16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP

REGULARIZED DICTIONARY LEARNING FOR SPARSE APPROXIMATION

M. Yaghoobi, T. Blumensath, M. Davies

Institute for Digital Communications,
Joint Research Institute for Signal and Image Processing,
University of Edinburgh, UK

ABSTRACT

Sparse signal models approximate signals using a small number of elements from a large set of vectors, called a dictionary. The success of such methods relies on the dictionary fitting the signal structure. Therefore, the dictionary has to be designed to fit the signal class of interest. This paper uses a general formulation that allows the dictionary to be learned form the data with some *a priori* information about the dictionary. In this formulation a universal cost function is proposed and practical algorithms are presented to minimize this cost under different constraints on the dictionary. The proposed methods are compared with previous approaches using synthetic and real data. Simulations highlight the advantages of the proposed methods over other currently available dictionary learning strategies.

1. INTRODUCTION

Signals can be approximated using overcomplete representations with more elementary functions (atoms) than the dimension of the signal. These representations are not unique for a given set of atoms. A sparse representation is an overcomplete representation that uses the minimal number of non-zero coefficients. For example, sparse representations have been used for low bitrate coding, denoising and source separation. Let $\mathbf{y} \in \mathbb{R}^d$ and $\mathbf{x} \in \mathbb{R}^N$ (where $d < N$) be the input and the coefficient vectors and let the matrix $\mathbf{D} \in \mathbb{R}^{d \times N}$ be the *dictionary*. One form of sparse approximation is to solve an unconstrained optimization problem,

$$\min_{\mathbf{x}} \Phi(\mathbf{x}) \; ; \; \Phi(\mathbf{x}) = ||\mathbf{y} - \mathbf{D}\mathbf{x}||^2 + \lambda ||\mathbf{x}||_0 \tag{1}$$

where $||\mathbf{x}||_0$ and λ are the sparsity measure (which counts the number of non-zero coefficients) and a constant multiplier respectively. This problem is NP-hard in general. Therefore various relaxed sparsity measures have been presented to make the problem tractable. A commonly used class of measures are $||\mathbf{x}||_p^p = \sum_i |x_i|^p$ with $0 < p \leq 1$.

When the generative model for the signals is unknown, appropriate dictionary learning algorithms can be used to adaptively find better dictionaries for a set of training samples. We are thus searching for a set of elementary functions that allow the set of training signals to be represented sparsely and with a small approximation error.

In this paper we consider the dictionary learning problem as a constrained optimization problem with two sets of parameters, coefficient matrix and dictionary. The constraints are generalizations of those in [1]. The proposed constrained optimization problem is converted into an unconstrained optimization problem using Lagrangian multipliers. We then present reasonably fast methods to update the dictionary. A comparison between the proposed method and other dictionary learning methods is presented.

2. DICTIONARY LEARNING METHODS

In dictionary learning, one often starts with some initial dictionary and finds sparse approximations of the set of training signals whilst keeping the dictionary fixed. This is followed by a second step in which the sparse coefficients are kept fixed and the dictionary is optimized. This algorithm runs for a specific number of alternating optimizations or until a specific approximation error is reached. The proposed method is based on such an alternating optimization (or block-relaxed optimization) method with some advantages over the current methods in the condition and speed of convergence.

If the set of training samples is $\{\mathbf{y}^{(i)} \; : \; 1 \leq i \leq L\}$, where L is the number of training vectors, then sparse approximations are often found (for all $i: \; 1 \leq i \leq L$) by,

$$\min_{\mathbf{x}^{(i)}} \Phi_i(\mathbf{x}^{(i)}) \; ; \; \Phi_i(\mathbf{x}) = ||\mathbf{y}^{(i)} - \mathbf{D}\mathbf{x}||^2 + \lambda ||\mathbf{x}||_p^p \tag{2}$$

An alternative to minimizing (2) individually on each vector is to find a joint sparse approximation of the matrix $\mathbf{Y} = [\mathbf{y}^{(1)} \; \mathbf{y}^{(2)} \; ... \; \mathbf{y}^{(L)}]$ by employing a sparsity measure in matrix form. The sparse matrix approximation problem can be formulated as,

$$\min_{\mathbf{X}} \Phi(\mathbf{X}) \; ; \; \Phi(\mathbf{X}) = ||\mathbf{Y} - \mathbf{D}\mathbf{X}||_F^2 + \lambda J_{p,p}(\mathbf{X}), \tag{3}$$

where $J_{p,q}(\mathbf{X})$ is defined as ([2]),

$$J_{p,q}(\mathbf{X}) = \sum_{i \in I} [\sum_{j \in J} |x_{ij}|^q]^{p/q}. \tag{4}$$

$||\mathbf{X}||_F = J_{2,2}^{1/2}(\mathbf{X})$ would be the Frobenius-norm. When $p = q$ all elements in $\mathbf{X}$ are treated equally.

