Sample Complexity of Dictionary Learning on Stationary Mixing Data

Li-juan Guo
Changsha Aeronautical Vocational and Technical College;
Changsha Hunan 410124

Abstract:
Dictionary learning is important for many pattern recognition and image processing. Some known jobs focus on the sample complexity of dictionary learning on the independent data for characterizing the performance of a learned dictionary. In this paper, the sample complexity of dictionary learning on the stationary mixing input sequence is considered because the stationary mixing input sequence appears in many applications. By discussing the sample complexity of learning dictionary on the $\beta$-mixing sequence, it has been shown that the better performance of a learned dictionary is a result of controlling the size of a learned dictionary, which means too large size of a learned dictionary will decrease the generalization of the learned dictionary.

Keywords:
Dictionary learning; Sample complexity; Stationary Mixing; $\beta$-mixing

1. Introduction
Dictionary learning which is motivated by the success of compressed sensing [1, 2, 3] is a popular tool for analyzing various data. Given a training set

$$s_n = \{x_1, x_2, \cdots, x_n\} \subset \mathbb{R}^d,$$

(1)

a set of sparse features $\Lambda$ could be generated by the following dictionary learning problem

$$(D^*, Y^*) = \arg \min_{D, \Lambda} \| X - DY \|^2_F + \gamma \| Y \|_1,$$

(2)

where the matrix $D^*$ is the learned dictionary, $Y^*$ is the set of sparse representations of the given training samples corresponding to the learned dictionary $D^*$, the matrix $X$ consists of all the training samples, $\gamma > 0$ is a regularized parameter, $\| \cdot \|_F$ means Frobenius norm, and $\| \cdot \|_1$ means $\ell_1$ norm. By using the dictionary $D^*$, the priori information contained in the training samples are learned. Meanwhile, the corresponding sparse representation of a given sample hints the most critical features related with the sample. Moreover, the sparsity of the representation vector $y_i \in Y$ emphasizes could be understood as a adaptive selection of optimal local features. Therefore, dictionary learning techniques could be thought as a method to combine the advantages of global features and efficient local representation.

Dictionary learning have been wildly applied in many real applications such as image denoising[4], face recognition[5], hyperspectral classification[6] and single-image superresolution [7]. For image denoising, the process of dictionary learning is used to find the basic image structures whose linear combinations could be used to reconstruct clean images. By using sparse representations of noised image patches, the most representable basic image structures are adaptively selected from the learned dictionary $D^*$, which means finding an optimal local bases from all of probable sets of local bases for reconstruct a clean image. Because the optimal local bases are used, the energy of an image could be well preserved even though part energy will be filtered when noised information is removed. Similar situations could be found in the other applications such as [5, 8, 7] though these applications adopt a supervised model to learn dictionary.

It could be found that the success of using dictionary
and sparse representations roots in the redundancy of the learned dictionary. Because the number of columns of the matrix \( D \) is generally much larger than the dimensional number \( d \) of \( \mathbb{R}^d \), a huge number of sets of bases could be generated from \( D \). It could be imagined that different set of bases will be good at representing different samples. Following the line of the idea, it seems that a large size of dictionary which could offer more chance to match the needs of representing some samples is better for applications. However, it contradicts another intuition that a learning machine should balance its complexity and training error. Thus, a problem how the size of dictionary effects the performance of a learned dictionary exists.

Some jobs in the field of machine learning [9] and statistical learning theory [10] devoted to this interesting question. For example, Maurer and Pontil [11] have shown generalization error bounds on the square representation error of a learned dictionary. Following their jobs, Vainsencher, Mannor and Bruckstein [12] offered some new results on the generalization error measured by \( \ell_1 \)-norm. All of these results have shown that too large size of dictionary will decrease the performance of a learned dictionary.

