Dimensionality-Dependent Generalization Bounds for k-Dimensional Coding Schemes

Tongliang Liu[1]**, Dacheng Tao**[1]**, Dong Xu**[2]

[1]QCIS, University of Technology Sydney. tliang.liu@gamil.com; dacheng.tao@uts.edu.au

[2]School of Computer Engineering, Nanyang Technological University. dongxu@ntu.edu.sg

Keywords: Generalization bound, Bennett type inequality, covering number, k-dimensional coding schemes, non-negative matrix factorization, dictionary learning, sparse coding, k-means clustering and vector quantization.

Abstract

The k-dimensional coding schemes refer to a collection of methods that attempt to represent data using a set of representative k-dimensional vectors, and include non-negative matrix factorization, dictionary learning, sparse coding, k-means clustering and vector quantization as special cases. Previous generalization bounds for the reconstruction error of the k-dimensional coding schemes are mainly dimensionality-independent. A major advantage of these bounds is that they can be used to analyze the generalization error when data is mapped into an infinite- or high-dimensional feature space. However, many applications use finite-dimensional data features. Can we obtain dimensionality-dependent generalization bounds for k-dimensional coding schemes that are tighter than dimensionality-independent bounds when data is in a finite-dimensional feature space? The answer is positive. In this paper, we address this problem and derive a dimensionality-dependent generalization bound for k-dimensional coding schemes by bounding the covering number of the loss function class induced by the reconstruction error. The bound is of order $\mathcal{O}\left((mk\ln(mkn)/n)^{\lambda_n}\right)$, where m is the dimension of features, k is the number of the columns in the linear implementation of coding schemes, n is the size of sample, $\lambda_n > 0.5$ when n is finite and $\lambda_n = 0.5$ when n is infinite. We show that our bound can be tighter than previous results, because it avoids inducing the worst-case upper bound on k of the loss function. The proposed generalization bound is also applied to some specific coding schemes to demonstrate that the dimensionality-dependent bound is an indispensable complement to the dimensionality-independent generalization bounds.

1 Introduction

The k-dimensional coding schemes (Maurer & Pontil, 2010) are abstract and general descriptions of a collection of methods, all of which encode a data point $x \in \mathcal{H}$ as a representative vector $y \in \mathbb{R}^k$ by a linear map T, where $\mathcal{H}$ denotes the Hilbert space.

These coding schemes can be formulated as follows:

$$\hat{y} = \arg\min_{y \in Y} \|x - Ty\|^2,$$

where $Y \subseteq \mathbb{R}^k$ is called the *codebook* and the linear map $T \in \mathbb{R}^{m \times k}$ is called the *implementation* of the codebook. The implementation projects the codebook back to the data source space. The dimension of a data point x can be either finite or infinite. In this paper, we consider the data as having finite dimensions of features, that is $\mathcal{H} = \mathbb{R}^m$.

Each data point in $\mathcal{H}$ can be exactly or approximately reconstructed by a *code* y in the codebook. The *reconstruction error* of a data point x is defined as

$$f_T(x) = \min_{y \in Y} \|x - Ty\|^2. \tag{1}$$

The function $f_T(x)$, whose variables are x and T, is also called the *loss function*. Non-negative matrix factorization (NMF) (see, e.g., Lee and Seung, 1999; Févotte et al., 2009), dictionary learning (see, e.g., Chen et al., 1999; Ivana & Pascal, 2011), sparse coding (see, e.g., Olshausen & Field, 1996; Amiri & Haykin, 2014), k-means clustering (see, e.g., MacQueen et al., 1967; Anderberg, 1973) and vector quantization (see, e.g., Gray, 1984; Schneider et al., 2009a) are specific forms of k-dimensional coding schemes, because they share the same form of the reconstruction error as equation (1). They have achieved great successes in the fields of pattern recognition and machine learning for their superior performances on a broad spectrum of applications (see, e.g., Pehlevan et al., 2015; Mairal et al., 2012; Hunt et al., 2012; Wright et al., 2009; Schneider et al., 2009b; Dhillon et al., 2007; Quiroga et al., 2004; Kanungo et al., 2002; Abbott & Dayan, 1999).

Any coding scheme should find a proper implementation T. A natural choice for T

is the one that minimizes the *expected reconstruction error*

$$R(T) = \int_x f_T(x) d\rho(x) = \int_x f_T(x) p(x) dx,$$

where $\rho(x)$ is a Borel measure of the data source, and $p(x)$ is the probability density function. However, in most cases, $p(x)$ is unknown, and $R(T)$ cannot be directly minimized. An alternative approach is the *empirical risk minimization* (ERM) method (Vapnik, 2000; Cucker & Smale, 2002). Given a finite number of independent and identically distributed observations $x_1, \ldots, x_n \in \mathbb{R}^m$, the *empirical reconstruction error* with respect to T is defined as

$$R_n(T) = \frac{1}{n} \sum_{i=1}^{n} f_T(x_i).$$

The ERM method searches for a T_n that minimizes $R_n(T)$, and in the hope that $R(T_n)$ has a small distance to the expected reconstruction error $R(T^*)$, where

$$T^* = \arg\min_{T \in \mathcal{T}} R(T),$$

and $\mathcal{T}$ denotes a particular class of linear operators T.

A probabilistic bound on the defect

$$\sup_{T \in \mathcal{T}} |R(T) - R_n(T)|$$

is called the *generalization (error) bound.* This paper focuses on this error bound in the framework of k-dimensional coding schemes. Although different restrictions are imposed on the choices of $\mathcal{T}$ and Y for different concrete forms of k-dimensional coding schemes (for example, NMF requires both $\mathcal{T}$ and Y to be non-negative, and sparse coding requires sparsity in Y), they are closely related. For example, Ding et al. (2005)

showed that NMF with orthogonal $(y_1, \ldots, y_n)^\top$ is identical to k-means clustering of $\{x_1, \ldots, x_n\}$. Since these different forms of k-dimensional coding schemes are closely related, analyzing the generalization bounds together in this context has the advantages of exploiting the common properties and mutual cross-fertilization.

1.1 Related work

Maurer & Pontil (2010) and Gribonval et al. (2015) have performed the only known theoretical analyses on the generalization error in the framework of k-dimensional coding schemes. Other works have concentrated only on specific k-dimensional coding schemes. Since some previous works have studied *consistency* performance, which considers the quantity $R(T_n)-R(T^*)$ of the related ERM-based algorithms, we demonstrate the relationship between the generalization error and consistency performance here:

$$\begin{aligned}
&R(T_n) - R(T^*) \\
&= R(T_n) - R_n(T_n) + R_n(T_n) - R_n(T^*) + R_n(T^*) - R(T^*) \\
&\leq R(T_n) - R_n(T_n) + R_n(T^*) - R(T^*) \\
&\leq 2 \sup_{T \in \mathcal{T}} |R(T) - R_n(T)|.
\end{aligned}$$

Thus, analyzing the generalization error provides an approach for analyzing the consistency performance, and the consistency performance provides directions to generalization error analysis. We review the generalization error and consistency performance of k-dimensional coding schemes together:

- Non-negative matrix factorization (NMF). The only known generalization bounds of NMF are developed by Maurer & Pontil (2010) and Gribonval et al. (2015).

- Dictionary learning. Maurer & Pontil (2010) have developped dimensionality-independent generalization bounds. Vainsencher et al. (2011) and Gribonval et al. (2015) have studied the dimensionality-dependent generalization bounds.

- Sparse coding. A generalization bound for sparse coding was first derived by Maurer & Pontil (2010), and subsequently extended by Xu & Lafferty (2012), Mehta & Gray (2013), Maurer et al. (2013), and Gribonval et al. (2015). Maurer et al. (2013) derived a faster convergence rate upper bound of the consistency performance in a transfer learning setting.

- K-means clustering and vector quantization. Consistency performances of k-means clustering and vector quantization have mostly been studied for $\mathcal{H} = \mathbb{R}^m$. Asymptotic and non-asymptotic consistency performances have been considered by Pollard (1982), Chou (1994), Linder et al. (1994), Bartlett et al. (1998), Linder (2000), Antos et al. (2005), Antos (2005) and Levrard et al. (2013). Recently, Biau et al. (2008), Maurer & Pontil (2010) and Levrard et al. (2015) developed dimensionality-independent generalization bounds for k-means clustering.

We are aware that these specific forms of k-dimensional coding schemes have many applications for finite-dimensional data, and only a few dimensionality-dependent methods have been developed to analyze the generalization bounds for all these coding schemes.

In this paper, we develop a dimensionality-dependent method to analyze the generalization bounds for the framework of k-dimensional coding schemes. Our method is based on Hoeffding's inequality (Hoeffding, 1963) and the Bennett type inequalities (Boucheron et al., 2013), and directly bounds the covering number of the loss function class induced by the reconstruction error, which avoids inducing the worst-case upper bound on k of the loss function. Our method allows a generalization bound of order $\mathcal{O}\left((mk\ln(mkn)/n)^{\gamma_n}\right)$, where γ_n is much bigger than 0.5 when n is small, which delicately describes the non-asymptotic behavior of the learning process. However, when n goes to infinity, γ_n approaches to 0.5. The obtained dimensionality-dependent generalization bound can be much tighter than the previous ones when the number k of columns of the implementation is larger than the dimensionality m, which could often happen for dictionary learning, sparse coding, k-means clustering and vector quantization. We therefore obtain state-of-the-art generalization bounds for NMF, dictionary learning, sparse coding, k-means clustering and vector quantization.

The remainder of the paper is organized as follows. We present our motivation in Section 2 and main results in Section 3. In Section 4, our results are applied to specific coding schemes and are empirically compared with state-of-the-art generalization bounds. We prove our results in Section 5 and conclude the paper in Section 6.

2 Motivation

We first introduce the dimensionality-independent generalization bounds and demonstrate why our dimensionality-dependent bound complements them.

