Relaxed Analysis Operator Learning

Mehrdad Yaghoobi
Institute for Digital Communications (IDCom)
University of Edinburgh
Edinburgh, EH9 3JL, UK
m.yaghoobi-vaighan@ed.ac.uk

Mike E. Davies
Institute for Digital Communications (IDCom)
University of Edinburgh
Edinburgh, EH9 3JL, UK
mike.davies@ed.ac.uk

Abstract

The problem of analysis operator learning can be formulated as a constrained optimisation problem. This problem has been approximately solved using projected gradient or geometric gradient descent methods. We will propose a relaxation for the constrained analysis operator learning in this paper. The relaxation has been suggested here to, a) reduce the computational complexity of the optimisation and b) include a larger set of admissible operators. We will show here that an appropriate relaxation can be useful in presenting a projection-free optimisation algorithm, while preventing the problem to become ill-posed. The relaxed optimisation objective is not convex and it is thus not always possible to find the global optimum. However, when a rich set of training samples are given, we empirically show that the desired synthetic analysis operator is recoverable, using the introduced sub-gradient descent algorithm.

1 Introduction to the Constrained Analysis Operator Learning

Low-dimensional signal modelling has been used frequently for inverse problems and system regularisation. In the last two decades, linear sparse modelling has been very popular, as it is strong enough to resolve the ambiguities of inverse problems, while it can fairly describe many sensing and imaging systems. Although the synthesis sparse model [1] has been investigated thoroughly in the last decade, its counterpart, *i.e.* analysis sparsity[2], has not been considered as much. Similarly, data modelling for synthesis sparsity, which is called the "dictionary learning" [3], has been considered much more than the "analysis operator learning" [4, 5, 6].

When a set of training samples is given, most of the dictionary/analysis operator learning algorithms are often based on the constrained optimisation of an objective. The necessary constraint in the dictionary learning problem is on the norm of each column of the generative model (dictionary), which is called an atom. This constraint has been handled in the learning problem by different techniques [7], including some convex relaxations of the constraint [8]. Unfortunately, the same techniques can not be used for the analysis operator learning, which make the signal modelling in this framework more challenging.

Let $\mathbf{X} = [\mathbf{x}_i]_i \in \mathbb{R}^{n\times l}$ be the training matrix, made by putting each training signal $\mathbf{x}_i$ as a column. We are looking for an overcomplete (linear) operator $\mathbf{\Omega} \in \mathbb{R}^{a\times n}$ which maps $\mathbf{X}$ into a higher dimensional space with some zeros or small value elements. This can mathematically be formulated[1] as,

$$\min_{\mathbf{\Omega}} \|\mathbf{\Omega X}\|_1, \text{ s.t. } \mathbf{\Omega} \in \mathcal{C}, \tag{1}$$

[1]We here investigate the noise-less scenario, which is similar to the setting in [5]. This means that the training samples are exactly following the parsimony model and there is no corrupting noise. Extension to a noise-aware setting can be done using a technique presented in [9].

where $\mathcal{C}$ is a constraint on the operators. We use $\|\cdot\|_1 = \sum_{i,j} |\{\cdot\}_{i,j}|$ as a measure of sparsity of the operand matrix. It is important that the optimal $\mathbf{\Omega}^*$ is orthogonal to many columns of $\mathbf{X}$. The number of j's such that $\mathbf{\Omega}^*_{j,:} \perp \mathbf{x}_i$, is equivalent to the number of zeros in $\mathbf{\Omega}^*\mathbf{x}_i$, which is called the *cosparsity* of $\mathbf{x}_i$, q. Similarly, we call the total number of zeros in $\mathbf{\Omega}^*\mathbf{X}$, the *cosparsity* of $\mathbf{X}$ and the goal here is to find an $\mathbf{\Omega}$, which cosparsifies $\mathbf{X}$. This problem without a constraint, *i.e.* $\mathcal{C} = \mathbb{R}^{a\times n}$, is ill-defined, as it has many trivial solutions like $\mathbf{\Omega}^* = \mathbf{0}$ [5, 10]. This emphasises the fact that the problem needs some extra constraints for regularisation. It is worth mentioning, although constrained optimisation of an objective is the most direct approach to learn an operator, it is not the only way to tackle the ill-posedness of the problem. There exist other methods to sparsify $\mathbf{X}$, while *indirectly* avoiding trivial solutions [4, 11]. Although some promising results are reported using these techniques, there is no clear understanding of how these methods avoid such trivial solutions and when we expect to learn a *good* operator. In this paper we only consider the problem of constrained analysis operator learning (CAOL).