It should be noticed that all of these mentioned results on the performance of a learned dictionary are considered on the assumption that all training samples are independent and identically distributed (i.i.d). However, many real applications such as market prediction, system diagnosis and speech recognition are not i.i.d processes [13]. Meanwhile, some theoretical results [14, 15, 16] have shown that the performance of a learning algorithm will be different when the training samples are not i.i.d. Thus, we focus on the problem of estimating the generalization bounds of a learned dictionary when the training samples are depended.

According to our theoretical results, the performance of a learned dictionary is related with the size of this dictionary when a \( \beta \)-mixing training sequence is considered. This result shows that too large size of a dictionary may decline the generalization of a learned dictionary when dependent training samples are used, which coincides with the situations of learning dictionary with an independent training set.

2. Preliminaries

2.1. \( \beta \)-mixing sequence

\( \beta \)-mixing sequence is a kind of strictly stationary sequence. Denote a strictly stationary sequence of random variables as \( \{ z_i \}_{i \in \mathbb{Z}} \) in which each variables follow the same distribution \( \rho \) on \( z \subseteq \mathbb{R}^d \). The \( \sigma \)-algebras generated by the random variables \( \{ z_i \}_{i \leq t} \) and \( \{ z_i \}_{i \geq k} \) are respectively denoted as

\[
\sigma_i = \sigma(z_1, z_2, \cdots, z_t),
\]

\[
\sigma_k = \sigma(z_k, z_{k+1}, \cdots).
\]

The following \( \beta \)-mixing coefficients characterize how close to independent a stationary sequence \( \{ z_i \}_{i \in \mathbb{Z}} \) is.

**Definition 1** For any sequence \( \{ z_i \}_{i \in \mathbb{Z}} \), the \( \beta \)-mixing coefficient is defined by

\[
\beta(n) = \sup_k \mathbb{E} \sup \{ \| P(A | \sigma_1^k) - P(A) \| : A \in \sigma_{k+n}^\infty \},
\]

where the expectation is taken with respect to \( \sigma_1^k \).

**Definition 2** A sequence is called \( \beta \)-mixing, if its \( \beta \)-coefficients satisfy
\[ \lim_{n \to \infty} \beta(n) = 0. \] (5)

Moreover, this sequence is called exponentially strongly \( \beta \)-mixing, if its \( \beta \)-coefficients satisfy
\[ \beta(n) \leq u_\epsilon e^{-v_n n^\epsilon}, n \geq 1, \] (6)
and it is algebraically strongly \( \beta \)-mixing, if
\[ \beta(n) \leq u_\epsilon n^{-\epsilon}, n \geq 1. \] (7)

### 2.2. Blocking decomposition technique

The purpose of the blocking decomposition technique is to represent approximatively a mixing sequence with an independent block sequence. Precisely, a mixing sequence \( Z_n = \{z_1, z_2, \cdots, z_n\} \) is divided into \( 2\mu_n \) blocks which consist of \( b_n \) members of \( Z_n \). If \( b_n \) is large enough, these blocks are dependent very weakly. If \( b_n \) is small enough, the odd-numbered and even-numbered blocks will share similar distribution to the original mixing sequence. By balancing the size of these blocks \( b_n \), the relation between expectation and average value could be bounded by the mixing coefficients.

**Lemma 1** ([17]) Let the distributions of \( Z_{b_n} \) and \( \Xi_{b_n} \) be \( Q \) and \( \tilde{Q} \) respectively where \( Z_{b_n} \) is the odd blocks of the original mixing sequence \( Z_n \) and \( \Xi_{b_n} \) is an independent blocks which shares the same distributions as \( Z_{b_n} \). For any measurable function \( f \) on \( \mathcal{R}^{2\mu_n} \) with bound \( M \),
\[ |Qf(Z_{b_n}) - \tilde{Q}f(\Xi_{b_n})| \leq M(\mu_n - 1)\beta(b_n). \] (8)

