



HAL
open science

K-SVD with a real L0 optimization: application to image denoising

Yuan Liu, Stéphane Canu, Paul Honeine, Su Ruan

► **To cite this version:**

Yuan Liu, Stéphane Canu, Paul Honeine, Su Ruan. K-SVD with a real L0 optimization: application to image denoising. Proc. 28th IEEE workshop on Machine Learning for Signal Processing (MLSP), 2018, Aalborg, Denmark. pp.1 - 6, 10.1109/MLSP.2018.8517064 . hal-01965904

HAL Id: hal-01965904

<https://hal.science/hal-01965904>

Submitted on 4 Jan 2019

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.

K-SVD WITH A REAL ℓ_0 OPTIMIZATION: APPLICATION TO IMAGE DENOISING*Yuan Liu, Stéphane Canu, Paul Honeine, Su Ruan*Normandie Univ, INSA Rouen, UNIROUEN, UNIHAVRE, LITIS,
76000 Rouen, France**ABSTRACT**

This paper deals with sparse coding for dictionary learning in sparse representations. Because sparse coding involves an ℓ_0 -norm, most, if not all, existing solutions only provide an approximate solution. Instead, in this paper, a real ℓ_0 optimization is considered for the sparse coding problem providing a global optimal solution. The proposed method reformulates the optimization problem as a Mixed-Integer Quadratic Program (MIQP), allowing then to obtain the global optimal solution by using an off-the-shelf solver. Because computing time is the main disadvantage of this approach, two techniques are proposed to improve its computational speed. One is to add suitable constraints and the other to use an appropriate initialization. The results obtained on an image denoising task demonstrate the feasibility of the MIQP approach for processing well-known benchmark images while achieving good performance compared with the most advanced methods.

Index Terms— Mixed-integer quadratic programming, sparse representation, sparse coding, dictionary learning, image denoising, K-SVD

1. INTRODUCTION

Learning sparse representations to model data, signals and images have been widely investigated since introduced 20 years ago by Olshausen and Field in [1]. In recent years, learning sparse representations have been successfully applied to signal and image processing, as well as computer vision tasks, such as image denoising, image inpainting, object recognition, face recognition and classification and many classification tasks (see [2] and references therein).

A sparse representation describes a given signal by a linear decomposition of few elements of a dictionary. Beyond predefined dictionaries, such as wavelets and many variants [3], data-driven constructed dictionaries allow to have well-adapted and more natural representations for the signals at hand. In practice, the dictionary learning in sparse representation is generally conducted in two alternating steps: dictionary updating and sparse coding. While the former can be easily tackled, sparse coding is the hardest to address since it involves an ℓ_0 -norm.

K-SVD, the most known dictionary learning algorithm, is a two-stage generalization of the k-means algorithm. The first stage operates sparse coding with a greedy scheme, for example matching pursuit (MP) or orthogonal matching pursuit (OMP). The second stage is dictionary updating by a singular value decomposition (SVD) algorithm. K-SVD has solid theoretical foundations, such as the projection theory [4], and good performance on image reconstruction. From this baseline, many variants of K-SVD have been developed in order to address different tasks, such as discriminative K-SVD [5].

In this paper, the sparse coding in its exact ℓ_0 -norm formulation is recast as a mixed-integer quadratic programming (MIQP), namely a mixed-integer programming (MIP) with a quadratic objective function. Even if the use of MIP for pattern recognition is not new [6], only very recently it has been investigated with success to obtain the sparse approximation of a signal [7]. However, due to its computation complexity, this preliminary study was restricted to tiny toy data (120-sample synthesized signals).

We demonstrate in this paper that dictionary learning with K-SVD for image processing can be naturally achieved with MIQP. To this end, we recast the sparse coding problem as a MIQP. We investigate recent theoretical progress in linear program and novel improvements of efficient implementation [8]. To provide an efficient resolution of the resulting MIQP, we propose two techniques to increase the convergence speed by reducing the searching time and decreasing the boundary as well: include appropriate constraints and initialize with the proximal method. These improvements confirm the feasibility of applying the algorithm to image processing. The relevance of these developments is demonstrated on well-known benchmark images frequently used in image processing. Conducted experiments on image denoising show the high tolerance to noise of the proposed MIQP algorithm, outperforming algorithms based on OMP and proximal methods.