A canonical constraint on $\mathbf{\Omega}$ is to fix its row ℓ_2-norms. As it has been reported in [5], and we can empirically see in this paper, this constraint is not enough, and we find some bad-conditioned matrices, with rows aliened to the orthogonal direction to dominant directions of $\mathbf{X}$. Applying a constraint on the smallest singular value of $\mathbf{\Omega}$ or enforcing to be Tight Frame (TF), does not solve the ill-definition of (1) [5]. We empirically demonstrate that when a TF constraint is applied, we find an orthobasis for sparsifying $\mathbf{X}$. Here, the aim is to find an overcomplete model, which can be found if we restrict the search space of (1) to the Uniform Normalised TF's (UNTF), defined by $\mathcal{C} = \{\mathbf{\Omega} \in \mathbb{R}^{a\times n} \ : \ \mathbf{\Omega}^T\mathbf{\Omega} = \mathbf{I},\ \forall i\ \|\omega_i\|_2 = \sqrt{\frac{n}{a}}\}$, where $\omega_i \in \mathbb{R}^n$ is the ith row of $\mathbf{\Omega}$. The constant norm $\sqrt{\frac{n}{a}}$ for each row ω_i is enforced by the fact that $\|\omega_i\|_2^2 = \frac{1}{a}\|\mathbf{\Omega}\|_F = \frac{1}{a}\sum_j \|\mathbf{\Omega}_{:,j}\|_2^2$, where $\mathbf{\Omega}_{:,j}$ is the j^{th} column of $\mathbf{\Omega}$, which has unit norm. It has been shown in [9] that if we have a set of exemplars, which sparsifies with an ideal synthetic analysis operator, it can be recovered, up to some sign-flips and row-permutations, solving (1). The set of UNTF's includes the union of orthobases-bases and some useful multi-scale overcomplete transforms, and we therefore hope to fit a better cosparsifying operator to the class of signals, which already have been shown to have parsimonious representations in the mentioned transforms.

1.1 Why the Relaxation is Useful?

We intend to introduce some relaxation to the problem of constrained analysis operator learning, to reduce the computation complexity and to generalise the framework. More specifically, the following issues in the standard AOL formulation and algorithm, motivated us for this work.

1. The optimisation program (1), when $\mathcal{C}$ is the set of UNTF's, is non-convex and how well we can solve it, depends on the optimisation algorithm. A projected sub-gradient type iterative algorithm has been introduced in [5] to approximately solve it. Although some promising results in the operator recovery and operator learning for a class of smooth images, have been reported, **the algorithm is not computationally scalable**. The reason is that the algorithm needs a full singular value decomposition (SVD) at each iteration. This is indeed very costly operation for high dimensional signals. A relaxation in the formulation can reduce the complexity of each algorithm iteration, if does not involve such a costly projection. We will introduce such a relaxation in the next subsection.

2. The CAOL algorithm [5] needs a full projection onto the set of UNTF's to derive a simple convergence to a fixed point. As there is no analytical method to do such a projection, there is no convergence proof for this method. If we relax the constraint such that there is no need for projection onto UNTF, the convergence proof may be possible. The Relaxed AOL algorithm is stable and it converges to a single or a continuum of fixed point(s).