**Lemma 2** ([17]) Suppose a sequence \( \{z_i\}_{i \geq 1} \) is a stationary \( \beta \)-mixing sequence with the mixing coefficients \( \beta(n) \). Then for any uniformly bounded class \( G_n \) of measurable functions, the following holds for any \( \epsilon_n > 0 \)
\[ P(\sup_{g \in G_n} 1 \leq \frac{1}{n} \sum_{i=1}^{n} g(z_i) - E g(z_i) \geq \epsilon_n) \]
\[ \leq 2P(\sup_{g \in G_n} \frac{1}{n} \sum_{i=1}^{b_n} V_{\Xi(O_i)}(g) \geq \frac{\epsilon}{3} + 4\mu_n \beta(b_n)), \] (9)
where \( V_{\Xi(O_i)} = \sum_{j \in O_i} (g(\xi_j) - E g(\xi_j)), \Xi(O_i) \) is the i-thblock of \( \Xi_{b_n} \), \( O_i \) is the set of indexes of the members of \( \Xi(O_i) \).

### 2.3. McDiarmid’s inequalities

**Lemma 3** ([18]) Let \( \xi_1, \xi_2, \cdots, \xi_n \) be independent random variables taking values in a set \( A \), and assume that \( f : A^n \to \mathbb{R} \) satisfies for \( 1 \leq i \leq n \),
\[ \sup_{\xi, \tilde{\xi}} |f(\xi_1, \xi_2, \cdots, \xi_n) - f(\tilde{\xi}_1, \xi_2, \cdots, \xi_n)| \leq c_i \]

Then for all \( \epsilon > 0 \),
\[ P(f(\xi_1, \xi_2, \cdots, \xi_n) - Ef(\xi_1, \xi_2, \cdots, \xi_n) \geq \epsilon) \]
\[ \leq \exp\left( \frac{-2\epsilon^2}{\sum_{i=1}^{n} c_i^2} \right) \] (10)

### 2.4. Bound of Radamacher complexity

**Lemma 4** (Proposition 3.2, [11]) Suppose that the probability measure \( \mu \) is supported on the unit ball
of \( H \), that \( \{ e_i \mid i = 1, 2, \cdots, K \} \) is an orthonormal basis of \( \mathfrak{R}^K \) and that \( \Gamma \) is a class of linear operators \( T : \mathfrak{R}^K \mapsto H \) with \( \| Te_i \| \leq c \) for \( 1 \leq i \leq K \), with \( c \geq 1 \). Let \( Y \) be a nonempty closed subset of the unit ball in \( \mathfrak{R}^K \) and \( F_\gamma = \{ x \in H \mapsto \min_{y \in Y} \| x - Ty \|_2 \mid T \in \Gamma \} \). Then

\[
\text{Re}(F_\gamma, \mu) \leq 6c^2 K^2 \frac{\sqrt{\pi}}{\sqrt{m}},
\]

where \( m \) is the number of training samples.

### 3. Main results

Given a \( \beta \)-mixing sequence \( \{ x_i \mid i = 1, 2, \cdots, n \} \), we consider the generalization of a learned dictionary

\[
\hat{D} = \arg\min_{D \in \mathcal{D}} L_n(D)
\]

\[
= \arg\min_{D \in \mathcal{D}_K} \frac{1}{n} \sum_{i=1}^n \| X - DY \|_F^2 + \gamma \| Y \|_1,
\]

where \( \mathcal{D}_K \subset \mathfrak{R}^{d \times K} \) is a set of possible dictionary set,

\[
\mathcal{Y}_M = \{ y \in \mathfrak{R}^{K \times n} \mid \| Y \|_1 \leq M \}
\]

is a set of possible sparse coding vectors.