The second step in dictionary learning is the optimization of the dictionary based on the current sparse approximation. The cost function in (3) can be thought of as an objective function with two parameters,

$$\Phi(\mathbf{D}, \mathbf{X}) = ||\mathbf{Y} - \mathbf{D}\mathbf{X}||_F^2 + \lambda J_{p,p}(\mathbf{X}) \tag{5}$$

Without additional constraints on the dictionary, minimizing the above objective function is an ill-posed problem. An obvious solution is $\mathbf{D} \to \infty, \mathbf{X} \to 0 \; s.t. \; \mathbf{D}\mathbf{X} = \mathbf{Y}$. By constraining the norm of $\mathbf{D}$ we can exclude these undesired solutions. Dictionaries with fixed column-norms or fixed Frobenius-norm have been used in different papers (for example [3] and [1]). We present the more general admissible sets assuming "'*bounded* column-norm" and "*bounded* Frobenius-norm".

In the Method of Optimal Directions (MOD) [3] the best D is found by using the pseudo inverse of X followed by re-normalizing each atom. The Maximum Likelihood Dictionary Learning algorithm (ML-DL), which is presented in [4], is similar to MOD but uses gradient optimization. If the update is done iteratively, we find the best possible dictionary update without any constraint (similar to MOD). This update is followed by normalizing atoms based on the variance of the corresponding coefficients. The dictionary normalization step in these methods may increase total approximation error. The Maximum *a Posteriori* dictionary learning algorithm (MAP-DL) [1] is based on the assumption that '*a priori*' information is available about the dictionary. By the use of an iterative

This work is funded by EPSRC grant number D000246

method, if the algorithm converges, it finds a dictionary consistent with this *a priori* information [1]. When a fixed column-norm constraint is used, the algorithm updates atom by atom, making the method too slow to be used for many real signals [5].

The K-SVD method presented in [5] is fundamentally different from these methods. In the dictionary update step, the supports of the coefficient vectors (the positions of the non-zero coefficients) is fixed and an update of each atom is found as the best normalized elementary function that matches the errors (calculated after representing the signals with all atoms except the currently selected atom).

The dictionary learning approach proposed in this paper has several similarities with the formulation used in MAP-DL. However, our approach is based on a joint cost function for both, the sparse approximation and the dictionary update and uses a new class of constraints on the desired dictionaries. Furthermore, the algorithms presented to solve the problem are different and are proven to converge. Because the proposed cost functions are not convex, using gradient based methods to update the dictionary will not in general find the global optimum and, like the other methods mentioned above, the algorithms presented in this paper are only guaranteed to find a local minimum.

3. REGULARIZED DICTIONARY LEARNING (RDL)

In this section we consider the dictionary learning problem as a constrained optimization problem.

$$\min_{\mathbf{D},\mathbf{X}} \Phi(\mathbf{D},\mathbf{X}) \; s.t. \; \mathbf{D} \in \mathscr{D} \; ; \; \Phi(\mathbf{D},\mathbf{X}) = ||\mathbf{Y}-\mathbf{D}\mathbf{X}||_F^2 + \lambda J_{p,p}(\mathbf{X}) \quad (6)$$

where $\mathscr{D}$ is some admissible set. In an iterative two-step optimization scheme, we find the optimum $\mathbf{X}$ with fixed $\mathbf{D}$ in one of the steps. In this paper we use iterative thresholding (IT) [6] for this optimization. In this algorithm a convex function is added to the objective function to decouple the optimization of the x_{ij}. Then the convex function is updated based on the current solution and the algorithm continues with the new objective function. The objective function in (6) and the added convex function have matrix valued parameters leading to a generalization for the IT method.

In every other step of the dictionary learning algorithm we update the dictionary. As noted in [1], two typical constraints are the unit Frobenius-norm and the unit column-norm constraints, both of which lead to non-convex solution sets. In addition to these constraints, the algorithms proposed in this paper can also be used to solve (6) if bounded norm constraints (defined later) are used. With these, the algorithms are guaranteed to find the global optimum within the dictionary update step. Note that (5) is a convex function of $\mathbf{D}$ (for fixed $\mathbf{X}$) and of $\mathbf{X}$ (for fixed $\mathbf{D}$), but it is not convex as a function of both, $\mathbf{X}$ and $\mathbf{D}$, so that the alternating optimization of (3) is not guaranteed to find a global optimum.

