SPARSE REPRESENTATIONS IN NESTED NON-LINEAR MODELS

Angélique Drémeau ⊛*, Patrick Héas* ⋇ *and Cédric Herzet* ⋇

⊛ CNRS and ESPCI ParisTech, 10 rue Vauquelin, UMR 7083 Gulliver, 75005 Paris, France
⋇ INRIA Centre Rennes - Bretagne Atlantique, Campus universitaire de Beaulieu, 35000 Rennes, France

ABSTRACT

Following recent contributions in non-linear sparse representations, this work focuses on a particular non-linear model, defined as the nested composition of functions. Recalling that most linear sparse representation algorithms can be straightforwardly extended to non-linear models, we emphasize that their performance highly relies on an efficient computation of the gradient of the objective function. In the particular case of interest, we propose to resort to a well-known technique from the theory of optimal control to evaluate the gradient. This computation is then implemented into the "ℓ_1-reweighted" procedure proposed by Candès *et al.*, leading to a non-linear extension of it.

Index Terms— Non-linear sparse representation, dynamic programming, ℓ_0-norm relaxation

1. INTRODUCTION

The sparse model assumes that a signal can be represented, exactly or approximatively, by a number of elements much smaller than its dimension. Exploited for more than twenty years, this model has proved to be a good prior for many types of signals in a variety of domains including audio [1] and image [2] processing and is at the heart of the recent compressive-sensing paradigm [3].

Standard sparse representation procedures focus on linear observation models, that is

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}, \tag{1}$$

where $\mathbf{y}$ is the observed data, say of dimension N, $\mathbf{x}$ is assumed to be sparse (*i.e.*, contains very few non-zero elements) in a larger space of dimension $M \geq N$ and $\mathbf{n}$ stands for some observation noise. Recovering $\mathbf{x}$ from $\mathbf{y}$ then requires to solve a problem of the form (or some variants thereof):

$$\hat{\mathbf{x}} = \underset{\mathbf{x}}{\operatorname{argmin}} \|\mathbf{x}\|_0 \quad \text{subject to} \quad \|\mathbf{y} - \mathbf{H}\mathbf{x}\|_2^2 \leq \epsilon, \tag{2}$$

where $\|.\|_0$ denotes the ℓ_0 pseudo-norm which counts the number of non-zero elements in its argument and $\epsilon > 0$ controls the reconstruction error.

In practice, the linear observation model may be poorly adapted to many situations. As an example, we can mention the compressive phase retrieval problem, which aims at recovering a sparse signal from the knowledge of the amplitudes of some complex linear measurements (see *e.g.*, [4, 5]). Hence, recent contributions have addressed the problem of exploiting sparse priors with non-linear observation models, that is

$$\mathbf{y} = h(\mathbf{x}) + \mathbf{n}, \tag{3}$$

for some non-linear observation operator[1] $h : \mathbb{R}^M \to \mathbb{R}^N$.

Extending the approach followed in the linear case, these contributions consider a generalized version of problem (2)

$$\hat{\mathbf{x}} = \underset{\mathbf{x}}{\operatorname{argmin}} \|\mathbf{x}\|_0 \quad \text{subject to} \quad J(\mathbf{x}) \leq \epsilon, \tag{4}$$

where $J(\mathbf{x}) : \mathbb{R}^M \to \mathbb{R}^+$ is some scalar function (*e.g.*, $J(\mathbf{x}) = \|\mathbf{y} - h(\mathbf{x})\|_2^2$) accounting for model (3).

