Learning Stable Multilevel Dictionaries for Sparse Representation of Images

Jayaraman J. Thiagarajan, Karthikeyan Natesan Ramamurthy, and Andreas Spanias

Abstract—Dictionaries adapted to the data provide superior performance when compared to predefined dictionaries in applications involving sparse representations. Algorithmic stability and generalization are desirable characteristics for dictionary learning algorithms that aim to build global dictionaries which can efficiently model any test data similar to the training samples. In this paper, we propose an algorithm to learn dictionaries for sparse representation of image patches, and prove that the proposed learning algorithm is stable and generalizable asymptotically. The algorithm employs a 1-D subspace clustering procedure, the K-lines clustering, in order to learn a hierarchical dictionary with multiple levels. Furthermore, we propose a regularized pursuit scheme for computing sparse representations using a multilevel dictionary. Using simulations, we demonstrate the stability and generalization characteristics of the proposed algorithm with natural image patches. Finally, we employ multilevel dictionaries for compressed recovery and demonstrate improvements in recovery performance using both random and optimized projections when compared to baseline K-SVD dictionaries.

Index Terms—dictionary learning, sparse representations, algorithmic stability, generalization, subspace clustering

1 INTRODUCTION

1.1 Dictionary Learning for Sparse Representations

The statistical structure of naturally occurring signals and images allows for their efficient representation as a sparse linear combination of patterns, such as edges, lines and other elementary features [1]. A finite collection of normalized features is referred to as a dictionary. The linear model used for general sparse coding is given by

$$y = \Psi a + n,$$

where $y \in \mathbb{R}^M$ is the data vector and $\Psi = [\psi_1, \psi_2, \ldots, \psi_K] \in \mathbb{R}^{M \times K}$ is the dictionary. Each column of the dictionary, referred to as an atom, is a representative pattern normalized to unit $\ell_2$ norm. $a \in \mathbb{R}^K$ is the sparse coefficient vector and $n$ is a noise vector whose elements are independent realizations from the Gaussian distribution $\mathcal{N}(0, \sigma^2)$.

The sparse coding problem can be stated as

$$\hat{a} = \arg\min_a ||a||_0 \text{ s.t. } ||y - \Psi a||_2^2 \leq \epsilon,$$

where $||.||_0$ indicates the $\ell_0$ norm, $||.||_2$ denotes the $\ell_2$ norm and $\epsilon$ is the error goal for the representation. However, exact $\ell_0$ minimization is a combinatorial problem and hence its convex surrogate, the $\ell_1$ norm, is often used. Some of the widely used methods for computing sparse representations include the Matching Pursuit (MP) [2], Orthogonal Matching Pursuit (OMP) [3], Basis Pursuit (BP) [4], FOCUSS [5] and iterated shrinkage algorithms [6], [7]. The sparse coding model has been successfully used for inverse problems in images [8]–[10], and also in machine learning applications such as classification and clustering [11]–[21].

Predefined dictionaries obtained using the discrete cosine transform (DCT), wavelet, and curvelet [22] bases have been used successfully for image reconstruction and compression. The dictionary $\Psi$ can also be designed from a union of orthonormal bases [23] or structured as an overcomplete set of individual vectors optimized to the training data [24], [25]. A wide range of dictionary learning algorithms have been proposed in the literature [26]–[32], some of which are tailored for specific applications. The conditions under which a dictionary can be identified from the training data using an $\ell_1$ minimization approach are derived in [33]. The joint optimization problem for dictionary learning and sparse coding with $\ell_0$ sparsity constraints can be expressed as [8], [34], [35]

$$\min_{\Psi, a} ||Y - \Psi A||_F^2 \text{ s.t. } ||a_i||_0 \leq S, \forall i, ||\psi_j||_2 = 1, \forall j,$$

where $Y = [y_1, y_2, \ldots, y_T]$ is a collection of $T$ training vectors, $A = [a_1, a_2, \ldots, a_T]$ is the coefficient matrix, $S$ is the sparsity of the coefficient vector and $||.||_F$ denotes the Frobenius norm. Learned dictionaries have been successfully applied to image compression, denoising and inpainting [9], [10].

In this paper, we propose a stable and generalizable learning algorithm for designing multilevel dictionaries that are particularly suited for sparse approximation of natural images. A simple example of learning a dictionary with two levels is demonstrated in Figure 1. The properties and performance of this learning algorithm will be analyzed in detail in this paper. The multilevel dictionary (MLD) learning algorithm is a hierarchical procedure where the dictionary atoms in each level are
obtained using a 1-D subspace clustering algorithm, which we refer to as K-lines clustering [36]. The proposed algorithm builds global dictionaries using a set of randomly chosen training patches obtained from a large collection of natural images that can generalize well to any test set of patches. For a learned dictionary to provide a good approximation, the test data must be similar to the data samples used for training. Since local regions of natural images have high redundancy and consistent statistical properties [39], learning global dictionaries from a random collection of natural image patches will provide a good representation for patches from images not in the training set. The effectiveness of such dictionaries have been demonstrated in denoising [9] and compressed recovery [40].

1.2 Stability and Generalization in Learning
A learning algorithm is a map from the space of training examples to the hypothesis space of functional solutions. Algorithmic stability characterizes the behavior of a learning algorithm with respect to the perturbations of its training set [41], and generalization ensures that the expected error of the learned function with respect to the novel test data will be close to the average empirical training error [42]. In clustering, the learned function is completely characterized by the cluster centers. Stability of a clustering algorithm implies that the cluster centroids learned by the algorithm are not significantly different when different sets of i.i.d. samples from the same probability space are used for training [43]. When there is a unique minimizer to the clustering objective with respect to the underlying data distribution, stability of a clustering algorithm is guaranteed [44] and this analysis has been extended to characterize the stability of K-means clustering in terms of the number of minimizers [45]. In [38], the stability properties of the K-lines clustering algorithm have been analyzed and they have been shown to be similar to those of K-means clustering. Note that all the stability characterizations depend only on the underlying data distribution and the number of clusters, and not on the actual training data itself. Generalization implies that the average empirical training error becomes asymptotically close to the expected error with respect to the probability space of data, as the number of training samples $T \to \infty$. In [46], the generalization bound for sparse coding in terms of the number of samples $T$, also referred to as sample complexity, is derived and in [47] the bound is improved by assuming a class of dictionaries that are nearly orthogonal.

The algorithmic stability of dictionary learning methods has not been discussed in the literature until now, to the best of our knowledge. Given a sufficiently large training set, a stable learning algorithm will result in global dictionaries that will depend only on the probability space to which the training samples belong and not on the actual samples themselves. Generalization ensures that such global dictionaries learned result in a good performance with test data. In other words, the asymptotic stability and generalization of a dictionary learning algorithm provides the theoretical justification for the uniformly good performance of global dictionaries learned from an arbitrary training set. We study the stability properties of the proposed MLD learning algorithm and prove that it is asymptotically stable. We also show that the proposed algorithm generalizes asymptotically.