The rest of the paper is organized as follows. The sparse representation problem is presented in Section 2. The proposed MIQP-based dictionary learning algorithm and the theoretical analysis are described in Section 3. In Section 4, experimental results on image denoising show that the proposed method can train the dictionary with high tolerance to noise. The last section concludes this paper.

2. PROBLEM STATEMENT

Considering a matrix $Y = [\mathbf{y}_1, \dots, \mathbf{y}_i, \dots, \mathbf{y}_\ell] \in \mathbb{R}^{n \times \ell}$ of ℓ signals of dimension n , a sparse representation of Y consists in finding a matrix $X = [\mathbf{x}_1, \dots, \mathbf{x}_\ell] \in \mathbb{R}^{p \times \ell}$ of decomposition coefficients, which is sparse over a learned dictionary $D = [\mathbf{d}_1, \dots, \mathbf{d}_p] \in \mathbb{R}^{n \times p}$. The columns of the latter, *i.e.*, \mathbf{d}_j for $j = 1, \dots, p$, are called atoms. In image processing, a given image (or a set of images) is fragmented into (often overlapping) patches, where each patch is unfolded to define a “signal” \mathbf{y}_i .

The sparse representation can be obtained by solving the following optimization problem:

$$\min_{D, X} \frac{1}{\ell} \sum_{i=1}^{\ell} \left(\frac{1}{2} \|\mathbf{y}_i - D\mathbf{x}_i\|_2^2 + \lambda \Omega(\mathbf{x}_i) \right), \quad (1)$$

The first term $\frac{1}{2} \|\mathbf{y}_i - D\mathbf{x}_i\|_2^2$ is the reconstruction error with $\|\cdot\|_2$ being the Euclidean norm. The second one includes the regularization term $\Omega(\mathbf{x}_i)$ to enforce sparsity. The regularization parameter $\lambda > 0$ controls the trade-off between data fitting and sparsity of X . For the sake of clarity of this paper, the reconstruction error is measured with the square loss. Generally, the regularization function Ω is associated to a norm that promotes sparsity and its formulation depends on the task at hand [9, 10]. A natural definition of Ω to promote sparsity is the ℓ_0 quasi-norm, *i.e.*, $\Omega(\mathbf{x}) = \|\mathbf{x}\|_0$, which refers to the number of non-zeros of \mathbf{x} .

The problem of estimating simultaneously X and D is non-convex and belongs to NP-hard problems. It is often solved via an alternating strategy:

- Fixing D and finding sparse coefficients X , the procedure is called sparse coding. For each column of X , this problem can be written as the following constrained optimization problem:

$$\min_{\mathbf{x} \in \mathbb{R}^p} \frac{1}{2} \|\mathbf{y} - D\mathbf{x}\|_2^2 \quad \text{subject to} \quad \|\mathbf{x}\|_0 \leq T, \quad (2)$$

where the sparsity parameter T is related to λ .

- Fixing X and finding the solution of D , this procedure is called dictionary updating.

Once convergence is reached, the obtained dictionary can be used in two different settings: in the “adapted dictionary” setting, the dictionary is used to denoise the initial matrix Y of signals, assumed noisy; in the “global” setting, it is used to denoise any new matrix \tilde{Y} of noisy signals.

While the dictionary updating yields a convex optimization problem, sparse coding is more difficult due to the sparsity constraint. For this reason, sparse coding has often been approximated, such as by substituting the ℓ_0 norm with the ℓ_1 norm, and greedy algorithms are often used (*e.g.* MP, OMP). In this paper, we introduce a new algorithm for sparse coding by solving the exact ℓ_0 optimization problem (2) with MIQP.

