COMPRESSIBLE DICTIONARY LEARNING FOR FAST SPARSE APPROXIMATIONS

Mehrdad Yaghoobi, Mike E. Davies

Institute for Digital Communications, Joint Research Institute for
Signal and Image Processing, The University of Edinburgh, EH9 3JL, UK.
yaghoobi@ieee.org, mike.davies@ed.ac.uk

ABSTRACT

By solving a linear inverse problem under a sparsity constraint, one can successfully recover the coefficients, if there exists such a sparse approximation for the proposed class of signals. In this framework the dictionary can be adapted to a given set of signals using dictionary learning methods. The learned dictionary often does not have useful structures for a fast implementation, i.e. fast matrix-vector multiplication. This prevents such a dictionary being used for the real applications or large scale problems. The structure can be induced on the dictionary throughout the learning progress. Examples of such structures are shift-invariance and being multi-scale. These dictionaries can be efficiently implemented using a filter bank. In this paper a well-known structure, called compressibility, is adapted to be used in the dictionary learning problem. As a result, the complexity of the implementation of a compressible dictionary can be reduced by wisely choosing a generative model. By some simulations, it has been shown that the learned dictionary provides sparser approximations, while it does not increase the computational complexity of the algorithms, with respect to the pre-designed fast structured dictionaries.

Index Terms— Sparse Approximation, Dictionary Learning, Compressed Sensing, Compressible Signal, Majorization Minimization.

1. INTRODUCTION

Sparse approximation methods have been successfully applied to various signal processing problems. In this framework we have a linear generative model which can be represented using a full-rank matrix $\mathbf{D} \in \mathbb{R}^{d\times N}$, called a dictionary, in the space of discrete signals. Let $\mathbf{y} \in \mathbb{R}^d$ and $\mathbf{x} \in \mathbb{R}^N$ respectively be the signal and the coefficient vectors. When $d \leq N$, the generative model is underdetermined and does not have a unique solution. By inducing the sparsity over $\mathbf{x}$ the sparse approximation problem, in a relaxed form, can be formulated as,

$$\arg\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{D}\mathbf{x}\|^2 + \lambda\mathcal{J}(\mathbf{x}), \tag{1}$$

where $\mathcal{J}(.)$ is the sparsity measure [1], and when it is selected to be the ℓ_1-norm, the objective becomes convex. The convexity of the objective not only helps us to find the global solution of (1), but also guarantees the uniqueness of the solution and, under some conditions, to find the ℓ_0 sparse approximation, where ℓ_0 is the number of non-zero components.

The success of sparse approximation of a given class of signals, is directly determined by choosing a right dictionary, which is often unavailable for the real signals. Various methods have therefore been introduced to select a suitable dictionary. There are two important methods to select a dictionary, which are called dictionary design and dictionary learning, see for example [2–4] and references therein. In this paper we only investigate the dictionary learning problem. A set of training signals $\mathscr{Y} = \{\mathbf{y}_i\}_{i\in\mathcal{I}}$ is given which makes the matrix of training signals $\mathbf{Y} \in \mathbb{R}^{d\times L}$, by putting $\mathbf{y}_i$ as the i^{th} column. The learned dictionary is often found by minimizing an objective based on both $\mathbf{D}$ and $\mathbf{X} \in \mathbb{R}^{N\times L}$ [2, 5, 6], where the latter is the coefficient matrix. In this framework one can find the dictionary by solving the following optimization problem,

$$\arg\min_{\mathbf{D}\in\mathcal{D}} \min_{\mathbf{X}} \|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_F^2 + \lambda\mathcal{J}(\mathbf{X}), \tag{2}$$

where $\mathcal{J}(.)$ is the sparsity measure, which is often column separable, and $\mathcal{D}$ is an admissible set in $\mathbb{R}^{d\times N}$. Different admissible sets have been used to resolve the scale-ambiguity[1] of the optimization problem, e.g. constrained column or Frobenius norms [7].

