COMPRESSIVE DICTIONARY LEARNING FOR IMAGE RECOVERY

Mohammad Aghagolzadeh, Student Member IEEE, Hayder Radha, IEEE Fellow

Department of Electrical and Computer Engineering
Michigan State University, East Lansing, MI, USA
{aghagol1, radha}@msu.edu

ABSTRACT

In this paper, we tackle real-time learning of a dictionary D from compressive measurements Y of an image X. Existing dictionary learning algorithms are inapplicable because compressive samples $Y = \Phi X$ are incomplete and can be arbitrary linear combinations of different pixels. Our strategy is to learn a dictionary of the form $D = \Psi\Theta$, which represents *compressible dictionaries* with respect to the base dictionary Ψ. We show that our method for learning dictionaries during compressive image recovery can improve the recovery results by up to 3 dBs for general random sampling matrices.

Index Terms— Compressed sensing; dictionary learning.

1. INTRODUCTION

Real-time learning and utilization of image dictionaries for denoising via sparse representation has shown promising results [1, 3]. These efforts employ the captured noisy image during the learning process, and consequently generate a dictionary that matches the underlying image attributes. Another area that could benefit from real-time learning of sparsifying dictionaries is Compressive Sampling (CS) [5] where images are recovered from low-dimensional random linear measurements from their sparse representations.

In this paper, we raise the following question: is it possible to learn dictionaries from compressive measurements of images? More specifically, suppose that m compressive measurements are acquired per each n-pixel patch, where $m < n$. Is it possible to learn a dictionary using this collection of compressive samples? Toward finding a solution to this problem, we propose an algorithm which we refer to by Compressive Dictionary Learning (CDL). There are two key challenges in applying dictionary learning over CS samples: (*a*) there are less samples than the size of the image ($m < n$) and (*b*) CS measurements can be linear combinations of different pixels. The combination of dimension deficiency and linearity of samples makes existing dictionary learning algorithms inapplicable. We should note that Mairal et al. [3] have addressed image inpainting which has property (*a*) (but not (*b*)) by mapping it into image denoising.

This work is supported in part by NSF Award CCF-1117709.

In our approach, the dictionary is constrained to have a sparse representation with respect to the base dictionary Ψ. Hence, $D = \Psi\Theta$, where Θ is a sparse matrix to be learned. The base dictionary is the starting point for the learning algorithm and can be an orthonormal as well as overcomplete dictionary. For example, Ψ could consist of overcomplete discrete cosines as in [1, 2]. The notion of sparse dictionary representation was introduced for the first time in [2] for learning sparse structured dictionaries from noisy data, as opposed to learning arbitrary dictionaries which are costly to deploy and vulnerable to overfitting.

Although we are not aware of other works that explicitly address the learning of dictionaries from $m(< n)$ compressive samples, there are some efforts with close directions. For example, the Best Basis CS [4] selects the dictionary from a predefined set of orthogonal transforms that are arranged in a tree structure. Clearly, this method offers only a finite number of choices for the dictionary. Our paper is organized as follows. In Section 2, we overview CDL. We present the CDL algorithm in Section 3. The simulation results are presented in Section 4 where we show the utility of CDL for compressive imaging. Finally, in Section 5, we conclude this paper.

1.1. Notation

Table 1 outlines the notations that are used in this paper. Note that $\sqrt{n} \times \sqrt{n}$ image patches are reshaped into n-dimensional column vectors. We denote the (i, j)'th element of matrix X by $x_{i,j}$. Also, we refer to the i'th column of X by x_i and to the j'th row by $\overline{x}_j$.

We define the matrix norm, similar to the vector norm, as $\|G\|_p = (\sum_{i,j} |G_{i,j}|^p)^{1/p}$. The ℓ_0-(pseudo)norm $\|.\|_0$ counts the number of non-zeros in a vector as well as a matrix.