3. OPTIMIZATION ALGORITHM

Sparse coding addresses intrinsically a bi-objective optimization problem, where both sparsity and reconstruction error need to be optimized. So far, sparse coding has been tackled using approximate algorithms, such as a number of greedy algorithms and descent-based iterative hard thresholding. However, when put aside the computational complexity and memory usage, approximate algorithms fail to obtain the exact solution.

In the following, we cast the sparse coding as a MIQP in order to address the exact ℓ_0 optimization problem, as defined in (2). To this end, we investigate recent theoretical and technical developments in Linear Programming. Moreover, modern hardware’s computational ability provide great speed enhancement [11]. By taking advantage of all these improvements, we demonstrate that applying MIQP to solve sparse coding becomes feasible in the field of image processing.

3.1. Sparse coding as a MIQP

In the following, the sparse coding problem is addressed in its formulation (2). We propose to rewrite this constrained optimization problem, with all the entries of the sparse vector \mathbf{x} indicated by a binary variable $\mathbf{z} \in \{0, 1\}^p$, which can be explained by the logical relation:

$$\begin{cases} z_i = 0, & \text{if } x_i = 0 \\ z_i = 1, & \text{if } x_i \neq 0 \end{cases},$$

where z_i and x_i indicate the i -th entry of the vectors \mathbf{z} and \mathbf{x} , $i = 1, \dots, p$. Since such logical relation cannot be easily integrated into the objective function, we recast the sparsity condition into a linear inequality by introducing a sufficient big value $M > 0$ which should ensure that $\|\hat{\mathbf{x}}\|_\infty < M$ for any desirable solution $\hat{\mathbf{x}}$, where $\|\cdot\|_\infty$ is the maximum norm. A too big M will result in increased feasible region which will make the problem less computational efficient. An appropriate value of M improves the performance. The method to provide a lower M to obtain tight bounds will be discussed in the following.

Now the indicative function of \mathbf{z} is ensured by satisfying the constraints:

$$-z_i M < x_i < z_i M, \quad \forall i \in \{1, \dots, p\}. \quad (3)$$

Then, the sparsity constraint $\|\mathbf{x}\|_0 \leq T$ in formulation (2) can be depicted by \mathbf{z} as:

$$\sum_{i=1}^p z_i \leq T. \quad (4)$$

As a consequence, the ℓ_0 -based sparse coding problem (2) can have a “big- M ” reformulation, that is, for a given M large

enough:

$$\begin{aligned} \min_{\mathbf{x} \in \mathbb{R}^p, \mathbf{z} \in \{0,1\}^p} \quad & \frac{1}{2} \|\mathbf{y} - D\mathbf{x}\|_2^2 \\ \text{subject to} \quad & -\mathbf{z}M < \mathbf{x} < \mathbf{z}M \\ & \mathbf{1}_p^T \mathbf{z} \leq T, \end{aligned} \quad (5)$$

where $\mathbf{1}_p$ is the column vector of size p with all elements equal to one. In this formulation, the optimization variables \mathbf{x} and \mathbf{z} are respectively continuous and integer. The problem is a mixed-integer program (MIP). Moreover, the objective function in (5) is quadratic, and all the constraints are linear. Hence, sparse coding can be interpreted as a MIQP.

We can reformulate our problem as a standard formulation of MIQP by combining the vectors \mathbf{x} and \mathbf{z} , that is, let

$$\mathbf{v} = (\mathbf{x}^T, \mathbf{z}^T)^T,$$

then

$$\begin{aligned} \min_{\mathbf{v}} \quad & \frac{1}{2} \mathbf{v}^T Q \mathbf{v} + \mathbf{c}^T \mathbf{v} \\ \text{subject to} \quad & A_{in} \mathbf{v} \leq \mathbf{b}_{in} \\ & v_j \in \{0, 1\} \quad \forall j \in \mathcal{I}, \end{aligned} \quad (6)$$

where Q is a matrix of size $2p \times 2p$ made up of four sub-matrices

$$Q = \begin{pmatrix} D^T D & \mathbf{0}_{p,p} \\ \mathbf{0}_{p,p} & \mathbf{0}_{p,p} \end{pmatrix},$$

with $\mathbf{0}_{p,q}$ is the zero matrix of size $p \times q$, \mathbf{c} is a column vector of size $2p$ with

$$\mathbf{c} = \begin{pmatrix} -D^T \mathbf{y} \\ \mathbf{0}_{p,1} \end{pmatrix}.$$