A new framework is introduced here for the dictionary learning, which its formulation is slightly different to (2), to find a compressible dictionary. The definition of the compressible dictionary is introduced in the next section, followed by some remarks on the features of the compressible dictionaries. In Section 3 a formulation is presented for the Compressible Dictionary Learning (CDL) problem, which is non-convex and difficult to solve exactly. A practical algorithm is then presented in Section 4 to solve the CDL problem *approximately*. By some simulations it has been demonstrated that although the CDL problem is non-convex, the proposed algorithm finds an acceptable sparse dictionary.

2. COMPRESSIBLE DICTIONARY

To impose the compressibility constraint to the dictionary learning problem, we need to introduce the concept of signal compressability [8]. A signal ψ is compressible when the entries obey a power law,

$$|\psi|_{(k)} \leq c_r k^{-r}, \tag{3}$$

where $|\psi|_{(k)}$ is the k^{th} largest value of ψ, $r \geq 1$ and c_r is a constant. In a similar way, we call a matrix $\mathbf{\Psi}$ to be compressible if its entries obey a power law. An important feature of the compressible

This work is supported by EPSRC grant number D000246/1 and EU FP7, FET-Open grant number 225913. MED acknowledges support of his position from the Scottish Funding Council and their support of the Joint Research Institute with the Heriot-Watt University as a component part of the Edinburgh Research Partnership.

[1] $\forall\alpha \in \mathbb{R}^+$, $(\mathbf{D}, \mathbf{X})$ and $(\alpha\mathbf{D}, 1/\alpha\mathbf{X})$ have the same approximation errors and the number of non-zero components in the coefficient matrices.

signals, also has been used in the compressed sensing [9], is that a K-sparse signal approximates a compressible signal with a good approximation. Let $\mathbf{\Psi}_K$ be the matrix of the K largest elements of $\mathbf{\Psi}$, and let the other elements be zero. $\mathbf{\Psi}_K$ is the best estimate for $\mathbf{\Psi}$, in ℓ^2 space, and the approximation error is upper-bounded by the following formula,

$$\|\mathbf{\Psi} - \mathbf{\Psi}_K\|_F \leq c'_r K^{-r+1/2}. \tag{4}$$

This property has been used in the sensing of a compressible signal by recovering the best K-sparse signal which is a good approximation for the original compressible signal [8].

***Definition* 2.1.** A dictionary $\mathbf{D} \in \mathbb{R}^{d\times N}$ is called *compressible* when for a given full-rank matrix $\mathbf{\Phi} \in \mathbb{R}^{d\times M}$, called the mother dictionary, $\mathbf{D}$ can be generated using the following linear model,

$$\mathbf{D} = \mathbf{\Phi}\mathbf{\Psi}, \tag{5}$$

where $\mathbf{\Psi} \in \mathbb{R}^{M\times N}$ is a compressible matrix and $M \geq d$.

The compressible dictionaries have two important features which are presented by following remarks,

***Remark* 2.1 (Complexity of approximations).** (4) indicates that the approximation error introduced by using $\Psi_{\mathbf{K}}$ is upper bounded. To approximate a compressible dictionary, given $\mathbf{\Phi}$, one can find the best K-sparse $\mathbf{\Psi}_K$. The approximation complexity of $\mathbf{D}$, in general, reduces from $d.N$ to K as a result.

***Proposition* 2.1.** Let $\mathbf{D}$ be a compressible dictionary with the generative model (5) and $|\psi|_{(k)} \leq c_r k^{-r}$. The approximation error of the generated K-sparse dictionary $\mathbf{D}_K = \mathbf{\Phi}\mathbf{\Psi}_K$ decays rapidly by increasing K. The upper-bound of approximation error is as follows,

$$\|\mathbf{D} - \mathbf{D}_K\|_F \leq c'_r \|\mathbf{\Phi}\| K^{-r+1/2},$$

where $\|.\|$ is the operator norm and c'_r is a constant defined in (4).

