Journal of Uncertain Systems
Vol.7, No.3, pp.203-207, 2013
Online at: www.jus.org.uk

# Why $\ell_1$ is a Good Approximation to $\ell_0$: A Geometric Explanation

Carlos Ramirez[1,*], Vladik Kreinovich[1,2], Miguel Argaez[1,3]

[1] *Computational Sciences Program, University of Texas at El Paso, El Paso, TX 79968, USA*

[2] *Department of Computer Science, University of Texas at El Paso, El Paso, TX 79968, USA*

[3] *Department of Mathematical Sciences, University of Texas at El Paso, El Paso, TX 79968, USA*

Received 1 April 2012; Revised 16 April 2012

**Abstract**

In practice, we usually have partial information; as a result, we have several different possibilities consistent with the given measurements and the given knowledge. For example, in geosciences, several possible density distributions are consistent with the measurement results. It is reasonable to select the simplest among such distributions. A general solution can be described, e.g., as a linear combination of basic functions. A natural way to define the simplest solution is to select one for which the number of the non-zero coefficients $c_i$ is the smallest. The corresponding "$\ell_0$-optimization" problem is non-convex and therefore, difficult to solve. As a good approximation to this problem, Candès and Tao proposed to use a solution to the convex $\ell_1$ optimization problem $\sum |c_i| \to \min$. In this paper, we provide a geometric explanation of why $\ell_1$ is indeed the best convex approximation to $\ell_0$.


**Keywords:** $\ell_0$-norm, $\ell_1$-norm, sparse representation, Ockham razor

## 1 Formulation of the Problem

**Need to select a solution from all solutions consistent with the observations.** In each practical situations, we usually have partial information; as a result, we have several different possibilities consistent with the given measurements and the given knowledge.

For example, in many practical situations, we want to find out how a certain quantity changes from one spatial location to another: in geophysics, we want to find out the density $\rho(x)$ at different spatial locations $x$; in meteorology, we want to find out the temperature, wind speed and wind direction at different spatial locations, etc. From the mathematical viewpoint, what we want to find out is a function.

To exactly describe a general function, we need to know the values of infinitely many parameters. For example, functions can be represented as a linear combination $\sum c_i \cdot e_i(x)$ of functions from some basis $\{e_i(x)\}$. Elements of this basis can be monomials (in Taylor series), sines and cosines (in Fourier series), etc. So, to determine a function, we must find all these parameters $c_i$ from the results of measurements and observations. At any given moment of time, we only have finitely many measurement and observation results. So, we only have finitely many constraints on infinitely many parameters. In general, when we have a system of equations in which there are more unknown than equations, this system allows multiple solutions; this is a well-known fact for generic linear systems, it is a known fact for generic non-linear systems as well. So, several different solutions are consistent with all the measurement results.

For example, in geosciences, several possible density distributions are consistent with the measurement results. Scientists usually want us not only to present them with the set of all possible solutions, but also want us to select one of these solutions as the most "reasonable" one – in some natural sense.

**Occam's razor: idea.** One of the ways to select a solution is to select the solution which is, in some reasonable sense, the simplest among all possible solutions consistent with all the observations.

*Corresponding author. Email: carlosrv19@gmail.com (C. Ramirez).

**$\ell_0$-solutions as a natural formalization of Occam's razor.** As we have mentioned, a general function can be described, e.g., as a linear combination of basic functions. In this representation, a natural way to define the simplest solution is to select one for which the number of the non-zero coefficients $c_i$ is the smallest. This number is known as the $\ell_0$-*norm* $\|c\|_0 \stackrel{\text{def}}{=} \#\{i : c_i \neq 0\}$.

**$\ell_0$-solutions are difficult to compute.** The $\ell_0$-norm is non-convex. It is known that non-convex optimization problems are computationally difficult to solve exactly; see, e.g., [8]. Not surprisingly, the $\ell_0$-optimization problem is also computationally difficult: it is known to be NP-hard; see, e.g., [2, 3, 4, 6].

**How to solve non-convex optimization problems.** The difficulty of solving non-convex optimization problems is caused by the non-convexity of the corresponding objective function. For convex objective function, there exist feasible optimization algorithms; see, e.g., [8]. Because of this, one of the possible ways to solve a non-convex optimization is to solve a similar *convex* optimization problem. This idea is known as *convex relaxation*.

**$\ell_1$-solutions as a good approximation to $\ell_0$.** For $\ell_0$-problems, as a good convex approximation, Candès and Tao proposed to use a solution to the corresponding convex $\ell_1$ optimization problem, i.e., to find the values of all the coefficients $c_i$ for which the $\ell_1$-norm $\|c\|_1 \stackrel{\text{def}}{=} \sum |c_i|$ is the smallest possible; see, e.g., [1, 2, 3, 4, 5, 7].

