Combining sparsity and dynamics: an efficient way

Angélique Drémeau[1], Patrick Héas[2] and Cédric Herzet[2]

[1]CNRS and ESPCI ParisTech, 10 rue Vauquelin, UMR 7083 Gulliver, 75005 Paris, France.

[2]INRIA Centre Rennes - Bretagne Atlantique, Campus universitaire de Beaulieu, 35000 Rennes, France.

***Abstract*— Most linear sparse representation algorithms can be straightforwardly extended to non-linear models. Their performance however, relies on an efficient computation of the gradient of the objective function. In this paper, we focus on a particular non-linear model, defined as the nested composition of functions and propose to resort to a well-known technique from the theory of optimal control to compute the gradient. As a proof of concept, this computation is then implemented into the optimization procedure proposed by Candès *et al.*, and applied to a geophysical dynamical model.**

1 Introduction

Recent contributions have addressed the problem of exploiting sparse priors with non-linear observation models, that is

$$\mathbf{y} = h(\mathbf{x}) + \mathbf{n}, \tag{1}$$

where $h : \mathbb{R}^M \to \mathbb{R}^N$ (with $M \geq N$) is a non-linear observation operator and $\mathbf{n}$ stands for an observation noise. Extending the approach followed in the linear case, these contributions propose also a generalization of the penalty function, leading to an optimization problem of the form (or some variants thereof)

$$\hat{\mathbf{x}} = \underset{\mathbf{x}}{\operatorname{argmin}} \|\mathbf{x}\|_0 \quad \text{subject to} \quad J(\mathbf{x}) \leq \epsilon, \tag{2}$$

where $J(\mathbf{x})$ is some scalar function (*e.g.*, $J(\mathbf{x}) = \|\mathbf{y} - h(\mathbf{x})\|_2^2$) accounting for discrepancies from model (1).

Noticing that many sparse representation algorithms dealing with linear observation models rely - implicitly or explicitly - on the computation of the gradient of the function $J(\mathbf{x})$, non-linear versions of them can be straightforwardly derived. Following this idea, the extensions of the well-known algorithms IHT [1], MP [2], OMP [3] and CoSaMP [4] have thus been proposed, respectively in [5], [6], [7] and [8] .

However, whereas in the linear case, the evaluation of the gradient of $J(\mathbf{x})$ only involves multiplications by the dictionary and its transpose, its computational cost can be prohibitive in some non-linear cases. In this paper, we elaborate on this problem for the particular family of cost functions $J(\mathbf{x})$ defined as the nested composition of some functions. Formally, we write

$$J(\mathbf{x}) = \sum_{l=1}^{L} J_l \circ f_l \circ \ldots \circ f_1(\mathbf{x}), \tag{3}$$

where $\{f_l\}_{l=1}^L$ are some differentiable functions 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 2.2). In particular, we emphasize that results from optimal control [9] 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 (3). We propose then a practical implementation of this computation into the optimization procedure proposed in [10].

2 Sparse Representations in Nested Non-Linear Models

In this section, we elaborate on the efficient evaluation of the gradient of $J(\mathbf{x})$ when structured as in (3). The methodology is then applied to a particular geophysical problem.

2.1 Efficient gradient computation

We use the following definitions and notations. Considering model (3), we set, $\forall l \in \{1, \ldots, L\}$, $\forall \mathbf{x} \in \mathbb{R}^M$,

$$\mathbf{s}_l \triangleq f_l \circ \ldots \circ f_1(\mathbf{x}). \tag{4}$$

We thus have $\forall l \in \{1, \ldots, L\}$,

$$\mathbf{s}_l = f_l(\mathbf{s}_{l-1}), \tag{5}$$

with the convention $\mathbf{s}_0 = \mathbf{x}$. We also define the gradient operator as

$$\nabla_{\mathbf{x}} \triangleq [\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_M}]^T, \tag{6}$$

so that $\nabla_{\mathbf{x}}$ applied to a vector $\mathbf{z} = [z_1, ..., z_N]^T$ results in the $M \times N$ matrix whose (i, j)-th element is $\frac{\partial z_j}{\partial x_i}$.

With these notations in mind, (3) evaluated at $\mathbf{x}^\star$ can be rewritten as

$$J(\mathbf{x}^\star) = \sum_{l=1}^{L} J_l(\mathbf{s}_l^\star), \tag{7}$$

where $\mathbf{s}_l^\star$ is defined as in (4) with $\mathbf{x} = \mathbf{x}^\star$. Therefore, using the chain rule of derivative, we obtain

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

and from the dependence between $\mathbf{s}_l$ and $\mathbf{s}_{l-1}$,

$$\begin{aligned} \nabla_{\mathbf{x}} f_l(\mathbf{s}_{l-1}^\star)^T &= \nabla_{\mathbf{x}} f_l(f_{l-1}(\mathbf{s}_{l-2}^\star))^T, \\ &= \nabla_{\mathbf{x}} f_{l-1}(\mathbf{s}_{l-2}^\star)^T \; \nabla_{\mathbf{s}_{l-1}} f_l(\mathbf{s}_{l-1}^\star)^T. \end{aligned} \tag{8}$$

Finally, applying this expression recursively, we have

$$\nabla_{\mathbf{x}} J(\mathbf{x}^\star) = \sum_{l=1}^{L} \left(\prod_{j=1}^{l} \nabla_{\mathbf{s}_{j-1}} f_j(\mathbf{s}_{j-1}^\star)^T \right) \; \nabla_{\mathbf{s}_l} J_l(\mathbf{s}_l^\star). \tag{9}$$