One can prove this proposition by using (4) and the definition of operator norm. Note that the operator norm of $\mathbf{D}_\Delta := \mathbf{D} - \mathbf{D}_K$ is upper-bounded by $\|\mathbf{D} - \mathbf{D}_K\|_F$. Therefore the error caused by the operator $\mathbf{D}_\Delta$ also tends to zero, when $K \to d.N$ at least with the decay rate presented in Proposition 2.1.

***Remark* 2.2 (Fast multiplications).** Any vector multiplication with $\mathbf{D}$ can be done in two steps, a multiplication with the sparse matrix $\mathbf{\Psi}_K$ followed by a multiplication with $\mathbf{\Phi}$. Multiplication with the sparse matrix $\mathbf{\Psi}_K$ is $\mathcal{O}(K)$. When $\mathbf{\Phi}$ has structures which provide fast matrix-vector multiplication, e.g. Fourier and wavelets, the matrix multiplication can be done in $\mathcal{O}(N \log N)$ or better. In the practical applications we are interested in the cases $K \leq N \log N$. Therefore the overall complexity of multiplication with $\mathbf{D}$ is reduced to $N \log N$. It is a significant improvement over the traditional non-structured dictionary multiplication, for example found by dictionary learning, where complexity is $d\ N$.

3. PROBLEM FORMULATION

Let the matrix of training samples $\mathbf{Y} \in \mathbb{R}^{d\times L}$ and the mother dictionary $\mathbf{\Phi} \in \mathbb{R}^{d\times N}$ be given. In the CDL problem, the sparse approximation $\mathbf{X}$ and the dictionary generator matrix $\mathbf{\Psi}$ are unknown. Like the standard dictionary learning problem (2), we can define an appropriate objective function based on $(\mathbf{X}, \mathbf{\Psi})$ and find the dictionary by minimizing the objective. Here we need to add a term to the objective in (2) to promote sparsity of $\mathbf{\Psi}$. Therefore the CDL can be formulated by the following non-convex optimization problem,

$$\begin{aligned} &\arg\min_{\mathbf{\Psi}}\{\min_{\mathbf{X}} \nu(\mathbf{\Psi}, \mathbf{X})\} \ : \\ &\nu(\mathbf{\Psi}, \mathbf{X}) = \|\mathbf{\Phi}\mathbf{\Psi}\mathbf{X} - \mathbf{Y}\|_F^2 + \lambda \mathcal{J}_p(\mathbf{X}) + \gamma \mathcal{J}_q(\mathbf{\Psi}), \end{aligned} \tag{6}$$

where $\mathcal{J}_\rho(\mathbf{\Theta}) = \sum_{i,j} |\theta_{i,j}|^\rho$, for $\rho \in \{p, q\} \leq 1$ and a matrix $\mathbf{\Theta} = \{\theta_{i,j}\}$, is the sparsity measure and $\lambda, \gamma \in \mathbb{R}^+$. Let $p = q = 1$ for simplicity. The sparsity measure $\mathcal{J}_\rho(.)$ is now ℓ_1-norm, which turns (6) into a *bi-convex* optimization problem. The parameters λ and γ control the sparsity of the coefficient and the dictionary generator matrices respectively.

The following lemma shows that (6) is a *well-defined* optimization problem.

***Lemma* 3.1.** *The solution set of the problem (6) is bounded.*

Proof. $\nu(\mathbf{\Psi}, \mathbf{X})$ is a continuous function. Let epigraph of $\nu(\mathbf{\Psi}, \mathbf{X})$ at $(\mathbf{\Psi}^\star, \mathbf{X}^\star)$ be $\mathrm{epi}(\nu, (\mathbf{\Psi}^\star, \mathbf{X}^\star))$. $\mathrm{epi}(\nu, (\mathbf{0}, \mathbf{0}))$ for a continuous function $\nu(\mathbf{\Psi}, \mathbf{X})$ is compact [10]. The solution set is a subset of $\mathrm{epi}(\nu, (\mathbf{0}, \mathbf{0}))$ and therefore bounded. □

The scale ambiguity in the standard dictionary learning is often resolved by constraining $\mathbf{D}$ to be in $\mathcal{D}$. Although the formulation (6) does not have scale ambiguity, it might have non-unique solutions.