3. The proposed constraint, *i.e.* UNTF, includes many interesting cosparsifying overcomplete transforms. We here broaden the class of operators being considered, while still avoiding the undesirable trivial solutions.

These points motivated us to introduce the relaxed formulation of next section. However, we should avoid extra-relaxation, which degrades the overall performance of the algorithm. We should remember that the constraint $\mathcal{C}$, has a fundamental role in avoiding uninteresting solutions. Extra-relaxation

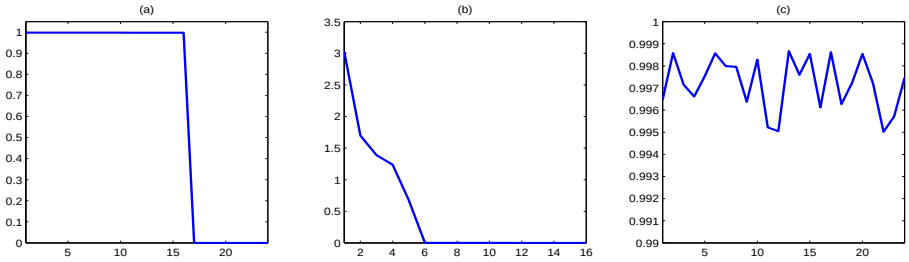

Figure 1: The sorted ℓ_2 norms of the learned operator with the TF constraint (a), singular values of the leaned operator with the UN constraint (b) and normalised inner-products between the rows of the synthetic ideal operator and the corresponding rows in the learned operator (c).

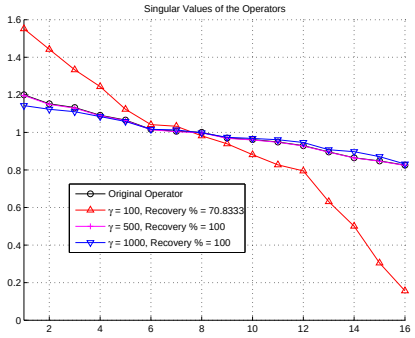

Figure 2: Singular values of the synthetic and leaned operators using the relaxed algorithm and $\lambda = 10^5$, $\gamma = [1\ 5\ 10] * 10^2$. Percentages of the exact operator recovery are also mentioned for different γ's.

of the constraint $\mathcal{C}$, will remake the problem ill-posed, which will be investigated in the simulation section 2.

1.2 Relaxed Analysis Operator Learning (RAOL)

The proposed relaxation includes two steps: a) extending the constraint $\mathcal{C}$ to include its neighbourhood and b) using some Lagrange multipliers to generate an unconstrained program. We will see in the simulation section that the RAOL is fairly robust to the selection of these parameters.

The UNTF constraint is the intersection of two manifolds, which can also be interpreted as two quadratic constraints, $\mathbf{\Omega}^T\mathbf{\Omega} - \mathbf{I} = \mathbf{0}$ and $\omega_i^T\omega_i - \frac{n}{a} = 0,\ \forall i \in [1, a]$. By extending these two constraints, to include the $\epsilon-$neighbourhood in an ℓ_2 sense, we can derive the following *quartic* constraints,

$$\begin{aligned} &\|\mathbf{\Omega}^T\mathbf{\Omega} - \mathbf{I}\|_F^2 \leq \epsilon_{TF}, \\ &\|\omega_i^T\omega_i - \frac{n}{a}\|_2^2 \leq \epsilon_{UN}, \quad \forall i \in [1, a], \end{aligned} \tag{2}$$

where $\|\cdot\|_F$ represents the Frobenius norm and $\epsilon_{TF},\ \epsilon_{UN}$ are some constants. The new constraint $\mathcal{C}_R$, which is generated using the constraints in (2), also includes the operators having singular values close to one and the row norms approximately $\sqrt{\frac{n}{a}}$. We control the size of neighbourhoods by selecting ϵ_{TF} and ϵ_{UN}.