In this paper, we are interested in a particular case of non-linearity, where the penalty function $J(\mathbf{x})$ is defined as the nested composition of some functions. Formally, we write

$$J(\mathbf{x}) = \sum_{l=1}^{L} J_l \circ f_l \circ \ldots \circ f_1(\mathbf{x}), \tag{5}$$

where $\{f_l\}_{l=1}^L$ and $\{J_l\}_{l=1}^L$ are some differentiable functions such that $f_1 : \mathbb{R}^M \to \mathbb{R}^P$, $f_l : \mathbb{R}^P \to \mathbb{R}^P$, $\forall l \in \{2, \ldots, L\}$ and $J_l : \mathbb{R}^T \to \mathbb{R}^+$, $\forall l \in \{1, \ldots, L\}$, and $\circ$ stands for the function-composition operator. This type of model is for instance of interest in the ubiquitous situations where one collects partial information on the state of a *dynamical system* whose initial condition admits a sparse decomposition (see section 3). In particular, we emphasize that results from optimal control [6] can be exploited to provide a fast implementation of any gradient-based algorithm by taking benefit of the special structure of the non-linear model (5). We propose then a practical implementation of this computation into the optimization procedure proposed in [7].

[1] We restrict here our exposition to the real case, but similar reasoning can be conducted in the complex case.

2. SPARSE REPRESENTATION AND GRADIENT EVALUATION

2.1. Sparsity-constrained linear models

In the literature addressing the standard sparse representation problem (2), many computationally-efficient procedures rely (often implicitly) on the fact that the evaluation of the gradient of $J(\mathbf{x}) = \|\mathbf{y} - \mathbf{H}\mathbf{x}\|_2^2$ involves a low computational burden. More specifically, letting $\nabla_{\mathbf{x}} \triangleq [\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_M}]^T$, we have that the gradient of $J(\mathbf{x})$ evaluated at some point $\mathbf{x}^\star$ can be written as

$$\nabla_{\mathbf{x}} J(\mathbf{x}^\star) = -2\mathbf{H}^T(\mathbf{y} - \mathbf{H}\mathbf{x}^\star). \tag{6}$$

We note that the evaluation of $\nabla_{\mathbf{x}} J(\mathbf{x}^\star)$ only requires multiplications by the dictionary $\mathbf{H}$ and its transpose $\mathbf{H}^T$. In the case of general dictionaries, the complexity associated with the evaluation of the gradient thus scales as $\mathcal{O}(MN)$. This constitutes one of the key ingredients of the success of several procedures efficiently tackling high-dimensional problems. Among others, we can mention the procedures based on a relaxation of the ℓ_0 pseudo-norm (*e.g.*, FISTA [8], reweighted ℓ_1 norm [7]), the family of thresholding algorithms (*e.g.*, IHT [9]), or the greedy procedures (*e.g.*, MP [10], OMP [11], CoSaMP [12]) which sequentially update a support estimate by including the element x_j leading to the highest local descent of $J(\mathbf{x})$ (that is the element with the largest absolute partial derivative $|\frac{\partial J(\mathbf{x}^{(k)})}{\partial x_j}|$, where $\mathbf{x}^{(k)}$ is the current estimate).

2.2. Sparsity-constrained non-linear models

It is noticeable that many algorithms mentioned above, when expressed in terms of the gradient of $J(\mathbf{x})$, can straightforwardly be applied to non-linear sparse representation problems. Following this idea, the principles underlying IHT, MP, OMP and CoSaMP have for example been extended to the non-linear setting in [13], [14], [15] and [16] respectively. Similarly, relaxed problems, more specifically based on a ℓ_1 relaxation of the ℓ_0 pseudo-norm, have been devised for constraints of particular form $J(\mathbf{x}) = ||\mathbf{y} - h(\mathbf{x})||_2^2$. In [17], the authors consider the noiseless case ($\epsilon = 0$ in (4)) and propose to approximate h by its Taylor expansion reducing the non-linear term to a quadratic expression and allowing then the use of lifting techniques. A similar idea is applied in [18] to a non-linear operator h defined as a nested composition of functions. The initial optimization problem is thus reformulate as a quadratic programming problem through a first-order linearization.

The tractability of these extensions is however highly dependent on the efficient evaluation of the gradient $\nabla_{\mathbf{x}} J(\mathbf{x})$. In this paper, we elaborate on this problem for the particular family of cost functions $J(\mathbf{x})$ defined in (5). Our exposition is based on the well-known theory of optimal control which traces back to the 70's (see *e.g.*, [6]).