**Theorem 1** Let the training data \( \{ x_i \mid \| x_i \|_2 \leq 1 \}_{i=1}^n \) comes from a existed but unknown stationary \( \beta \)-mixing sequence with \( \beta \)-coefficients \( \beta(m), m \in \mathcal{N} \). Then, for any \( D \in \mathcal{D}_K \) where each elements \( d_i \in D \) satisfy \( \| d_i \|_2 \leq 1 \) and \( \delta > 0 \), it holds with probability at least \( 1 - (2\delta + 4\mu_n \beta(b_n)) \) that

\[
L(D) = L_n(D) \leq \frac{33c^2 K^2 + (1 + 2\sqrt{KM})}{\sqrt{\mu_n}},
\]

where \( L(D) = EL_n(D) \), and \( c \geq 1 \).

### 3.1. Proof of Theorem 1

Let \( U = \sup_{D \in \mathcal{D}_K} | L(D) - L_n(D) | \). By using the blocking decomposition technique of Lemma 2, there exists an independent block sequence \( \Xi_{b_n} = \{ \Xi(O_j) \}_{j=1}^{\mu_n} \) which satisfies the inequality

\[
P \{ U \geq \epsilon \} 
\leq 2P \left\{ \sup_{D \in \mathcal{D}_K} \left| \frac{1}{n} \sum_{j=1}^{\mu_n} V_{\Xi(O_j)}(D) \right| \geq \frac{1}{3} \epsilon \right\} + 4\mu_n \beta(b_n),
\]

where \( V_{\Xi(O_j)}(D) = \sum_{i \in O_j} \| x_i - Dy_i \|_2^2 + \gamma \| y_i \|_1 - E(\| x_i - Dy_i \|_2^2 + \gamma \| y_i \|_1 ) \), \( x_i \in \Xi(O_j) \).
Because of the independency of the block sequence 
\{ \Xi(O_j) \}_{j=1}^{n_b} \), the probability

\[ P \left\{ \sup_{D \in \mathcal{D}_K} \left| \frac{1}{n} \sum_{j=1}^{n_b} \left( \sum_{j=1}^{n_b} V \Xi(O_j)(D) \right) \right| \geq \frac{1}{3} \varepsilon \right\} \]

could be bounded according to Lemma 3.

To apply Lemma 3, it should firstly estimate the error bounds of changing one sample

\[ | F(\Xi_{b_n}) - F(\Xi_{b_{n_j}}) | \]

\[ \leq \frac{1}{n} \left( \sup_{D \in \mathcal{D}_K} \left| \sum_j V \Xi(O_j) - \left( \sum_j V \Xi(O_j) \right) \right| \right) \]

\[ \leq \frac{1}{n} \sup_{D \in \mathcal{D}_K} \left| \sum_j V \Xi(O_j) - V \Xi(O_j) \right| \]

\[ \leq \frac{2}{n} \sup_{D \in \mathcal{D}_K} \left( \max \left\{ \| X_{O_j} \|_F^2, \| X_{O_j} \|_F^2 \right\} \right) \]

\[ + \left\| X_{O_j} - X_{O_j} \right\|_F^2, \| Y \|_1 \left|, (15) \right. \]

where \( \Xi_{b_{n_j}} = \{ \Xi(O_1), \cdots, \Xi(O_j), \cdots, \Xi(O_{n_b}) \} \).

\( \Xi(O_j) \) is an independent identically distributed copy of

\( \Xi(O_j) \in \Xi_{b_n} \), and

\[ F(\Xi_{b_n}) = \sup_{D \in \mathcal{D}_K} \left| \frac{1}{n} \sum_{j=1}^{n_b} V \Xi(O_j)(D) \right| \]

Noticed the conditions in Theorem 1 including

\[ \| d_i \|_2 \leq 1, \| x_i \|_2 \leq 1, \] \[ \max \left\{ \| X_{O_j} \|_F^2, \| X_{O_j} \|_F^2 \right\} \leq \| X \|_F^2 \leq b_n \]

\[ \| X_{O_j} - X_{O_j} \|_F \leq 2, \| X \|_F \leq 2b_n \]

\[ \| D \|_F \leq \| D \|_F \leq \sqrt{K} \]

\[ \| Y \|_1 \leq M \]. \quad (16) \]