Note that if the sparsity measure in the sparse approximation step penalizes coefficients based on their magnitudes (for example $l_p : 0 < p \leq 1$), it is easy to show that the fixed points of the algorithm are on the boundary of the convex sets.

3.1 Constrained Frobenius-Norm Dictionaries

In this section we derive an algorithm for the case in which we constrain the Frobenius-norm of $\mathbf{D}$. An advantage of using a constraint on the Frobenius-norm is that the dictionary size can be reduced during dictionary learning by pruning out atoms whose norm becomes small. Another advantage is that the learned dictionary will have atoms with different norms as used in the weighted-pursuit framework [7]. Atoms with large norm then have more chance of appearing in the approximation. It has been shown that the average performance of the sparse approximation increases when the weights are chosen correctly for the class of signals under study [7].

The admissible set for the *bounded* Frobenius-Norm dictionaries is,

$$\mathscr{D} = \{\mathbf{D}_{d\times N} : ||\mathbf{D}||_F \leq c_F^{1/2}\} \quad (7)$$

where c_F is a constant. With the help of a Lagrangian multiplier γ we turn this into an unconstrained optimization problem,

$$\min_{\mathbf{D}} \Phi_\gamma(\mathbf{D},\mathbf{X}), \quad (8)$$

where $\Phi_\gamma(\mathbf{D},\mathbf{X})$ is defined as,

$$\Phi_\gamma(\mathbf{D},\mathbf{X}) = ||\mathbf{Y}-\mathbf{D}\mathbf{X}||_F^2 + \lambda J_{p,p}(\mathbf{X}) + \gamma(||\mathbf{D}||_F^2 - c_F). \quad (9)$$

The solution to the above minimization problem is a global minimum if the solution satisfies the K.K.T conditions [8, Theorem 28.1]. The admissible set is convex, so any minimum of $\Phi_\gamma(\mathbf{D},\mathbf{X})$ is an optimal solution if $\gamma(||\mathbf{D}||_F^2 - c_F) = 0$. Therefore if $||\mathbf{D}||_F^2 \neq c_F$ then γ must be zero. The objective function is differentiable in $\mathbf{D}$. Therefore its minimum is a point with zero gradient. For fixed $\mathbf{X}$,

$$\begin{aligned} d\Phi_\gamma(\mathbf{D},\mathbf{X}) &= d\,tr\{\mathbf{X}^T\mathbf{D}^T\mathbf{D}\mathbf{X} - \mathbf{X}^T\mathbf{D}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{D}\mathbf{X} \\ &\quad + \mathbf{Y}^T\mathbf{Y}\} + \gamma.\, d\,tr\{\mathbf{D}^T\mathbf{D}\} \\ &= (2\mathbf{X}\mathbf{X}^T\mathbf{D}^T - 2\mathbf{X}\mathbf{Y}^T + 2\gamma\mathbf{D}^T)d\mathbf{D} \end{aligned}$$

$$\Rightarrow \frac{d}{d\mathbf{D}}\Phi_\gamma(\mathbf{D},\mathbf{X}) = 2\mathbf{X}\mathbf{X}^T\mathbf{D}^T - 2\mathbf{X}\mathbf{Y}^T + 2\gamma\mathbf{D}^T = \mathbf{0}$$

$$\Rightarrow \mathbf{D} = \mathbf{Y}\mathbf{X}^T(\mathbf{X}\mathbf{X}^T + \gamma\mathbf{I})^{-1} \quad (10)$$

$\Phi_\gamma(\mathbf{D},\mathbf{X})$ is a non-negative convex function of $\mathbf{D}$ and this solution is minimal. To find the appropriate γ satisfying the K.K.T condition, we note that $\Phi_\gamma(\mathbf{D},\mathbf{X})$ is a continuous function of γ (in the regions in which $(\mathbf{X}\mathbf{X}^T + \gamma\mathbf{I})$ is not singular). Therefore if $\mathbf{D}$ as calculated by (10) and with $\gamma = 0$ is admissible, this $\mathbf{D}$ is the optimum solution. If (10) does not give an admissible solution, we can use a line-search method to find a $\gamma \neq 0$ such that $||\mathbf{D}||_F = c_F^{1/2}$ (by changing γ in the direction which reduces $|||\mathbf{D}||_F - c_F^{1/2}|$). Interestingly, MOD uses $\mathbf{D} = \mathbf{Y}\mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}$, whilst our update uses a *regularized* pseudo inverse.