Assume that data points are drawn from a Hilbert space $\mathcal{H}$ with distribution μ. For any $r \geq 0$, let $\mathcal{P}(r)$ denote the set of probability distributions on $\mathcal{H}$ supported on the closed ball of radius r centered at the origin. In other words, $\mu \in \mathcal{P}(r)$ means that $P\{\|x\| \leq r\} = 1$. Let $\mathcal{T}$ be bounded in the operator norm, that is for every $T \in \mathcal{T}$, it holds that $\|Tv\| \leq c$ for all v with $\|v\| \leq 1$. Then, we also have that the columns of T are bounded as $\|Te_i\| \leq c, i = 1, \ldots, k$, where $\{e_i | 1 \leq i \leq k\}$ is the orthonormal basis of $\mathbb{R}^k$.

The following two theorems are equivalent to the main theorems proved by Maurer & Pontil (2010), but are represented in a different way. They are dimensionality-independent generalization bounds obtained in the frame of the k-dimensional coding schemes. They exploited the Rademacher complexity technique (Bartlett & Mendelson, 2003) which is suitable for deriving dimensionality-independent bounds (see Biau et al., 2008).

Theorem 1 *Assume that $\mu \in \mathcal{P}(r)$ and Y is a closed subset of the unit ball of $\mathbb{R}^k$, and that there is $c \geq 0$ such that for all $T \in \mathcal{T}$, $\|Te_i\| \leq c, i = 1, \ldots, k$. Suppose that the reconstruction error functions f_T for $T \in \mathcal{T}$ have a range contained in $[0, b]$. For any $\delta \in (0, 1)$, with probability at least $1 - \delta$ in the independently observed data $x_1, \ldots, x_n \sim \mu$, we have*

$$\sup_{T \in \mathcal{T}} |R(T) - R_n(T)| \leq (4crk + 2c^2k^2)\sqrt{\frac{\pi}{n}} + b\sqrt{\frac{8 \ln 2/\delta}{n}}.$$

Remark 1 *The dimensionality-independent generalization bound in Theorem 1 is valuable because it shows a convergence rate of order $O(\sqrt{1/n})$.*

Theorem 2 *Assume that $\mu \in \mathcal{P}(r)$ and $\|\mathcal{T}\|_Y = \sup_{T \in \mathcal{T}} \sup_{y \in Y} \|Ty\|$, and that the reconstruction error functions f_T for $T \in \mathcal{T}$ have a range contained in $[0, b]$. For any $\delta \in (0, 1)$, with probability at least $1 - \delta$ in the independently observed data*

$x_1, \ldots, x_n \sim \mu$*, we have*

$$\sup_{T\in\mathcal{T}} |R(T) - R_n(T)| \\ \leq b\sqrt{\frac{\ln 2/\delta}{2n}} + \frac{bk}{2}\sqrt{\frac{\ln\left(16n\|\mathcal{T}\|_Y^2\right)}{n}} + \frac{4 + 4\|\mathcal{T}\|_Y + \sqrt{8\pi}rk\|\mathcal{T}\|_Y}{\sqrt{n}}.$$

If $\mathcal{H}$ *is finite dimensional, the above result will be improved to*

$$\sup_{T\in\mathcal{T}} |R(T) - R_n(T)| \\ \leq b\sqrt{\frac{\ln 2/\delta}{2n}} + \frac{b}{2}\sqrt{\frac{mk\ln\left(16n\|\mathcal{T}\|_Y^2\right)}{n}} + \frac{4 + 4\|\mathcal{T}\|_Y + \sqrt{8\pi}rk\|\mathcal{T}\|_Y}{\sqrt{n}}.$$

Remark 2 *The condition that* Y *is a closed subset of the unit ball of* $\mathbb{R}^k$ *can be easily achieved by controlling the upper bound of columns of* T*, because there is a trade-off between the bounds of columns of* T *and the entries of* $y \in Y$*.*

Remark 3 *We note that Theorems 1 and 2 are more complicated than the original results presented in (Maurer & Pontil, 2010). This is because we have removed the restrictions that* $c \geq 1$ *and* $\|\mathcal{T}\|_Y \geq 1$*, which are required to simplify their results, to reveal the intrinsic relationships between the order of* k *and the Rademacher complexities (discussed below). The proof methods of Theorems 1 and 2 in this paper are exactly the same as those presented by Maurer & Pontil (2010).*

We note that if y is in the unit ball of $\mathbb{R}^k$, then

$$\begin{aligned} f_T(x) &= \min_{y\in\mathbb{R}^{k\times 1}} \|x - Ty\|^2 \leq \min_{y\in\mathbb{R}^{k\times 1}} \left(\|x\|^2 + \|Ty\|^2\right) \leq r^2 + \min_{y\in\mathbb{R}^{k\times 1}} \|Ty\|^2 \\ &= r^2 + \min_{y\in\mathbb{R}^{k\times 1}} \sum_{i,j}^{k} \langle y_i Te_i, y_j Te_j\rangle \leq r^2 + \min_{y\in\mathbb{R}^{k\times 1}} \sum_{i,j}^{k} \|y_i Te_i\|\|y_j Te_j\| \\ &\leq r^2 + c^2k^2, \end{aligned}$$

where r, the upper bound of the data point, can be reduced by normalization. However, k is a fixed integer, whose value is usually large in practice. Thus, c^2k^2 is the dominant factor in the upper bound of f_T. It is evident that f_T has the worst-case upper bound on k of order $\mathcal{O}(k^2)$, *i.e.,* the dependency *w.r.t.* k of the upper bound of f_T has the worst case order $\mathcal{O}(k^2)$. However, for some special forms of k-dimensional coding schemes, the upper bound of f_T has a very small order about k. Taking NMF as an example, the order about k is zero because

$$f_T(x) = \min_{y \in \mathbb{R}_+^k} \|x - Ty\|^2 \leq \|x\|^2 + \|T0\|^2 \leq r^2.$$

It is evident that the term $2c^2k^2\sqrt{\pi/n}$ in Theorem 1 has the same order as that of the worst-case upper bound on k of f_T. It will therefore be loose for some specific k-dimensional coding schemes. Maurer & Pontil (2010) introduced the proof method of Theorem 2 to overcome this problem; however, the term $rk\|\mathcal{T}\|_Y\sqrt{8\pi/n}$ implies that the problem is only partially solved, because rk represents the worst-case upper bound on k of $\sqrt{f_T}$ (details can be found in the proof therein). For example, in NMF, the term $rk\|\mathcal{T}\|_Y\sqrt{8\pi/n}$ is of order $\mathcal{O}(\sqrt{k^3/n})$ (discussed below in Remark 4). The dimensionality-dependent bound in Theorem 2 faces the same problem because the proof method computes the Rademacher complexity, corresponding to which part the obtained bound is dimensionality-independent and involves the worst-case upper bound on k of $\sqrt{f_T}$.

We try to avoid the aforementioned worst case by employing a covering number method to measure the complexity of the induced loss function class $F_{\mathcal{T}} = \{f_T | T \in \mathcal{T}\}$. However, in our setting, the dimensionality m of data space must be finite.

3 Main results

Before presenting our main results, we first introduce the definition of *covering number* $\mathcal{N}_p(F, \epsilon, n)$ (T. Zhang, 2002).

Definition 1 *Let $\mathcal{B}$ be a metric space with metric d. Given observations $X = \{x_1, \ldots, x_n\}$, and vectors $f(X) = \{f(x_1), \ldots, f(x_n)\} \in \mathcal{B}^n$, the covering number in p-norm, denoted as $\mathcal{N}_p(F, \xi, X)$, is the minimum number m of a collection of vectors $v_1, \ldots, v_m \in \mathcal{B}^n$, such that $\forall f \in F, \exists v_j$:*

$$\|d(f(X), v_j)\|_p = \left[\sum_{i=1}^{n} d(f(x_i), v_j^i)^p\right]^{1/p} \leq n^{1/p}\xi,$$

where v_j^i is the i-th component of vector v_j. We also define $\mathcal{N}_p(F, \xi, n) = \sup_X \mathcal{N}_p(F, \xi, X)$.

Let $\mathcal{T} = \mathbb{R}^{m \times k}$. We can upper bound the covering number of the induced loss function class of any k-dimensional coding scheme.

Lemma 1 *Let $F_{\mathcal{T}} = \{f_T | T \in \mathcal{T}, \mathcal{T} = \mathbb{R}^{m \times k}\}$ be the loss function class induced by the reconstruction error for a k-dimensional coding scheme. We have*

$$\ln \mathcal{N}_1(F_{\mathcal{T}}, \xi', n) \leq mk \ln\left(\frac{4(r + ck)\sqrt{m}ck}{\xi'}\right).$$

By employing Hoeffding's inequality (Hoeffding, 1963), we can derive a dimensionality-dependent generalization bound for k-dimensional coding schemes.

Theorem 3 (main result one) *Assume that $\mu \in \mathcal{P}(r)$ and Y is a closed subset of the unit ball of $\mathbb{R}^k$, and that there is $c \geq 0$ such that for all $T \in \mathcal{T}$, $\|Te_i\| \leq c, i = 1, \ldots, k$, and that the functions f_T for $T \in \mathcal{T}$ have a range contained in $[0, b]$. For any $\delta \in (0, 1)$, with probability at least $1 - \delta$, we have*

$$\sup_{T \in \mathcal{T}} |R(T) - R_n(T)| \leq \frac{2}{n} + b\sqrt{\frac{mk \ln(4(r + ck)\sqrt{m}ckn) + \ln 2/\delta}{2n}}.$$

Our result is dimensionality-dependent. Compared to the bound in Theorem 2, our bound could be tighter if $m \ln m \leq k\|\mathcal{T}\|_Y^2$.

