 1993.

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

[6] A. Torralba, R. Fergus and Y.Weiss, "Small codes and large databases for recognition," in CVPR, June 2008.

[7] Y.Weiss, A. Torralba and R. Fergus, "Small codes and large databases for recognition," in NIPS, 2008.