1.3 Contributions
In this paper, we propose the MLD learning algorithm to design global representative dictionaries for image patches. We show that, for a sufficient number of levels, the proposed algorithm converges and also demonstrate that a multilevel dictionary with a sufficient number of atoms per level exhibits energy hierarchy. Furthermore, we develop a Regularized Multilevel OMP (RM-OMP) procedure for computing the sparse codes of test data using the proposed dictionary. Some preliminary algorithmic details and results obtained using MLD have been reported in [37].

Using the fact that the K-lines clustering algorithm is stable, we perform stability analysis of the MLD algorithm. For any two sets of i.i.d. training samples from the same probability space, as the number of training samples $T \to \infty$, we show that the dictionaries learned become close to each other asymptotically. When there is a unique minimizer to the objective in each level of learning, this holds true even if the training sets are completely disjoint. However, when there are multiple minimizers for the objective in at least one level, we prove that the learned dictionaries are asymptotically
close when the difference between their corresponding training sets is $o(\sqrt{T})$. Instability of the algorithm when the difference between two training sets is $\Omega(\sqrt{T})$, is also shown for the case of multiple minimizers. Furthermore, we prove the asymptotic generalization of the learning algorithm.

The stability characteristics of MLD learning are experimentally demonstrated using natural image data. We show that, the stability in terms of the learned dictionaries improves as the difference between their corresponding training sets becomes small and as the number of training samples increases. We train a global multilevel dictionary from a set of patches chosen randomly from a corpus of natural images and study its generalization behavior using several simulations. For comparison, we use a dictionary learned using the K-SVD algorithm, with similar training parameters, for the same training data set. We observe that the error in sparse approximation for the training and test data sets become comparable as the size of the training set increases. When compared to the K-SVD, the proposed algorithm exhibits much improved generalization by providing reduced test error even with a small number of training samples. Extensive simulations with compressed sensing of natural image patches and standard images, using both random and optimized measurement matrices, show that the learned MLD exhibits better recovery performance when compared to the K-SVD dictionary.

1.4 Organization of the Paper

The rest of this paper is organized as follows. In Section 2, we briefly discuss the K-lines clustering algorithm and its stability characteristics. Section 3 presents the motivation for learning a multilevel dictionary and describes the proposed algorithm in detail. The stability analysis and generalizability of the proposed dictionary learning algorithm are discussed in Section 4 and the experimental results are presented in Section 5. The concluding remarks are provided in Section 6.

2 Background

In this section, we describe the K-lines clustering, a 1-D subspace clustering procedure proposed in [36], which forms a building block of the proposed dictionary learning algorithm. Furthermore, we briefly discuss the results for stability analysis of K-means and K-lines algorithms reported in [43] and [38] respectively. The ideas described in this section will be used in Section 4 to study the stability characteristics of the proposed dictionary learning procedure.

2.1 K-lines Clustering Algorithm

The K-lines clustering algorithm is an iterative procedure that performs a least squares fit of $K$ 1-D linear subspaces to the training data [36]. Note that the K-lines clustering is a special case of general subspace clustering methods proposed in [48]–[50], when the subspaces are 1-dimensional and constrained to pass through the origin. In contrast with K-means, K-lines clustering allows each data sample to have an arbitrary coefficient value corresponding to the centroid of the cluster it belongs to. Furthermore, the cluster centroids are normalized to unit $\ell_2$ norm. Given the set of $T$ data samples $Y = \{y_i\}_{i=1}^T$ and the number of clusters $K$, K-lines clustering proceeds in two stages after initialization: the cluster assignment and the cluster centroid update. In the cluster assignment stage, training vector $y_i$ is assigned to a cluster $j$ based on the minimum distortion criteria, $d(y_i, \psi_j)$, where the distortion measure is

$$d(y, \psi) = \|y - \psi(y^T\psi)\|_2^2. \quad (4)$$

In the cluster centroid update stage, we perform singular value decomposition (SVD) of $Y_j = \{y_i\}_{i \in C_j}$, where $C_j = \{i | H(y_i) = j\}$ contains indices of training vectors assigned to the cluster $j$. The left singular vector corresponding to the largest singular value of the decomposition, is the centroid of cluster $j$. The different strategies for initialization of cluster centroids and estimation of the number of hyperlines are described in [36].

2.2 Stability Analysis of Clustering Algorithms

Analyzing the stability of unsupervised clustering algorithms can be valuable in terms of understanding their behavior with respect to perturbations in the training set. These algorithms extract the underlying structure in the training data and the quality of clustering is determined by an accompanying cost function. As a result, any clustering algorithm can be posed as a Empirical Risk Minimization (ERM) procedure, by defining a hypothesis class of loss functions to evaluate the possible cluster configurations and to measure their quality [51]. For example, K-lines clustering can be posed as an ERM problem over the distortion function class

$$G_K = \left\{ g_\Psi(y) = d(y, \psi_j), j = \arg\max_{l \in \{1, \ldots, K\}} |y^T\psi_l| \right\}. \quad (5)$$

The class $G_K$ is obtained by taking functions $g_\Psi$ corresponding to all possible combinations of $K$ unit length vectors from the $\mathbb{R}^M$ space for the set $\Psi$. Let us define the probability space for the data in $\mathbb{R}^M$ as $(\mathcal{Y}, \Sigma, P)$, where $\mathcal{Y}$ is the sample space and $\Sigma$ is a sigma-algebra on $\mathcal{Y}$, i.e., the collection of subsets of $\mathcal{Y}$ over which the probability measure $P$ is defined. The training samples, $\{y_i\}_{i=1}^T$, are $T$ i.i.d. realizations from the probability space.

Ideally, we are interested in computing the cluster centroids $\Psi$ that minimize the expected distortion $\mathbb{E}[g_\Psi]$ with respect to the probability measure $P$. However, the underlying distribution of the data samples is not known.
and hence we resort to minimizing the average empirical distortion with respect to the training samples \( \{y_i\}_{i=1}^T \) as

\[
g_\Psi = \arg\min_{g \in \mathcal{G}_K} \frac{1}{T} \sum_{i=1}^T g_\Psi(y_i). \tag{6}
\]

When the empirical averages of the distortion functions in \( \mathcal{G}_K \) uniformly converge to the expected values over all probability measures \( P \),

\[
\lim_{T \to \infty} \sup_P \left( \sup_{g \in \mathcal{G}_K} \left| \mathbb{E}[g_\Psi] - \frac{1}{T} \sum_{i=1}^T g_\Psi(y_i) \right| > \delta \right) = 0, \tag{7}
\]

for any \( \delta > 0 \), we refer to the class \( \mathcal{G}_K \) as uniform Glivenko-Cantelli (uGC). In addition to being uGC, if the class also satisfies a version of the central limit theorem, it is defined as uniform Donsker [41]. In order to determine if \( \mathcal{G}_K \) is uniform Donsker, we have to verify if the covering number of \( \mathcal{G}_K \) with respect to the supremum norm, \( N_\infty(\gamma, \mathcal{G}_K) \), grows polynomially in the dimensions \( M \) [43]. Here, \( \gamma \) denotes the maximum \( L_\infty \) distance between an arbitrary distortion function in \( \mathcal{G}_K \), and the function that covers it. For K-lines clustering, the covering number is upper bounded by [38, Lemma 2.1]

\[
N_\infty(\gamma, \mathcal{G}_K) \leq \left( \frac{8R^3K + \gamma}{\gamma} \right)^{MK}, \tag{8}
\]

where we assume that the data lies in an \( M \)-dimensional \( \ell_2 \) ball of radius \( R \) centered at the origin. Therefore, \( \mathcal{G}_K \) belongs to the uniform Donsker class.