The $(2p+1) \times 2p$ matrix

$$A_{in} = \begin{pmatrix} -I_p & -MI_p \\ I_p & MI_p \\ \mathbf{0}_p^T & \mathbf{1}_p^T \end{pmatrix},$$

with I_p the identity matrix of size $p \times p$, and the $(2p+1) \times 1$ column vector $\mathbf{b}_{in} = (\mathbf{0}_{2p}^T, T)^T$, are both obtained according to the inequality of formulation (5). Finally, the set \mathcal{I} in (6) indicates the integer components in the MIQP, namely

$$\mathcal{I} = \{p+1, p+2, \dots, 2p\}.$$

In practice, the variable's type is indicated as continuous or binary in the input for the solver at hand.

To solve this MIQP problem, various optimization software packages can be explored, for example CPLEX and Gurobi Optimizer. The developed tools make it possible to apply MIQP into image processing, but by considering its computational complexity, some effort can be done to improve it as described next.

3.2. Additive constraints

The developments of the MIQP solvers have been following the progress in Linear Programming theory. The advanced-start capabilities of simplex algorithms in the branch-and-bound [12] (or now more correctly, branch-and-cut [13]) search tree are well exploited by MIQP solvers. No matter which optimization technique is used, the search process remains the main time consumption factor. The searching time heavily relies on the feasible region determined by the constraints. Hence, the effort on getting a good formulation of the constraints do help to accelerate the resolution of the optimization problem.

Hoffman and Ralphs have proven in [14] that, if a feasible solution is obtained by a relaxation, then it must also be optimal solution to the original problem. Especially, in the ideal case, if the convex envelop is found, a mixed integer programming will be transformed to the classical linear programming. However, it is an NP-hard problem to find constraints defining the convex envelop. The viable strategy is to create a convex envelop of the continuous variables

$$C = \left\{ \mathbf{x} \in \mathbb{R}^p \mid \mathbf{z} \in \{0, 1\}^p, \sum_{j=1}^p z_j \leq T, |\mathbf{x}_j| \leq z_j T, \right\},$$

by adding the constraint about ℓ_1 -norm and ℓ_∞ -norm of \mathbf{x} :

$$\begin{cases} \sum_{i=1}^p |x_i| < TM \\ |x_i| < M \quad \forall i = 1, \dots, p. \end{cases}$$

However, the absolute value is difficult to be formulated as linear programs. To overcome this difficulty, we replace each unrestricted variable x_i , for $i = 1, \dots, p$, with the difference of two restricted variables $x_i = x_i^+ - x_i^-$, with $x_i^+, x_i^- \geq 0$. Then the absolute value of x_i in the above constraints can be represented in the linear program as $|x_i| = x_i^+ + x_i^-$. Thus, the constraints for MIQP can be summarized as:

$$\begin{cases} \sum_{i=1}^p x_i^+ + x_i^- < TM \\ -\mathbf{z}M < \mathbf{x}^+ - \mathbf{x}^- < \mathbf{z}M \\ 0 \leq x_i^+, x_i^- < M \\ \mathbf{1}_p^T \mathbf{z} \leq T \end{cases} \quad (7)$$

With the new constraints, MIQP can be reformulated as the standard formulation by introducing as updated optimization variable $\mathbf{v} = (\mathbf{x}^{+T}, \mathbf{x}^{-T}, \mathbf{z}^T)^T$, and updated model components Q , \mathbf{c} , A_{in} , \mathbf{b}_{in} , \mathbf{l}_b and \mathbf{u}_b correspondingly.

3.3. Initialization by the proximal method

MIQP solvers are based on the search tree theory [13]. The MIQP problem, represented by the root of the tree, is partitioned into subproblems. And the feasible region is also divided into subregions. The objective value of any feasible solution to a subproblem provides an upper bound on the global optimal value. The optimal solution is produced when the

global lower bound and global upper bound are equal. Usually, a global bound is needed to make the algorithm more efficient. Hence, a good initialization or tight bounds can both help to improve the performance. In the following, the proximal method will be applied to give a good initialization and an optimized value for M which forms the global bound of the problem.