2.3. Efficient gradient computation

Considering (5), the gradient of $J(\mathbf{x})$ evaluated at some point $\mathbf{x}^\star$ is written as

$$\nabla_{\mathbf{x}} J(\mathbf{x}^\star) = \sum_{l=1}^{L} \nabla_{\mathbf{x}} \left(J_l \circ f_l \circ \ldots \circ f_1\right)(\mathbf{x}^\star), \tag{7}$$

by virtue of the linearity of the operator $\nabla_{\mathbf{x}}$.

Let us make two remarks. First, the composed function $J_l \circ f_l \circ \ldots \circ f_1$ does not have any simple analytical expression in many situations; in such cases, we have therefore to resort to the chain rule of derivatives of composed functions to evaluate its gradient. Second, the functions $\{f_l\}_{l=1}^L$ appear in each term of $J(\mathbf{x})$ in a structured manner and this fact should be taken into account in any efficient evaluation of $\nabla_{\mathbf{x}} J$. This is the goal of the procedure described hereafter.

Let us define, $\forall l \in \{1, \ldots, L\}$, $\forall \mathbf{x} \in \mathbb{R}^M$,

$$\mathbf{s}_l \triangleq f_l \circ \ldots \circ f_1(\mathbf{x}), \tag{8}$$

and at the particular point of interest $\mathbf{x}^\star$: $\mathbf{s}_l^\star \triangleq f_l \circ \ldots \circ f_1(\mathbf{x}^\star)$. Clearly, with this definition, $\forall l \in \{1, \ldots, L\}$, $\mathbf{s}_l = f_l(\mathbf{s}_{l-1})$, and (5) evaluated at $\mathbf{x}^\star$ can be rewritten as

$$J(\mathbf{x}^\star) = \sum_{l=1}^{L} J_l(\mathbf{s}_l^\star). \tag{9}$$

Therefore, using the chain rule of derivatives, we obtain[2]

$$\nabla_{\mathbf{x}} J(\mathbf{x}^\star) = \sum_{l=1}^{L} \nabla_{\mathbf{x}} f_l(\mathbf{s}_{l-1}^\star)^T \; \nabla_{\mathbf{s}_l} J_l(\mathbf{s}_l^\star),$$

with the convention $\mathbf{s}_0 = \mathbf{x}$ (resp. $\mathbf{s}_0^\star = \mathbf{x}^\star$), and from the dependence between $\mathbf{s}_l$ and $\mathbf{s}_{l-1}$,

$$\begin{aligned} \nabla_{\mathbf{x}} f_l(\mathbf{s}_{l-1}^\star)^T &= \nabla_{\mathbf{x}} f_l(f_{l-1}(\mathbf{s}_{l-2}^\star))^T, \\ &= \nabla_{\mathbf{x}} f_{l-1}(\mathbf{s}_{l-2}^\star)^T \; \nabla_{\mathbf{s}_{l-1}} f_l(\mathbf{s}_{l-1}^\star)^T. \end{aligned} \tag{10}$$

Applying this expression recursively, we finally have

$$\nabla_{\mathbf{x}} J(\mathbf{x}^\star) = \sum_{l=1}^{L} \left(\prod_{j=1}^{l} \nabla_{\mathbf{s}_{j-1}} f_j(\mathbf{s}_{j-1}^\star)^T \right) \; \nabla_{\mathbf{s}_l} J_l(\mathbf{s}_l^\star). \tag{11}$$

On the one hand, we note that the latter expression does no longer involve the derivation of some composed functions but is exclusively based on the derivative of each component function. On the other hand, some care has to be taken in order to avoid unnecessary computational burden in the evaluation of (11). This gives rise to the following backward-forward procedure:

[2] If $f_l(\mathbf{s}_{l-1}^\star) \triangleq [f_{l,1}, ..., f_{l,P}]^T$, the operator $\nabla_{\mathbf{x}}$ applied to $f_l(\mathbf{s}_{l-1}^\star)^T$ results in the $M \times P$ matrix whose (i,j)-th element is $\frac{\partial f_{l,j}}{\partial x_i}$.