The general idea behind stability of a clustering algorithm is that the algorithm should produce cluster centroids that are not significantly different when different i.i.d. training sets from the same probability space are used for training [43]–[45]. Stability is characterized based on the number of minimizers to the clustering objective with respect to the underlying data distribution. A minimizer corresponds to a function \( g_\Psi \in \mathcal{G}_K \) with the minimum expectation \( \mathbb{E}[g_\Psi] \). Stability analysis of the K-means algorithm has been reported in [43], [45].

Though the geometry of K-lines clustering is different from that of K-means, the stability characteristics of the two clustering algorithms have been found to be similar [38]. Given two sets of cluster centroids \( \Psi = \{\psi_1, \ldots, \psi_K\} \) and \( \Lambda = \{\lambda_1, \ldots, \lambda_K\} \) learned from training sets of \( T \) i.i.d. samples each realized from the same probability space, let us define the \( L_1(P) \) distance between the corresponding clusterings as

\[
\|g_\Psi - g_\Lambda\|_{L_1(P)} = \int |g_\Psi(y) - g_\Lambda(y)|dP(y). \tag{9}
\]

When \( T \to \infty \), and \( \mathcal{G}_K \) is uniform Donsker, stability in terms of the distortion functions is expressed as

\[
\|g_\Psi - g_\Lambda\|_{L_1(P)} \overset{P}{\to} 0, \tag{10}
\]

where \( \overset{P}{\to} \) denotes convergence in probability. This holds true even for \( \Psi \) and \( \Lambda \) learned from completely disjoint training sets, when there is a unique minimizer to the clustering objective. When there are multiple minimizers, (10) holds true with respect to a change in \( o(\sqrt{T}) \) samples between two training sets and fails to hold with respect to a change in \( \Omega(\sqrt{T}) \) samples [38]. The distance between the cluster centroids themselves is defined as [43]

\[
\Delta(\Psi, \Lambda) = \max_{1 \leq j \leq K} \min_{1 \leq i \leq \ell} \left[ (d(\psi_j, \lambda_i))^{1/2} + (d(\psi_j, \lambda_j))^{1/2} \right]. \tag{11}
\]

**Lemma 2.1** ( [38] ): If the \( L_1(P) \) distance between the distortion functions for the clusterings \( \Psi \) and \( \Lambda \) is bounded as \( \|g_\Psi - g_\Lambda\|_{L_1(P)} < \mu \), for some \( \mu > 0 \), and \( dP(y)/dy > C \), for some \( C > 0 \), then \( \Delta(\Psi, \Lambda) \leq 2\sin(\rho) \) where

\[
\rho \leq 2\sin^{-1}\left[ \frac{1}{r} \left( \frac{\mu}{C_{C,M}} \right)^{\frac{1}{M+1}} \right]. \tag{12}
\]

Here the training data is assumed to lie outside an \( M \)-dimensional \( \ell_2 \) ball of radius \( r \) centered at the origin, and the constant \( C_{C,M} \) depends only on \( C \) and \( M \).

When the clustering algorithm is stable according to (10), for admissible values of \( \rho \), Lemma 2.1 indicates that the cluster centroids become arbitrarily close to each other, \( \Delta(\Psi, \Lambda) \overset{P}{\to} 0 \), which implies stability in terms of cluster centroids. From (12), it is also clear that the K-lines clustering cannot be stable if some training vectors have a norm close enough to 0, (i.e.) \( r \to 0 \).

### 3 Multilevel Dictionary Learning

In this section, we motivate and develop a multilevel dictionary learning approach for sparse representations, whose algorithmic stability and generalizability will be proved in Section 4. Furthermore, we propose the R-OMP algorithm, that can be used to obtain sparse codes for a test image using the multilevel dictionary.

#### 3.1 Motivation for Multilevel Learning

Our motivation for learning an MLD is two-fold. Firstly we require a global dictionary that can exploit, (a) the redundancy observed across local regions in natural images and, (b) the hierarchy of patterns found in training image patches. Secondly, the learning procedure must be provably stable, with respect to the notion of algorithmic stability, and generalizable.

The generative model in (1) is well suited for natural signals and images as they can be represented using a sparse linear combination of elementary features chosen from a dictionary [24]. The redundancy in the local regions of natural images [39] allows for the design of global dictionaries that can generalize well to a wide range of images. Global dictionaries learned from a set of randomly chosen patches from natural images have been successfully used for denoising [9], compressed sensing [40] and classification [52]. In addition to exhibiting redundancy, the natural image patches typically
contain either geometric patterns or stochastic textures or a combination of both. This fact is demonstrated in [53], where the authors define two types of atomic subspaces to model image patches: subspaces of low dimensions (explicit manifolds) for primitive geometric patterns and subspaces of high dimensions (implicit manifolds) for stochastic textures. Since the image patches can contain both geometric and stochastic structures, a hybrid combination of explicit and implicit manifolds can be used for modeling them [53]. The proposed MLD algorithm learns global representative patterns in multiple levels, according to the order of their energy contribution. Since the geometric patterns usually are of higher energy when compared to stochastic textures in images, geometric patterns are learned in the first few levels and stochastic textures are learned in the last few levels.

Considering the dictionary learning formulation in (3), it can be seen that clustering algorithms such as the K-means and the K-lines can be obtained by constraining the desired sparsity to be 1. Since the stability characteristics of clustering algorithms are well understood, employing similar tools to analyze the more general dictionary learning can be beneficial. Note that the proposed algorithm poses dictionary learning as performing K-lines clustering in multiple levels and hence in this case we can use the stability characteristics of the clustering algorithm to study the stability of multilevel learning. Furthermore, by exploiting the fact that the distortion function class for each level of learning is uniform Donsker, the generalizability of the algorithm can also be proved. Note that multilevel learning is different from the work in [54], where multiple sub-dictionaries are designed and one of them is chosen for representing a group of patches.