The proximal method is based on the first order approximation method. It produces a reasonable approximate solution by minimization a succession of upper bound of the objective function. It is a powerful tool to tackle non-smooth, constrained, large-scale, and distributed optimization problems. The proximal operator is expressed as

$$\text{prox}_t^h(\mathbf{u}) = \arg \min_{\mathbf{x}} \left(h(\mathbf{u}) + \frac{t}{2} \|\mathbf{x} - \mathbf{u}\|^2 \right),$$

where h defines a proper and lower semi-continuous function, and $t > 0$ is a step size parameter. See [15] for more details.

For our problem, let $H(\mathbf{x})$ denote the quadratic objective function in the optimization problem (6), and $h(\mathbf{x})$ the function that makes sure that the feasible region is in the space \mathcal{S} of T -sparsity, that is

$$h(\mathbf{x}) = \begin{cases} 0 & \text{if } \|\mathbf{x}\|_0 \leq T \\ \infty & \text{otherwise.} \end{cases} \quad (8)$$

The proximal operator boils down to the projection onto the sparse space \mathcal{S} :

$$P_{\mathcal{S}}(\mathbf{u}) := \arg \min_{\mathbf{x} \in \mathcal{S}} (\|\mathbf{x} - \mathbf{u}\|^2).$$

The solution to this problem can be easily obtained by keeping T biggest absolute value components of \mathbf{u} and setting the rest to zeros:

$$P_{\mathcal{S}}(\mathbf{u}) = \begin{cases} u_j & \text{if } j \in \{(1), \dots, (T)\} \\ 0 & \text{otherwise} \end{cases}$$

where j is the index of the sequences that $|u_{(1)}| \geq |u_{(2)}| \geq \dots \geq |u_{(p)}|$. By applying a proximal algorithm, the sparse representation problem can be solved through a serial update process [9]:

$$\mathbf{x}^{k+1} \in P_{\mathcal{S}}(\mathbf{x}^k - t \nabla H(\mathbf{x}^k)), \quad (9)$$

where $\nabla H(\mathbf{x}) = D^T D \mathbf{x} - D^T \mathbf{y}$. After a finite number of iterations n_{iter} , the $\mathbf{x}^{n_{iter}}$ will be much approximate to the optimal solution of the MIQP problem. By considering the definition of ‘big- M ’, the constraints in the problem (7) related to M can be well determined by an approximation of the optimal solution. A simple method to determine an appropriate value for M can be:

$$M = \mu \|\mathbf{x}^{n_{iter}}\|_{\infty}. \quad (10)$$

4. EXPERIMENTAL RESULTS

Experiments on image denoising are conducted in this section to evaluate the proposed sparse coding algorithm relying on MIQP, and compare it with OMP and the proximal method for sparse coding. The dictionary updating method is always SVD.

4.1. Design of experiments

We choose well-known images in the USC-SIPI Image Database¹ for experiments. The dataset contains five frequently used images in image processing, as presented in the Figure 1. The images, of size 512×512 , 256×256 or 580×720 , are corrupted with an additive zero-mean white Gaussian noise of standard deviation $\sigma = 50$, which corresponds to the highest noise level in denoising benchmarks [16].

Gurobi Optimizer v7.0.2 is chosen to solve the MIQP optimization problem. We run the software in the Matlab environment on a server with 4 Intel[®] Xeon[®] processors with a CPU clocked at 2.4 GHz. The parameters settings of Gurobi are: TimeLimit 50 and IterationLimit 200. For initialization by the proximal method, the number of iterations is also set to 200. The coefficient to decrease M is set to $\mu = 2.5$. For comparative purpose, the popular dictionary learning algorithms K-SVD with OMP [16], its proximal version [9], and the proposed method are executed under the same conditions, namely the total number of iterations in K-SVD is set to 30.