**Challenge.** The idea of replacing the original $\ell_0$-problem with the corresponding $\ell_1$-problem was based on the result – described in [2, 3, 4] – that under certain conditions, $\ell_1$-optimization provides us with the solution to the original $\ell_0$-problem.

However, in practice, the $\ell_1$-approximation to the original $\ell_0$-problem is used way beyond these conditions. As a result, we often get a solution which is not exactly minimizing the original $\ell_0$-norm, but which provides much smaller values of the $\ell_0$-norm than other known techniques.

In such situations, the use of $\ell_1$ norm is purely heuristic, not justified by any proven results. It is therefore desirable to provide a mathematical explanation for the success of $\ell_1$-approximation to $\ell_0$-optimization.

**What we do in this paper.** In this paper, we provide a geometric explanation for the empirical success of $\ell_1$-approximation to the $\ell_0$-problems.

## 2 Geometric Justification of $\ell_1$-Norm

**Important observation: we need $\ell_\varepsilon$, not $\ell_0$.** Intuitively, if we can decrease the absolutely value $|c_i|$ of one of the coefficients without changing other coefficients, we get a simpler sequence. The original $\ell_0$-norm does not capture this difference, since, e.g., sequences $(10, 10, 0, \ldots, 0)$ and $(10, 1, 0, \ldots, 0)$ have the exact same $\ell_0$-norm equal to 2. To capture this difference, it is reasonable to use an $\ell_\varepsilon$-norm $\|c\|_\varepsilon \stackrel{\text{def}}{=} \sum |c_i|^\varepsilon$ for some small $\varepsilon > 0$.

This new norm captures the above difference: e.g., $\|(10, 1, 0, \ldots, 0)\|_\varepsilon = 10^\varepsilon + 1 < \|(10, 10, 0, \ldots, 0)\|_\varepsilon = 2 \cdot 10^\varepsilon$. On the other hand, when $\varepsilon \to 0$, we have $|c_i|^\varepsilon \to 0$ when $c_i = 0$ and $|c_i|^\varepsilon \to 1$ when $c_i \neq 0$, so $\|c\|_\varepsilon \to \|c\|_0$. Thus, for sufficiently small $\varepsilon$, the value $\|c\|_\varepsilon$ is practically indistinguishable from $\|c\|_0$. Because of this, in practice, instead of the $\ell_0$-norm, a $\ell_\varepsilon$-norm corresponding to some small $\varepsilon > 0$ is actually used.

**Towards formalizing the problem.** Our objective is to select, among all possible combinations $c = (c_1, \ldots, c_n)$ which are consistent with observations, the one which is, in some sense, most reasonable. In other words, we need to be able, given any two combinations $c = (c_1, \ldots, c_n)$ and $c' = (c'_1, \ldots, c'_n)$, to decide which combination is better. In precise terms, we need to describe a total (= linear) *pre-ordering* relation $\leq$ on the set $\mathbb{R}^n$ of possible combinations, i.e., a transitive relation for which for every $c$ and $c'$, either $c \leq c'$ ($c$ is better or of the same quality as $c'$) or $c' \leq c$. As usual, we will use the notation $c < c'$ when $c \leq c'$ and $c' \not\leq c$, and $c \equiv c'$ when $c \leq c'$ and $c' \leq c$.

If we use an objective function $f(c_1, \ldots, c_n)$, then the relation $(c_1, \ldots, c_n) \leq (c'_1, \ldots, c'_n)$ takes the form $f(c_1, \ldots, c_n) \leq f(c'_1, \ldots, c'_n)$.

**Natural requirements.** As we have mentioned, if we decrease one of the absolute values $c_i$, we should get a better solution. It also makes sense to require that the quality does not depend on permutations and that the relative quality of two combination does not change if we simply use different measuring units (i.e., replace $c = (c_1, \ldots, c_n)$ with $\lambda \cdot c = (\lambda \cdot c_1, \ldots, \lambda \cdot c_n)$).

It is also reasonable to require that the relation is *Archimedean* in the sense that for every two combinations $c \neq 0$ and $c' \neq 0$, there exists a $\lambda > 0$ for which $\lambda \cdot c \equiv c'$. Indeed, when $\lambda = 0$, we have $\lambda \cdot c = 0 \leq c'$; for very large $\lambda$, we have $c' \leq \lambda \cdot c$; thus, intuitively, there should be an intermediate value $\lambda$ for which $\lambda \cdot c$ and $c'$ are equivalent.