### 3.2 Proposed MLD Learning Algorithm

We denote the MLD as \( \Psi = [\Psi_1, \Psi_2, ..., \Psi_L] \), and the coefficient matrix as \( A = [A_1^T, A_2^T, ..., A_L^T]^T \). Here, \( \Psi_l \) is the sub-dictionary and \( A_l \) is the coefficient matrix for level \( l \). The approximation in level \( l \) is expressed as

\[
R_{l-1} = \Psi_l A_l + R_l, \quad \text{for } l = 1, ..., L, \tag{13}
\]

where \( R_{l-1}, R_l \) are the residuals for the levels \( l-1 \) and \( l \) respectively and \( R_0 = Y \), the matrix of training image patches. This implies that the residual matrix in level \( l-1 \) serves as the training data for level \( l \). Note that the sparsity of the representation in each level is fixed at 1. Hence, the overall approximation for all levels is

\[
Y = \sum_{l=1}^{L} \Psi_l A_l + R_L. \tag{14}
\]

MLD learning can be interpreted as a block-based dictionary learning problem with unit sparsity per block, where the sub-dictionary in each block can allow only a 1-sparse representation and each block corresponds to a level. The sub-dictionary for level \( l \), \( \Psi_l \), is the set of clustern centroids learned from the training matrix for that level, \( R_{l-1} \), using K-lines clustering. MLD learning can be formally stated as an optimization problem that proceeds from the first level until the stopping criteria is reached. For level \( l \), the optimization problem is

\[
\underset{\Psi_l, A_l}{\text{argmin}} \| R_{l-1} - \Psi_l A_l \|_F^2 \quad \text{subject to } \| a_{l,i} \|_0 \leq 1, \quad \text{for } i = 1, ..., T, \tag{15}
\]

along with the constraint that the columns of \( \Psi_l \) have unit \( l_2 \) norm, where \( a_{l,i} \) is the \( i \)-th column of \( A_l \) and \( T \) is the number of columns in \( A_l \). We adopt the notation \( \{ \Psi_l, A_l \} = \text{KLC}(R_{l-1}, K_l) \) to denote the problem in (15) where \( K_l \) is the number of atoms in the sub-dictionary \( \Psi_l \). The stopping criteria is provided either by imposing a limit on the residual representation error or the maximum number of levels (\( L \)). Note that the total number of levels is the same as the maximum number of non-zero coefficients (sparsity) of the representation. The error constraint can be stated as, \( \| R_{l,i} \|_2^2 \leq \epsilon, \forall i = 1, ..., T \) for some level \( l \), where \( \epsilon \) is the error goal.

Table 1 lists the MLD learning algorithm with sparsity and error constraints. We use the notation \( A_l(j) \) to denote the \( j \)-th element of the set \( A_l \) and \( r_{l,i} \) denotes the \( i \)-th column vector in the matrix \( R_l \). The set \( A_l \) contains the indices of the residual vectors of level \( l \) whose norm is greater than the error goal. The residual vectors indexed by \( A_l \) are stacked in the matrix, \( R_l \), which in turn serves as the training matrix for the next level, \( l+1 \). In MLD learning, for a given level \( l \), the residual \( r_{l,i} \) is orthogonal.
to the representation $\Psi_l a_{l,i}$. This implies that
\[ \| r_{l-1,i} \|^2 = \| \Psi_l a_{l,i} \|^2 + \| r_{l,i} \|^2. \] (16)
Combining this with the fact that $y_i = \sum_{l=1}^{L} \Psi_l a_{l,i} + r_{L,i}$, $a_{l,i}$ is 1-sparse, and the columns of $\Psi_l$ are of unit $\ell_2$ norm, we obtain the relation
\[ \| y_i \|^2 = \sum_{l=1}^{L} \| a_{l,i} \|^2 + \| r_{L,i} \|^2. \] (17)
Equation (17) states that the energy of any training vector is equal to the sum of squares of its coefficients and the energy of its residual. From (16), we also have that
\[ \| R_{l-1} \|^2 = \| \Psi_l A_l \|^2 + \| R_l \|^2. \] (18)

In our implementation of MLD learning, we include an additional step where the residual at each level is orthogonalized to the dictionary atoms chosen so far, and the coefficients are recomputed. Note that this does not affect any other behavior of the algorithm that is discussed in this section.

The training vectors for the first level of the algorithm, $r_{0,i}$, lie in the ambient $\mathbb{R}^M$ space and the residuals, $r_{1,i}$, lie in a finite union of $\mathbb{R}^{M-1}$ subspaces. This is because, for each dictionary atom in the first level, its residual lies in an $M-1$ dimensional space orthogonal to it. In the second level, the dictionary atoms can possibly lie anywhere in $\mathbb{R}^M$ and hence the residuals can lie in a finite union of $\mathbb{R}^{M-2}$ dimensional subspaces. Hence, we can generalize that the dictionary atoms for all levels lie in $\mathbb{R}^M$, whereas the training vectors of level $l \geq 2$, lie in finite unions of $\mathbb{R}^{M-1}, \ldots, \mathbb{R}^{M-l+1}$ dimensional subspaces of the $\mathbb{R}^M$ space.

### 3.3 Convergence

The convergence of MLD learning and the energy hierarchy in the representation obtained using an MLD can be shown by providing two guarantees. The first guarantee is that for a fixed number of atoms per level, the algorithm will converge to the required error within a sufficient number of levels. This is because the K-lines clustering makes the residual energy of the representation smaller than the energy of the training matrix at each level (i.e., $\| R_{l} \|^2 < \| R_{l-1} \|^2$). This follows from (18) and the fact that $\| \Psi_l A_l \|^2 > 0$.

The second guarantee is that for a sufficient number of atoms per level, the representation energy in level $l - 1$ will be less than the representation energy in level $l$. To show this, we first state that for a sufficient number of dictionary atoms per level, $\| \Psi_l A_l \|^2 > \| R_l \|^2$. This means that for every $l$
\[ \| R_l \|^2 < \| \Psi_l A_l \|^2 < \| R_{l-1} \|^2, \] (19)
because of (18). This implies that $\| \Psi_l A_l \|^2 < \| \Psi_{l-1} A_{l-1} \|^2$, i.e., the energy of the representation in each level reduces progressively from $l = 1$ to $l = L$.

### 3.4 Sparse Approximation using an MLD

Sparse approximation for any test data can be performed by stacking all the levels of an MLD together into a single dictionary, and using any standard pursuit algorithm on $\Psi$. Though this implementation is straightforward, it does not exploit the energy hierarchy observed in the learning process. Hence, we propose to perform reconstruction using a Multilevel Orthogonal Matching Pursuit (M-OMP) procedure which evaluates a 1-sparse representation for each level using the dictionary atoms from that level, and orthogonalizes the residual to the dictionary atoms chosen so far. Obtaining sparse representations using this scheme is computationally cheaper than using OMP on a flat dictionary. Since the complexity of orthogonalizing the residual to the chosen dictionary atoms is the same for both the pursuit methods, we compare only the computations involved in choosing the dictionary atoms for an $S$-sparse representation. Consider the case where we have $L$ levels in the MLD with each level having $K/L$ atoms, such that the total number of atoms is $K$. To choose $S$ dictionary atoms using the OMP algorithm, the complexity is of order $SKM$, where $M$ is the dimensionality of the dictionary atom. Whereas, M-OMP based reconstruction incurs computations of the order $SKM/L$.

Though M-OMP is computationally cheaper, imposing the energy hierarchy observed in the training process to any test data might result in poor generalization, as observed from our experiments. Hence, there is a need to regularize this procedure such that there is more flexibility in choosing dictionary atoms for representing the test data. We propose to build a sub-dictionary with atoms selected from the current level as well as the $u$ immediately preceding and following levels, $\Phi_l = [\Phi_{l-u} \Phi_{l-(u-1)} \ldots \Phi_{l+(u-1)} \Phi_{l+u}]$, in every step of the pursuit algorithm. In our implementation, we fix $u = 2$ and also reduce the size of the sub-dictionary appropriately when $l < u$ and $l > L - u$. The dictionary $\Phi_l$ is used to compute a 1-sparse representation for that step of the pursuit. It can be observed from the results presented in the following sections that, this RM-OMP scheme achieves improvements in performance when compared to M-OMP. Evaluating sparse representation for a test image using the RM-OMP scheme incurs computations of the order $S(2u+1)KM/L$. If $u$ is chosen to be much smaller than $L$, we have $(2u+1)/L < 1$ and hence the savings in computations is still significant in comparison to the standard OMP algorithm.