Table 1. List of variable notations

$X_{n\times\eta}$	Set of η n-dimensional image patches
$Y_{m\times\eta}$	Set of η m-dimensional measurement vectors
$\Phi_{m\times n}$	The sampling matrix
$D_{n\times k}$	The dictionary of k atoms
$A_{k\times\eta}$	Set of η k-dimensional coefficient vectors

2. THE CDL FRAMEWORK

2.1. Overview

In patch-based Compressive Sampling (CS) of an image consisting of η non-overlapping $\sqrt{n}\times\sqrt{n}$ patches, m linear measurements are acquired (via a CS device) from each patch [1]:

$$\forall i \in \{1, 2, \ldots, \eta\} : y_i = \Phi x_i \tag{1}$$

If the sampling matrix Φ is $m \times n$ with $m < n$, the CS solution for x_i in (1) is one with the sparsest representation. By linear representation of x_i we mean expressing it as the weighted summation of $k(\geq n)$ predefined functions, called atoms. The sparsest representation of x_i corresponds to the vector of weight coefficients $a_i = [a_{1i}, \ldots, a_{ki}]^T$ with the smallest number of non-zeros. The k n-dimensional atoms constitute an $n \times k$ dictionary D. The typical expression for the sparsest representation of x_i is:

$$a_i = \arg\min_{a_i} \|a_i\|_0 \text{ subject to } x_i = Da_i = \sum_{j=1}^{k} d_j a_{ji}$$

The CS approach is to use convex optimization tools, such as Basis Pursuit [5], to find the coefficient vector a_i with the smallest ℓ_1-norm - the closest convex approximation to the non-convex ℓ_0-(pseudo)norm of a_i [5]:

$$\hat{x}_i = D.\arg\min_{\hat{a}_i} \|\hat{a}_i\|_1 \text{ subject to } \Phi D\hat{a}_i = y_i \tag{2}$$

Although exact recovery is guaranteed for inherently sparse signals [5], natural images tend to have wide ranges of details that are specific to each image. As a result, learning a global dictionary from a corpus of example images could fail to sparsely represent every image. This fact motivated us to investigate the possibility of adapting the dictionary to an image, given a set of compressive patch measurements.

2.2. Conventional dictionary learning methods

The general form of dictionary learning problem can be expressed as:

$$\min_{A,D}\{\frac{1}{2}\|X - DA\|_2^2 + \lambda_A\|A\|_1\} \tag{3}$$

Leading examples of dictionary learning algorithms, e.g. [1–3], consist of recursively updating the sparse coefficients A and the dictionary D. First, given a batch of signals X, a coarse sparse approximation with respect to the dictionary $D^{(t-1)}$ from step $t-1$ is computed by solving the following LASSO [7] problem with regularization parameter λ_A:

$$\min_{A^{(t)}}\{\frac{1}{2}\|X - D^{(t-1)}A^{(t)}\|_2^2 + \lambda_A\|A^{(t)}\|_1\} \tag{4}$$

[1] In this paper, we assume the same sampling matrix is applied to all image patches. Such sampling allows grouping of the compressive samples into a single matrix notation as $Y = \Phi X$

The dictionary is subsequently updated to minimize the representation error while $A^{(t)}$ is fixed:

$$D^{(t)} = \arg\min_{D^{(t)}}\{\frac{1}{2}\|X - D^{(t)}A^{(t)}\|_2^2\} \tag{5}$$

In this work, we are interested in the case, where only a projection of X onto a rectangular sampling matrix Φ is available. In conjunction with the above iterative optimization (regularization and dictionary update), $A^{(t)}$ is computed using compressive samples as we explain later. However, without further manipulation, the dictionary update step (5) becomes the following underdetermined problem - underdetermined because $\Phi^T\Phi$ is singular:

$$\min_{D^{(t)}}\{\frac{1}{2}\|Y - \Phi D^{(t)}A^{(t)}\|_2^2\} \text{ where } Y = \Phi X \tag{6}$$

which does not have a unique solution for $D^{(t)}$ for a compressive sampling matrix having less rows than columns. In the next subsection, we explain the additional structural constraints that help us solve (6).

