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 pa-per, 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 β-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; β-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

$$s_n = \{x_1, x_2, \cdots, x_n\} \subset \Re^d \quad , \tag{1}$$

a set of sparse features Λ could be generated by the following dictionary learning problem

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

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 ℓ_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 $\Re^d$,a huge number of sets of bases could be generated from D^*.It could be imaged 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 ℓ_1 -norm. All of these results have shown that too large size of dictionary will decease 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 β -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. β -mixing sequence

β -mixing sequence is a kind of strictly stationary sequence.Denote a strictly stationary sequence of random variables as $\{z_i\}_{i\geq 1}$ in which each variables follow the same distribution ρ on $z \subseteq \Re^d$. The σ -algebras generated by the random variables $\{z_i\}_{1\leq i\leq l}$ and $\{z_i\}_{i\geq k}$ are respectively denoted as

$$\begin{aligned}\sigma_1^l &= \sigma(z_1, z_2, \cdots, z_l),\\ \sigma_k^\infty &= \sigma(z_k, z_{k+1}, \cdots).\end{aligned} \tag{3}$$

The following β -mixing coefficients characterize how close to independent a stationary sequence $\{z_i\}_{i\geq 1}$ is.

Definition 1 For any sequence $\{z_i\}_{i\geq 1}$, the β -mixing coefficient is defined by

$$\beta(n) = \sup_k E \sup\{|P(A\,|\,\sigma_1^k) - P(A)|\,|\, A \in \sigma_{k+n}^\infty\}, \tag{4}$$

where the expectation is taken with respect to σ_1^k .

Definition 2 A sequence is called β -mixing, if its β -coefficients satisfy

$$\lim_{n\to\infty}\beta(n)=0. \tag{5}$$

Moreover, this sequence is called exponentially strongly β-mixing, if its β-coefficients satisfy

$$\beta(n)\le u_e \exp(-v_e n^{r_e}), n\ge 1, \tag{6}$$

and it is algebraically strongly β-mixing, if

$$\beta(n)\le u_a n^{-r_a}, n\ge 1. \tag{7}$$

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 Ξ_{b_n} be Q and $\tilde{Q}$ respectively where Z_{b_n} is the odd blocks of the original mixing sequence Z_n and Ξ_{b_n} is an independent blocks which shares the same distributions as Z_{b_n}. For any measurable function f on $\Re^{\mu_n b_n}$ with bound M,

$$|Qf(Z_{b_n})-\tilde{Q}f(\Xi_{b_n})|\le M(\mu_n-1)\beta(b_n). \tag{8}$$

Lemma 2 ([17]) Suppose a sequence $\{z_i\}_{i\ge1}$ is a stationary β-mixing sequence with the mixing coefficients $\beta(n)$. Then for any uniformly bounded class G n of measurable functions, the following holds for any $\varepsilon_n>0$

$$P(\sup_g \in G_n\,|\frac{1}{n}\sum_{i=1}^{n}g(z_i)-Eg(z_i)|\ge\varepsilon_n)$$

$$\le 2P(\sup_{g\in G_n}|\frac{1}{n}\sum_{i=1}^{\mu_n}V_{\Xi(\mathrm{O})_i}(g)|\ge\frac{\varepsilon}{3}+4\mu_n\beta(b_n), \tag{9}$$

where $V_{\Xi(\mathrm{O}_i)}=\sum_{j\in \mathrm{O}_i}(g(\xi_j)-Eg(\xi_j))$, $\Xi(\mathrm{O}_i)$ is the i-thblock of Ξ_{b_n}, O_i is the set of indexes of the members of $\Xi(\mathrm{O}_i)$.

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 $f:A^n\mapsto\Re$ satisfies for $1\le i\le n$,

$$\sup_{\xi_i,(\hat{\xi})_i\in A}|f(\xi_1,\xi_2,\cdots,\xi_n)-f(\xi_1,\xi_2,\cdots,\hat{\xi}_i,\cdots,\xi_n)|\le c_i$$

.

Then for all ϵ > 0,

$$P(f(\xi_1,\xi_2,\cdots,\xi_n)-Ef(\xi_1,\xi_2,\cdots,\xi_n)\ge\varepsilon)$$

$$\le\exp(\frac{-2\varepsilon^2}{\sum_{i=1}^{n}c_i^2}) \tag{10}$$

2.4. Bound of Radamacher complexity

Lemma 4 (Proposition 3.2, [11]) Suppose that the probability measure μ is supported on the unit ball of H, that $\{e_i \mid i=1,2,\cdots,K\}$ is an orthonormal basis of $\Re^K$ and that Γ is a class of linear operators $T:\Re^K \mapsto H$ with $\| Te_i \| \le c$ for $1 \le i \le K$, with $c \ge 1$. Let Y be a nonempty closed subset of the unit ball in $\Re^K$ and $F_Y = \{x \in H \mapsto \min_{y\in Y} \| x - \mathrm{T}y \|_2^2 \mid T \in \Gamma\}$. Then

$$\Re(F_Y, \mu) \le 6c^2K^2\sqrt{\frac{\pi}{m}}, \tag{11}$$

where m is the number of training samples.