The experiments are conducted using two different settings: large-scale (global) dictionary learning and adapted dictionary learning, as detailed below. For denoising some given \tilde{Y} , two reconstruction models are investigated for this purpose, the direct one uses the conventional direct reconstruction model

$$\hat{Y} = DX, \quad (11)$$

and the one proposed by Elad and Aharon in [16] with

$$\hat{Y} = \left(\lambda I + \sum_{ij} R_{ij}^T R_{ij} \right)^{-1} \left(\lambda \tilde{Y} + \sum_{ij} R_{ij}^T D \mathbf{x}_{ij} \right), \quad (12)$$

where the matrix R_{ij} is the matrix extracting the (i, j) -th block from the image, and λ is set to $30/\sigma$ as recommended in [16]. To assess the quality of the denoising, the peak signal-to-noise ratio (PSNR) is considered.

4.2. Large-scale (global) dictionary learning

In this setting, a single global dictionary is learned for all the original images. More than $\ell \approx 3.5 \times 10^4$ overlapping patches of size $n = 8 \times 8$ from all the images are extracted to get a single training dataset which forms the signals matrix Y . The

¹<http://sipi.usc.edu/database/database.php?volume=misc>

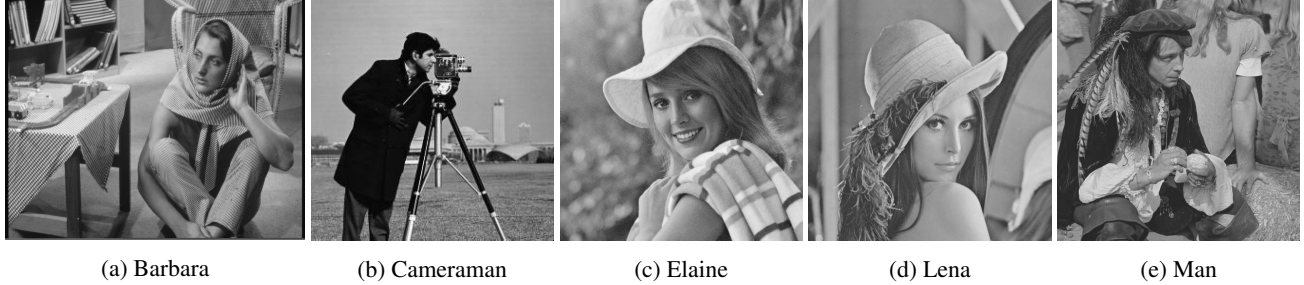


Fig. 1: Images under study from the USC-SIPI Image Database

number of the atoms is set to $p = 100$ and the sparsity level is $T = 20$ (these parameters are determined by preliminary experiments). We do not take any assumption on the noise level. To assess the relevance of the obtained dictionary, tests are conducted on noisy images \tilde{Y} .

4.3. Adapted dictionary learning

In this setting, the dictionary is trained on the corrupted image under scrutiny. By considering the time consumption, a segment of size 121×121 in the images is selected. As in the first experiments, the signal matrix is created in the same way using overlapping patches. For each corrupted image, an adapted dictionary is trained and used for denoising it.

This setting is exactly the one proposed in [16], namely the knowledge about the noise level σ is used for restricting the reconstruction error in the dictionary training, as presented with the constraint in

$$\min_{\mathbf{x} \in \mathbb{R}^p} \|\mathbf{x}\|_0 \quad \text{subject to} \quad \|\mathbf{y} - D\mathbf{x}\|_2^2 \leq \epsilon,$$

where $\epsilon = cn\sigma^2$ with $c = 1.15$ for OMP, as recommended in [16]. These values, initially optimized for OMP, are used here for proximal method and MIQP, thus putting our method in a non favorable situation. Indeed, it is easy to make OMP satisfies the constraint by measuring the reconstruction error each time after adding a non-zero entry; The proximal method searches for the Pareto optimal when the sparsity level varies [17]; MIQP keeps all the signals in the constraint based on the decided sparsity of initialization obtained by the proximal method. As in the first setting, T is set during training to 20 for the proximal and MIQP methods. The number of atoms is chosen from a set of 14 candidate values $\{50, 55, \dots, 110, 150, 200, 256, 300\}$, enclosing the most used values in the literature, obtaining 256 for OMP as suggested in [16] where extensive experiments were conducted.