If we use an equality in the definition of (7) to get the fixed Frobenius-norm constraint, the set becomes non-convex so that we might only find a local minimum, in which case γ could become negative.

3.2 Constrained Column-Norm Dictionaries

The admissible set for the *bounded* column-norm dictionary is defined as,

$$\mathscr{D} = \{\mathbf{D}_{d\times N} : ||\mathbf{d}_i||_2 \leq c_C^{1/2}\}, \quad (11)$$

where $\mathbf{d}_i$ is the i$^{\text{th}}$ column of the dictionary and c_C is a constant. This admissible set is again a convex set. However, now we need N (number of columns in $\mathbf{D}$) Lagrangian multipliers (equal to the number of constraints) and the unconstrained optimization turns to,

$$\min_{\mathbf{D}} \Phi_{\mathbf{\Gamma}}(\mathbf{D},\mathbf{X}), \quad (12)$$

where $\Phi_{\mathbf{\Gamma}}(\mathbf{D},\mathbf{X})$ is defined as,

$$\Phi_{\mathbf{\Gamma}}(\mathbf{D},\mathbf{X}) = ||\mathbf{Y}-\mathbf{D}\mathbf{X}||_F^2 + \lambda J_{p,p}(\mathbf{X}) + \sum_{i=1}^{N} \gamma_i(\mathbf{d}_i^T\mathbf{d}_i - c_C) \quad (13)$$

With this formulation, the K.K.T conditions are,

$$\forall i : 1 \leq i \leq N, \qquad \gamma_i(\mathbf{d}_i^T\mathbf{d}_i - c_C) = 0\,. \quad (14)$$

This means that for each i if $\mathbf{d}_i^T\mathbf{d}_i$ is not equal to c_C then γ_i should be zero. (12) can be rewritten as

$$\Phi_{\mathbf{\Gamma}}(\mathbf{D},\mathbf{X}) = ||\mathbf{Y}-\mathbf{D}\mathbf{X}||_F^2 + \lambda J_{p,p}(\mathbf{X}) + tr\{\mathbf{\Gamma}(\mathbf{D}^T\mathbf{D} - c_C\mathbf{I})\}, \quad (15)$$

where Γ is a diagonal matrix with the γ_is as the diagonal elements. If we use a similar method as before we get an optimum at,

$$\mathbf{D} = \mathbf{Y}\mathbf{X}^T(\mathbf{X}\mathbf{X}^T + \mathbf{\Gamma})^{-1} \tag{16}$$

Even though the minimum seems to be similar to (10), finding $\mathbf{\Gamma}$ is now more difficult as we can no longer use a line search.

Instead of optimizing the original objective function (15) directly we can use an iterative method. By adding a convex function of $\mathbf{D}$ to (15) we get the surrogate function,

$$\Phi^S_{\mathbf{\Gamma}}(\mathbf{D},\mathbf{B},\mathbf{X}) = \Phi_\Gamma(\mathbf{D},\mathbf{X}) + c_s||\mathbf{D}-\mathbf{B}||_F^2 - ||\mathbf{D}\mathbf{X}-\mathbf{B}\mathbf{X}||_F^2 \tag{17}$$

where $\mathbf{B}$ is a $d \times N$ matrix that is set to the previous solution of $\mathbf{D}$ ($\mathbf{D}^{[n-1]}$) in each iteration. c_s is a constant such that $||\mathbf{X}^T\mathbf{X}||_2 < c_s$. To minimize the surrogate function we set the gradient to zero.

$$\frac{d}{d\mathbf{D}}\Phi^S_{\mathbf{\Gamma}}(\mathbf{D},\mathbf{D}^{[n-1]},\mathbf{X}) = -2\mathbf{X}\mathbf{Y}^T + 2\mathbf{X}\mathbf{X}^T\mathbf{D}^{[n-1]^T} + 2c_s\mathbf{D}^T - 2c_s\mathbf{D}^{[n-1]^T} + 2\mathbf{\Gamma}\mathbf{D}^T = \mathbf{0}$$

$$\Rightarrow \mathbf{D}^{[n]} = (\mathbf{Y}\mathbf{X}^T + \mathbf{D}^{[n-1]}(c_s\mathbf{I} - \mathbf{X}\mathbf{X}^T))(\mathbf{\Gamma} + c_s\mathbf{I})^{-1} \tag{18}$$

