



## **CEL0: a continuous alternative to l0 penalty**

Emmanuel Soubies, Laure Blanc-Féraud, Gilles Aubert

► **To cite this version:**

Emmanuel Soubies, Laure Blanc-Féraud, Gilles Aubert. CEL0: a continuous alternative to l0 penalty. Signal Processing with Adaptive Sparse Structured Representations (SPARS), Jul 2015, Cambridge, United Kingdom. <<http://sigproc.eng.cam.ac.uk/SPARS2015>>. <hal-01167192>

**HAL Id: hal-01167192**

**<https://hal.inria.fr/hal-01167192>**

Submitted on 24 Jun 2015

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# CEL0: a continuous alternative to $\ell_0$ penalty.

Emmanuel Soubies and Laure Blanc-Féraud  
 UNS, CNRS I3S, UMR 7271, Sophia Antipolis, France.  
 {soubies,blancf}@i3s.unice.fr

Gilles Aubert  
 UNS, CNRS J.A.D, UMR 7351, Nice, France.  
 gaubert@unice.fr

**Abstract**—This paper presents a new way to address the NP-hard combinatorial  $\ell_2$ - $\ell_0$  problem by minimizing a continuous relaxed functional preserving the minimizers of the initial energy. We propose the *Continuous Exact  $\ell_0$  penalty (CEL0)*, an approximation of the  $\ell_0$  norm leading to a *tight continuous relaxation of the  $\ell_2$ - $\ell_0$  criteria whose global minimizers contain those of the  $\ell_0$  penalized least-squares functional*. Links between local minimizers of these two functionals are also investigated. This short communication summarizes the main results of our recent work [1].

## I. CONTEXT

In this work we deal with the following  $\ell_0$  penalized least squares problem

$$\hat{x} = \arg \min_{x \in \mathbb{R}^N} G_{\ell_0}(x) := \frac{1}{2} \|Ax - d\|_2^2 + \lambda \|x\|_0, \quad (1)$$

where  $A \in \mathbb{R}^{M \times N}$ ,  $M \ll N$ ,  $d \in \mathbb{R}^M$ ,  $\|\cdot\|_0$  denotes the  $\ell_0$ -counting “norm” and  $\lambda > 0$  is an hyperparameter allowing a trade-off between data fidelity and sparsity. This NP-hard combinatorial problem and its constrained form (not equivalent) are of fundamental importance in many applications such as coding, compressed sensing, source separation or variable selection.

During the last decades, many researchers proposed methods and algorithms, with some theoretical guaranties in terms of convergence, to find a good approximate solution of (1). The most popular are the  $\ell_1$  convex relaxation [2], the greedy algorithms [3] and the continuous nonsmooth nonconvex penalties widely used to replace the  $\ell_0$ -norm in (1). Particularly, some authors have recently proposed an exact reformulation of  $\ell_0$  regularized problems as DC programs [4]. The following work comes within this framework of exact reformulation were a new penalty, called *Continuous Exact  $\ell_0$  (CEL0)*, is proposed. This penalty leads to a *tight continuous relaxation* preserving the minimizers of  $G_{\ell_0}$ .

## II. MAIN CONTRIBUTIONS

Let us first introduce some notations used in the sequel:

- $\mathbb{I}_N = \{1, \dots, N\}$ ,
- $a_i \in \mathbb{R}^M$  denotes the  $i$ th column of  $A \in \mathbb{R}^{M \times N}$ ,
- $\sigma(x) := \{i \in \mathbb{I}_N; x_i \neq 0\}$  defines the support of  $x \in \mathbb{R}^N$ ,
- $\sigma^-(x) := \left\{ i \in \sigma(x) : |x_i| < \sqrt{2\lambda} / \|a_i\| \right\}$  a part of the support.

### A. Orthogonal case analysis

When the matrix  $A$  is orthogonal, one can compute analytically the biconjugate  $G_{\ell_0}^{**}$  which is the convex hull of  $G_{\ell_0}$ . Simple calculations [1, §3] lead to,

$$G_{\ell_0}^{**}(x) = \frac{1}{2} \|Ax - d\|_2^2 + \Phi_{\text{CEL0}}(x), \quad (2)$$

where  $\Phi_{\text{CEL0}}$  denotes the CEL0 penalty defined by

$$\Phi_{\text{CEL0}}(x) := N\lambda \sum_{i \in \mathbb{I}_N} \frac{\|a_i\|^2}{2} \left( |x_i| - \frac{\sqrt{2\lambda}}{\|a_i\|} \right)^2 \mathbb{1}_{\{|x_i| \leq \frac{\sqrt{2\lambda}}{\|a_i\|}\}}. \quad (3)$$

Therefore, in the case of orthogonal matrices, the convex hull of  $G_{\ell_0}$  is obtained by replacing the noncontinuous  $\ell_0$ -norm by the continuous CEL0 penalty given in (3). In this case, since  $G_{\ell_0}^{**}$  is convex, all minimizers are global and we can deduce from them the sparsest solution of (1) which is given by thresholding the entries of  $A^T d$  [1, §4.3]. The resulting thresholding rule can be seen as a generalization of the well-known hard thresholding rule [5].