2.3. Sparse structured dictionaries

We are interested in structured dictionaries that can be compactly represented as:

$$D = \Psi\Theta \tag{7}$$

where $\Psi_{n\times n}$ is the basis and $\Theta_{n\times k}$ is an array of coefficients that describe the dictionary structure. Θ can take various forms. Clearly, if Θ was the identity matrix then D would be an ordinary basis. If we choose to add $\delta = k - n$ columns to the basis, Θ would have the following form:

$$\Theta = [I_{n\times n}|\Sigma_{n\times\delta}] \tag{8}$$

Θ is constrained to be a sparse matrix so that each atom of D would be a linear combination of only a few columns of Ψ. This constraint keeps the learned dictionary from taking arbitrary forms and make it less susceptible to overfitting. For further discussion about sparse dictionary representation we refer the reader to [2, 8]. From now on, in this paper, we use 'sparse dictionaries' to refer to dictionaries of the form $D = \Psi\Theta$ with sparse Θ [2].

2.4. Compressive dictionary sensing

The notion of sparsity in the dictionary was used in [2]. Yet, a property of dictionary sparsity has arguably been unnoticed: the sufficiency of low-dimensional samples for learning the

[2] It is important to note that sparse projection matrices for CS have been proposed by other works to reduce the complexity of CS recovery. However, similar to ordinary CS projections, these sparse projections are fixed. In other words, prior sparse CS projections have no connection with the sparse dictionaries that we propose to learn during the recovery stage of compressed sensing.

sparse dictionary. One might conclude by induction that, similar to compressive sampling of sparse signals, sparse dictionaries may be learned from low-dimensional measurements. By incorporating the sparse dictionary constraint $D = \Psi\Theta$ with a sparse Θ, the compressive dictionary learning problem based on CDL framework can be expressed as:

$$\min_{A,\Theta}\{\frac{1}{2}\|X - \Psi\Theta A\|_2^2 + \lambda_A\|A\|_1 + \lambda_\Theta\|\Theta\|_1\} \quad (9)$$

The above form suggests using LASSO for the dictionary update step. However, special care should be taken since the optimization is with respect to Θ and in the quadratic cost function $\|X - \Psi\Theta A\|_2^2$ the order of Θ and A should be reversed to project the LASSO form. Restructuring the above cost function to the LASSO form is explained in Section 3.

Similar to iterative dictionary learning approaches, there are two steps in each cycle t of the CDL algorithm. In the sparse coding step, sparse coefficients for compressive samples are computed using the dictionary from the previous cycle $D^{(t-1)}$:

$$\min_{A^{(t)}}\{\frac{1}{2}\|Y - \Phi D^{(t-1)}A^{(t)}\|_2^2 + \lambda_A\|A^{(t)}\|_1\} \quad (10)$$

In the dictionary update step, the dictionary is updated to minimize the projected representation error while adhering to the sparsity constraint. Therefore, the following dictionary update problem substitutes (6):

$$\min_{\Theta^{(t)}}\{\frac{1}{2}\|Y - \Phi\Psi\Theta^{(t)}A^{(t)}\|_2^2 + \lambda_\Theta\|\Theta^{(t)}\|_1\} \quad (11)$$

Note that the sparsity constraint (on the dictionary) prevents overfitting of the learned dictionary to the sampling matrix. The conversion of (11) into the LASSO will be explained in Section 3. Numeric values of the regularization parameters are provided in Section 4 for the simulations.

3. THE CDL ALGORITHM

3.1. The sparse coding step

At the beginning of each loop, sparse coefficients $A^{(t)}$ are computed by solving LASSO for all samples:

$$\min_{A^{(t)}}\{\frac{1}{2}\|Y - \Phi\Psi\Theta^{(t-1)}A^{(t)}\|_2^2 + \lambda_A\|A^{(t)}\|_1\} \quad (12)$$

The value of the regularization parameter, λ_A, depends on the noise level of compressive samples or generally on the complexity of the underlying image. When a large value is set for λ_A, only important and essential components of the patches are absorbed into $A^{(t)}$ which is appropriate for high noise cases. A small value for λ_A results in less sparse coefficients that contain too many details for each patch. Large variety of details can confuse the learning process and cause the algorithm to diverge. We will further discuss the values used in our work for λ_A in the simulation section. At the end of this step, we define the following crucial error parameter: the *projected* estimation error associated with atom d_j:

$$E_{d_j}^{(t)} := Y - \sum_{i=1, i\neq j}^{k} \Phi d_i^{(t-1)}\overline{a}_i(t) \quad (13)$$