### 4 Stability and Generalization

In this section, the behavior of the proposed dictionary learning algorithm is considered from the viewpoint of algorithmic stability: the behavior of the algorithm with respect to the perturbations in the training set. It will be shown that the dictionary atoms learned by the algorithm from two different training sets whose samples are realized from the same probability space,
become arbitrarily close to each other, as the number of training samples $T \to \infty$. Since the proposed MLD learning is equivalent to learning K-lines cluster centroids in multiple levels, the stability analysis of K-lines clustering [38], briefly discussed in Section 2.2, will be utilized in order to prove its stability. For each level of learning, the cases of single and multiple minimizers to the clustering objective will be considered. Proving that the learning algorithm is stable will show that the global dictionaries learned from the data depend only on the probability space to which the training samples belong and not on the actual samples themselves, as $T \to \infty$. We also show that the MLD learning generalizes asymptotically, i.e., the difference between expected error and average empirical error in training approaches zero, as $T \to \infty$. Therefore, the expected error for novel test data, drawn from the same distribution as the training data, will be close to the average empirical training error.

The stability analysis of the MLD algorithm will be performed by considering two different dictionaries $\Psi$ and $\Lambda$ with $L$ levels each. Each level consists of $K_l$ dictionary atoms and the sub-dictionaries in each level are indicated by $\Psi_l$ and $\Lambda_l$ respectively. Note that the sub-dictionaries $\Psi_l$ and $\Lambda_l$ are the cluster centers learned using K-lines clustering on the training data for level $l$. The steps involved in proving the overall stability of the algorithm are: (a) showing that each level of the algorithm is stable in terms of $L_1(P)$ distance between the distortion functions, defined in (9), as the number of training samples $T \to \infty$ (Section 4.1), (b) proving that stability in terms of $L_1(P)$ distances indicates closeness of the centers of the two clusterings (Section 4.2), in terms of the metric defined in (11), and (c) showing that level-wise stability leads to overall stability of the dictionary learning algorithm (Section 4.3).

### 4.1 Level-wise Stability

Let us define a probability space $(\mathcal{Y}_l, \Sigma_l, P_l)$ where $\mathcal{Y}_l$ is the data that lies in $\mathbb{R}^M$, and $P_l$ is the probability measure. The training samples for the sub-dictionaries $\Psi_l$ and $\Lambda_l$ are two different sets of $T$ i.i.d. realizations from the probability space. We also assume that the $\ell_2$ norm of the training samples is bounded from above and below (i.e., $0 < r \leq \|y\|_2 \leq R < \infty$). Note that, in a general case, the data will lie in $\mathbb{R}^M$ for the first level of dictionary learning and in a finite union of lower-dimensional subspaces of $\mathbb{R}^M$ for the subsequent levels. In both cases, the following argument on stability will hold. This is because when the training data lies in a union of lower dimensional subspaces of $\mathbb{R}^M$, we can assume it to be still lying in $\mathbb{R}^M$, but assign the probabilities outside the union of subspaces to be zero.

In each level, $\Psi_l$ and $\Lambda_l$ are learned using the K-lines clustering algorithm on two different i.i.d. sets of training data. The distortion function class for the clusterings, defined similar to (5), is uniform Donsker because the covering number with respect to the supremum norm grows polynomially, according to (8). When a unique minimizer exists for the clustering objective, the distortion functions corresponding to the different clusterings $\Psi_l$ and $\Lambda_l$ become arbitrarily close, $\|g_{\Psi_l} - g_{\Lambda_l}\|_{L_1(P_l)} \to 0$, even for completely disjoint training sets, as $T \to \infty$. However, in the case of multiple minimizers, $\|g_{\Psi_l} - g_{\Lambda_l}\|_{L_1(P_l)} \to 0$ holds only with respect to a change of $o(\sqrt{T})$ training samples between the two clusterings, and fails to hold when there is a change of $\Omega(\sqrt{T})$ samples [38], [43].

### 4.2 Distance between Cluster Centers for a Stable Clustering

For each cluster center in the clustering $\Psi_l$, we pick the closest cluster center from $\Lambda_l$, in terms of the distortion measure (4), and form pairs. Let us indicate the $j^{th}$ pair of cluster centers by $\psi_{l,j}$ and $\lambda_{l,j}$. Let us define $\tau$ disjoint sets $\{A_l\}_{l=1}^\tau$, in which the training data for the clusterings exist, such that $P(\cup_{l=1}^\tau A_l) = 1$. By defining such disjoint sets, we can formalize the notion of training data lying in a union of subspaces of $\mathbb{R}^M$. The intuitive fact that the cluster centers of two clusterings are close to each other in $\mathbb{R}^M$ space, given that their distortion functions are close, is proved in the lemma below.

**Lemma 4.1:** Consider two sub-dictionaries (clusterings) $\Psi_l$ and $\Lambda_l$ with $K_l$ atoms each obtained using the $T$ training samples that exist in the $\tau$ disjoint sets $\{A_l\}_{l=1}^\tau$ in the $\mathbb{R}^M$ space, with $0 < r \leq \|y\|_2 \leq R < \infty$, and $dP_l(y)/dy > C$ in each of the sets. When the distortion functions become arbitrarily close to each other, $\|g_{\Psi_l} - g_{\Lambda_l}\|_{L_1(P_l)} \to 0$ as $T \to \infty$, the smallest angle between the subspaces spanned by the cluster centers becomes arbitrarily close to zero, i.e.,

$$\angle(\psi_{l,j}, \lambda_{l,j}) \to 0, \forall j \in 1, \ldots, K_l.$$ (20)

**Proof:** Denote the smallest angle between the subspaces represented by $\psi_{l,j}$ and $\lambda_{l,j}$ as $\angle(\psi_{l,j}, \lambda_{l,j}) = \rho_{l,j}$, and define a region $S(\psi_{l,j}, \rho_{l,j}/2) = \{y|\angle(\psi_{l,j}, y) \leq \rho_{l,j}/2, 0 < r \leq \|y\|_2 \leq R < \infty\}$. If $y \in S(\psi_{l,j}, \rho_{l,j}/2)$, then $y^T(I - \psi_{l,j}^T \psi_{l,j})y \leq y^T(I - \lambda_{l,j}^T \lambda_{l,j})y$. An illustration...
of this setup for a 2-D case is given in Figure 2. In this figure, the arc $q_1q_3$ is of radius $r$ and represents the minimum value of $\|y\|_2$. By definition, the $L_1(P_i)$ distance between the distortion functions of the clusterings for data that exists in the disjoint sets $\{A_i\}_{i=1}^\infty$ is

$$\|g_{l,i,j} - g_{l,i} \|_{L_1(P_i)} = \sum_{i=1}^\infty \|g_{l,i}(y) - g_{l,i}(y)\|dP(y).$$