However this result is false when  $A$  is not orthogonal. Indeed, in this case, replacing the  $\ell_0$  norm in (1) by the CEL0 penalty (3) leads to a nonconvex functional denoted  $G_{\text{CEL0}}$ . Nevertheless, this functional has interesting properties which are analyzed in the following.

### B. Links between minimizers of $G_{\text{CEL0}}$ and $G_{\ell_0}$

Let  $A$  be an arbitrary matrix of  $\mathbb{R}^{M \times N}$ . Based on the description of the minimizers of  $G_{\ell_0}$  given in [6], the two following theorems characterize the links between minimizers of  $G_{\ell_0}$  and  $G_{\text{CEL0}}$ . Proofs can be found in [1].

*Theorem 1:* Let  $d \in \mathbb{R}^M$  and  $\lambda > 0$ ,

- the set of global minimizers of  $G_{\ell_0}$  is included in the set of global minimizers of  $G_{\text{CEL0}}$ ,

$$\arg \min_{x \in \mathbb{R}^N} G_{\ell_0}(x) \subseteq \arg \min_{x \in \mathbb{R}^N} G_{\text{CEL0}}(x) \quad (4)$$

- conversely if  $\hat{x} \in \mathbb{R}^N$  is a global minimizer of  $G_{\text{CEL0}}$ , let  $\hat{x}^0$  be defined by

$$\forall i \in \mathbb{I}_N, \quad \hat{x}_i^0 = \hat{x}_i \mathbb{1}_{\{|\hat{x}_i| \geq \frac{\sqrt{2\lambda}}{\|a_i\|}\}}, \quad (5)$$

then  $\hat{x}^0$  is a global minimizer of  $G_{\ell_0}$  and

$$G_{\text{CEL0}}(\hat{x}) = G_{\text{CEL0}}(\hat{x}^0) = G_{\ell_0}(\hat{x}^0). \quad (6)$$

*Theorem 2:* Let  $d \in \mathbb{R}^M$ ,  $\lambda > 0$ , and  $G_{\text{CELO}}$  have a local minimum (not global) at  $\hat{x} \in \mathbb{R}^N$ . Then  $\hat{x}^0$  (defined by (5)) is a local minimizer (not global) of  $G_{\ell_0}$  and (6) is verified.

Theorem 1 gives an “equivalence” between global minimizers of the two functionals while Theorem 2 partially extends this result to local minimizers: from all local minimizers of  $G_{\text{CELO}}$  we can easily extract a local minimizer of  $G_{\ell_0}$ . However the converse is false and we observed experimentally that an important amount of strict local minimizers of  $G_{\ell_0}$  are not critical point of  $G_{\text{CELO}}$  [1, §4.2]. In particular  $G_{\text{CELO}}$  eliminates the strict local minimizers  $\hat{x}$  of  $G_{\ell_0}$  such that  $\sigma^-(\hat{x}) \neq \emptyset$ .

Note that, since from [6, Theorem 4.4 (i)], the set of global minimizers of  $G_{\ell_0}$  is nonempty, Theorem 1 (i) ensures the existence of global minimizers for  $G_{\text{CELO}}$ .

Then from Theorem 1 and 2 we conclude that it is preferable to address problem (1) by minimizing the *continuous* functional  $G_{\text{CELO}}$  instead of  $G_{\ell_0}$  since the global minimizers of  $G_{\text{CELO}}$  *contain* those of  $G_{\ell_0}$  and that  $G_{\text{CELO}}$  has “*less*” local minimizers than  $G_{\ell_0}$ .

### C. How to minimize $G_{\text{CELO}}$ ?

The continuity of  $G_{\text{CELO}}$  allows to use nonsmooth nonconvex algorithms (e.g. [7], [8]) for minimizing  $G_{\text{CELO}}$  and thus  $G_{\ell_0}$ . Usually, such algorithms converge to a *critical point* of the minimized functional. Consequently, they cannot ensure the convergence to a minimizer of  $G_{\text{CELO}}$ . However, the following lemma provides a relation between some critical points of  $G_{\text{CELO}}$  and minimizers of  $G_{\ell_0}$ .

*Lemma 1:* Let  $\hat{x} \in \mathbb{R}^N$  be a critical point of  $G_{\text{CELO}}$  verifying  $\sigma^-(\hat{x}) = \emptyset$ . Then it is a (local) minimizer of  $G_{\ell_0}$  and  $G_{\text{CELO}}(\hat{x}) = G_{\ell_0}(\hat{x})$ .

Therefore, from *any* state of the art algorithms verifying a *sufficient decrease condition* and the *convergence to a critical point* of  $G_{\text{CELO}}$  (e.g. [7], [8]), we can define a *macro algorithm* [1, Algorithm 1] which adds an outer loop to move iteratively from a critical point of  $G_{\text{CELO}}$  to another one while decreasing the cost function and converging to a point  $\hat{x}$  such that  $\sigma^-(\hat{x}) = \emptyset$  [1, Theorem 5.1]. From Lemma 1, such a point is a (local) minimizer of  $G_{\ell_0}$ .