- The sequence $\{\mathbf{s}_l^\star\}_{l=1}^L$ is evaluated via the forward recursion

$$\mathbf{s}_0^\star = \mathbf{x}^\star, \tag{12}$$
$$\mathbf{s}_l^\star = f_l(\mathbf{s}_{l-1}^\star). \tag{13}$$

- All multiplications by a same matrix $\nabla_{\mathbf{s}_{l-1}} f_l(\mathbf{s}_{l-1}^\star)^T$ are gathered in one single operation. This is done through the backward recursion

$$\mathbf{p}_L = \nabla_{L-1} f_L(\mathbf{s}_{L-1}^\star)^T \, \nabla_{\mathbf{s}_L} J_L(\mathbf{s}_L^\star), \tag{14}$$
$$\mathbf{p}_l = \nabla_{l-1} f_l(\mathbf{s}_{l-1}^\star)^T \, (\nabla_{\mathbf{s}_l} J_l(\mathbf{s}_l^\star) + \mathbf{p}_{l+1}), \tag{15}$$

  leading finally to $\mathbf{p}_0 = \nabla_{\mathbf{x}} J(\mathbf{x}^\star)$. In that way, the multiplication by each matrix $\nabla_{\mathbf{s}_{l-1}} f_l(\mathbf{s}_{l-1}^\star)^T$ is only performed once during the whole recursion.

This backward-forward procedure is widely used in geophysical applications (*e.g.*, [19]). However, to the best of our knowledge, the explicit (and motivated) use of this technique into contexts of sparsity-constrained problems has never been considered. In particular, in [18] which focuses on a similar non-linear model, this efficient computation of the gradient is not proposed.

3. SPARSE REPRESENTATIONS IN DYNAMICAL MODELS

We emphasize that the structure of the cost function in (5) is well-suited to the characterization of dynamical systems with partial state information. Let us indeed consider a dynamical system characterized by a generic state evolution equation

$$\mathbf{s}_l = f_l(\mathbf{s}_{l-1}) \qquad \forall l \in \{2, \ldots, L\}. \tag{16}$$

Assume moreover, that noisy partial observations of the states are collected at each time, that is

$$\mathbf{y}_l = g_l(\mathbf{s}_l) + \mathbf{n}, \tag{17}$$

where $\{g_l\}_{l=1}^L$ are some differentiable functions. A typical problem encountered in many domains of applications consists in recovering the sequence of $\{\mathbf{s}_l\}_{l=1}^L$ from the collected observations $\{\mathbf{y}_l\}_{l=1}^L$. In the quite common case where the dimension of the collected data is inferior to the number of unknowns, one has to include an extra constraint on the sought vector in order to hope achieving a consistent estimation. Hereafter, we will assume that the initial state is sparse in some redundant dictionary $\mathbf{H}$, that is

$$\mathbf{s}_1 = \mathbf{H}\mathbf{x}, \tag{18}$$

for some sparse vector $\mathbf{x}$. One possible formulation of the state estimation problem is therefore as follows

$$\min_{\mathbf{x}} \|\mathbf{x}\|_0 \text{ subject to } \left\{ \begin{array}{l} \sum_{l=1}^L \|\mathbf{y}_l - g_l(\mathbf{s}_l)\|_2^2 \leq \epsilon, \\ \mathbf{s}_l = f_l(\mathbf{s}_{l-1}), \quad \mathbf{s}_1 = \mathbf{H}\mathbf{x}. \end{array} \right. \tag{19}$$

Obviously, this problem can be reformulated as (4) with a cost function satisfying (5) by setting $J_l(\mathbf{z}) \triangleq \|\mathbf{y}_l - g_l(\mathbf{z})\|_2^2$. The methodology described in this paper is therefore well-suited for gradient evaluation in this type of setup.

It is noticeable that many dynamical models typically evolve in high-dimensional spaces, leading in turn to sparse-representation problems of very high dimensions. In such settings, an efficient evaluation of the gradient of $J(\mathbf{x})$ turns out to be crucial for the tractable search of a solution of (4). In particular, any attempt to evaluate $\nabla_{\mathbf{x}} J(\mathbf{x})$ by any standard numerical means (*e.g.*, finite differences) is computationally intractable.