***Remark* 3.1.** Let $(\mathbf{\Psi}^*, \mathbf{X}^*)$ be a non-zero solution of (6) and $\alpha := \gamma\mathcal{J}_1(\mathbf{\Psi}^*)/\lambda\mathcal{J}_1(\mathbf{X}^*)$. If $\alpha \neq 1$ then $(\frac{1}{\alpha}\mathbf{\Psi}^*, \alpha\mathbf{X}^*)$ is another solution of (6).

It is worth mentioning the similarity between CDL and the sparse dictionary learning framework [11]. Rubinstein et. al. induced a k-sparsity constraint over each atom and used ℓ_0 as the sparsity measure. In CDL the sparsity is induced over the dictionary, which provides more flexibility in finding sparser dictionary generator matrix. A greedy method has been used in [11] to *approximately* find sparse approximations and dictionary updates. Although no convergence issue has been reported, the mathematical analysis of the algorithm is very difficult. In contrast CDL is guaranteed not only to be stable but also to converge to a set of local minima.

4. CDL ALGORITHM

The problem proposed in Section 3 is non-convex and non-differentiable. The difficulty of the problem can be reduced with the block-relaxation method which has been used for the standard dictionary learning [2]. In this framework, we minimize $\nu(\mathbf{\Psi}, \mathbf{X})$ with respect to $\mathbf{\Psi}$ or $\mathbf{X}$ each time, when the other parameter is kept fixed. In the other words, by starting from an initial solution $(\mathbf{\Psi}^{[0]}, \mathbf{X}^{[0]})$, the algorithm refines the solution by $\mathbf{\Psi}^{[n]} \to \mathbf{\Psi}^{[n+1]}$ or $\mathbf{X}^{[n]} \to \mathbf{X}^{[n+1]}$ to reduce $\nu(\mathbf{\Psi}, \mathbf{X})$. When we reduce such a positive objective at each step, the algorithm is stable due to the Lyapunov's second theorem. Due to the continuity of $\nu(\mathbf{\Psi}, \mathbf{X})$, the convergence of the algorithm, to a set of fixed points, can easily be driven using Proposition B.3 of [2].

In the setting introduced in Section 3, $\nu(\mathbf{\Psi}, \mathbf{X})$ is bi-convex and each step of the block-relaxed minimization can be done using a convex optimization method. The majorization minimization method [12] has been chosen to optimize $\nu(\mathbf{\Psi}, \mathbf{X})$ with respect to each parameter. This method is parallelizable and only needs matrix-matrix multiplications, and therefore it is applicable to large scale optimization problems like dictionary learning [2]. A majorizing

objective, which is easier to be optimize, is minimized at each step of this method. Recall a function g majorizes f when it satisfies the following conditions,

$$f(\omega) \leq g(\omega, \xi),\ \forall \omega, \xi \in \Upsilon$$
$$f(\omega) = g(\omega, \omega),\ \forall \omega \in \Upsilon,$$

where Υ is the admissible set. The majorizing function has an extra parameter ξ. At each iteration, we first choose this parameter as the current value of ω and find the optimal update for ω.

$$\omega_{new} = \arg\min_{\omega \in \Upsilon} g(\omega, \xi)$$

We then update ξ with ω_{new}. The algorithm continues until we find an accumulation point. In practice the algorithm is terminated when ω and ω_{new} are very close.

The majorizing functions for the $\nu(\mathbf{\Psi}, \mathbf{X})$, when $\mathbf{\Psi}$ or $\mathbf{X}$ is kept fixed, are derived in the next subsections. The majorizing objectives are convex with respect to the corresponding parameters. By letting zero be in the subgradient of the objectives, the update formulas are derived.