Using the *Iteratively Reweighted  $\ell_1$  (IRL1)* [7] or *Forward-Backward Splitting (FBS)* [8] as inner algorithm within the proposed macro algorithm, numerical experiments [1, §5.1] compare its performances with the *Iterative Hard Thresholding (IHT)* algorithm which is also ensured to converge to a (local) minimizer of problem (1) (see [5], [8]). A part of the experiments conducted in [1, §5.1] are reported on Fig 1. We can see the interesting behaviour of the proposed macro algorithm which converges to (local) minimizers of  $G_{\ell_0}$  with a lower function value than those obtained with IHT. This shows that the macro algorithm is more “robust” against local minimizers of  $G_{\ell_0}$  than IHT.

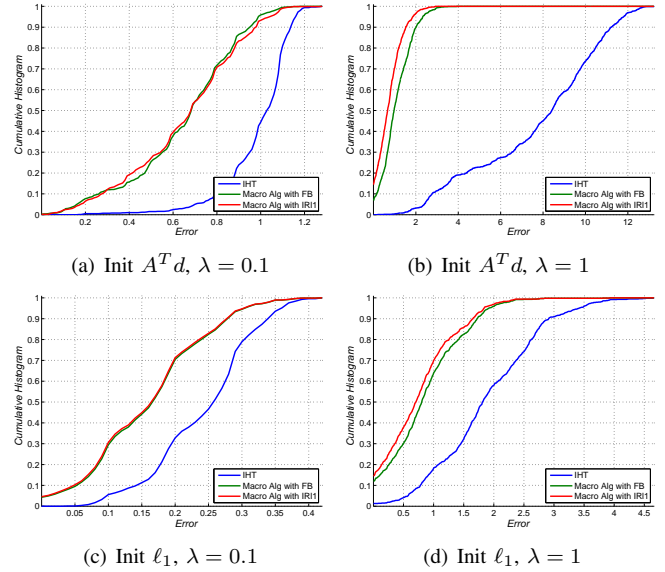


Fig. 1. Cumulative histograms of the error  $|G_{\ell_0}(\hat{x}) - G_{\ell_0}(x^*)|$  where  $\hat{x}$  is the estimated solution and  $x^*$  is a global minimizer of  $G_{\ell_0}$ . The histograms are computed from 1000 random matrices  $A \in \mathbb{R}^{7 \times 15}$  and  $d \in \mathbb{R}^7$  generated from a uniform distribution (the random seed is fixed at the same value for the four configurations (Init,  $\lambda$ ) in order to generate the same sequence of problems). Two different initializations are considered:  $x^0 = A^T d$  (Init  $A^T d$ ) and  $x^0 = x_{\ell_1}$  (Init  $\ell_1$ ) the solution of the  $\ell_1$  relaxed problem. The experiment is repeated for two values of  $\lambda$  (0.1 and 1). For each configuration the estimation is performed using the IHT algorithm (blue) and the macro algorithm combined with IRL1 (red) or FBS (green).

The theoretical analysis of the proposed *tight continuous relaxation*  $G_{\text{CELO}}$  of  $G_{\ell_0}$  and the numerical experiments conducted on low dimensional examples are promising for the development of new algorithms to deal with problem (1) taking benefit from the nice properties of the  $G_{\text{CELO}}$  functional.

### REFERENCES

- [1] E. Soubies, L. Blanc-Féraud, and G. Aubert, “A Continuous Exact  $\ell_0$  penalty (CELO) for least squares regularized problems,” *Preprint hal-01102492*, 2015.
- [2] D. L. Donoho, “For most large underdetermined systems of linear equations the minimal  $\ell_1$ -norm solution is also the sparsest solution,” *Communications on Pure and Applied Mathematics*, vol. 59, no. 6, pp. 797–829, 2006.
- [3] V. N. Temlyakov, “Greedy approximation,” *Acta Numerica*, vol. 17, pp. 235–409, 2008.
- [4] H. A. Le Thi, H. M. Le, and T. P. Dinh, “Feature selection in machine learning: an exact penalty approach using a difference of convex function algorithm,” *Machine Learning*, pp. 1–24, 2014.
- [5] T. Blumensath and M. E. Davies, “Iterative thresholding for sparse approximations,” *Journal of Fourier Analysis and Applications*, vol. 14, no. 5–6, pp. 629–654, 2008.
- [6] M. Nikolova, “Description of the minimizers of least squares regularized with  $\ell_0$ -norm. Uniqueness of the global minimizer,” *SIAM Journal on Imaging Sciences*, vol. 6, no. 2, pp. 904–937, 2013.
- [7] P. Ochs, A. Dosovitskiy, T. Brox, and T. Pock, “On Iteratively Reweighted Algorithms for Nonsmooth Nonconvex Optimization in Computer Vision,” *SIAM Journal on Imaging Sciences*, vol. 8, no. 1, pp. 331–372, 2015.
- [8] H. Attouch, J. Bolte, and B. F. Svaiter, “Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized gaussseidel methods,” *Mathematical Programming*, vol. 137, no. 1–2, pp. 91–129, 2013.