All γ_is are non-negative and $(\mathbf{\Gamma} + c_s\mathbf{I})$ is a diagonal matrix. Therefore $(\mathbf{\Gamma} + c_s\mathbf{I})$ is invertible. In equation (18) by changing γ_i we multiply the corresponding column of $\mathbf{Y}\mathbf{X}^T + \mathbf{D}^{[n-1]}(c_s\mathbf{I} - \mathbf{X}\mathbf{X}^T)$ by a scalar and we can *regulate* the norm of each column in $\mathbf{D}$ by the corresponding γ_i. We start with all $\gamma_i = 0$ and for any column of $\mathbf{D}$ for which the norm is more than one, we find the smallest value for γ_i that normalizes that column. In other words, we find $\mathbf{D}_\sharp = \mathbf{Y}\mathbf{X}^T + \mathbf{D}^{[n-1]}(c_s\mathbf{I} - \mathbf{X}\mathbf{X}^T)$ and then project $\mathbf{D}_\sharp$ onto the admissible set to find $\mathbf{D}^{[n]}$. The algorithm starts with the dictionary $\mathbf{D}^{[0]} = \mathbf{D}_i$ and iteratively reduces the surrogate objective function. We can run the algorithm for a specific number of iterations or stop based on the distance between the dictionaries in two consecutive iterations ($||\mathbf{D}^{[n]} - \mathbf{D}^{[n-1]}||_F < \xi$), for a small positive constant ξ). This iterative method can be shown to converge to the minimum of the original objective function (15) ($\mathbf{X}$ fixed). Alternatively, we can again set the constraint set to have fixed column-norm ($\mathscr{D} = \{\mathbf{D}_{d\times N} : ||\mathbf{d}_i||_2 = c_c^{1/2}\}$). Here the algorithm may find a local minimum and some of the γ_i might become negative.

4. SIMULATIONS

We evaluate the proposed methods with synthetic and real data. Using synthetic data with random dictionaries helps us to examine the ability of the proposed methods to recover dictionaries exactly (to within an acceptable squared error). We generated the synthetic data and dictionaries as proposed in [1] and [5]. To evaluate the performance on real data, we chose an audio signal, which has been shown to have some sparse structure. We then used the learned dictionary for audio coding and show some improvements in the Rate-Distortion performance.

4.1 Synthetic Data

A 20×40 matrix $\mathbf{D}$ was generated by normalizing a matrix with i.i.d. uniform random entries. The number of non-zero elements was selected between 3 and 7 to generate different sparse coefficient vectors. The locations of the non-zero coefficients were selected uniformly at random. For the unit column-norm dictionary learning, we generated 1280 training samples where the absolute values of the non-zero coefficients were selected uniformly between 0.2 and 1. Iterative Thresholding (IT) [6] was used to optimize (3) using the ℓ_1 measure. This was followed by orthogonal projection onto the selected sub-spaces (to find the best representation in that subspace). The stopping criteria for IT was the distance between two consecutive iterations ($\delta = 3 \times 10^{-4}$) and λ was set to 0.4 . The termination conditions for the iterative dictionary learning methods (RDL and MAP-DL) was set to ($||\mathbf{D}^{[n]} - \mathbf{D}^{[n-1]}||_F < 10^{-7}$).

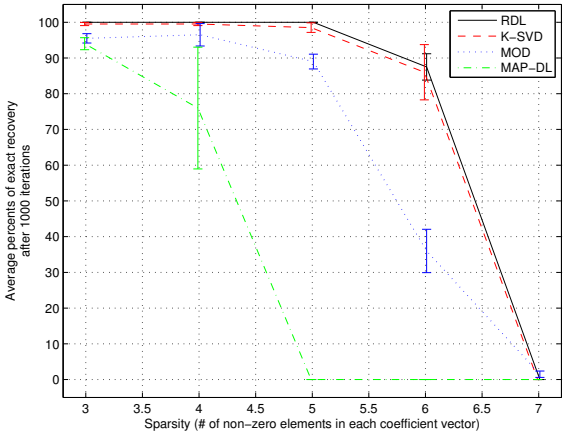

Figure 1: Exact recovery with fixed column-norms dictionary learning.

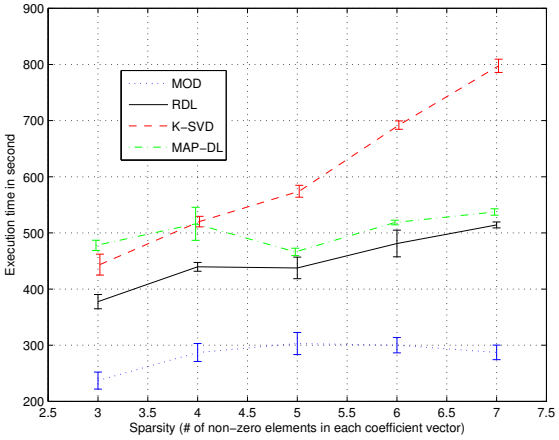

Figure 2: Computation cost of the fixed column-norm dictionary learnings algorithms.