In the second step, we generate an unconstrained objective by penalising the errors corresponding to the constraints in (2). The objective now can be presented as follows,

$$f(\mathbf{\Omega}) = \|\mathbf{\Omega}\mathbf{X}\|_1 + \frac{\gamma}{4}\|\mathbf{\Omega}^T\mathbf{\Omega} - \mathbf{I}\|_F^2 + \frac{\lambda}{4}\sum_i \left\{\|\omega_i^T\omega_i - \frac{n}{a}\|_2^2\right\} \tag{3}$$

where γ and λ are some Lagrange multipliers. In the new objective $f(\mathbf{\Omega})$, γ and λ controls the approximation and when $\gamma \to \infty$ and $\lambda \to \infty$, optimisation of $f(\mathbf{\Omega})$ provides the solution of (1), as the new terms are characteristic function of the constraint $\mathcal{C}$. The new penalty terms in $f(\mathbf{\Omega})$ are of order four and non-convex in general. However the following proposition shows that we can restrict the domain of $\mathbf{\Omega}$, such that f is convex over the domain.

Table 1: Simulation Parameters

$\mathbf{X}$	$\mathbf{\Omega}$	q	α_{ini}	μ	K
$\mathbb{R}^{16\times 576}$	$\mathbb{R}^{24\times 16}$	10	1×10^{-6}	2	10^5

***Proposition* 1** *$f(\mathbf{\Omega}) : \mathcal{C}_c \to \mathbb{R}_+$ is convex, if $\mathcal{C}_c = \{\mathbf{\Omega} : \mathbf{\Omega}^T\mathbf{\Omega} - \mathbf{I} \succeq \mathbf{0},\ \forall i,\ (\omega_i^T\omega_i - \frac{n}{a}) \geq 0\}$.*

This proposition can be proved by calculating the Taylor series of the quartic terms in $f(\mathbf{\Omega})$ and assuring that the curvature is non-negative for each term. The domain $\mathcal{C}_c$ is not convex and we can not thus have a convex program, if we restrict the optimisation to $\mathcal{C}_c$. As there is currently no evidence that such a constraining of the optimisation program helps, we here minimise (3) over the domain $\mathbb{R}^{a\times n}$, to not increase the computational costs.

1.3 A Practical Optimisation Algorithm

The sparsity penalty in $f(\mathbf{\Omega})$ is not differentiable. We thus calculate the subgradient of $f(\mathbf{\Omega})$, which is subgradient of the first term plus the gradients of the second two terms, as follows,

$$\partial f(\mathbf{\Omega}) = \overline{\operatorname{sgn}}(\mathbf{\Omega}\mathbf{X})\mathbf{X}^T + \gamma\left(\mathbf{\Omega}\mathbf{\Omega}^T - \mathbf{I}\right)\mathbf{\Omega} + \lambda\left[\omega_i\left(\omega_i^T\omega_i - \frac{n}{a}\right)\right]_i^T, \tag{4}$$

where $\overline{\operatorname{sgn}}$ is the extended sign function [5]. We can iterate the subgradient descent algorithm, *i.e.* $\mathbf{\Omega}^{[k+1]} = \mathbf{\Omega}^{[k]} - \alpha\partial f(\mathbf{\Omega}^{[K]})$, for some variable stepsize α, found by a line search technique, until it converges. We here start with $\alpha = \alpha_{ini}$ and shrink it by a factor of μ when the step-size is large. As the line-search guarantees the objective decreasing monotically and the sublevel set of (3) is compact, the algorithm converges to a local minma or it gets arbitrary close to a *bounded* continuum of local minimums of $f(\mathbf{\Omega})$.