Remark 4 *Let us take NMF for example to show how our method avoids inducing the worst-case upper bound on k of the loss function compared to those of Theorems 1 and 2. Regarding NMF,*

$$\|\mathcal{T}\|_Y = \sup_{T \in \mathcal{T}} \sup_{y \in Y} \|Ty\| = \sup_{T \in \mathcal{T}} \sup_{y \in Y} \left\| \sum_{i=1}^{k} y_i T e_i \right\| = \sup_{y \in Y} c \sum_{i=1}^{k} \|y_i e_i\| = c\sqrt{k}.$$

If we only consider the order of m, k and n, our bound is of order $\mathcal{O}(\sqrt{km \ln{(mkn)}/n})$ while Theorem 1 has order $\mathcal{O}(\sqrt{k^4/n})$ and Theorem 2 is of order $\mathcal{O}(\sqrt{k^3/n}+\sqrt{k^2 \ln(kn)/n})$. Our bound is tighter when $m \ln m \leq k^2$.

Remark 5 *For dictionary learning, sparse coding, k-means clustering and vector quantization, the number k of the columns of the linear implementation may be larger than the dimensionality m. If $k > m$, our bound will be much tighter than the dimensionality-independent generalization bound.*

Remark 6 *According to the proofs of Lemma 1 and Theorem 3, our result is based on the estimation of the Lipschitz constant of the loss function $f_T(x)$ w.r.t. the implementation T. Particularly, we proved the property $|f_T(x) - f_{T'}(x)| \leq L|T - T'|$ for all T and T' in $\mathcal{T}$, where L is a constant depending on a specific k-dimensional coding scheme. Similar to our idea, Gribonval et al. (2015) also developed dimensionality-dependent generalization bounds for k-dimensional coding schemes. However, their method is different from ours. Their results are essentially based on the property that $|f_T(x) - f_{T'}(x)| \leq L'\|T - T'\|_{1 \to 2}$ for all T and T' in $\mathcal{T}$, where L' is also a constant* *and the operator norm* $\|\cdot\|_{1\to 2}$ *of an* $m \times k$ *matrix* $A = [A_1, \dots, A_k]$ *is defined as* $\|A\|_{1\to 2} = \sup_{\|\alpha\|_1 \le 1} \|A\alpha\|_2$. *As a result, under some assumptions (see assumptions A1-A4, B1-B3 and C1-C2 therein) and with high probability, they have that*

$$\begin{aligned}\sup\nolimits_{T\in\mathcal{T}} |R(T) - R_n(T)| &\le 3c\sqrt{\frac{mk \cdot \max(\ln \frac{2L'C}{c}, 1) \ln n}{n}} \\ &+c\sqrt{\frac{mk \cdot \max(\ln \frac{2L'C}{c}, 1) + \ln 2/\delta}{n}},\end{aligned}$$

where c, C, T *are constants depending on a specific* k*-dimensional coding scheme. Note that in most applications,* $\ln \frac{2L'C}{c} > 1$ *and* $\ln n > 1$. *Their bound could be looser than the derived bound in Theorem 3 because in the cases, it holds that* $\ln \frac{2L'C}{c} \ln n > \ln \frac{2L'C}{c} + \ln n$. *Detailed comparisons are presented in Section 4.*

The result in Theorem 3 can be improved by exploiting Bennett type inequalities. We can make the upper bound to have either a smaller constant or a faster convergence rate as follows.

By employing Bernstein's inequality, we show that a tighter generalization bound of k-dimensional coding schemes than that in Theorem 3 can be derived.

Theorem 4 (main result two) *Assume that* $\mu \in \mathcal{P}(r)$ *and* Y *is a closed subset of the unit ball of* $\mathbb{R}^k$*, and that there is* $c \ge 0$ *such that for all* $T \in \mathcal{T}$*,* $\|Te_i\| \le c, i = 1, \dots, k$*, and that the functions* f_T *for* $T \in \mathcal{T}$ *have a range contained in* $[0, 1]$. *For any* $\delta \in (0, 1)$*, with probability at least* $1 - \delta$*, we have*

$$\begin{aligned}\sup\nolimits_{T\in\mathcal{T}} |R(T) - R_n(T)| &\le \frac{2}{n} + \frac{5\left(mk \ln\left(4(r+ck)\sqrt{m}ckn\right) + \ln 2/\delta\right)}{n} \\ &+\sqrt{\frac{2R_n(T)\left(mk \ln\left(4(r+ck)\sqrt{m}ckn\right) + \ln 2/\delta\right)}{n}}.\end{aligned}$$

Remark 7 *The upper bound in Theorem 4 can be much tighter than that in Theorem 3. The dominant term in the upper bound of Theorem 4 is* $\sqrt{\frac{2R_n(T)\left(mk\ln\left(4(r+ck)\sqrt{m}ckn\right)+\ln 2/\delta\right)}{n}}$. *Since the empirical reconstruction error* $R_n(T)$ *is no bigger and sometimes much smaller than* 1*, the upper bound in Theorem 4 can therefore be much tighter than that in Theorem 3.*

We can represent the result by using the inequlaity that for all $a, b, \lambda > 0$, $\sqrt{2ab} < \lambda a + \lambda^{-1}b/4$.

Proposition 1 *Assume that* $\mu \in \mathcal{P}(r)$ *and* Y *is a closed subset of the unit ball of* $\mathbb{R}^k$*, and that there is* $c \geq 0$ *such that for all* $T \in \mathcal{T}$*,* $\|Te_i\| \leq c, i = 1, \ldots, k$*, and that the functions* f_T *for* $T \in \mathcal{T}$ *have a range contained in* $[0, 1]$*. For any* $T \in \mathcal{T}$*, any* $\lambda > 0$ *and any* $\delta \in (0, 1)$*, with probability at least* $1 - \delta$*, we have*

$$R(T) \leq (1+\lambda)R_n(T) + \frac{2}{n} + \left(\frac{1}{4\lambda} + 5\right)\frac{(mk\ln(4(r+ck)\sqrt{m}ckn) + \ln 2/\delta)}{n}.$$

We have claimed that Theorem 4 and Proposition 1 can be tighter than Theorem 3 by saying that $R_n(T)$ can be very small. However, sometimes, such a term could be large. If $R_n(T) > 1/4$ (note that the reconstruction error function $f_T \in [0, 1]$), Theorem 4 and Proposition 1 will be looser than Theorem 3.

The following theorem implies that by employing Bennett's type inequality, the generalization bound can be improved no matter what the value of $R_n(T)$ is.

Theorem 5 (main result three) *Assume that* $\mu \in \mathcal{P}(r)$ *and* Y *is a closed subset of the unit ball of* $\mathbb{R}^k$*, and that there is* $c \geq 0$ *such that for all* $T \in \mathcal{T}$*,* $\|Te_i\| \leq c, i = 1, \ldots, k$*, and that the functions* f_T *for* $T \in \mathcal{T}$ *have a range contained in* $[0, 1]$*. For any* $\delta \in (0, 1)$*,* *with probability at least* $1-\delta$ *it holds for all* $T \in \mathcal{T}$ *that*

$$|R(T) - R_n(T)| \leq \frac{2}{n} + \left(\frac{mk \ln\left(4(r+ck)\sqrt{m}ckn\right) + \ln \frac{2}{\delta}}{\beta n} \right)^{\frac{1}{2 - \frac{\ln(8\beta V/3)}{\ln |R(T)-R_n(T)|}}} ,$$

when V *satisfies that* $|R(T) - R_n(T)| \leq V \leq 3/8\beta$ *and* β *is any positive constant.*

Remark 8 *Since* $f_T(x) \leq 1$ *in Theorem 5, we have that* $\frac{\ln(8\beta V/3)}{\ln|R(T)-R_n(T)|} \geq 0$ *if the condition* $8\beta V < 3$ *holds. Let simply set* $\beta = 2$*. If we further have that* $|R(T) - R_n(T)| \leq V \leq 3/16$*, the upper bound in Theorem 5 will be the same as that in Theorem 3 except for a faster convergence rate. Thus, the upper bound in Theorem 5 can be much tighter than that in Theorem 3 in the sense that it converges much faster.*

Remark 9 *The generalization bound in Theorem 3 is of order* $\mathcal{O}\left((mk\ln(mkn)/n)^{\frac{1}{2}}\right)$*; while the generalization bound in Theorem 5 is of order* $\mathcal{O}\left((mk\ln(mkn)/n)^{\gamma_n}\right)$*, where* $\gamma_n > 1/2$ *when* n *is finite. The generalization bound in Theorem 5, derived by employing Bennett's inequality, converges faster when the sample size* n *is small, which is often the case in practice and more detailedly describes the non-asymptotic behavior of the learning process. More empirical discussions can be found in C. Zhang (2013). However, when the sample size* n *goes to infinity, the term* $\frac{1}{2-\frac{\ln(8\beta V/3)}{\ln|R(T)-R_n(T)|}}$ *will approach to* $\frac{1}{2}$*, which means that the upper bounds in Theorems 5 and 3 describe the same asymptotic behavior of the learning process.*

Remark 10 *Theorem 5 looks complex, since the exponent in the convergence rate depents itself on the sample size in an implicit way. Here we show the superiority of Theorem 5 by comparing it with Theorem 3. From the proof of Theorem 5, we can see that the theorem depends on the following inequality (15):*

$$P\left\{|R(T) - R_n(T)| \geq \epsilon\right\} \leq 2\exp\left(-nVh\left(\frac{\epsilon}{V}\right)\right) \leq 2\exp\left(-\beta n \epsilon^{2-\frac{\ln(8\beta V/3)}{\ln \epsilon}}\right),$$

where $\epsilon \leq V$. Note that for Hoeffding's inequality, with any β we also have

$$P\left\{|R(T) - R_n(T)| \geq \epsilon\right\} \leq 2\exp\left(-2n\epsilon^2\right) = 2\exp\left(-\beta n\epsilon^{2-\frac{\ln(\beta/2)}{\ln\epsilon}}\right).$$

Thus, according to Hoeffding's inequality and the prove method of Theorem 5, for all $T \in \mathcal{T}$, with probability at least $1-\delta$ it holds that

$$|R(T) - R_n(T)| \leq \frac{2}{n} + \left(\frac{mk\ln\left(4(r+ck)\sqrt{m}ckn\right) + \ln\frac{2}{\delta}}{\beta n}\right)^{\frac{1}{2-\frac{\ln(\beta/2)}{\ln|R(T)-R_n(T)|}}}.$$

Comparing the above bound with that in Theorem 5, we can see that, if we interpret Theorem 3 with a faster convergence rate, the upper bound therein is looser than that in Theorem 5 when $V \leq 3/16$.