4. RESULTS: APPLICATION TO SQG MODEL

As an example of the methodology presented in this paper, we consider a non-linear sparse-representation problem in a particular geophysical application. More specifically, we focus on the characterization of the state of the ocean by exploiting the Surface Quasi-Geostropic (SQG) dynamical model [20].

The SQG model assumes that some geophysical quantity $s(\mathbf{u}, t)$ (the so-called "buoyancy") obeys the following partial differential equation

$$\frac{\partial}{\partial t} s(\mathbf{u}, t) + (\nabla^\perp \nabla^{-1/2} s(\mathbf{u}, t))^T \, \nabla_{\mathbf{u}} s(\mathbf{u}, t) = 0, \tag{20}$$

in which $\mathbf{u} \in \mathbb{R}^2$ and $t \in \mathbb{R}$ play the respective roles of spatial and temporal variables, and $\nabla^\perp \nabla^{-1/2}$ is a vectorial differential operator whose definition can be found in [20]. In the sequel, we will consider a discretized version of (20), of the form of the state equation (16). This discretized model is built by means of a standard 4th-order Runge-Kutta numerical integration scheme [21].

Satellites collect partial information $\{\mathbf{y}_l\}_{l=1}^L$ about the buoyancy at different time instants. We assume hereafter that each observation $\mathbf{y}_l$ is related to the state of the system $\mathbf{s}_l$ (but not directly to $\mathbf{x}$) by a noisy linear observation model:

$$\mathbf{y}_l = \mathbf{G}_l \, \mathbf{s}_l + \mathbf{n}, \tag{21}$$

where $\forall l \in \{1, \ldots, L\}$, $\mathbf{G}_l \in \mathbb{R}^{\frac{N}{L} \times P}$ with $N \leq PL$.

The goal is then to recover the buoyancy from the low-dimensional information provided by the satellite by exploiting: *i)* the geophysical model (20), nesting the buoyancy at different time instants; *ii)* the sparse decomposition of the initial condition $\mathbf{s}_1$ in some redundant dictionary $\mathbf{H}$, see (18).

At this point, the question is posed about the choice of the optimization procedure. As emphasized in section 2.2, providing an efficient evaluation of the gradient of the cost function, different well-known sparse optimization algorithms can be considered. Here, we propose to formulate the optimization problem as

$$\min_{\mathbf{x}} \sum_{l=1}^L \|\mathbf{y}_l - \mathbf{G}_l(f_l \circ \ldots f_2(\mathbf{H}\mathbf{x}))\|_2^2 + \lambda \, r(\mathbf{x}), \tag{22}$$

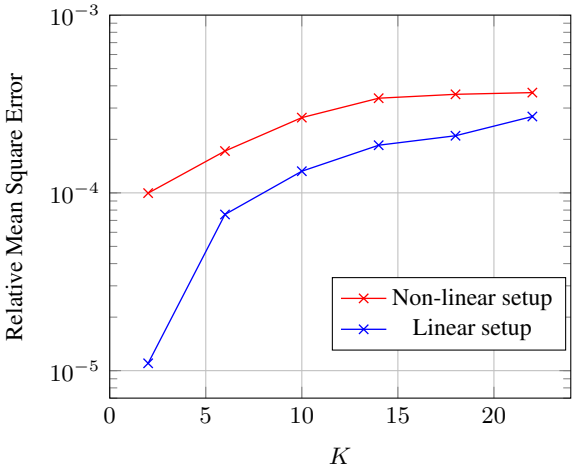

Fig. 1. Relative MSE versus the number of non-zero coefficients K in the sparse vector.