4.1. CDL with the majorization method

The objective $\nu(\mathbf{\Psi}, \mathbf{X})$ is an additive combination of the quadratic part $\|\mathbf{\Phi\Psi X} - \mathbf{Y}\|_F^2$, which has bounded curvatures when $\mathbf{\Psi}$ or $\mathbf{X}$ are fixed, and the sparsity measures. A majorizing function can be derived using Taylor series in the matrix form. This operation can simply be done by adding an appropriate strictly convex function to $\nu(\mathbf{\Psi}, \mathbf{X})$, see [2] for more details.

Two distinctive majorizing functions are derived for updating $\mathbf{X}$ and $\mathbf{\Psi}$, for fixed $\mathbf{\Psi}$ and $\mathbf{X}$ respectively. These are followed by deriving the update formulas for each case.

4.1.1. Deriving the update formula for $\mathbf{X}$*:*

Let $\nu_{\mathbf{\Psi}}(\mathbf{X}) : \mathbb{R}^{N \times L} \rightarrow \mathbb{R}^+$ be $\nu(\mathbf{\Psi}, \mathbf{X})$ at a fixed $\mathbf{\Psi}$. The majorizing function is found by adding $\nu_{\mathbf{\Psi}}(\mathbf{X})$ and $\pi_{\mathbf{\Psi}}(\mathbf{X}, \mathbf{X}^{[n]})$, which is found by,

$$\pi_{\mathbf{\Psi}}(\mathbf{X}, \mathbf{X}^{[n]}) = c_{\mathbf{\Phi}} c_{\mathbf{\Psi}} \|\mathbf{X} - \mathbf{X}^{[n]}\|_F^2 - \|\mathbf{\Phi\Psi X} - \mathbf{\Phi\Psi X}^{[n]}\|_F^2,$$

where $c_{\mathbf{\Phi}} > \|\mathbf{\Phi}^T\mathbf{\Phi}\|$ and $c_{\mathbf{\Psi}} > \|\mathbf{\Psi}^T\mathbf{\Psi}\|$. The majorizing objective $\mu_{\mathbf{\Psi}}(\mathbf{X}, \mathbf{X}^{[n]})$ is then found by,

$$\mu_{\mathbf{\Psi}}(\mathbf{X}, \mathbf{X}^{[n]}) = tr\{c_{\mathbf{\Phi}} c_{\mathbf{\Psi}} \mathbf{X}^T\mathbf{X} - 2\mathbf{X}^T(\mathbf{\Psi}^T\mathbf{\Phi}^T(\mathbf{Y} - \mathbf{\Phi\Psi X}^{[n]}) + c_{\mathbf{\Phi}} c_{\mathbf{\Psi}} \mathbf{X}^{[n]})\} + \lambda \mathcal{J}_1(\mathbf{X}) + const,$$

where $const$ presents the terms which are constant with respect to $\mathbf{X}$. $\mu_{\mathbf{\Psi}}$ is a non-differentiable convex function. The matrix $\mathbf{0}$ is then in the subgradient of $\mu_{\mathbf{\Psi}}$ at the minimum. We know that $\mathbf{X}^{[n+1]} = \arg\min_{\mathbf{X}} \mu(\mathbf{X}, \mathbf{X}^{[n]})$. Therefore $\mathbf{X}^{[n+1]}$ should satisfy,

$$0 \in \partial\mu_{\Psi}(\mathbf{X}^{[n+1]}, \mathbf{X}^{[n]}),$$
$$\partial\mu_{\Psi}(\mathbf{X}, \mathbf{X}^{[n]}) = 2c_{\mathbf{\Phi}} c_{\mathbf{\Psi}} \mathbf{X} - 2(\mathbf{\Psi}^T\mathbf{\Phi}^T(\mathbf{Y} - \mathbf{\Phi\Psi X}^{[n]}) + c_{\mathbf{\Phi}} c_{\mathbf{\Psi}} \mathbf{X}^{[n]}) + \lambda \partial\mathcal{J}_1(\mathbf{X}).$$