3.2. The dictionary update step

We should note that, given compressive samples, only a projection of the representation error $X - D^{(t)}A^{(t)}$ onto Φ, i.e. $\Phi(X - D^{(t)}A^{(t)}) = Y - \Phi D^{(t)}A^{(t)}$, is obtainable. Also note that similar to the projection of the original data X onto Φ, the projection of the error, i.e. $\Phi(X - D^{(t)}A^{(t)})$, has, in general, infinite solutions. Therefore, extra information is required to update the dictionary so that the actual representation error, i.e. $X - D^{(t)}A^{(t)}$, is minimized. Instead, by adding the sparsity constraint on the dictionary, as in (11), it can be shown that the dictionary update step of CDL framework can be solved for each column σ_j of the structure matrix Σ by the following LASSO-like problem [7, 8]:

$$\sigma_j^{(t)} = \arg\min_{\sigma_j^{(t)}}\{\frac{1}{2}\|E_{\sigma_j}^{(t)}\overline{a}_j^{(t)^T} - \Phi\Psi\sigma_j^{(t)}\|_2^2 + \lambda_\Theta\|\sigma_j^{(t)}\|_1\} \quad (14)$$

λ_Θ controls the degree of sparsity of the dictionary atoms.

Due to space limitations, the full derivation for the above CDL optimization formula is provided in [8]. Furthermore, there are other key challenges that need to be addressed during the iterative steps of the CDL algorithm. These challenges are also covered in [8].

A summary of the CDL algorithm is presented in Algorithm 1. We have also developed a simple extension to the baseline CDL Algorithm 1 to handle noise [8]. (Details are omitted due to space limitations.) The algorithm may be terminated before T iterations, if $\Theta^{(t)}$ does not change much over iterations of CDL.

Algorithm 1 noiseless CDL

Require: Ψ, Y, Φ, λ_A, λ_Θ, T, ϵ_{stop}
  Initialization $t \leftarrow 1$, $\Sigma^{(0)} \leftarrow I_{n\times n}$
  while $t \leq T$ and $\|\Sigma^{(t)} - \Sigma^{(t-1)}\|_2 > \epsilon_{stop}$ **do**
    Update $A^{(t)}$ using (12).
    Update $\Sigma^{(t)}$ using (14).
    Scale $\Sigma^{(t)}$ so columns of $D^{(t)} = \Psi\Theta^{(t)}$ have unit norm.
    Discard inactive columns of $\Sigma^{(t)}$.
    $t \leftarrow t + 1$
  end while
  $D = \Psi\Theta = [\Psi\ \Psi\Sigma]$
  Compute $\hat{A}$: $\min_{\hat{A}}\|\hat{A}\|_1$ s.t. $Y = \Phi\hat{X}$, $\hat{X} = D\hat{A}$
  return D, $\hat{X}$

4. SIMULATION RESULTS

In this section, we simulate the application of CDL for the single-pixel camera [6] with a random ± 1 sampling matrix (for other cases of CS with noise, we refer the reader to [8]). A dictionary is trained for each image of 512×512 pixels where 8×8 non-overlapping patches are sampled using the same sampling operator Φ (random ± 1). We plot the PSNR results for different sampling ratios for $\lambda_A = \lambda_\Theta = 0.1$. The test images are shown in Fig. 1. Orthogonal 8×8 discrete cosines were selected for the base dictionary.

The top-right graph in Fig. 1 shows the average improvements for the test images. Specifically, images with periodic patterns such as the fingerprint image benefit more from CDL. The bottom-right graph shows the PSNR curves only for the fingerprint image. The most significant performance improvements due to CDL can be observed more clearly at intermediate values of m where there is sufficient samples to learn the dictionary and there is room for improving the CS results. Improvements as high as 3 dBs can be observed over different ranges of m. When m drops below $n/4$, the information bound does not allow to learn meaningful dictionaries and improvement is less apparent. Nevertheless, even over these low values of m improvements in the vicinity of 1 dBs to 1.7 dBs are attainable.