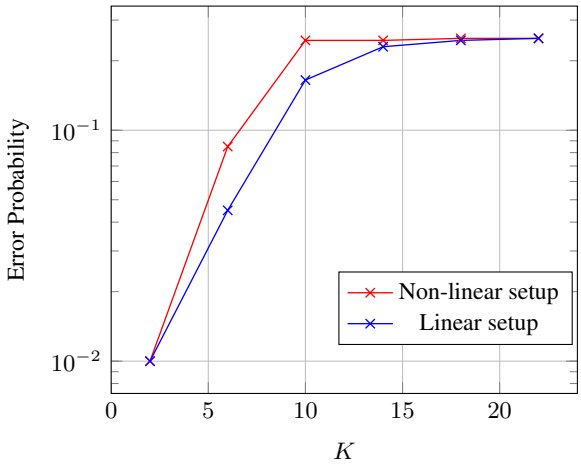

Fig. 2. Error probability versus the number of non-zero coefficients K in the sparse vector.

where $\lambda > 0$ and $r(\mathbf{x})$ is some sparsity-enforcing regularizing function. In our simulation, we chose $r(\mathbf{x}) = \sum_m \log(x_m + \epsilon)$, with $\epsilon = 10^{-1}$. Then our optimization procedure follows the majorization-minimization technique exposed in [7]; at each iteration an upper bound on the goal function is constructed by majorizing $r(\mathbf{x})$ by a weighted ℓ_1 norm. We look for the minimum of each of these majorizing functions by means of descent procedures involving the gradient of $J(\mathbf{x})$ evaluated as exposed in section 2.3.

Particularized to the SQG model, the evaluation of the forward and backward recursions (13)-(15) have a complexity of order $\mathcal{O}(ML)$. By comparison, using a finite-difference scheme to evaluate the gradient requires to run (at least) two forward recursions by element of $\mathbf{x}$, leading to an overall complexity of $\mathcal{O}(M^2L)$. This order of complexity thus precludes us from using this type of approach in moderate-to-high dimensional problems.

The simulation setup considered in this paper is as follows. The state vectors $\mathbf{s}_l$ are assumed to live in 256-dimensional space. The initial condition is supposed to have a sparse decomposition $\mathbf{x}$ in a dictionary $\mathbf{H} \in \mathbb{R}^{256\times512}$ made up of sine and cosine functions. These dimensions have been chosen for the sake of running extensive simulation. We note that in practical SQG setups, the dimension of $\mathbf{x}$ is of the order of 512^2 or 1024^2. The positions of the non-zero coefficients in $\mathbf{x}$ are randomly chosen. Their amplitudes are drawn from a zero-mean Gaussian distribution with variance $\sigma_x^2 = 10$. The observations $\mathbf{y}_l \in \mathbb{R}^{16}$ are collected at $L = 4$ different time instants and the observation matrices $\mathbf{G}_l$ correspond to random subsampling operators. The ratio between the number of observations $N = 16 \times 4$ and the dimension of $\mathbf{x}$ is therefore equal to $64/512 = 1/8$. Finally, we consider a small observation noise, drawn from zero-mean Gaussian distributions with variance $\sigma^2 = 10^{-9}$ (quasi-noiseless scenario). In Fig. 1, we represent the relative mean-square error (MSE) $||\hat{\mathbf{x}} - \mathbf{x}||_2^2/||\mathbf{x}||_2^2$ achieved by the minimization of (22) via the majorization-minimization procedure described above. As a point of comparison, we run the same algorithm on a linear sparse representation problem having the same problem dimensions (namely $\mathbf{y} = \mathbf{GHx}$ where $\mathbf{G}$ is a rate-$1/4$ random subsampling matrix). For each data point, we average the performance over 50 trials.

We can notice that the considered procedure can achieve an acceptable relative mean square error over a wide range of sparsity levels. We note also that the non-linear setup suffers from a reasonable degradation with respect to the linear setup. This tendency is confirmed in Fig. 2 which illustrates the probability of making at least one error on the support of the sought sparse vector.

5. CONCLUSION

In this paper, we address the problem of sparse representations in a non-linear setting. We emphasize that the computation of the gradient of the cost function may be a bottleneck for the extension of standard estimation procedures. We show that this computation may be handled efficiently, by applying principles from the theory of optimal control, as long as the cost function satisfies some desirable structural property. Our derivations are illustrated on a particular example dealing with the estimation of the state of a geophysical system from partial observations.