For any $l$ and $j$ with a non-empty $B_{l,i,j} = S(\psi_{l,j}, \rho_{l,j}/2) \cap A_i$ we have,

$$\|g_{l,i,j} - g_{l,i} \|_{L_1(P_i)} \geq \int_{B_{l,i,j}} \|g_{l,i}(y) - g_{l,i}(y)\|dP(y),$$

$$= \int_{B_{l,i,j}} [y^T (I - \lambda_{l,i,j}^T y - \sum_{k=1}^\infty y^T (1 - \psi_{l,k} \psi_{l,k}^T) y] \|dP(y),$$

$$\geq \int_{B_{l,i,j}} [y^T (1 - \lambda_{l,i,j}^T y - \sum_{k=1}^\infty y^T (1 - \psi_{l,k} \psi_{l,k}^T) y] \|dP(y),$$

$$\geq C \int_{B_{l,i,j}} [(y^T \psi_{l,i,j})^2 - (y^T \lambda_{l,i,j})^2] dy. \quad (25)$$

We have $g_{l,i}(y) = y^T (I - \lambda_{l,i,j} \lambda_{l,i,j}^T) y$ in (23), since $\lambda_{l,i,j}$ is the closest cluster center to the data in $S(\psi_{l,j}, \rho_{l,j}/2) \cap A_i$ in terms of the distortion measure (4). Note that $I$ is the indicator function and (25) follows from (24) because $dP(y)/dy > C$. Since by assumption, $\|g_{l,i,j} - g_{l,i} \|_{L_1(P_i)} \overset{P}{\to} 0$, from (25), we have

$$\|y\|_2^2 (\cos^2 \theta_{l,i,j} - \cos^2 \theta_{l,i,j}) \overset{P}{\to} 0, \text{ for all } y. \text{ By definition of the region } B_{l,i,j}, \text{ we have } \theta_{l,i,j} \leq \theta_{l,i,j}. \text{ Since } \|y\|_2 \text{ is bounded away from zero and infinity, if } (\cos^2 \theta_{l,i,j} - \cos^2 \theta_{l,i,j}) \overset{P}{\to} 0 \text{ holds for all } y \in B_{l,i,j}, \text{ then we have } \angle(\psi_{l,i,j}, \lambda_{l,i,j}) \overset{P}{\to} 0. \text{ This is true for all cluster center pairs as we have shown this for an arbitrary } i \text{ and } j.$$  \hfill \Box

### 4.3 Stability of the MLD Algorithm

The stability of the MLD algorithm as a whole, is proved in Theorem 4.3 from its levelwise stability by using an induction argument. The proof will depend on the following lemma which shows that the residuals from two stable clusterings belong to the same probability space.

**Lemma 4.2:** When the training vectors for the subdictionaries (clusterings) $\Phi_i$ and $\Lambda_i$ are obtained from the probability space $(\gamma_i, \Sigma_i, P_i)$, and the cluster center pairs become arbitrarily close to each other as $T \to \infty$, the residual vectors from both the clusterings belong to an identical probability space $(\gamma_{i+1}, \Sigma_{i+1}, P_{i+1})$.

**Proof:** For the $j$th cluster center pair $\psi_{l,j}, \lambda_{l,j}$, define $\tilde{\psi}_{l,j}$ and $\tilde{\lambda}_{l,j}$ as the projection matrices for their respective orthogonal complement subspaces $\psi_{l,j}^\perp$ and $\lambda_{l,j}^\perp$. Define the sets $D_{\psi_{l,j}} = \{y \in \psi_{l,j}(\beta + d\beta) + \psi_{l,j} \alpha\}$ and $D_{\lambda_{l,j}} = \{y \in \lambda_{l,j}(\beta + d\beta) + \lambda_{l,j} \alpha\}$, where $-\infty < \alpha < \infty$, $\beta$ is an arbitrary fixed vector, not orthogonal to both $\psi_{l,j}$ and $\lambda_{l,j}$, and $d\beta$ is a differential element. The residual vector set for the cluster $\psi_{l,j}$, when $y \in D_{\psi_{l,j}}$ is given by $r_{\psi_{l,j}} \in \{\tilde{\psi}_{l,j} y | y \in D_{\psi_{l,j}}\}$, or equivalently $r_{\psi_{l,j}} \in \{\tilde{\psi}_{l,j}(\beta + d\beta)\}$. Similarly for the cluster $\lambda_{l,j}$, we have $r_{\lambda_{l,j}} \in \{\tilde{\lambda}_{l,j}(\beta + d\beta)\}$. For a 2-D case, Figure 3 shows the 1-D subspace $\psi_{l,j}$, its orthogonal complement $\psi_{l,j}^\perp$, the set $D_{\psi_{l,j}}$ and the residual set $\{\tilde{\psi}_{l,j}(\beta + d\beta)\}$.

In terms of probabilities, we also have that $P_i(y \in D_{\psi_{l,j}}) = P_{i+1}(r_{\psi_{l,j}} \in \{\tilde{\psi}_{l,j}(\beta + d\beta)\})$, because the residual set $\{\tilde{\psi}_{l,j}(\beta + d\beta)\}$ is obtained by a linear transformation of $D_{\psi_{l,j}}$. Here $P_i$ and $P_{i+1}$ are probability measures defined on the training data for levels $l$ and $l+1$ respectively. Similarly, $P_i(y \in D_{\psi_{l,j}}) = P_{i+1}(r_{\lambda_{l,j}} \in \{\tilde{\lambda}_{l,j}(\beta + d\beta)\})$. When $T \to \infty$, the cluster center pairs become arbitrarily close to each other, i.e., $\angle(\psi_{l,j}, \lambda_{l,j}) \overset{P}{\to} 0$, by assumption. Therefore, the symmetric difference between the sets $D_{\psi_{l,j}}$ and $D_{\lambda_{l,j}}$ approaches the null set, which implies that $P_i(y \in D_{\psi_{l,j}}) - P_i(y \in D_{\lambda_{l,j}}) \overset{P}{\to} 0$. This implies

$$P_{i+1}(r_{\psi_{l,j}} \in \{\tilde{\psi}_{l,j}(\beta + d\beta)\}) - P_{i+1}(r_{\lambda_{l,j}} \in \{\tilde{\psi}_{l,j}(\beta + d\beta)\}) \to 0, \quad (27)$$

for an arbitrary $\beta$ and $d\beta$, as $T \to \infty$. This means that the residuals of $\psi_{l,j}$ and $\lambda_{l,j}$ belong to a unique but identical probability space. Since we proved this for an arbitrary $l$ and $j$, we can say that the residuals of clusterings $\Phi_i$ and $\Lambda_i$ belong to an identical probability space given by $(\gamma_{i+1}, \Sigma_{i+1}, P_{i+1})$. \hfill \Box

**Theorem 4.3:** Given that the training vectors for the first level are generated from the probability space
learning are stable and the MLD learning is stable as to that level. Hence all the levels of the dictionary that the training vectors for level and are residuals of the clusterings \( \Psi \) belong to a probability space \((vectors of both the dictionaries for any level \( l \) are realizations from the probability space \( Y_1, \Sigma_1, P_1 \)). Lemma 4.1 proved that the residuals of level \( l \) \( \Lambda \) belong to an identical probability space \( Y_1, \Sigma_1, P_1 \) as we supply the algorithm with training vectors from both dictionaries for any level \( l \). However, when learning the dictionary, this is true only for the first level, as we supply the algorithm with training vectors from the probability space \((Y_1, \Sigma_1, P_1)\).