Our main results in Theorems 3, 4, and 5 apply to all the k-dimensional coding schemes because the covering number in Lemma 1 measures the complexity of the loss function class that includes all the possible loss functions of k-dimensional coding schemes. However, for some specific k-dimensional coding schemes, the complexity of the corresponding induced loss function class can be refined. We discuss the details in the next section[1].

4 Applications

In this section, we apply our proof methods to specific k-dimensional coding schemes. We show that our methods provide state-of-the-art dimensionality-dependent generalization bounds.

[1] Even though the faster convergence interpretation in Theorem 5 is interesting, it looks complicated and the upper bound is almost the same tight as that of Theorem 4. Therefore, we do not disscuss its applicaitons for specific k-dimensional codeing schemes.

4.1 Non-negative matrix factorization

NMF factorizes a data matrix $X \in \mathbb{R}_+^{m\times n}$ into two non-negative matrices $T \in \mathbb{R}_+^{m\times k}$ and $Y \in \mathbb{R}_+^{k\times n}$, where $k < \min(m, n)$. NMF has been widely exploited since Lee and Seung (1999) provided a powerful psychological and physiological interpretation as a parts-based factorization and an efficient multiplicative update rule for obtaining a local solution. Many fast and robust algorithms are then followed (see, e.g., Gillis & Vavasis, 2014). In all applications, both the data points and the vectors $Te_i, i = 1, \ldots, k$ are contained in the positive orthant of a finite-dimensional space. In this case, our method for deriving dimensionality-dependent generalization bounds is likely to be superior to the method for obtaining dimensionality-independent results.

Letting $X = (x_1, \ldots, x_n) \in \mathbb{R}_+^{m\times n}$, NMF can be formulated as follows:

$$\begin{aligned} \min\nolimits_{T,Y} \quad & \|X - TY\|_F^2, \\ \text{s.t.} \quad & T \in \mathbb{R}_+^{m\times k}, Y \in \mathbb{R}_+^{k\times n} \end{aligned}$$

where $\|\cdot\|_F$ is the matrix Frobenius norm.

Because $TY = TQ^{-1}QY$ if Q is a scaling matrix, we can normalize T without changing the optimization problem by choosing

$$Q = \begin{pmatrix} \|T_1\| & & & \\ & \|T_2\| & & \\ & & \ddots & \\ & & & \|T_k\| \end{pmatrix}.$$

If we restrict $\mu \in \mathcal{P}(r)$ and normalize T, columns of Y will also be upper bounded by r. This can be seen in the following lemma, which generalizes Lemma 2 in Maurer & Pontil (2010):

Lemma 2 *For NMF with normalized T, if $\mu \in \mathcal{P}(r)$, then every column of Y is upper bounded by r; that is $\|y\| \leq r$ for all $y \in Y$.*

For a fixed T, Y is determined by a convex problem. Thus, the reconstruction error for NMF is

$$f_T(x) = \min_{y \in \mathbb{R}_+^k} \|x - Ty\|^2,$$

and the generalization error of NMF can be analyzed under the framework of the k-dimensional coding schemes.

Using the same proof method as that of Lemma 1, we have the following lemma.

Lemma 3 *Let $\mu \in \mathcal{P}(1)$ and $F_{\mathcal{T}} = \{f_T | T \in \mathcal{T}, \mathcal{T} = \mathbb{R}_+^{m \times k}\}$ be the loss function class induced by the reconstruction error of NMF. We have*

$$\ln \mathcal{N}_1(F_{\mathcal{T}}, \xi', n) \leq mk \ln \left(\frac{2(1+k)\sqrt{m}k}{\xi'} \right).$$

Then, according to the proof methods of Theorems 3, 4 and 5, we have the following dimensionality-dependent generalization bounds for NMF.

Theorem 6 *For NMF, assume that $\mu \in \mathcal{P}(1)$ and that $\mathcal{T}$ is normalized. For any $\delta \in (0,1)$, with probability at least $1 - \delta$ it holds for all $T \in \mathcal{T}$ that*

$$\begin{aligned}
&|R(T) - R_n(T)| \\
&\leq \frac{2}{n} + \min\left\{ \sqrt{\frac{mk \ln(2(1+k)\sqrt{m}kn) + \ln 2/\delta}{2n}}, \right. \\
&\left. \frac{5\,(mk \ln(2(1+k)\sqrt{m}kn) + \ln 2/\delta)}{n} + \sqrt{\frac{2R_n(T)\,(mk \ln(2(1+k)\sqrt{m}kn) + \ln 2/\delta)}{n}} \right\}.
\end{aligned}$$

Since the value of $R_n(T)$ is unknown in this paper (it is usually known in an optimization procedure), in the rest of the paper, we will only compare the bound in Theo-

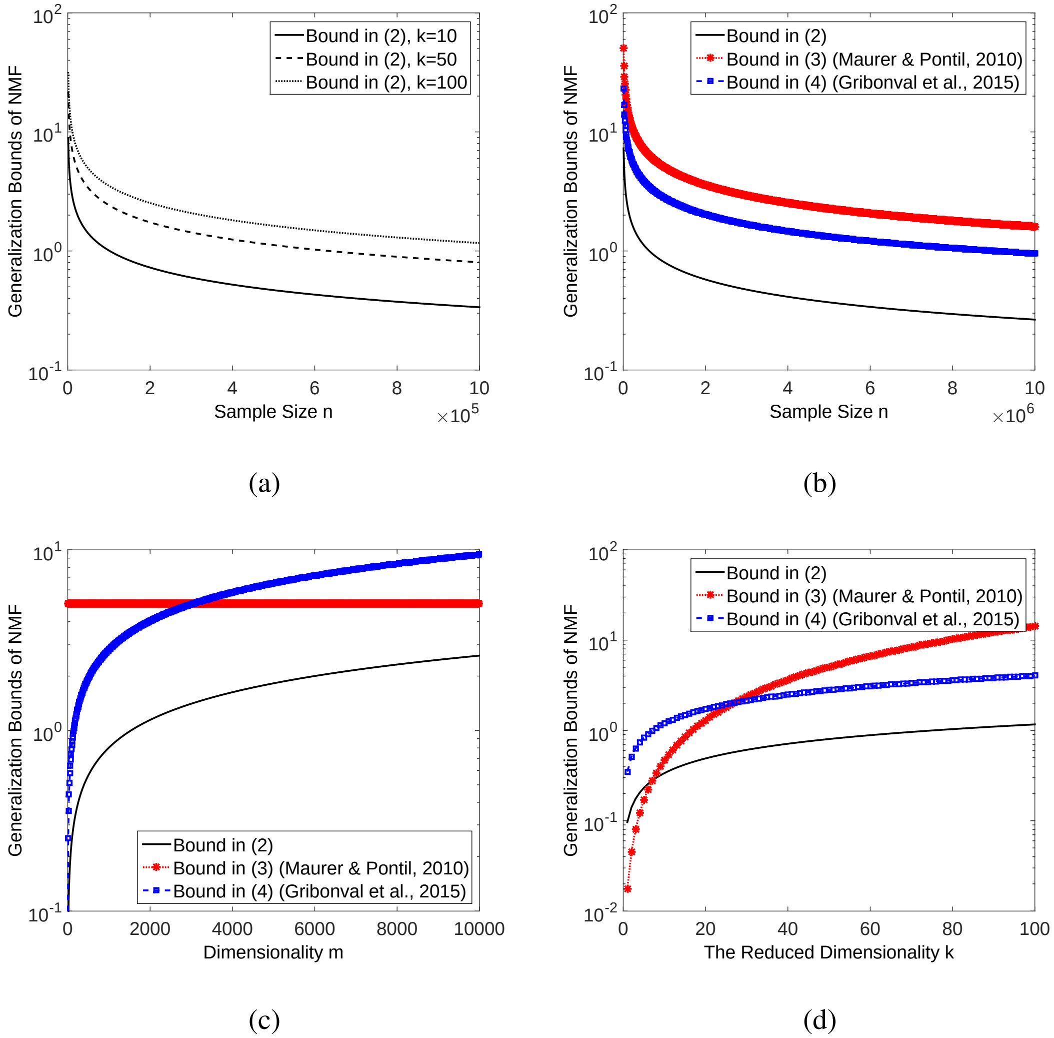


Figure 1: Comparisons of the generalization bounds of NMF. (a) The convergence of the bound in (2), where $m = 1000$. (b) Comparing the convergence with state-of-the-art generalization bounds, where $k = 50, m = 1000$. (c) Comparing the generalization bound with state-of-the-art generalization bounds in terms of the parameter m, where $k = 50, n = 10^6$. (d) Comparing the generalization bound with state-of-the-art generalization bounds in terms of the parameter k, where $m = 10^3, n = 10^6$.

rem 3 with state-of-the-art bounds. Theorem 3 gives the following bound for NMF

$$\frac{2}{n} + \sqrt{\frac{mk \ln\left(2(1+k)\sqrt{m}kn\right) + \ln 2/\delta}{2n}}. \tag{2}$$

Under the setting of Theorem 6, Theorem 2 yields the following bound

$$\frac{k}{\sqrt{n}}\left(14\sqrt{k} + \frac{1}{2}\sqrt{\ln(16nk)}\right) + \sqrt{\frac{\ln 2/\delta}{2n}}; \tag{3}$$

Gribonval et al. (2015)'s result gives the following bound

$$\frac{3}{\sqrt{8}}\sqrt{\frac{mk\ln(12\sqrt{8}mk)\ln n}{n}} + \frac{1}{\sqrt{8}}\sqrt{\frac{mk\ln(12\sqrt{8}mk) + \ln 2/\delta}{n}}. \tag{4}$$

We then carefully compare the above generalization bounds. For NMF problems, the dimensionality m is usually very large compared to the reduced dimensionality k. We set $m = 1000, k = 50, \delta = 0.01$. The comparisons are illustrated in Figure 1. The figure shows that in most cases, the derived generalization bound is tighter than state-of-the-art bounds. In Figure 1d, the bound in (3) is tighter than the derived bound in a small range because it is dimensionality-independent and $m = 1000$ is set to be much larger than the corresponding reduced dimensionality k.