3. Main results

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

$$\begin{aligned}\hat{D} &= \arg\min_{D\in\mathcal{D}} L_n(D) \\ &= \arg\min_{D\in\mathcal{D}_K} \min_{Y\in\mathcal{Y}_M} \frac{1}{n}\|X - DY\|_F^2 + \gamma\|Y\|_1,\end{aligned} \tag{12}$$

where $\mathcal{D}_K \subset \Re^{d\times K}$ is a set of possible dictionary set, $\mathcal{Y}_M = \{y \in \Re^{K\times n} \mid \| Y \|_1 \le M\}$ is a set of possible sparse coding vectors.

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

$$L(D) - L_n(D) \le \frac{33c^2K^2 + (1+2\sqrt{K}M)}{\sqrt{\mu_n}}, \tag{13}$$

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

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 $\Xi_{b_n} = \{\Xi(\mathrm{O}_j)\}_{j=1}^{\mu_n}$ which satisfies the inequality

$$\begin{aligned}&P\{U \ge \epsilon\} \\ \le\ & 2P\left\{\sup_{D\in\mathcal{D}_K} \left|\frac{1}{n}\sum_{j=1}^{\mu_n} V_{\Xi(\mathcal{O}_j)}(D)\right| \ge \frac{1}{3}\epsilon\right\} + 4\mu_n\beta(b_n),\end{aligned} \tag{14}$$

Where $V_{\Xi(\mathrm{O}_j)}(D) = \sum_{i\in\mathrm{O}_j}(\| x_i - Dy_i \|_2^2 + \gamma \| y_i \|_1 - E(\| x_1 - Dy_1 \|_2^2 + \gamma \| y_1 \|_1)), x_1 \in \Xi(\mathrm{O}_j)$.

Because of the independency of the block sequence $\{\Xi(\mathrm{O}_j)\}_{j=1}^{\mu_n}$ the probability $P\left\{\sup_{D\in\mathcal{D}_K} |\frac{1}{n}\sum_{j=1}^{\mu_n} V_{\Xi(\mathcal{O}_j)}(D)| \geq \frac{1}{3}\epsilon\right\}$ could be bounded according to Lemma 3.

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

$$|F(\Xi_{b_n})-F(\Xi_{b_n,j})|$$

$$\leq \frac{1}{n}(\sup_{D\in\mathcal{D}_K}|\sum_j V_{\Xi(\mathcal{O}_j)}| - \sup_{D\in\mathcal{D}_K}|\sum_j V_{\Xi(\mathcal{O}_j')}|)$$

$$\leq \frac{1}{n}\sup_{D\in\mathcal{D}_K}|\sum_j V_{\Xi(\mathcal{O}_j)} - V_{\Xi(\mathcal{O}_j')}|$$

$$\leq \frac{2}{n}\sup_{D\in\mathcal{D}_K}(\max\left\{\|X_{\mathcal{O}_j}\|_F^2, \|X_{\mathcal{O}_j'}\|_F^2\right\}$$

$$+\|X_{\mathrm{O}_j}-X_{\mathrm{O}_j}\|_F\|D\|_\infty\|Y\|_1), \tag{15}$$

where $\Xi_{b_b,j}=\{\Xi(\mathrm{O}_1),\cdots,\tilde{\Xi}(\mathrm{O}_j),\cdots,\Xi(\mathrm{O}_{\mu_n})\}$, $\tilde{\Xi}(\mathrm{O}_j)$ is an independent identically distributed copy of $\Xi(\mathrm{O}_j)\in\Xi_{b_n}$,and $F(\Xi_{b_n})=\sup_{D\in\mathcal{D}_K}|\frac{1}{n}\sum_{i=1}^{\mu_n}V_{\Xi(\mathcal{O}_i)}(D)|$.