We note that the training vectors for level \( l + 1 \) are residuals of the clusterings \( \Psi_1 \) and \( \Lambda_1 \). Lemma 4.2 showed that the residuals of level \( l \) for both the clusterings belong to an identical probability space \((Y_{l+1}, \Sigma_{l+1}, P_{l+1})\), given that the training vectors of level \( l \) are realizations from the probability space \((Y_1, \Sigma_1, P_1)\) and \( T \to \infty \). By induction, this along with the fact that the training vectors for level 1 belong to the same probability space \((Y_1, \Sigma_1, P_1)\), shows that all the training vectors of both the dictionaries for any level \( l \) indeed belong to a probability space \((Y_1, \Sigma_1, P_1)\) corresponding to that level. Hence all the levels of the dictionary learning are stable and the MLD learning is stable as a whole.

If there is a unique minimizer to the clustering objective in all levels of MLD learning, then the MLD algorithm is stable even for completely disjoint training sets, as \( T \to \infty \). However, if there are multiple minimizers in at least one level, the algorithm is stable only with respect to a change of \( o(\sqrt{T}) \) training samples between the two clusterings. In particular, a change in \( \Omega(\sqrt{T}) \) samples makes the algorithm unstable.

### 4.4 Generalization Analysis

Since our learning algorithm consists of multiple levels, and cannot be expressed as an ERM on a whole, the algorithm can be said to generalize asymptotically if the sum of empirical errors for all levels converge to the sum of expected errors, as the number of training samples \( T \to \infty \). This can be expressed as

\[
\left| \frac{1}{T} \sum_{i=1}^{L} \sum_{t=1}^{T} g_{\Psi_i}(y_{1,i}) - \sum_{i=1}^{L} \mathbb{E}_{P_1}[g_{\Psi_i}] \right| \overset{P}{\to} 0, \tag{28}
\]

where the training samples for level \( l \) given by \( \{y_{1,i}\}_{i=1}^{T} \) are obtained from the probability space \((Y_1, \Sigma_1, P_1)\). When (28) holds and the learning algorithm generalizes, it can be seen that the expected error for test data which is drawn from the same probability space as that of the training data, is close to the average empirical error. Therefore, when the cluster centers for each level are obtained by minimizing the empirical error, we are guaranteed that the expected test error will also be small.

In order to show that (28) holds, we use the fact that each level of MLD learning is obtained using K-lines clustering. Hence, from (7), the average empirical distortion in each level converges to the expected distortion as \( T \to \infty \),

\[
\left| \frac{1}{T} \sum_{i=1}^{T} g_{\Psi_i}(y_{1,i}) - \mathbb{E}_{P_1}[g_{\Psi_i}] \right| \overset{P}{\to} 0. \tag{29}
\]

The validity of the condition in (28) follows directly from the triangle inequality,

\[
\left| \frac{1}{T} \sum_{i=1}^{L} \sum_{t=1}^{T} g_{\Psi_i}(y_{1,i}) - \sum_{i=1}^{L} \mathbb{E}_{P_1}[g_{\Psi_i}] \right| \\
\leq \sum_{i=1}^{L} \left| \frac{1}{T} \sum_{t=1}^{T} g_{\Psi_i}(y_{1,i}) - \mathbb{E}_{P_1}[g_{\Psi_i}] \right|. \tag{30}
\]

If the M-OMP coding scheme is used for test data, and the training and test data for level 1 are obtained from the probability space \((Y_1, \Sigma_1, P_1)\), the probability space for both training and test data in level 1 will be \((Y_1, \Sigma_1, P_1)\). This is because, both the M-OMP coding scheme and the MLD learning associate the data to a dictionary atom using the maximum absolute correlation measure and create a residual that is orthogonal to the atoms chosen so far. Hence, the assumption that training and test data are drawn from the same probability space in all levels hold and the expected test error will be similar to the average empirical training error.
5 Results and Discussion

Simulation results presented in this section were obtained with dictionaries learned using 50,000 randomly chosen patches of size $8 \times 8$, extracted from the grayscale images in the training set of the Berkeley segmentation dataset (BSDS) \cite{BSDS}, unless stated otherwise. Note that we did not observe any significant performance variations when we increased the size of the training set further. The mean value of each training patch was removed as a preprocessing step. For MLD learning, the number of atoms was fixed at 16 per level and the number of levels was fixed at 16, which leads to a total of 256 atoms. For comparison, a global K-SVD dictionary of size $64 \times 256$ atoms was learned using the MATLAB toolbox available online \cite{K-SVD}. In this case, the desired sparsity, which refers to the number of non-zero coefficients ($S$), was fixed at 16. Initial dictionary atoms for the K-SVD algorithm and for each level of MLD learning were obtained using the K-means. The test dataset used for evaluating the performance of the two dictionaries consisted of 120,000 non-overlapping $8 \times 8$ patches extracted from images in the BSDS test dataset, and a set of standard images (Barbara, Lena, Boat, Man).

Figure 4 illustrates the multilevel dictionary designed using the algorithm in Table 1. Note that no noise was added to the image patches during learning. As it can be observed, the learned MLD contains geometric patterns in the first few levels, stochastic textures in the last few levels and a combination of both in the middle levels. The representation energy, $\|\mathbf{A}^l\|_2^2$, captured across all the levels in MLD is shown in Figure 5(a), where the energy hierarchy in learning can be clearly seen.

Given a multilevel dictionary, an $S$-sparse representation for a test sample can be evaluated using the M-OMP or the RM-OMP procedures described in Section 3.4. For the learned K-SVD and multilevel dictionaries, we computed the sparse codes for patches from the BSDS test dataset, by varying the desired sparsity. The illustration in Figure 5(b) shows the mean squared error (MSE) of the representation as a function of the number of non-zero coefficients. For the case of MLD, the results obtained using both the M-OMP and the RM-OMP schemes are shown. The OMP algorithm was employed to compute the sparse coefficients with the K-SVD dictionary. It can be observed that the MSE obtained using the M-OMP procedure is higher in all cases of sparsity, when compared to RM-OMP. Since the RM-OMP procedure considers dictionary atoms from the neighboring levels when computing a coefficient, it results in an improved performance. When compared to K-SVD, multilevel dictionaries lead to a more accurate reconstruction when the sparsity level $S \geq 4$, which is the range typically used in several applications. Note that, although we proved in Section 4.4 that M-OMP procedure leads to test error that is similar to training error, we observe that RM-OMP generalizes better when training the dictionary with finite number of samples. Hence, the results for MLD reported in Sections 5.1 and 5.2 were obtained using only the RM-OMP algorithm.