4.2 Dictionary learning

Dictionary learning tries to find a dictionary such that all observed data points can be approximated by linear combinations of atoms in the dictionary. Let the columns of T be the atoms of the dictionary; for an observation $x \in \mathbb{R}^m$, the dictionary learning method will represent x by a linear combination of columns of T as

$$x' = \sum_{i=1}^{k} \alpha_i T_i, \alpha_i \in \mathbb{R}, i = 1, \ldots, k.$$

Thus, the reconstruction error of dictionary learning is the same as those of k-dimensional coding schemes.

Vainsencher et al. (2011) provided notable dimensionality-dependent generalization bounds for dictionary learning by considering two types of constraints on coefficient selection, respectively. For the ℓ_0-norm regularized coefficient selection, where every signal is approximated by a combination of, at most, p dictionary atoms, the generalization bound (Theorem 14 therein) is of order $\mathcal{O}(\sqrt{mk\ln(np)/n})$ under an approximate orthogonality assumption on the dictionary. For the ℓ_1-norm regularized coefficient selection, the generalization bound (Theorem 7 therein) is of order $\mathcal{O}(\sqrt{mk\ln(n\lambda)/n})$ under the requirements that λ, which is the upper bound of the ℓ_1-norm of the coefficient, is larger than $e/4$, and that the signal x is mapped onto the $(m-1)$-sphere. Our result on k-dimensional coding scheme can also be applied to dictionary learning and provides a more general bound, which does not require x to be on the $(m-1)$-sphere or the near-orthogonality requirement and directly applies to all dictionary learning problems.

Theorem 7 *For dictionary learning, assume that $\mu \in \mathcal{P}(1)$ and that Y is a closed subset of the unit ball of $\mathbb{R}^k$, and that every atom $T_i, i = 1, \ldots, k$ is bounded by $\|T_i\| \leq c, i = 1, \ldots, k$. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$ it holds for all $T \in \mathcal{T}$ that*

$$\begin{aligned}
&|R(T) - R_n(T)| \\
&\leq \frac{2}{n} + \min\left\{\sqrt{\frac{mk\ln\left(4(1+ck)\sqrt{m}ckn\right) + \ln 2/\delta}{2n}},\right. \\
&\left.\frac{5\left(mk\ln\left(4(1+ck)\sqrt{m}ckn\right) + \ln 2/\delta\right)}{n} + \sqrt{\frac{2R_n(T)\left(mk\ln\left(4(1+ck)\sqrt{m}ckn\right) + \ln 2/\delta\right)}{n}}\right\}.
\end{aligned}$$

The proof of Theorem 7 is the same as that of Theorem 6.

Remark 11 *If we substitute an upper bound* $\lambda \leq \sqrt{k}$ *into the bound in Vainsencher et al. (2011), the bound in Theorem 7 therein will be of order* $\mathcal{O}(\sqrt{mk\ln(kn)/n})$*, which has the same order as term* $\sqrt{\frac{mk\ln\left(4(1+ck)\sqrt{m}ckn\right)+\ln 2/\delta}{2n}}$*. However, our bound in Theorem 5 also shows a faster convergence rate.*

Remark 12 *The method Vainsencher et al. (2011) used to upper bound the covering number of the induced loss function class is very different from ours. To upper bound the covering number of the induced loss function class for dictionary learning, Vainsencher et al. (2011) used the knowledge that a uniform* L *Lipschitz mapping between metric spaces converts* ξ/L *covers into* ξ *covers. Then, they focused on analyzing the Lipschitz property of the reconstruction error function that maps a dictionary into a reconstruction error, i.e,* $\Psi_\lambda : D \mapsto h_{R_\lambda,D}, R_\lambda = \{a : \|a\|_1 \leq \lambda\}$*, as shown in Lemma 7 therein. Also note that to upper bound the Lipschitz constant of the mapping* $\Phi_k : D \mapsto h_{H_k,D}, H_k = \{a : \|a\|_0 \leq k\}$*, they introduced the approximate orthogonality condition (a bound on the Babel function) on the dictionary.*

Remark 13 *Analyzing the Lipschitz properties of the induced loss functions is essential for upper bounding the generalization error of* k*-dimensional coding schemes. Different form the method used in Vainsencher et al. (2011), Maurer & Pontil (2010) employed Slepian's Lemma to exploit the Lipschitz property; while in this paper, we also proposed a novel method as presented in the proof of Theorem 3.*

The comparisons of the generalization bounds of dictionary learning are similar to that of NMF because NMF can be regarded as dictionary learning in the positive orthant. We therefore omit the comparison. Many algorithms used in applications require sparsity in Y, because sparsity has advantages, such as for computation and storage. We therefore analyze sparsity in the next subsection.

4.3 Sparse coding

Sparse coding requires sparsity in the codebook. We use the hard constraint discussed in Maurer & Pontil (2010), that is $\mathcal{T} = \{T : \mathbb{R}^k \to \mathbb{R}^m | \|Te_i\| \leq c, i = 1, \ldots, k\}$, $Y = \{y | y \in \mathbb{R}^k, \|y\|_p \leq s\}$, and $1/p + 1/q = 1, 2 \leq p \leq \infty$. Thus, we have

$$\begin{aligned} \|Ty\| &= \left\| \sum_{i=1}^k y_i T e_i \right\| \leq \sum_{i=1}^k |y_i| \|Te_i\| \\ &\quad \text{(Using Hölder's inequality)} \\ &\leq s \left(\sum_{i=1}^k \|Te_i\|^q \right)^{1/q} \leq sck^{1/q} = sck^{1-1/p}. \end{aligned}$$

The following generalization bound for sparse coding is also from the work of Maurer & Pontil (2010), derived using the proof method of Theorem 2.

Theorem 8 *For sparse coding, assume that $\mu \in \mathcal{P}(1)$. Let $Y = \{y | y \in \mathbb{R}^k, \|y\|_p \leq s\}$ where $1 \leq p \leq \infty$. Let also assume that for all $T \in \mathcal{T}$, $\|Te_i\| \leq 1, i = 1, \ldots, k$. Then, for any $\delta \in (0,1)$, with probability at least $1 - \delta$ it holds for all $T \in \mathcal{T}$ that*

$$|R(T) - R_n(T)| \leq \frac{k}{2} \sqrt{\frac{\ln\left(16ns^2 2k^{2-2/p}\right)}{n}} + \sqrt{\frac{\ln 2/\delta}{2n}} + \frac{4 + 4sk^{1-1/p} + \sqrt{8\pi} sk^{2-1/p}}{\sqrt{n}}.$$

We now consider the generalization bound of sparse coding using our method. The following lemma is proved in Section 5.7.

Lemma 4 *Follow the setting of Theorem 8. Let $F_{\mathcal{T}}$ be the loss function class of sparse* *coding. We have*

$$\ln \mathcal{N}_1(F_{\mathcal{T}}, \xi', n) \leq mk \ln \left(\frac{4(s + s^2 k^{1-1/p})\sqrt{m} k^{1-1/p}}{\xi'} \right).$$

Then, we have the generalization bounds for sparse coding as follows:

Theorem 9 *Follow the setting of Theorem 8. For any $\delta \in (0,1)$, with probability at least $1-\delta$ it holds for all $T \in \mathcal{T}$ that*

$$\begin{aligned} &|R(T) - R_n(T)| \\ &\leq \min \left\{ \frac{2}{n} + \sqrt{\frac{\Delta + \ln 2/\delta}{2n}},\ \frac{2}{n} + \frac{5\,(\Delta + \ln 2/\delta)}{n} + \sqrt{\frac{2R_n(T)\,(\Delta + \ln 2/\delta)}{n}} \right\}, \end{aligned}$$

where $\Delta = mk \ln \left(4(s + s^2 k^{1-1/p})\sqrt{m} k^{1-1/p} n\right)$.

The proof of Theorem 9 is the same as that of Theorem 6.