**Definition 1.** *A linear pre-ordering relation $\leq$ on $\mathbb{R}^n$ is called:*

- natural *if for all values $c_1, \ldots, c_{i-1}, c_i, c_i', c_{i+1}, \ldots, c_n$, if $|c_i| < |c_i'|$, then*

$$(c_1, \ldots, c_{i-1}, c_i, c_{i+1}, \ldots, c_n) < (c_1, \ldots, c_{i-1}, c_i', c_{i+1}, \ldots, c_n);$$

- permutation-invariant *if $(c_1, \ldots, c_n) \equiv (c_{\pi(1)}, \ldots, c_{\pi(n)})$ for every $c$ and for every permutation $\pi$;*
- scale-invariant *if $c \leq c'$ implies $\lambda \cdot c \leq \lambda \cdot c'$;*
- Archimedean *if for every $c \neq 0$ and $c' \neq 0$, there exist a real number $\lambda > 0$ for which $\lambda \cdot c \equiv c'$.*

It turns out that to describe each such pre-order can be uniquely determined by a set:

**Proposition 1.** *A natural Archimedean pre-order $\leq$ is uniquely determined by the set*

$$B_{\leq} \stackrel{\text{def}}{=} \{c : c \leq (1, 0, \ldots, 0)\}.$$

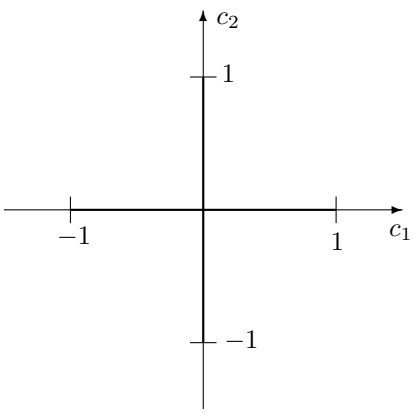
**Proof.** Indeed, since $\leq$ is Archimedean, for every vector $c$, there exists a value $\lambda(c)$ for which

$$c \equiv \lambda(c) \cdot (1, 0, \ldots, 0) = (\lambda(c), 0, \ldots, 0).$$

Due to the naturalness property, the smaller $\lambda(c)$, the better the corresponding vector $(\lambda(c), 0, \ldots, 0)$ and thus, the better the combination $c$: $c \leq c' \Leftrightarrow \lambda(c) \leq \lambda(c')$. Thus, to determine the pre-order $\leq$, it is sufficient to know the value $\lambda(c)$ for all $c$. One can easily see that this value, in turn, can be uniquely determined from the set $B_{\leq}$, as $\min\{k : c/k \in B_{\leq}\}$. The proposition is proven.

**Proposition 2.** *For a natural permutation-invariant scale-invariant Archimedean pre-order $\leq$, the set $B_{\leq}$ contains the set*

$$B_0 \stackrel{\text{def}}{=} \{(c_1, 0, \ldots, 0) : |c_1| \leq 1\} \cup \cdots \{(0, \ldots, 0, c_i, 0 \ldots, 0) : |c_i| \leq 1\} \cup \cdots \cup \{(0, \ldots, 0, c_n) : |c_n| \leq 1\}.$$



**Proof.** Let us show that every element of the set $B_0$ indeed belongs to $B_{\le}$.

Indeed, due to naturalness, when $|c_1| \le 1$, we have $(c_1, 0, \ldots, 0) \le (1, 0, \ldots, 0)$ and thus, $(c_1, 0, \ldots, 0) \in B_{\le}$. Due to permutation-invariance, for every $i$ and for every $c_i$ with $|c_i| \le 1$, we have $(0, \ldots, 0, c_i, 0, \ldots, 0) \equiv (c_i, 0, \ldots, 0)$. So, from $(c_i, 0, \ldots, 0) \le (1, 0, \ldots, 0)$, we conclude that $(0, \ldots, 0, c_i, 0, \ldots, 0) \le (1, 0, \ldots, 0)$ and thus, $(c_1, 0, \ldots, 0) \in B_{\le}$. The proposition is proven.

**How to describe approximation accuracy?** The $\ell_0$-norm is not convex, and we want to approximate it by a convex one, i.e., by a convex objective function $f(c_1, \ldots, c_n)$ which is, in some reasonable sense, the "most accurate" approximation to the $\ell_0$-norm. How can we describe approximation accuracy? According to Proposition 1, each pre-order $\le$ is uniquely determined by the corresponding set $B_{\le}$. Thus, it is reasonable to use the difference between the corresponding sets to gauge the approximation accuracy.