The update formula for $\mathbf{X}$ can be found by,

$$\mathbf{X}^{[n+1]} = \mathcal{S}_{\lambda/2}\left[\frac{1}{c_{\mathbf{\Phi}} c_{\mathbf{\Psi}}}(\mathbf{\Psi}^T\mathbf{\Phi}^T(\mathbf{Y} - \mathbf{\Phi\Psi X}^{[n]}) + c_{\mathbf{\Phi}} c_{\mathbf{\Psi}} \mathbf{X}^{[n]})\right],$$

where $\mathcal{S}_{\lambda/2}$ is the soft-shrinkage operator [13] and $\alpha = \lambda/2$,

$$\mathcal{S}_{\alpha}(\mathbf{A}) = \begin{cases} a_{i,j} - \alpha/2\ sign(a_{i,j}) & \alpha/2 < |a_{i,j}| \\ 0 & otherwise. \end{cases} \quad (7)$$

Algorithm 1 : $\mathcal{CDL}(\mathbf{X}_0, \mathbf{\Psi}_0)$

initialization: $c_{\Phi} > \|\mathbf{\Phi}^T\mathbf{\Phi}\|$, $K_X, K_{\Psi} \in \mathbb{N}$
for $t = 0$ **to** T **do**
  $c_{\mathbf{\Psi}} > \|\mathbf{\Psi}^T\mathbf{\Psi}\|$, $\mathbf{X}^{[0]} = \mathbf{X}_t$
  for $n = 0$ **to** $K_X - 1$ **do**
    $\mathbf{X}^{[n+1]} = \mathcal{S}_{\lambda/2}\ [\frac{1}{c_{\mathbf{\Phi}} c_{\mathbf{\Psi}}}(\mathbf{\Psi}^T\mathbf{\Phi}^T(\mathbf{Y} - \mathbf{\Phi\Psi X}^{[n]}) + c_{\mathbf{\Phi}} c_{\mathbf{\Psi}} \mathbf{X}^{[n]})]$
  end for
  $\mathbf{X}_{t+1} = \mathbf{X}^{[K_X]}$
  $c_{\mathbf{X}} > \|\mathbf{X}\mathbf{X}^T\|$, $\mathbf{\Psi}^{[0]} = \mathbf{\Psi}_t$
  for $n = 0$ **to** $K_{\Psi} - 1$ **do**
    $\mathbf{\Psi}^{[n+1]} = \mathcal{S}_{\gamma/2}\ [\frac{1}{c_{\mathbf{\Phi}} c_{\mathbf{X}}}(\mathbf{\Phi}^T(\mathbf{Y} - \mathbf{\Phi\Psi}^{[n]}\mathbf{X})\mathbf{X}^T + c_{\mathbf{\Phi}} c_{\mathbf{X}} \mathbf{\Psi}^{[n]})]$
  end for
  $\mathbf{\Psi}_{t+1} = \mathbf{\Psi}^{[K_{\Psi}]}$
end for
output: $\mathbf{\Psi}_T$

4.1.2. Deriving the update formula for $\mathbf{\Phi}$*:*

Let $\nu_{\mathbf{X}}(\mathbf{\Psi}) : \mathbb{R}^{M \times N} \rightarrow \mathbb{R}^+$ be $\nu(\mathbf{\Psi}, \mathbf{X})$ at a fixed $\mathbf{X}$. A technique, similar to what was used in 4.1.1, can be used to generate the majorizing function for $\nu_{\mathbf{X}}(\mathbf{\Psi})$. Here, $\pi_{\mathbf{X}}(\mathbf{\Psi}, \mathbf{\Psi}^T)$ is calculated by,

$$\pi_{\mathbf{X}}(\mathbf{\Psi}, \mathbf{\Psi}^{[n]}) = c_{\mathbf{\Phi}} c_{\mathbf{X}} \|\mathbf{\Psi} - \mathbf{\Psi}^{[n]}\|_F^2 - \|\mathbf{\Phi\Psi X} - \mathbf{\Phi\Psi}^{[n]}\mathbf{X}\|_F^2.$$