Theorem 9 gives the following bound for sparse coding

$$\frac{2}{n} + \sqrt{\frac{mk \ln \left(4(s + s^2 k^{1-1/p})\sqrt{m} k^{1-1/p} n\right) + \ln 2/\delta}{2n}}. \tag{5}$$

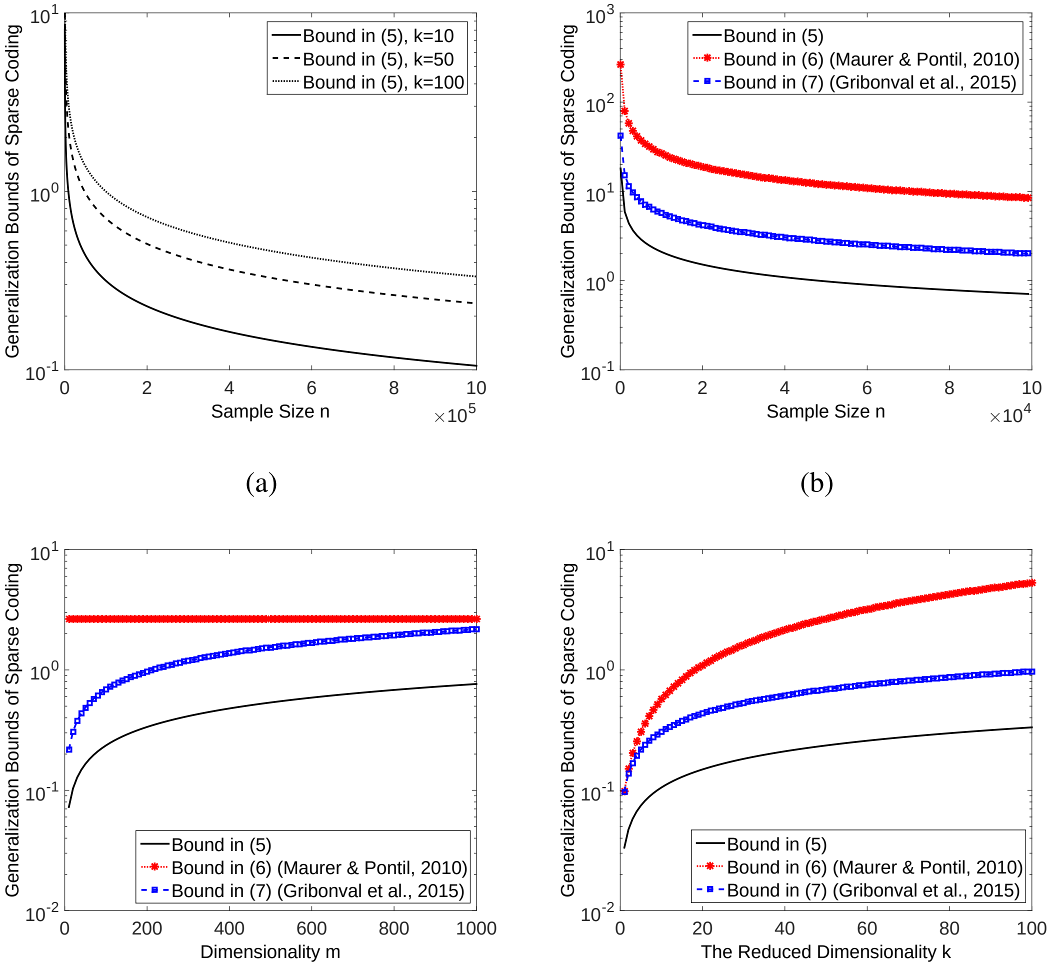
The upper bound for sparse coding derived by Maurer & Pontil (2010) is presented in Theorem 8:

$$\frac{k}{2}\sqrt{\frac{\ln\left(16ns^2 2k^{2-2/p}\right)}{n}} + \sqrt{\frac{\ln 2/\delta}{2n}} + \frac{4 + 4sk^{1-1/p} + \sqrt{8\pi} sk^{2-1/p}}{\sqrt{n}}. \tag{6}$$

Gribonval et al. (2015)'s result gives the following bound for sparse coding.

$$\frac{1}{\sqrt{8}}\left(3\sqrt{\frac{mk \max\left(\ln\left(6\sqrt{8} sk^{1-1/p}\right), 1\right) \ln n}{n}} + \sqrt{\frac{mk \max\left(\ln\left(6\sqrt{8} sk^{1-1/p}\right), 1\right) + \ln 2/\delta}{n}}\right). \tag{7}$$

We then compare the above generalization bounds of sparse coding in Figure 2 by setting $m = 100, k = 50, \delta = 0.01, p = 1$, and $s = 10$. The comparisons show that the derived generalization bound is tighter than state-of-the-art bounds.



(a)

(b)

(c)

(d)

Figure 2: Comparisons of the generalization bounds of sparse coding. (a) The convergence of the bound in (5), where $m = 100$. (b) Comparing the convergence with state-of-the-art generalization bounds, where $k = 50, m = 100$. (c) Comparing the generalization bound with state-of-the-art bounds in terms of the parameter m, where $k = 50, n = 10^6$. (d) Comparing the generalization bound with state-of-the-art bounds in terms of the parameter k, where $m = 100, n = 10^6$.

4.4 Vector quantization and k-means clustering

The k-means clustering (or vector quantization) method aims to find k cluster centers such that observations can be partitioned into k clusters and represented by the k cluster centers with a small reconstruction error. Taking every column of T as a cluster center and setting Y as the standard bases $\{e_1, \ldots, e_k\}$, we see that solving a k-means clustering problem is equal to finding an implementation T. The corresponding reconstruction error is

$$f_T(x) = \min_{i \in \{1,\ldots,k\}} \|x - Te_i\|^2.$$

So, the reconstruction error of k-means clustering and vector quantization is also within the framework of the reconstruction error of k-dimensional coding schemes.

The following lemma is essential for proving our dimensionality-dependent generalization bounds.

Lemma 5 *Assume that $\mu \in \mathcal{P}(1)$. Let $F_{\mathcal{T}}$ be the loss function class of k-means clustering and vector quantization. Then*

$$\ln \mathcal{N}_1(F_{\mathcal{T}}, \xi', n) \leq mk \ln \left(\frac{8\sqrt{m}}{\xi'} \right).$$

Theorem 10 *For k-means clustering and vector quantization, assume that $\mu \in \mathcal{P}(1)$, and that the functions f_T for $T \in \mathcal{T}$ have a range contained in $[0,1]$. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$ it holds for all $T \in \mathcal{T}$ that*

$$|R(T) - R_n(T)| \leq \frac{2}{n} + \min \left\{ \sqrt{\frac{mk \ln (8\sqrt{m}n) + \ln 2/\delta}{2n}}, \right.$$
$$\left. \frac{5\left(mk \ln (8\sqrt{m}n) + \ln 2/\delta\right)}{n} + \sqrt{\frac{2R_n(T)\left(mk \ln (8\sqrt{m}n) + \ln 2/\delta\right)}{n}} \right\}.$$

The proof of Theorem 10 is the same as that of Theorem 6.

Theorem 10 gives the following bound for k-means clustering and vector quantization

$$\frac{2}{n}+\sqrt{\frac{mk\ln\left(8\sqrt{m}n\right)+\ln 2/\delta}{2n}}. \tag{8}$$

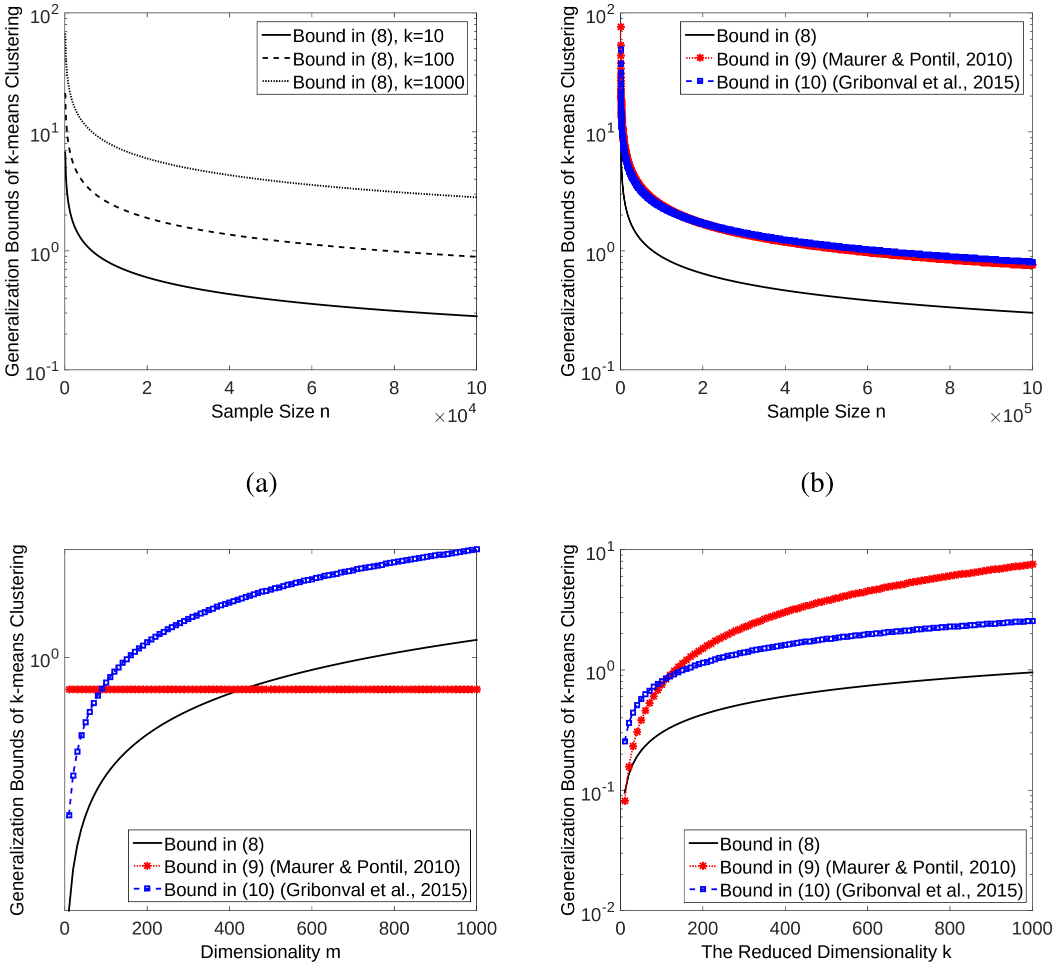
Maurer & Pontil (2010) derived the following bound

$$\frac{3\sqrt{2\pi}kr^2}{\sqrt{n}}+r^2\sqrt{\frac{8\ln 1/\delta}{n}}. \tag{9}$$

Gribonval et al. (2015) provided the following bound

$$\frac{3}{\sqrt{8}}\sqrt{\frac{mk\ln(12\sqrt{8})\ln n}{n}}+\frac{1}{\sqrt{8}}\sqrt{\frac{mk\ln(12\sqrt{8})+\ln 2/\delta}{n}}. \tag{10}$$