We started from a normalized random $\mathbf{D}$ and used 1000 iterations. The learning parameter (γ) in MAP-DL was selected as described in [1]. We down-scaled γ by a factor of 2^{-j} ($j > 1$) when the algorithm was diverging. To have a fair comparison, we did the simulations for 5 different trials. If the squared error between a learned and true dictionary element was below 0.01, it was classified as correctly identified. The average percentages and standard deviations are shown in Figure 1. It can be seen that in all cases, RDL and K-SVD recovered nearly the same number of atoms and more than the other methods (although for the signals with less than 6 non-zero coefficients, RDL recovered all desired atoms, performance of K-SVD was very close to it). The MAP-DL algorithm did not perform well in this simulation. We guess the reason for this is slow convergence of the approach and the use of more iterations might improve the performance.

In Fig.2 we compare the computation time of the algorithms for the above simulations. Simulations ran on the Intel Xeon 2.66 GHz dual-core processor machines and both cores were used by Matlab. In this graph the total execution time of the algorithms (sparse approximations plus dictionary updates for 1000 iterations) is shown. MOD was fastest followed by our RDL.

We have a larger admissible set when fixing the Frobenius-norm

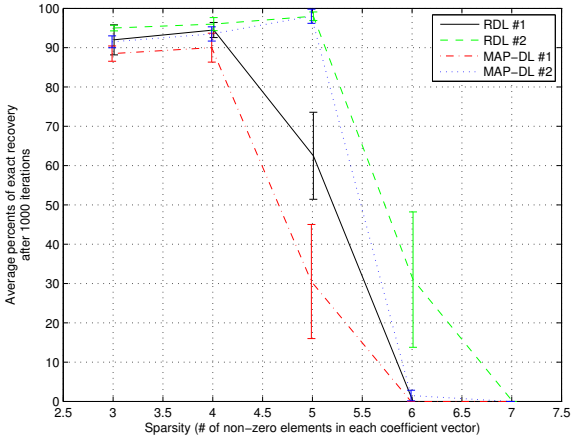

Figure 3: Exact recovery with fixed Frobenius-norm dictionary learning. 1: Desired dictionary had fixed Frobenius-norm. 2: Desired dictionary had fixed column-norms.

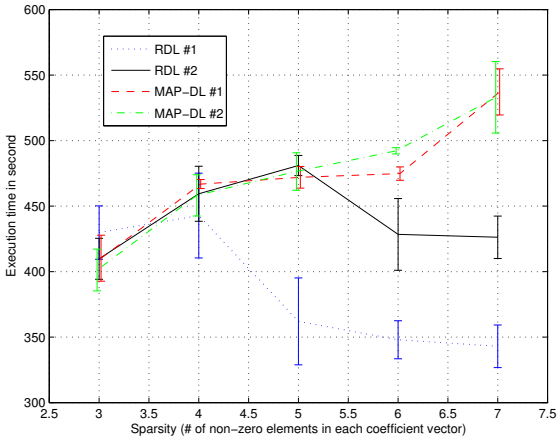

Figure 4: Computation cost of the fixed Frobenius-norm dictionary learning algorithms.

of the dictionary, which makes the problem of exact recovery more complicated and we expect to have less exact recovery for the same sparse signals. For this part we started with normalized random dictionaries, normalized to have either fixed Frobenius-norm or fixed column-norm.

The simulations were repeated for 5 trials and the averages and standard deviations of the atom recovery are shown in the Fig. 3. In these simulations RDL performed slightly better than MAP-DL. The other observation in this figure is that when the desired dictionaries have fixed column-norms, performance of the algorithms increase but do not reach the performance observed when using the more restricted (and appropriate) admissible set. Computation times of the algorithms, on the machines described formerly, are shown in the Fig.4. An interesting observation is the decrease in the computation time of RDL for less sparse signals, when the algorithm could barely recover the correct atoms.