2 Preliminary Simulation Results

We report two sets of simulation results on synthetic data to demonstrate how the relaxation, achieves some of the goals considered in the subsection 1.1. The general setting for the simulations is presented in table 1. In the first experiment, a random operator $\mathbf{\Omega}$ is generated and iteratively projected onto the sets of TF's and UN's to converge to an operator close to a UNTF. Some training samples were generated to be q cosparse, with respect to the ideal operator $\mathbf{\Omega}_I$, using a technique, which can be found for example in [5]. Now, we examine how much we can relax the constraints, using a set of parameters $(\gamma, \lambda) = \{(1, 0),\ (0, 1),\ (.01, 1)\} * 10^5$. The first pair, only has a penalty to enforce $\mathbf{\Omega}$ be close to the TF's. The sorted norms of the rows of the learned operator are plotted in figure 1.a. We see that, we roughly get an orthobasis, which is clearly not overcomplete. The second pair of parameters, only penalises the errors in the norms of rows, in comparison with $\sqrt{\frac{n}{a}}$. The singular values of the learned operator are shown in figure 1.b. The singular values are far from the desired value, *i.e.* 1, and the operator is actually close to singular. These two cases demonstrate that we should not extra-relax, as the problem becomes ill-posed. With the third pair of parameters, we demonstrate that the RAOL can recover a UNTF operator. The normalised inner-products between the rows of the synthetic operator and the corresponding rows in the learned operator, are shown in figure 1.c. This plot shows that the error in the recovery of original operator, is very small.

In the second experiment we try to show that the approximately UNTF operators can be robustly recovered, *i.e.* with some bounded errors, using RAOL. We now set $\lambda = 10^5$ and choose $\gamma = [100,\ 500,\ 1000]$. The ideal operator $\mathbf{\Omega}_I$ is generated randomly and then alternatingly projected onto the UN and the approximate TF sets. The projection onto approximate TF's is here done by scaling the singular values, to be in 1 ± 0.2 and q-cosparse training signals were generated as explained before. The singular values of the learned operators and the percentages of the approximate recovery, *i.e.* when the normalised inner-product error between corresponding rows is less than 0.01, are shown in figure 2. The RAOL can approximately recover the ideal operator, which *is not UNTF*, when γ is 500 or 1000. This demonstrates that the recoverability of the operators is fairly robust to the algorithm parameter selection.

In these experiments, we observed that the proposed algorithm is faster than the projected subgradient type algorithm for CAOL, in each iteration. We left a comprehensive study and a full speed comparison for a future work.

References

[1] S.S. Chen, D.L. Donoho, and M.A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 33–61, 1998.

[2] M. Elad, P. Milanfar, and R. Rubinstein, "Analysis versus synthesis in signal priors," *Inverse Problems*, vol. 23, no. 3, pp. 947–968, 2007.

[3] I. Tosic and P. Frossard, "Dictionary learning," *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 27–38, March 2011.

[4] G. Peyré and J. Fadili, "Learning analysis sparsity priors," in *Proc. of Sampta'11*, 2011.

[5] M. Yaghoobi, S. Nam, R. Gribonval, and M. Davies, "Analysis operator learning for overcomplete cosparse representations," in *EUSIPCO*, August 2011.

[6] B. Ophir, M. Elad, N. Bertin, and M.D. Plumbley, "Sequential minimal eigenvalues an approach to analysis dictionary learning," in *Proceedings of EUSIPCO*, September 2011.

[7] M. Yaghoobi, T. Blumensath, and M. Davies, "Regularized dictionary learning for sparse approximation," in *EUSIPCO*, 2008.

[8] M. Yaghoobi, T. Blumensath, and M. Davies, "Dictionary learning for sparse approximations with the majorization method," *IEEE Trans. on Signal Processing*, vol. 57, no. 6, pp. 2178–2191, 2009.

[9] M. Yaghoobi, S. Nam, R. Gribonval, and M. Davies, "Noise aware analysis operator learning for approximately cosparse signals," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2012.

[10] S. Hawe, M. Kleinsteuber, and K Diepold, "Analysis operator learning and its application to image reconstruction," submitted.

[11] M. Elad and R. Rubinstein, "K-SVD dictionary-learning for analysis sparse models," in *Workshop on signal processing with adaptive sparse structured representations, SPARS11*, 2011.