where $c_{\mathbf{X}} > \|\mathbf{X}\mathbf{X}^T\|$. The majorizing objective $\mu_{\mathbf{X}}(\mathbf{\Psi}, \mathbf{\Psi}^{[n]})$ is now found to be,

$$\mu_{\mathbf{X}}(\mathbf{\Psi}, \mathbf{\Psi}^{[n]}) = tr\{c_{\mathbf{\Phi}} c_{\mathbf{X}} \mathbf{\Psi}^T\mathbf{\Psi} - 2\mathbf{\Psi}^T(\mathbf{\Phi}^T(\mathbf{Y} - \mathbf{\Phi\Psi}^{[n]}\mathbf{X})\mathbf{X}^T + c_{\mathbf{\Phi}} c_{\mathbf{X}} \mathbf{\Psi}^{[n]})\} + \lambda \mathcal{J}_1(\Psi) + const,$$

The matrix $\mathbf{0}$ should be in the subgradient of $\mu_{\mathbf{X}}(\mathbf{\Psi}, \mathbf{\Psi}^{[n]})$ at the minimum, $\mathbf{\Psi}^{[n+1]}$. This provides the following update formula,

$$\mathbf{\Psi}^{[n+1]} = \mathcal{S}_{\gamma/2}\left[\frac{1}{c_{\mathbf{\Phi}} c_{\mathbf{X}}}(\mathbf{\Phi}^T(\mathbf{Y} - \mathbf{\Phi\Psi}^{[n]}\mathbf{X})\mathbf{X}^T + c_{\mathbf{\Phi}} c_{\mathbf{X}} \mathbf{\Psi}^{[n]})\right],$$

where $\mathcal{S}_{\gamma/2}$ is again the soft-shrinkage operator (7), with $\alpha = \gamma/2$. Algorithm 1 presents a pseudocode for the CDL method. In this pseudocode, the outer loop switches between the optimizing parameters. The inner loops are for updating each parameter, for a given number of iterations, before switching to the other parameter. It is also possible to choose different methods for switching between optimizing parameters.

5. SIMULATIONS

The CDL has been used to learn a dictionary for sparse audio coding in this section. Table 1 shows the parameters have been used in this simulation. The training matrix $\mathbf{Y}$ was generated by randomly selecting blocks of an audio signal recorded from BBC Radio 3, which often plays classical music. The parameters c_{Φ}, c_{Ψ} and c_X are chosen to be larger than, but close to, the corresponding operator norms to speed up convergence of the CDL.

The generation of a selected atom in the learned dictionary $\mathbf{D}$ is schematically demonstrated in Figure 1. ψ_i is plotted in part (a). This sparse vector, by multiplying to $\mathbf{\Phi}$, generates atom $\mathbf{d}_i$. Therefore the atoms of $\mathbf{\Phi}$, which are related to the non-zero coefficients

Table 1. The parameters of CDL for the sparse audio coding.

d	M = N	L	λ	γ	T	$\mathbf{\Psi}_0$	$\mathbf{X}_0$
256	512	8192	0.02	0.01	1000	$\mathcal{N}(0,1)$	$\mathbf{0}$

of ψ_i, contribute to generate $\mathbf{d}_i$. The plots (b), (c) and (d) demonstrate, respectively, the contributing atoms of $\mathbf{\Phi}$, scaled version of these atoms and $\mathbf{d}_i$.

Now that we have found the learned dictionary, we can show its advantages in the sparse approximation of the audio signals. We chose 4096 different random blocks of samples from the same audio sample. The iterative thresholding method has been used for sparse matrix approximation, using $\lambda = 0.02$. An extra step of CDL is re-normalizing the learned dictionary to the initial Frobenius-norm, to make further comparisons fair. Figure 2 shows the phase-plot of the algorithm. In this phase-plot the horizontal and vertical axes are $\mathcal{J}_1(\mathbf{X})$ and approximation error, respectively. The result shows that for an approximation error, the approximation by using learned dictionary is sparser (has less ℓ_1).