Noticed the conditions in Theorem 1 including $\|d_i\|_2\leq 1, \|x_i\|_2\leq 1$, it holds that

$$\max\{\|X_{\mathrm{O}_J}\|_F^2, \|X_{\mathrm{O}_J'}\|_F^2\leq\|X\|_F^2\leq b_n$$

$$\|X_{\mathrm{O}_j}-X_{\mathrm{O}_j'}\|_F\leq 2\|X\|_F\leq 2b_n$$

$$\|D\|_\infty\leq\|D\|_F\leq\sqrt{K}$$

$$\|Y\|_1\leq M. \tag{16}$$

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

$$|F(\Xi_{b_n})-F(\Xi_{b_n,j})|\leq\frac{2b_n}{n}(1+2\sqrt{K}M) \tag{17}$$

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

$$P\{|F(\Xi_{b_n})-E\Xi_{b_n}|\geq\varepsilon\}$$

$$\leq\exp(-\frac{2\varepsilon^2}{\mu_n(\frac{2b_n(1+2\sqrt{K}M)}{n})^2}$$

$$\leq\exp(-\frac{n\varepsilon^2}{2b_nC_{K,M}}), \tag{18}$$

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

Let $\delta=\exp(-\frac{n\varepsilon^2}{2b_nC_{K,M}})$, it holds that

$$F(\Xi_{b_n})\leq EF(\Xi_{b_n})+\sqrt{\frac{-2b_nC_{K,M}\ln\delta}{n}}. \tag{19}$$

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 $\|De_j\|\leq c$. Thus, the following bound could be generated by using Lemma 4.

$$EF(\Xi_{b_n})\leq\mathcal{R}(\mathcal{D}_K)\leq 6c^2K^2\sqrt{\frac{\pi}{\mu_n}}. \tag{20}$$

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

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

$$\leq\frac{11c^2K^2+(1+2\sqrt{K}M)}{\sqrt{\mu_n}}. \tag{21}$$

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

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

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, asmallsize K is welcome when a dictionary is learned with a β-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

[1] Emmanuel Candès, Justin Romberg, and Terence Tao, “Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information,”IEEE Transactions on Information Theory, vol. 52, no.2, pp. 489–509, 2006.

[2] Emmanuel Candès and Justin Romberg, “Quantitative robust uncertainty principles and optimally sparse decompositions,” Foundations of Comput. Math., vol. 6, no. 2,pp. 227–254, 2006.

[3] Emmanuel Candès and Terence Tao, “Near optimal signal recovery from random projections: Universal encoding strategies,” IEEE Transactions on Information Theory,vol. 52, no. 12, pp. 5406–5425, 2006.

[4] M. Aharon, M. Elad, and A. Bruckstein, “K-SVD:an algorithm for designing overcomplete dictionaries for sparse representation,” IEEE Transactions on Signal Processing, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[5] John Wright, Allen Yang, Arvind Ganesh, Shankar Shastry, and Yi Ma, “Robust face recognition via sparse representation,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 2, pp. 210–227, 2009.

[6] Yi Chen, Nasser M. Nasrabadi, and Trac D. Tran, “Hyperspectral image classification using dictionary-based sparse representation,” IEEE Transactions on Geoscience and Remote Sensing, vol. 49, no. 10, pp. 3973–3985,2011.

[7] J. Yang, J. Wright, T. Huang, and Y. Ma, “Image superresolution as sparse representation of raw image patches,”in Proc. IEEE CVPR’08, Aug. 2008, pp. 1–8.

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

images via spatial translation-invariant wavelet-based sparse representation," IEEE Transactions on Geoscience and Remote Sensing, vol. 53, no. 5, pp. 2696–2712, 2015.

[9] J. Shawe-Taylor and N. Cristianini, Kernel Methods for Pattern Analysis, Cambridge Univeersity Press, Cambridge, 2004.

[10] V. Vapnik, Statistical Learning theory, John Wiley and Sons, NY, 1998.

[11] Andreas Maurer and Massimiliano Pontil, "K-dimensional coding schemes in hilbert spaces," IEEE Transactions on Information Theory, vol. 56, no. 11, pp. 5839–5846, 2010.

[12] Daniel Vainsencher, Shie Mannor, and Alfred M. Bruckstein, "The sample complexity of dictionary learning,"Journal of Machine Learning Research, vol. 12, pp.3259–3281, 2011.

[13] I. Steinwart, D. Hush, and C. Scovel, "Learning from dependent observations," Journal of Multivariate Analysis,vol. 100, no. 1, pp. 175–194, 2009.

[14] Bin Zou and Luoqing Li, "The performance bounds of learning machines based on exponetially strongly mixing sequence," Journal of Computational and Applied Mathematics, vol. 53, no. 7, pp. 1050–1058, 2007.

[15] Bin Zou, Luoqing Li, and Zongben Xu, "The generalization performance of ERM algorithm with strongly mixing observations," Machine Learning, vol. 75, no. 3, pp. 275–279, 2009.

[16] Yi Ding and Yi Tang, "Sparsity-regularized support vector machine with stationay mixing input sequence,"in Proceedings of the 2010 International Conference on Wavelet Analysis and Pattern Recognition, 2010, pp. 195–200.

[17] Bin Yu, "Rate of convergence for empirical process of stationary mixig sequences," Annals of Probability, vol.22, pp. 94–116, 1994.

[18] J. Shaw-Taylor and N. Crisitianini, Kernel Methods for Pattern Analysis, Cambridge University Press, 2014.