4.4. Results

The reconstruction accuracy is given in Table 1. These results show that the proposed method outperforms OMP and

proximal methods in a high noise level. For the large-scale (global) dictionary learning setting, MIQP provides important enhancements, with an average improvement of 1.79 with regard to the proximal method and 3.73 with regard to the OMP algorithm. For the adapted dictionary learning setting, the average improvement is 0.44 with regard to OMP, which is important in the logarithmic-scale PSNR (*e.g.* it is greater than the refinement of (12) compared with (11)). It is worth noting that the enhancement is significant since, the parameters were optimized for OMP. According to the experiments, we can conclude that MIQP can reconstruct the signal more accurately with a high tolerance to noise.

In despite of its performance, the proposed MIQP method has a high computational complexity in implementation. Since we have different sizes of the training data in each setting (global or adapted), training time is not comparable. In the following, we focus on the average time of reconstruction of a single image. Comparing with a couple of minutes for the reconstruction of an image by OMP or the proximal method, MIQP needs about one hour. However, recent advances in MIQP solvers allow to reduce this gap, as explained in next section.

5. CONCLUSION AND FUTURE WORK

In this paper, the K-SVD algorithm was revisited by proposing the exact optimization method MIQP for sparse coding. Thanks to recent advances in linear programming techniques, as well as more powerful hardware, the speed of computation of MIQP has been greatly improved. Furthermore, by introducing additive constraints and an appropriate initialization, it was proved that it is feasible to use MIQP for sparse coding to redefine the K-SVD algorithm, and apply it in image processing. Though, the MIQP method had much more computational cost compared with the approximate methods, the feasibility of the method was proved on well-known images. Moreover, the image denoising experiments showed the advantage of the proposed MIQP method. Furthermore, the high noise-tolerance of our method was demonstrated on both the large-scale global and the adapted dictionary learning settings.

Table 1: Accuracy of the reconstruction in terms of the PSNR (the higher, the better)

Image	Barbara			Cameraman			Elaine			Lena			Man			Average over the five images		
	OMP	Prox	MIQP	OMP	Prox	MIQP	OMP	Prox	MIQP	OMP	Prox	MIQP	OMP	Prox	MIQP	OMP	Prox	MIQP
Large-scale dictionary learning, reconstruction with (11)	19.71	20.71	22.73	19.46	21.11	22.30	19.73	22.87	24.20	19.79	22.12	24.20	19.68	21.26	23.59	19.67	21.61	23.40
Large-scale dictionary learning, reconstruction with (12)	20.03	21.03	23.05	19.78	21.43	22.62	20.05	23.19	24.52	20.11	22.44	24.52	20.00	21.58	23.91	19.99	21.93	23.72
Adapted dictionary learning (ϵ optimized for OMP)	22.04	22.54	22.59	22.54	22.49	22.58	23.00	23.29	23.39	22.48	23.08	23.09	21.23	21.70	21.86	22.25	22.62	22.70

The major contribution of this paper is demonstrating that one can solve the exact ℓ_0 optimization problem in dictionary learning for real images. While the computational complexity remains its Achilles heel, great improvements are being carried out these days on MIQP solvers; the new Gurobi Optimizer v8.0, released a couple of days prior to the submission of this paper, is more than 220% faster² on MIQP problems than the one used in this paper. Moreover, an implementation on GPU, currently unavailable in off-the-shelf solvers (e.g. Gurobi, CPLEX), will certainly provide great improvements.

As for future work, we will address the problem of computational complexity by using recently proposed convex reformulation [18]. Furthermore, we will extend this work beyond K-SVD to deal with classification and segmentation.