6. CONCLUSION

A novel dictionary model was introduced. Dictionaries that obey this model, called compressible dictionaries, appear to be more suitable for implementation. An optimization problem is then formulated to find a compressible dictionary. This optimization objective is non-convex and non-differentiable. A practical algorithm was introduced to find an approximate solution (local minimum). A compressible dictionary was learned for the audio signals. It was shown that the sparser approximations of the evaluation samples are in average yielded, using the learned dictionary. Further investigations on the recoverability, the convergence proof, the parameter selection and the complexity of implementations have been left for future work.

7. REFERENCES

[1] K. Kreutz-Delgado, B. D. Rao, K. Engan, T.W. Lee, and T. J. Sejnowski, "Convex/schur-convex (csc) log-priors and sparse coding," in *Joint Symposium on Neural Computation*, 1999.

[2] M. Yaghoobi, T. Blumensath, and M. Davies, "Dictionary learning for sparse approximations with the majorization method," to appear in IEEE Trans. on Signal Processing, 2009.

[3] 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.

[4] M. Yaghoobi, L. Daudet, and M. Davies, "Parametric dictionary design for sparse coding," submitted, 2009.

[5] M.S. Lewicki and T.J. Sejnowski, "Learning overcomplete representations," *Neural Comp*, vol. 12, no. 2, pp. 337–365, 2000.

[6] 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.

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

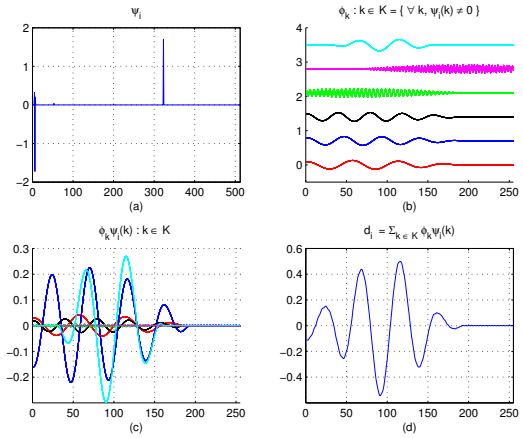

Fig. 1. The atom generation in the CDL framework: (a) i^{th} column of $\mathbf{\Psi}$, ψ_i, (b) The atoms ϕ_k's which are related to the non-zero values of selected ψ_i, $\{\phi_k : \psi_i(k) \neq 0\}$, (c) $\phi_k \psi_i(k) : \psi_i(k) \neq 0$, (d) The i^{th} atom of $\mathbf{D} = \mathbf{\Phi}\mathbf{\Psi}$.

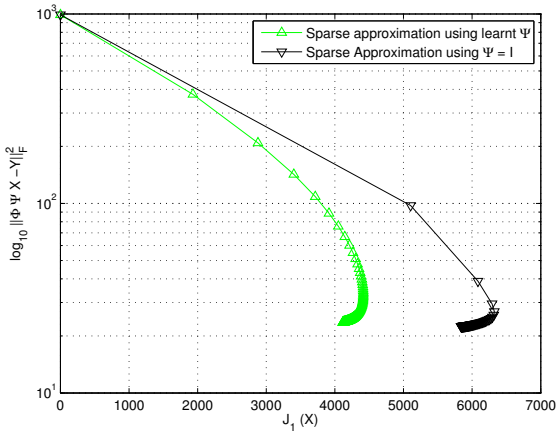

Fig. 2. The phase plots (representation error v.s. ℓ_1) of 4096 evaluation signals.

[8] E. Candes, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Mathematics*, vol. 59, no. 8, pp. 1207–1223, 2006.

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

[10] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[11] R. Rubinstein, M. Zibulevsky, and M. Elad, "Learning sparse dictionaries for sparse signal representation," submitted, 2008.

[12] K Lange, *Optimization*, Springer-Verlag, 2004.

[13] D.L. Donoho and J.M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, no. 3, pp. 425–455, 1994.