5.1 Stability and Generalization

In Section 4, we showed that the proposed multilevel dictionary learning algorithm is asymptotically stable and can generalize well to test data obtained from the same probability space as that of the training data. Both these characteristics are crucial for building effective global dictionaries to model natural image patches. Although it is not possible to demonstrate the asymptotic behavior experimentally, we study the changes in the behavior of the learning algorithm with increase in the number of samples used for training.

In order to illustrate the stability characteristics of
MLD learning, we setup an experiment where we consider a multilevel dictionary of 4 levels, with 8 atoms in each level. We extracted patches of size $4 \times 4$ from the BSDS training images and trained multilevel dictionaries using different number of training patches $T$. As we showed in Section 4, asymptotic stability is guaranteed when the training set is changed by not more than $o(\sqrt{T})$ samples. In other words, the inferred dictionary atoms will not vary significantly, if this condition is satisfied.

We fixed the size of the training set at different values $T = \{1000, 5000, 10000, 50000, 100000\}$ and learned an initial set of dictionaries using the proposed algorithm. The second set of dictionaries were obtained by replacing different number of samples from the original training set. For each case of $T$, the number of replaced samples was varied between 100 and $T$. For example, when $T = 10000$, the number of replaced training samples were 100, 1000, 5000, and 10000. The amount of change between the initial and the second set of dictionaries was quantified using the minimum Frobenius norm of their difference with respect to permutations of their columns and sign changes. In Figure 6(a), we plot this quantity for different values of $T$ as a function of the number of samples replaced in the training set. For each case of $T$, the difference between the dictionaries increases as we increase the replaced number of training samples. Furthermore, for a fixed number of replaced samples (say 100), the difference reduces with the increase in the number of training samples, since it becomes closer to asymptotic behavior.

Generalization of a dictionary learning algorithm guarantees a small approximation error for a test data sample, if the training samples are well approximated by the dictionary. In order to demonstrate the generalization characteristics of MLD learning, we designed dictionaries using different number of training image patches, of size $8 \times 8$, from the BSDS training dataset and evaluated the sparse codes for patches in the BSDS test dataset. The dictionaries were of size $64 \times 256$ with 16 atoms per level. Figure 6(b) shows the approximation error (MSE) for both the training and test datasets obtained using multilevel dictionaries. Furthermore, the corresponding MSE for the case of similarly designed K-SVD dictionaries are included for comparison. In all cases, the sparsity in training and testing was fixed at $S = 16$. As it can be observed, with MLD, the difference between the MSE for training and test data is small even for a small training set. However, the K-SVD dictionaries resulted in much higher MSE difference for a small training set, although the MSE with training data is similar for both MLD and KSVD. Note that, in both cases, the approximation error for the test data reduces with the increase in the size of the training set.

5.2 Compressed Recovery

In this section, we demonstrate the performance of multilevel dictionaries in compressed sensing of images using the BSDS dataset and a set of standard images, with random and optimized measurement systems. Sensing and recovery were performed on a patch-by-patch basis, on non-overlapping patches of size $8 \times 8$. MLD and K-SVD dictionaries were learned with 50000 patches using the parameters mentioned in the previous simulation. The measurement process can be expressed as

$$x = \Phi \Psi a + \eta,$$

where $\Psi$ is the dictionary (MLD or K-SVD in our case), $\Phi$ is the measurement or projection matrix, $\eta$ is the
Fig. 7. Compressed recovery performance of MLD and K-SVD dictionaries with BSDS test dataset for various measurement noise levels and number of measurements \( (N) \), (a) Using random measurement matrices, (b) Using optimized measurement matrices [40].

AWGN vector added to the measurement process, \( x \) is the output of the measurement process and \( a \) is the sparse coefficient vector such that \( y = \Psi a \). The size of the data vector \( y \) is \( M \times 1 \), that of \( \Psi \) is \( M \times K \), that of the measurement matrix \( \Phi \) is \( N \times M \), where \( N < M \), and that of the measured vector \( x \) is \( N \times 1 \). Besides performing sensing with random measurement matrices, that are used extensively in compressed sensing, we also used measurement matrices optimized to dictionaries that can provide much better performance than random matrices [40], [57]. We adopt the strategy proposed in [40], and compute optimized measurement matrices for the pre-existing K-SVD and multilevel dictionaries.

The entries in the random measurement matrices were independent realizations from a standard normal distribution. For the case of random projections, we present average results from 100 trial runs, each time with a different measurement matrix. Figure 7(a) illustrates the compressed recovery performance of the BSDS test dataset using random measurements and Figure 7(b) shows the results obtained using optimized measurements at different measurement noise levels. We observe for most cases in random projections and all cases in optimized projections that, reconstruction using RM-OMP with the multilevel dictionary achieves lesser MSE when the number of measurements is small. This is also evident from the results with standard images shown in Figure 8, where both K-SVD and MLD dictionaries were used with optimized projections. Small regions of the images are zoomed in to demonstrate the improved reconstruction quality obtained using the MLD. The proposed algorithm provides a PSNR improvement of 1.84 dB and 2.17 dB for the Man and Lena images respectively, when the number of measurements \( N = 16 \), and the measurement SNR = 15 dB.

In order to demonstrate that the MLD learning generalizes well under incomplete data conditions, we performed compressed recovery of different standard images using \( N = 16 \) measurements, fixing the measurement SNR at 15 dB. Similar to the generalization experiment in Section 5.1, we learned different dictionaries by varying the number of training patches, obtained from the BSDS dataset, between 5000 and 50000. We employed both random and optimized measurements for recovery and evaluated the PSNR of the recovered images in all cases. Figure 9, shows the PSNR obtained for Barbara, Boat, Lena and Man using the different dictionaries. The proposed dictionaries provided an improved recovery performance for both random and optimized measurements, even for small training sets.

6 Conclusions

We presented a multilevel learning algorithm to design global dictionaries that exploit the redundancy and energy hierarchy found in natural image patches. The proposed algorithm employs K-lines clustering to learn atoms for each level of the dictionary. We showed that the algorithm converges for a sufficient number of levels and that energy hierarchy is exhibited for a sufficient number of atoms per level. We also showed that the dictionaries learned using different sets of training data, from the same probability space, are arbitrarily close to each other, for a sufficiently large number of data samples. Furthermore, we proved the asymptotic generalization characteristics, and demonstrated the stability and generalization behavior using simulations. Simulation results for compressed sensing clearly demonstrated that the learned multilevel dictionaries provide superior performance when compared to K-SVD dictionaries. Possible future work includes analyzing the sample complexity of multilevel learning and developing an online learning procedure for large training sets.

References

Fig. 8. Compressed recovery of images using optimized measurements ($N = 16$, SNR of measurement process $= 15$ dB). Only small portions of the images are displayed for visualizing the differences in the quality of recovery.

(a) Man image - K-SVD recovery (PSNR = 28.47 dB)  (b) Man image - MLD recovery (PSNR = 30.31 dB)

(c) Lena image - K-SVD recovery (PSNR = 30.43 dB)  (d) Lena image - MLD recovery (PSNR = 32.6 dB)
Fig. 9. Compressed recovery performance of MLD and K-SVD dictionaries on standard images ($N = 16$ and measurement SNR = 15 dB). In each case, we show the PSNR (dB) obtained using both random measurements (RP) and optimized measurements (OP).