Remark 14 *The bound in (9) has order $\mathcal{O}(k/\sqrt{n})$, which is the same as the bound obtained by Biau et al. (2008). The term $\sqrt{\frac{mk\ln\left(8\sqrt{m}nr^2\right)+\ln 2/\delta}{2n}}$ in Theorem 10 has order $\mathcal{O}(\sqrt{mk\ln{(mn)}/n})$. If $m\ln{(mn)}\leq k$, our bound can be tighter than that of Maurer & Pontil (2010) and the result in Biau et al. (2008). The generalization bounds derived by Maurer & Pontil (2010) and Biau et al. (2008) also have an advantage that they converge faster. As discussed in Bartlett et al. (1998), Linder et al. (1994), and Devroye et al. (1996), the factor $\sqrt{\ln n}$ in Theorem 10 can be removed by the sophisticated uniform large-deviation inequalities of Alexander (1984) or Talagrand (1994). However, Devroye et al. (1996) proved that (Theorem 12.10 therein) the fast convergence upper bound has an astronomically large constant. The corresponding convergence bound is therefore loose. Our generalization bound, which is derived by exploiting Bennett's inequality, will be tighter if the empricial reconstruction error $R_n(T)$ is small.*



(a)

(b)

(c)

(d)

Figure 3: Comparisons of the generalization bounds of k-means clustering and vector quantization. (a) The convergence of the bound in (8), where $m = 100$. (b) Comparing the convergence with state-of-the-art generalization bounds, where $k = m = 100$. (c) Comparing the generalization bound with state-of-the-art bounds in terms of the parameter m, where $k = 100, n = 10^{6}$. (d) Comparing the generalization bound with state-of-the-art bounds in terms of the parameter k, where $m = 100, n = 10^{6}$.

We compare the above generalization bounds of k-means clustering and vector quantization in Figure 3 by setting $k = m = 100$. For k-means clustering and vector quantization problems, the dimensionality m can be independent of the reduced dimensionality k. Figure 3 shows that when k is not very large, the derived bound is tighter than state-of-the-art generalization bounds.

5 Proofs

In this section we prove the main results in Section 2 and some of the results presented in Section 3.

5.1 Concentration inequalities

In this subsection, we introduce the concentration inequalities that will be used to prove our assertions.

We first present Hoeffding's inequality (Hoeffding, 1963), which is widely used for deriving generalization bounds.

Theorem 11 (Hoeffding's inequality) *Let $X = \{x_1, \ldots, x_n\} \in \mathcal{H}^n$ be a sample set of independent random variables such that $x_i \leq B$ for some $B > 0$ almost surely for all $i \leq n$. Then for any $X \in \mathcal{H}^n$ and $\epsilon > 0$, the following inequality holds:*

$$P\left\{\left|E\frac{1}{n}\sum_{i=1}^{n} x_i - \frac{1}{n}\sum_{i=1}^{n} x_i\right| \geq \epsilon\right\} \leq 2\exp\left(\frac{-2n\epsilon}{B^2}\right).$$

We will also use Bernstein's inequality and Bennett's inequality (Boucheron et al., 2013; C. Zhang, 2013) to derive generalization bounds.

Theorem 12 (Bernstein's inequality) *Let $X = \{x_1, \ldots, x_n\} \in \mathcal{H}^n$ be a sample set of independent random variables such that $x_i \leq B$ for some $B > 0$ and Ex_i^2 is no bigger than V for some $V > 0$ almost surely for all $i \leq n$. Then for any $X \in \mathcal{H}^n$ and $\epsilon > 0$, the following inequality holds:*

$$P\left\{\left|E\frac{1}{n}\sum_{i=1}^{n} x_i - \frac{1}{n}\sum_{i=1}^{n} x_i\right| \geq \epsilon\right\} \leq 2\exp\left(\frac{-n\epsilon^2}{2(V + B\epsilon/3)}\right).$$

Theorem 13 (Bennett's inequality) *Let $X = \{x_1, \ldots, x_n\} \in \mathcal{H}^n$ be a sample set of independent random variables such that $x_i \leq B$ for some $B > 0$ and Ex_i^2 is no bigger than V for some $V > 0$ almost surely for all $i \leq n$. Then for any $X \in \mathcal{H}^n$ and $\epsilon > 0$, the following inequality holds:*

$$P\left\{\left|E\frac{1}{n}\sum_{i=1}^{n} x_i - \frac{1}{n}\sum_{i=1}^{n} x_i\right| \geq \epsilon\right\} \leq 2\exp\left(-\frac{nV}{B^2}h\left(\frac{B\epsilon}{V}\right)\right),$$

where $h(x) = (1 + x)\ln(1 + x) - x$ for $x > 0$.

5.2 Proof of Lemma 1

Proof. We will bound the covering number of the loss function class $F_{\mathcal{T}}$ by bounding the covering number of the implementation class $\mathcal{T}$. Cutting the subspace $[-c, c]^m \subset \mathbb{R}^m$ into small m-dimensional regular solids with width ξ, there are a total of

$$\left\lceil\frac{2c}{\xi}\right\rceil^m \leq \left(\frac{2c}{\xi} + 1\right)^m \leq \left(\frac{4c}{\xi}\right)^m$$

such regular solids. If we pick out the centers of these regular solids and use them to make up T, there are

$$\left\lceil\frac{2c}{\xi}\right\rceil^{mk} \leq \left(\frac{4c}{\xi}\right)^{mk}$$

choices, denoted by $\mathcal{S}$. Then $|\mathcal{S}|$ is the upper bound of the ξ-cover of the implementation class $\mathcal{T}$.

We will prove that for every T, there exists a $T' \in \mathcal{S}$ such that

$$\sup_x |f_T(x) - f_{T'}(x)| \leq \xi',$$

where $\xi' = (r + ck)\sqrt{m}k\xi$. The proof is as follows:

$$\begin{aligned}
&|f_T(x) - f_{T'}(x)| \\
&= \left| \min_y \|x - Ty\|^2 - \min_y \|x - T'y\|^2 \right| \\
&= \left| \min_y \|x - Ty\|^2 + \max_y \left(-\|x - T'y\|^2 \right) \right| \\
&\leq \left| \max_y \left(\|x - Ty\|^2 - \|x - T'y\|^2 \right) \right| \\
&\leq \left| \max_y 2x^\top Ty - 2x^\top T'y \right| + \left| \max_y \|Ty\|^2 - \|T'y\|^2 \right| \\
&= \left| \max_y \sum_{i=1}^{k} y_i \left\langle 2x, (T - T')e_i \right\rangle \right| + \left| \max_y \sum_{i,j}^{k} y_i y_j \left\langle (T + T')e_i, (T - T')e_j \right\rangle \right| \\
&\quad \text{(Using Hölder's inequality)} \\
&\leq \left| \sum_{i=1}^{k} | \left\langle 2x, (T - T')e_i \right\rangle | \right| + \left| \sum_{i,j}^{k} | \left\langle (T + T')e_i, (T - T')e_j \right\rangle | \right| \\
&\quad \text{(Using Cauchy-Schwarz inequality)} \\
&\leq \left| \sum_{i=1}^{k} \|2x\| \|(T - T')e_i\| \right| + \left| \sum_{i,j}^{k} \|(T + T')e_i\| \, \|(T - T')e_j\| \right| \\
&\leq \left| \sum_{i=1}^{k} \|2x\| \left\| \frac{\xi}{2}\mathbf{1} \right\| \right| + \left| \sum_{i,j}^{k} \|(T + T')e_i\| \left\| \frac{\xi}{2}\mathbf{1} \right\| \right| \\
&\leq \sqrt{m}rk\xi + \sqrt{m}ck^2\xi \\
&= (r + ck)\sqrt{m}k\xi = \xi'.
\end{aligned}$$

The last inequality holds because of the triangle inequality. We have

$$\sum_{i,j}^{k} \|(T + T')e_i\| \leq \sum_{i,j}^{k} \left(\|Te_i\| + \|T'e_i\| \right) \leq \sum_{i,j}^{k} 2c = 2ck^2.$$

Let $F_{\mathcal{T}}$ denote the loss function class for the algorithms when searching for implementations $T \in \mathcal{T}$ and the metric d be the metric that $d(f_T(x), f_{T'}(x)) = \sup_x |f_T(x) - f_{T'}(x)|$. According to Definition 1, for $\forall f_T \in F_{\mathcal{T}}$, there is a $T' \in \mathcal{S}$ such that

$$\|d(f_T(X), f_{T'}(X))\|_1 = \left[\sum_{i=1}^{n} d(f_T(x_i), f_{T'}(x_i))\right] \leq n\xi'.$$

Thus,

$$\mathcal{N}_1(F_{\mathcal{T}}, \xi', n) \leq |\mathcal{S}| \leq \left(\frac{4c}{\xi}\right)^{mk} = \left(\frac{4(r+ck)\sqrt{m}ck}{\xi'}\right)^{mk}.$$

Taking log on both sides, we have

$$\ln \mathcal{N}_1(F_{\mathcal{T}}, \xi', n) \leq mk \ln\left(\frac{4(r+ck)\sqrt{m}ck}{\xi'}\right).$$

■

5.3 Proof of Theorem 3

We first prove the following theorem, which is useful to prove Theorem 3.