Combining the equations (15) and (16), there exists

\[ | F(\Xi_{b_n}) - F(\Xi_{b_{n,j}}) | \leq \frac{2b_n}{n} (1 + 2\sqrt{KM}) \quad (17) \]

According to Lemma 3 and the bound (17), it holds that

\[ P \{ | F(\Xi_{b_n}) - E \Xi_{b_n} | \geq \varepsilon \} \leq \exp \left( \frac{-2\varepsilon^2}{\mu_n \frac{2b_n(1 + 2\sqrt{KM})}{n}} \right) \]

\[ \leq \exp \left( \frac{-n\varepsilon^2}{2b_n C_{K,M}} \right), \quad (18) \]

where \( C_{K,M} = (1 + 2\sqrt{KM})^2 \).

Let \( \delta = \exp \left( \frac{-n\varepsilon^2}{2b_n C_{K,M}} \right) \), it holds that

\[ F(\Xi_{b_n}) \leq EF(\Xi_{b_n}) + \sqrt{\frac{-2b_n C_{K,M} \ln \delta}{n}} \cdot (19) \]

Here, \( EF(\Xi_{b_n}) \) could be bounded by Lemma 4.

According to the conditions that \( \| d_i \|_2 \leq 1 \) and \( \| Y \|_1 \leq M \), there exists \( c \geq 1 \) satisfies that

\[ \| D e_j \| \leq c \). Thus, the following bound could be generated by using Lemma 4.

\[ EF(\Xi_{b_n}) \leq \mathcal{R}(D_K) \leq 6c^2 K^2 \sqrt{\frac{n}{\mu_n} \pi} \cdot (20) \]

Combining Eqs. (19) and (20), it holds with the probability at least \( 1 - \delta \) that

\[ F(\Xi_{b_n}) \leq 6c^2 K^2 \sqrt{\frac{n}{\mu_n} \pi} + \sqrt{\frac{-2b_n C_{K,M} \ln \delta}{n}} \]

\[ \leq \frac{11c^2 K^2 + (1 + 2\sqrt{KM})}{\sqrt{\mu_n}} \cdot (21) \]

\

316
Therefore, let $\varepsilon = \frac{33c^2K^2}{\mu_n} + (1 + 2\sqrt{KM})$, it holds that

$$P\{U \geq \frac{33c^2K^2}{\mu_n} + (1 + 2\sqrt{KM})\} \leq 2\delta + 4\mu_n\beta(b_n).$$ (22)

The proof is complete.

**Remark 1** According to the bound (13), it could be found that the upper bound of generalization of a learned dictionary is directly related with its size $K$. When $K$ is too large, the difference between the expected error $L(D)$ and its empirical error $L_n(D)$ is unnecessarily small. Therefore, a small size $K$ is welcome when a dictionary is learned with a $\beta$-mixing sequence.

**Remark 2** Noticed the denominator of the right hand of Eq.(13), it is clear that the rate of convergence is determined by the number of blocks. This result hints that the performance of learning a dictionary will get better when more dependent training samples are available. Therefore, there exists a problem if it is a good idea to learn a dictionary with a small dependent training set.

**4. Conclusion**

In this theoretical research, we analyze the topic of sample complexity of dictionary learning on stationary mixing data from the perspective of the machine learning and pattern recognition. Some known jobs focus on the sample complexity of dictionary learning on the independent data for characterizing the performance of a learned dictionary. Following the line of the idea, it seems that a large size of dictionary which could offer more chance to match the needs of representing some samples is better for applications. However, it contradicts another intuition that a learning machine should balance its complexity and training error. Our proposed algorithm solves the mentioned challenges successfully. In the future, more corresponding theoretical analysis and numerical simulation will be conducted to optimize the current approach.

**References**


[8] Lin He, Yuanqing Li, Xiaoxin Li, and Wei Wu, “Spectral-spatial classification of hyperspectral