4.1. Complexity analysis of CDL

The CDL algorithm consists of a maximum of T iterations where in each iteration a LASSO (of size η) is solved to find $A^{(t)}$, a LASSO (of size δ) is solved for $\Sigma^{(t)}$ plus the estimation errors $E_{\sigma_j}^{(t)}$. In the simulations, we run CDL for $T = 32$ iterations. The SPAMS software package developed for [3] solves η LASSO problems in parallel with the complexity of $O(k^2 n + \eta(kn + ks^2))$ where s denotes sparsity. Assuming $\eta \gg n, k$, the total complexity of CDL can be approximated as follows:

$$O(T(\eta + \delta)(kn + ks^2))$$

This is roughly T times the complexity of normal CS using LASSO.

5. CONCLUSION

In this paper, we showed that given compressive samples of images, it is possible to adapt the sparsifying dictionary to the underlying image in real-time. The simulation results show the image recovery PSNR values can be enhanced up to 3 dBs in some cases. The convergence of the CDL algorithm can be improved by applying a mini-batch modification [3].

6. REFERENCES

[1] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," IEEE *Transactions on Image Processing*, vol.15, no.12, pp.3736-3745, Dec 2006.

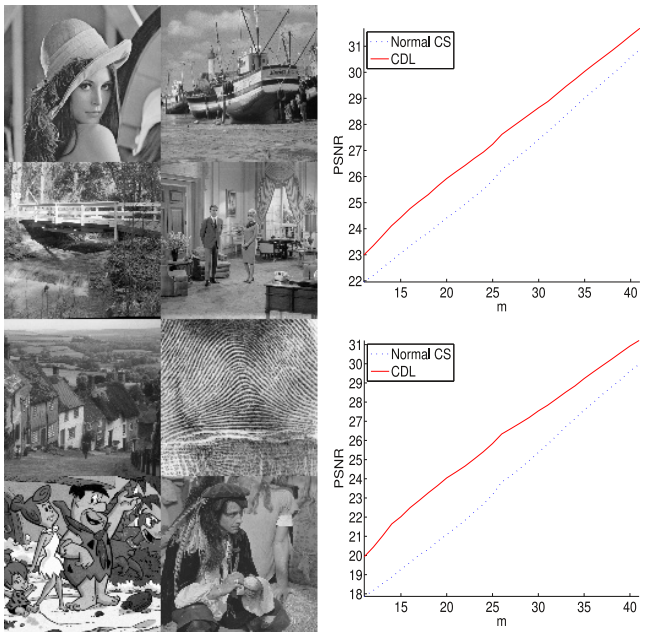

Fig. 1. Left: Test images. Top-right: average PSNR results. Bottom-right: PSNR results for the fingerprint image.

[2] R. Rubinstein, M. Zibulevsky, M. Elad, "Double sparsity: learning sparse dictionaries for sparse signal approximation," IEEE *Transactions on Signal Processing*, vol.58, no.3, pp.1553-1564, March 2010.

[3] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," In Proceedings of the 26th *Annual International Conference on Machine Learning* (ICML '09), ACM, pp.689-696, New York, NY, USA, 2009.

[4] G. Peyre, "Best basis compressed sensing," IEEE *Transactions on Signal Processing*, vol.58, no.5, pp.2613-2622, May 2010.

[5] E. Candes, "Compressive sampling," In Proceedings of the *International Congress of Mathematicians*, Madrid, Spain, 2006.

[6] M. Duarte, M. Davenport, D. Takhar, J. Laska, T. Sun, K. Kelly, and R. Baraniuk, "Single-pixel imaging via compressive sampling," IEEE *Signal Processing Magazine*, 2008.

[7] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Annals of Statistics*, 32, 407-499, 2004.

[8] M. Aghagolzadeh, H. Radha, "COLD: Compressive Online Learning of Dictionaries," *Technical Report*, Jan 2012, available at http://www.egr.msu.edu/~aghagol1/TR11.pdf.