Instead of constraining the dictionaries to have fixed norms, we can use the bounded-norm constraints. To show the possible advantage of these constraints, we repeated the simulations above. The results achieved with these constraints are shown in Fig. 5 We here did the simulations with and without orthogonal projections on the selected spaces found by sparse approximation method. It can be

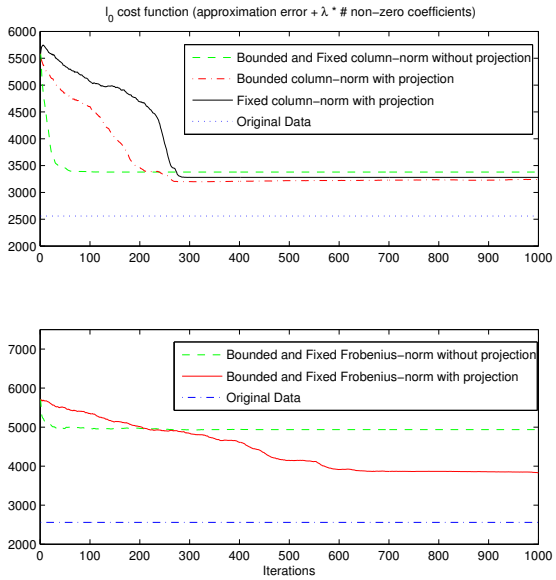

Figure 5: l_0 cost functions of the constrained Frobenius and column -norms dictionary learning algorithms respectively on top and bottom plots.

seen that using bounded-norm admissible set improves performance slightly when constraining the column-norm but it does not change performance of the other method. These plots also show that the orthogonal projection onto the selected spaces can improve overall performances.

4.2 Dictionary Learning for Sparse Audio Representations

In this part we demonstrate the performance of the proposed dictionary learning methods on real data. An audio sample of more than 8 hours was recorded from BBC radio 3, which plays mostly classical music. The audio sample was summed to mono and down-sampled by a factor of 4. From this 12kHz audio signal, we randomly took 4096 blocks of 256 samples each.

In the first experiment we used fixed column-norm and fixed Frobenius-norm dictionary admissible sets. The set of dictionaries with the column-norms equal to c_C is a subset of a larger set of fixed Frobenius-norm dictionaries, when $c_F = Nc_C$. We chose unit column-norm and fixed Frobenius-norm ($c_F = N$) dictionary learning algorithms. We initialized the dictionary with a 2 times overcomplete random dictionary and used 1000 iterations of alternative sparse approximation (using ℓ_1) and dictionary updates. The cost function against iteration, for two different values of λ, are shown in the Fig. 6. This figure shows that the optimal fixed Frobenius-norm dictionaries are better solutions for the objective functions.

As a second experiment we looked at an audio coding example. We used the RDL method with the fixed Frobenius-norm constraint to learn a dictionary based on a training set of 8192 blocks, each 256 samples long. The audio could be modeled using sinusoid, harmonic and transient components. We chose a 2 times overcomplete sinusoid dictionary (frequency oversampled DCT) as the initialization point and ran the simulations with different lambda values for 250 iterations. The number of appearances of each atom ($\lambda = .006$) are sorted based on their ℓ_2 norms and are shown in the Fig. 7. To design an efficient encoder we only used atoms that were used frequently in the representations and therefore shrunk the dictionary. In this test we chose a threshold of 40 (out of 8192) as the selection criteria. This dictionary was used to find the sparse approximations of 4096 different random blocks, each of 256 samples, from the recorded audio. We then coded the location (significant bit map)

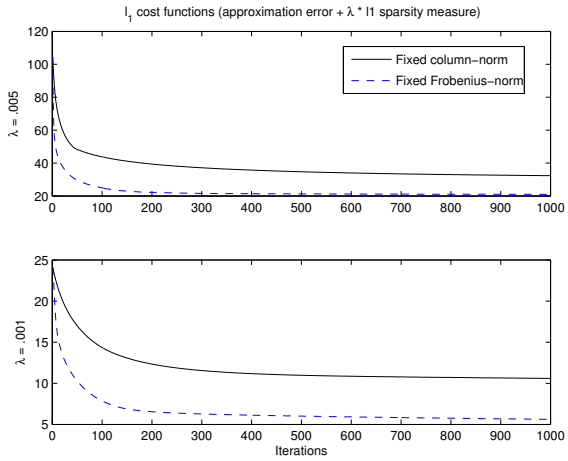

Figure 6: ℓ_1 cost functions for two different Lagrangian multipliers (λ) .005 (top) and .001 (bottom).