6. ACKNOWLEDGEMENTS

This work has been supported in part by the ERC under the European Union's 7th Framework Programme Grant Agreement 307087-SPARCS and by the CNRS/INSU through the LEFE funding program.

7. REFERENCES

[1] C. Févotte, L. Daudet, S. J. Godsill, and B. Torrésani, "Sparse regression with structured priors: application to audio denoising," *IEEE Trans. On Audio, Speech and Language Processing*, vol. 16, no. 1, pp. 174–185, 2008.

[2] R. M. Figueras i Ventura, P. Vandergheynst, and P. Frossard, "Low rate and scalable image coding with redundant representations," Tech. Rep., TR-ITS-03.02, June 2003.

[3] D. L. Donoho, "Compressed sensing," *IEEE Trans. On Information Theory*, vol. 52, no. 4, pp. 1289–1306, April 2006.

[4] H. Ohlsson, A. Y. Yang, R. Dong, and S. S. Sastry, "Compressive phase retrieval from squared output measurements via semidefinite programming," *IFAC Symposium on System Identification*, vol. 16, no. 1, pp. 89–94, March 2012.

[5] Y. Schechtman, A. Beck, and Y. C. Eldar, "Gespar: Efficient phase retrieval of sparse signals," Available on arXiv:1301.1018, 2013.

[6] M. Cannon, C. Cullum, and E. Polak, *Theory of Optimal Control and Mathematical Programming*, New York, 1970.

[7] E. J. Candes, M. B. Wakin, and S. Boyd, "Enhancing sparsity by reweighted l1 minimization," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5, 2008.

[8] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal Imaging Sciences*, vol. 2, no. 1, 2009.

[9] 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, December 2008.

[10] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. On Signal Processing*, vol. 41, no. 12, pp. 3397–3415, December 1993.

[11] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Proc. Asilomar Conference on Signals, Systems, and Computers*, 1993, pp. 40–44.

[12] D. Needell and J. A. Tropp, "Cosamp: iterative signal recovery from incomplete and inaccurate samples," *Applied and Computational Harmonic Analysis*, vol. 26, no. 3, pp. 301–321, May 2009.

[13] T. Blumensath, "Compressed sensing with nonlinear observations and related nonlinear optimization problems," *IEEE Trans. On Information Theory*, vol. 59, no. 6, pp. 3466–3474, June 2013.

[14] A. Beck and Y. C. Eldar, "Sparsity constrained nonlinear optimization: optimality conditions and algorithms," Available on arXiv:1203.4580, 2013.

[15] T. Blumensath and M. E. Davies, "Gradient pursuit for non-linear sparse signal modelling," in *European Signal Processing Conference (EUSIPCO)*, April 2008.

[16] S. Bahmani, B. Raj, and P. Boufounos, "Greedy sparsity-constrained optimization," *Journal of Machine Learning Research*, vol. 14, pp. 807–841, 2013.

[17] H. Ohlsson, A. Y. Yang, and S. S. Sastry, "Nonlinear basis pursuit," Available on arXiv:1304.5802, 2013.

[18] A. M. Ebtehaj, M. Zupanski, G. Lerman, and E. Foufoula-Georgiou, "Variational data assimilation via sparse regularization," *EGU General Assembly*, p. 14147, 2013.

[19] S. O. Ba, T. Corpetti, B. Chapron, and R. Fablet, "Variational data assimilation for missing data interpolation in sst images," in *IEEE Geoscience and Remote Sensing Symposium (IGARSS)*, July 2010.

[20] J. Isern-Fontanet, G. Lapeyre, P. Klein, B. Chapron, and M. W. Hecht, "Three-dimensional reconstruction of oceanic mesoscale currents from surface information," *Journal of Geophysical Research*, vol. 113, no. C9, September 2008.

[21] K. A. Atkinson, *An Introduction to Numerical Analysis (2nd ed.)*, New York, 1989.