The latter expression is exclusively based on the derivative of each function component. Its evaluation can then be performed through the following forward-backward procedure:

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

$$\mathbf{s}_l^\star = f_l(\mathbf{s}_{l-1}^\star), \tag{10}$$
$$\mathbf{s}_0^\star = \mathbf{x}^\star. \tag{11}$$

- 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

$$\mathbf{p}_L = \nabla_{L-1} f_L(\mathbf{s}_{L-1}^\star)^T \, \nabla_{\mathbf{s}_L} J_L(\mathbf{s}_L^\star), \tag{12}$$
$$\mathbf{p}_l = \nabla_{l-1} f_l(\mathbf{s}_{l-1}^\star)^T \, (\nabla_{\mathbf{s}_l} J_l(\mathbf{s}_l^\star) + \mathbf{p}_{l+1}), \tag{13}$$

  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 forward-backward procedure is widely used in geophysical applications (*e.g.*, [11]). 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 [12] which focuses on a similar non-linear model, this efficient computation of the gradient is not proposed.

2.2 Application to super-resolution in SQG dynamical model

As a practical example of the proposed methodology, we focus on a super-resolution problem in a geophysical context, namely the high-resolution characterization of the state of the ocean by exploiting: *i)* the Surface Quasi-Geostropic (SQG) dynamical model [13]; *ii)* a sparse prior on the initial condition of the dynamical model; *iii)* low-resolution satellite images.

We assume that the SQG model, of the form of (5), rules the evolution of some state variable $\mathbf{s}_l$. The definition of f_l depends on the considered numerical integration scheme (here a 4th-order Runge-Kunta method) but is not specified hereafter for conciseness. Moreover, as a prior knowledge, the initial state is supposed to be sparse in some redundant dictionary $\mathbf{H}$,

$$\mathbf{s}_1 = \mathbf{H}\mathbf{x}, \tag{14}$$

for some sparse vector $\mathbf{x}$. We are interested in recovering the value of $\{\mathbf{s}_l\}_{l=1}^L$ from the observation of low-dimensional images $\{\mathbf{y}_l\}_{l=1}^L$ (*e.g.*, collected by satellites), with

$$\mathbf{y}_l = \mathbf{G}_l \, \mathbf{s}_l + \mathbf{n}, \tag{15}$$

where $\mathbf{G}_l$ is some known observation matrix and $\mathbf{n}$ is an unknown corrupting noise.

In order to solve this inverse problem, we address the following optimization problem

$$\min_{\mathbf{x}} \sum_{l=1}^{L} \|\mathbf{y}_l - \mathbf{G}_l(f_l \circ \dots f_2(\mathbf{H}\mathbf{x}))\|_2^2 + \lambda \, r(\mathbf{x}), \tag{16}$$

where $\lambda > 0$ and $r(\mathbf{x}) = \sum_m \log(x_m + \epsilon)$, $\epsilon = 10^{-1}$, is some sparsity-enforcing regularizing function. A solution to (16) is searched by using the majorization-minimization optimization technique exposed in [10]; 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 $J(\mathbf{x})$ (corresponding here to the first term in (16)) evaluated as presented in section 2.1.

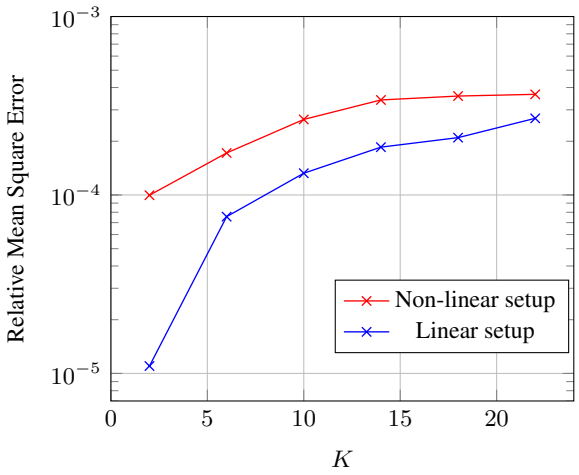

Figure 1: Relative MSE versus the number of non-zero coefficients K in the sparse vector.

Particularized to the SQG model, the evaluation of the forward-backward recursions (10)-(13) 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 $\mathcal{O}(M^2 L)$. 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 in a dictionary $\mathbf{H} \in \mathbb{R}^{256\times512}$ made up of sine and cosine functions. The observations $\mathbf{y}_l \in \mathbb{R}^{32}$ are collected at four different time instants and the observation matrices $\mathbf{G}_l$ correspond to random subsampling operators. The ratio between the number of observations and the dimension of $\mathbf{x}$ is therefore equal to $(32 \times 4)/512 = 1/4$. 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 (16) 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 $\mathbf{y} = \mathbf{G}\mathbf{H}\mathbf{x}$ where $\mathbf{G}$ is a rate-$1/2$ 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.

3 Conclusion

In this paper, we address the problem of sparse representations in a non-linear setting. While a high computational cost of the gradient of the goal function may prevent the use of standard estimation procedures, we show that it can be overcome 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.

4 Acknowledgments

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.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

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

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

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

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

[10] 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.

[11] 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.

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

[13] 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.