One can see that when $\varepsilon \to 0$, the set $\{c : \|c\|_\varepsilon \le \|(1, 0, \ldots, 0)\|_\varepsilon = 1\}$ tends to the above-defined set $B_0$. For the relation $\le$ corresponding to a convex function, the set

$$B_{\le} = \{c : c \le (1, 0, \ldots, 0)\} = \{c : f(c_1, \ldots, c_n) \le f(1, 0, \ldots, 0)\}$$

is also convex. In these terms, our goal is to find a convex set $B$ approximating the set $B_0$.

In general, the sets $S$ and $S'$ are equal when each element of the set $S$ belongs to $S'$ and each element of $S'$ belongs to $S$. Thus, the difference between two sets $S$ and $S'$ comes from elements which belong to $S$ but not to $S'$ (these elements form the difference $S - S'$) and the elements which belong to $S'$ but not to $S$ (these elements form the difference $S' - S$).

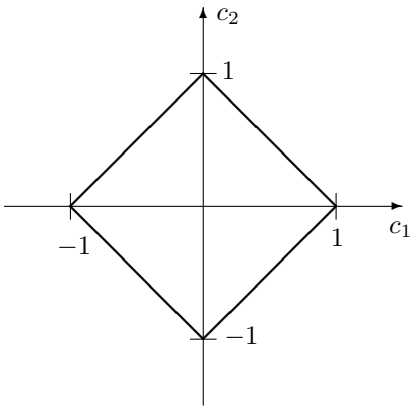
In our case, due to Proposition 2, each element of the set $B_0$ belongs to $B_{\le}$. Thus, the only difference between the sets $B_0$ (corresponding to $\ell_0$-norm) and the desired convex approximation $B$ is the difference $B - B_0$. So, we arrive at the following definition.

**Definition 2.** *We say that a convex set $B \supseteq B_0$ is a* better approximation *of the set $B_0$ than a convex set $B' \supseteq B_0$ if $B - B_0 \subset B' - B_0$ (and $B - B_0 \ne B' - B_0$).*

**Discussion.** The relation defined in Definition 2 is only a partial order, so it is not a priori clear that there is a convex set which is the best according to this criterion. However, the following result shows that such an optimal approximation does exist.

**Proposition 3.** *Out of all convex sets $B$ containing the set $B_0$, the best approximation to $B_0$ is the convex hull of $B_0$:*

$$B_1 = \mathrm{Conv}(B_0) = \{(c_1, \ldots, c_n) : \sum_{i=1}^{n} |c_i| = 1\}.$$




**Proof.** The proof is straightforward: every convex set containing $B_0$ contains its convex hull, so the convex hull is indeed the best approximation in the sense of Definition 2.

**Discussion.** Using the technique described in the proof of Proposition 1, we can see that the pre-order corresponding to the set $B_1$ is equivalent to minimizing the $\ell^1$-norm. Thus, *$\ell_1$-norm is indeed the best convex approximation to the $\ell_0$-norm.*

## Acknowledgments

This work was supported in part by the National Science Foundation grants HRD-0734825 and HRD-1242122 (Cyber-ShARE Center of Excellence) and DUE-0926721, by Grants 1 T36 GM078000-01 and 1R43TR000173-01 from the National Institutes of Health, and by a grant on F-transforms from the Office of Naval Research.

## References

[1] Argáez, M., C. Ramirez, and R. Sanchez, An $\ell_1$ algorithm for underdetermined systems and applications, *IEEE Proceedings of the 2011 Annual Conference on North American Fuzzy Information Processing Society*, pp.1–6, 2011.

[2] Candès, E.J., J. Romberg, and T. Tao, Stable signal recovery from incomplete and inaccurate measurements, *Communications on Pure and Applied Mathematics*, vol.59, pp.1207–1223, 2006.

[3] Candès, E.J., and T. Tao, Decoding by linear programming, *IEEE Transactions on Information Theory*, vol.51, no.12, pp.4203–4215, 2005.

[4] Donoho, D.L., Compressed sensing, *IEEE Transactions on Information Theory*, vol.52, no.4, pp.1289–1306, 2006.

[5] Elad, M., *Sparse and Redundant Representations*, Springer Verlag, 2010.

[6] Natarajan, B.K., Sparse approximate solutions to linear systems, *SIAM Journal on Computing*, vol.24, pp.227–234, 1995.

[7] Sánchez, R., M. Argáez, and P. Guillén, Sparse representation via $l^1$-minimization for underdetermined systems in classification of tumors with gene expression data, *Proceedings of the IEEE 33rd Annual International Conference of the Engineering in Medicine and Biology Society*, pp.3362–3366, 2011.

[8] Vavasis, S.A., *Nonlinear Optimization: Complexity Issues*, Oxford University Press, New York, 1991.