6. REFERENCES

- [1] Bruno A Olshausen and David J Field, “Sparse coding with an overcomplete basis set: A strategy employed by v1?,” *Vision research*, vol. 37, no. 23, pp. 3311–3325, 1997.
- [2] Julien Mairal, Francis Bach, Jean Ponce, et al., “Sparse modeling for image and vision processing,” *Foundations and Trends® in Computer Graphics and Vision*, vol. 8, no. 2-3, pp. 85–283, 2014.
- [3] Marc Antonini, Michel Barlaud, Pierre Mathieu, and Ingrid Daubechies, “Image coding using wavelet transform,” *IEEE Transactions on image processing*, vol. 1, no. 2, pp. 205–220, 1992.
- [4] Karin Schnass, “On the identifiability of overcomplete dictionaries via the minimisation principle underlying K-SVD,” *Applied and Computational Harmonic Analysis*, vol. 37, no. 3, pp. 464–491, 2014.
- [5] Avi Ben-Cohen, Eyal Klang, Michal Amitai, and Hayit Greenspan, “Sparsity-based liver metastases detection using learned dictionaries,” in *Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on*. IEEE, 2016, pp. 1195–1198.
- [6] M Rebollo and Laureano F Escudero, “A mixed integer programming approach to multi-spectral image classification,” *Pattern Recognition*, vol. 9, no. 1, pp. 47–57, 1977.
- [7] Sébastien Bourguignon, Jordan Ninin, Hervé Carfantan, and Marcel Mongeau, “Exact sparse approximation problems via mixed-integer programming: Formulations and computational performance,” *IEEE Transactions on Signal Processing*, vol. 64, no. 6, pp. 1405–1419, 2016.
- [8] Christian Blik, Pierre Bonami, and Andrea Lodi, “Solving mixed-integer quadratic programming problems with IBM-CPLEX: a progress report,” in *Proceedings of the Twenty-Sixth RAMP Symposium*, 2014, pp. 16–17.
- [9] Chenglong Bao, Hui Ji, Yuhui Quan, and Zuwei Shen, “ ℓ_0 norm based dictionary learning by proximal methods with global convergence,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3858–3865.
- [10] Michal Aharon, Michael Elad, and Alfred Bruckstein, “K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation,” *IEEE Transactions on signal processing*, vol. 54, no. 11, pp. 4311–4322, 2006.
- [11] Robert E Bixby, “A brief history of linear and mixed-integer programming computation,” *Documenta Mathematica*, pp. 107–121, 2012.
- [12] Oskar von Stryk and Markus Glocker, “Decomposition of mixed-integer optimal control problems using branch and bound and sparse direct collocation,” in *Proc. ADPM 2000 - The 4th International Conference on Automation of Mixed Processes Hybrid Dynamic Systems, Dortmund, September 18-1, 2000*, pp. 99–104.
- [13] Arnold Neumaier and Oleg Shcherbina, “Safe bounds in linear and mixed-integer linear programming,” *Mathematical Programming*, vol. 99, no. 2, pp. 283–296, 2004.
- [14] Karla L Hoffman and Ted K Ralphs, “Integer and combinatorial optimization,” in *Encyclopedia of Operations Research and Management Science*, pp. 771–783. Springer, 2013.
- [15] Jérôme Bolte, Shoham Sabach, and Marc Teboulle, “Proximal alternating linearized minimization for nonconvex and nonsmooth problems,” *Mathematical Programming*, vol. 146, no. 1-2, pp. 459–494, 2014.
- [16] Michael Elad and Michal Aharon, “Image denoising via sparse and redundant representations over learned dictionaries,” *IEEE Transactions on Image processing*, vol. 15, no. 12, pp. 3736–3745, 2006.
- [17] Charles Soussen, Jérôme Idier, Junbo Duan, and David Brie, “Homotopy based algorithms for L0-regularized least-squares,” *IEEE Transactions on Signal Processing*, vol. 63, no. 13, pp. 3301–3316, 2015.
- [18] Alain Billionnet, Sourour Elloumi, and Amélie Lambert, “Exact quadratic convex reformulations of mixed-integer quadratically constrained problems,” *Mathematical Programming*, vol. 158, no. 1-2, pp. 235–266, 2016.

²<http://www.gurobi.com/>