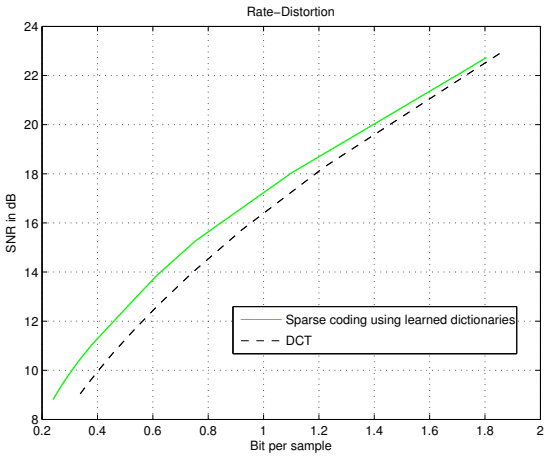

Figure 8: Estimated Rate-Distortion of the audio samples with sparse approximation using learned dictionary and DCT.

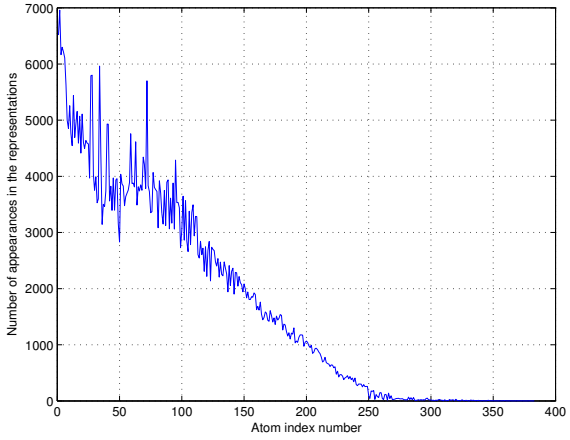

Figure 7: Number of appearances in the representations of the training samples (of size 8192).

and magnitude of the non-zero coefficients separately. In this paper we used a uniform scalar quantizer with a double zero bin. We calculated the entropy of the coefficients to approximate the required coding cost. To encode the significant bit map, we assumed an i.i.d. distribution for the location of the non-zero atoms. The same coding strategy was used to code the DCT coefficients of the same data. The performance is compared in Fig. 8. The convex hull of the rate-distortion performance calculated with different learned dictionaries, each optimized for a different bit-rates, is shown in this figure. Using the learned dictionaries is superior to using the DCT for the range of bit-rates shown, but the advantage is more noticeable for lower rates.

5. CONCLUSIONS

We have formulated the dictionary learning problem as a constrained minimization of a joint cost function. This allowed the derivation of a stable algorithm for dictionary learning, which was shown to perform well on several test data sets. The derived methods differ from most of the previously proposed approaches, such as K-SVD and MAP-DL with unit column-norm *a priori* information, which are based on atom-wise dictionary updates. The proposed methods update the whole dictionary at once. The computation cost of the algorithms were compared and it was found that the proposed methods performed better than, or similar to other competitors. Another simulation showed that using a bounded norm constraint was slightly better or at least as good as a fixed norm constraint. However, more simulations are needed. An alternative to the proposed method, when the constraint set is convex, is iterative gradient projection. This method is similar to the method that was used in Section 3.2 but with a different, and sometimes adaptive, c_s. The overall performance comparison of these methods is the next step of this project.

REFERENCES

[1] K. Kreutz-Delgado, J.F. Murray, B.D. Rao, K. Engan, T.W. Lee, and T.J. Sejnowski, "Dictionary learning algorithms for sparse representation," *Neural Comp.*, vol. 15, pp. 349–396, 2003.

[2] S.F. Cotter, B.D. Rao, K. Engan, and K. Kreutz Delgado, "Sparse solutions to linear inverse problems with multiple measurement vectors," *IEEE Trans. on Signal Processing*, vol. 53, no. 7, pp. 2477–2488, 2005.

[3] K. Engan, S.O. Aase, and J. Hakon-Husoy, "Method of optimal directions for frame design," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1999.

[4] B.A. Olshausen and D.J. Field, "Sparse coding with an overcomplete basis set: a strategy employed by V1?," *Vision Research*, vol. 37, no. 23, pp. 3311–3325, 1997.

[5] M. Aharon, E. Elad, and A.M. Bruckstein, "K-SVD: an algorithm for designining of overcomplete dictionaries for sparse representation," *IEEE Trans. on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, 2006.

[6] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Comm. Pure Appl. Math*, vol. 57, pp. 1413–1541, 2004.

[7] O. Divorra Escoda, L. Granai, and P. Vandergheynst, "On the use of a priori information for sparse signal approximations," *IEEE Trans. on Signal Processing*, vol. 54, no. 9, pp. 3468–3482, 2006.

[8] R.T. Rockafellar, *Convex Analysis*, Princeton University Press, 1970.