Theorem 14 *Let $X = \{x_1, \ldots, x_n\} \sim \mu^n$ be a set of independent random variables such that $f_T(x_i) \leq b$ for some $b > 0$ almost surely for all $f_T \in F_{\mathcal{T}}$ and $i \leq n$. Then for any $X \sim \mu^n$ and $\delta \in (0,1)$, with probability at least $1-\delta$, we have*

$$\sup_{f_T \in F_{\mathcal{T}}} |R(T) - R_n(T)| \leq \frac{2}{n} + b\sqrt{\frac{\ln \mathcal{N}_1(F_{\mathcal{T}}, 1/n, n) + \ln 2/\delta}{2n}},$$

where $R_n(T) = \frac{1}{n}\sum_{i=1}^{n} f_T(x_i)$ and $R(T) = E_x R_n(T)$.

Proof. Since $F_T(X) = \{f_T(x_1), \ldots, f_T(x_n)\}$ is a set of independent random variables, according to Hoeffding's inequality, for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have

$$|R(T) - R_n(T)| \leq b\sqrt{\frac{\ln 2/\delta}{2n}}.$$

Let $F_{\mathcal{T},\epsilon}$ be a minimal ϵ-cover of F_T. Then, $|F_{\mathcal{T},\epsilon}| = \mathcal{N}_1(F_{\mathcal{T}}, \epsilon, n)$. By a union bound of probability, we have that with probability at least $1-\delta$, the following holds

$$\sup_{f_T \in F_{\mathcal{T},\epsilon}} |R(T) - R_n(T)| \leq b\sqrt{\frac{\ln 2\mathcal{N}_1(F_{\mathcal{T}}, \epsilon, n)/\delta}{2n}} = b\sqrt{\frac{\ln \mathcal{N}_1(F_{\mathcal{T}}, \epsilon, n) + \ln 2/\delta}{2n}}. \quad (11)$$

It can be easily verified that

$$\sup_{f_T \in F_T} |R(T) - R_n(T)| \leq 2\epsilon + \sup_{f_T \in F_{T,\epsilon}} |R(T) - R_n(T)|. \quad (12)$$

Combine inequalities (11) and (12), and let $\epsilon = 1/n$, we have that with probability at least $1-\delta$, the following holds

$$\sup_{f_T \in F_{\mathcal{T}}} |R(T) - R_n(T)| \leq \frac{2}{n} + b\sqrt{\frac{\ln \mathcal{N}_1(F_{\mathcal{T}}, 1/n, n) + \ln 2/\delta}{2n}},$$

which concludes the proof. ■

Theorem 3 can be proven by combining Theorem 14 and Lemma 1. We can also prove Proposition 1 using the same method as that of Theorem 3.

5.4 Proof of Theorem 4

According to Bernstein's inequality, we have the following theorem, which is useful to prove Theorem 4.

Theorem 15 *Let $X = \{x_1, \ldots, x_n\} \sim \mu^n$ be a set of independent random variables such that $f_T(x_i) \leq 1$ almost surely for all $f_T \in F_{\mathcal{T}}$ and $i \leq n$. Then for any $X \sim \mu^n$ and $\delta \in (0,1)$, with probability at least $1-\delta$, we have*

$$\sup_{f_T \in F_{\mathcal{T}}} |R(T) - R_n(T)|$$
$$\leq \frac{2}{n} + \frac{5\left(\ln \mathcal{N}_1(F_{\mathcal{T}}, 1/n, n) + \ln 2/\delta\right)}{n} + \sqrt{\frac{2R_n(T)\left(\ln \mathcal{N}_1(F_{\mathcal{T}}, 1/n, n) + \ln 2/\delta\right)}{n}}.$$

Proof. Since $F_T(X) = \{f_T(x_1), \ldots, f_T(x_n)\}$ is a set of independent random variables, according to Bernstein's inequality, for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have

$$|R(T) - R_n(T)| \leq \frac{2\ln 2/\delta}{3n} + \sqrt{\frac{2V\ln 2/\delta}{n}}. \tag{13}$$

We also have that $V \leq R(T)$ because $Ef_T(x_i)^2 \leq Ef_T(x_i) = R(T)$. Collecting the terms in $R(T)$, completing the square and solving for $\sqrt{R(T)}$ shows that with probability at least $1-\delta$, we have

$$\sqrt{R(T)} \leq \sqrt{R_n(T)} + 3\sqrt{\frac{\ln 2/\delta}{n}}. \tag{14}$$

Straightforward substitution of inequality (14) into inequality (13) shows that with probability at least $1-\delta$, we have

$$|R(T) - R_n(T)| \leq \frac{5\ln 2/\delta}{n} + \sqrt{\frac{2R_n(T)\ln 2/\delta}{n}}.$$

Similar to the proof of Theorem 14, by a union bound of probability, we then have that with probability at least $1-\delta$, the following holds

$$\begin{aligned}&\sup_{f_T \in F_\mathcal{T}} |R(T) - R_n(T)| \\ &\leq \frac{2}{n} + \frac{5\left(\ln \mathcal{N}_1(F_\mathcal{T}, 1/n, n) + \ln 2/\delta\right)}{n} + \sqrt{\frac{2R_n(T)\left(\ln \mathcal{N}_1(F_\mathcal{T}, 1/n, n) + \ln 2/\delta\right)}{n}}.\end{aligned}$$

which concludes the proof. ■

Theorem 4 can be proven by combining Theorem 15 and Lemma 1.

5.5 Proof of Theorem 5

The following theorem, derived by exploiting Bennett's inequality, is essential to prove Theorem 5.

Theorem 16 *Let $X = \{x_1, \ldots, x_n\} \sim \mu^n$ be a set of independent random variables such that $f_T(x_i) \leq 1$ almost surely for all $f_T \in F_{\mathcal{T}}$ and $i \leq n$. Then for any $X \sim \mu^n$ and $\delta \in (0, 1)$, with probability at least $1 - \delta$ it holds for all $T \in \mathcal{T}$ that*

$$|R(T) - R_n(T)| \ \leq \frac{2}{n} + \left(\frac{\ln \mathcal{N}_1(F_{\mathcal{T}}, 1/n, n) + \ln 2/\delta}{\beta n}\right)^{\frac{1}{2-\frac{\ln(8\beta V/3)}{\ln|R(T)-R_n(T)|}}}$$

when V is no smaller than $|R(T) - R_n(T)|$ and there is a positive constant β such that $8\beta V < 3$.

Theorem 16 can be easily proven by using Berenstain's inequality. However, to show the faster convergence propery, we propose a new method to prove Berenstain's inequlity, which needs the following lemma.

Lemma 6 *For $\epsilon \in (0, 1]$ and $V \geq \epsilon$, there exists some $\beta > 0$ and $0 < \gamma < 2$ such that the following holds*

$$-Vnh\left(\frac{\epsilon}{V}\right) \leq -\beta n \epsilon^\gamma \leq O\left(-n\epsilon^2\right).$$

Let $\{x_1, \ldots, x_n\}$ be i.i.d. variables such that $x_i \leq 1$, $Ex_i^2 \leq V$ and $|R(T) - R_n(T)| \leq V$ are almost surely for all $i \leq n$. Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$, we have

$$|R(T) - R_n(T)| \leq \left(\frac{\ln 2/\delta}{\beta n}\right)^{\frac{1}{2-\frac{\ln(8\beta V/3)}{\ln|R(T)-R_n(T)|}}}.$$

Proof. We prove the first part. We have

$$\begin{aligned}
& -Vnh\left(\frac{\epsilon}{V}\right) \leq -\beta n \epsilon^\gamma \\
\iff \quad & V\left(\left(1 + \frac{\epsilon}{V}\right)\ln\left(1 + \frac{\epsilon}{V}\right) - \frac{\epsilon}{V}\right) \geq \beta \epsilon^\gamma \\
& \text{(Because that } \epsilon < 1\text{)} \\
\iff \quad & \gamma \geq \frac{\ln\left(\frac{V}{\beta}\left(\left(1 + \frac{\epsilon}{V}\right)\ln\left(1 + \frac{\epsilon}{V}\right) - \frac{\epsilon}{V}\right)\right)}{\ln \epsilon}.
\end{aligned}$$

It holds that

$$\begin{aligned}
&\frac{\ln\left(\frac{V}{\beta}\left(\left(1+\frac{\epsilon}{V}\right)\ln\left(1+\frac{\epsilon}{V}\right)-\frac{\epsilon}{V}\right)\right)}{\ln\epsilon}\\
&\quad\left(\text{Because } (1+x)\ln(1+x)\geq\frac{1}{2+\frac{2x}{3}}x^2+x \text{ for } x\geq 0\right)\\
&\leq\frac{\ln\left(\frac{V}{\beta}\frac{3}{6+\frac{2\epsilon}{V}}\left(\frac{\epsilon}{V}\right)^2\right)}{\ln\epsilon}=\frac{\ln\left(\frac{3\epsilon^2}{\beta(6V+2\epsilon)}\right)}{\ln\epsilon}\\
&=2-\frac{\ln\left(2\beta(V+\frac{\epsilon}{3})\right)}{\ln\epsilon}\\
&\leq 2,\ \text{when } \epsilon\leq V \text{ and } 8\beta V<3.
\end{aligned}$$

Thus, there are many pairs of (β,γ) such that the first part of Lemma 6 holds.

We then prove Berenstain's inequality and the second part. According to Bennett's inequality, we have

$$\begin{aligned}
P\left\{|R(T)-R_n(T)|\geq\epsilon\right\} &\leq 2\exp\left(-nVh\left(\frac{\epsilon}{V}\right)\right)\\
&\leq 2\exp\left(-\beta n\epsilon^{2-\frac{\ln\left(2\beta(V+\frac{\epsilon}{3})\right)}{\ln\epsilon}}\right)\\
&=2\exp\left(\frac{-n\epsilon^2}{2(V+\frac{\epsilon}{3})}\right),
\end{aligned}\tag{15}$$

which is the Berenstain's inequality.

To prove the second part, let $\epsilon<V$. We have

$$\begin{aligned}
P\left\{|R(T)-R_n(T)|\geq\epsilon\right\} &\leq 2\exp\left(\frac{-n\epsilon^2}{2(V+\frac{\epsilon}{3})}\right)\\
&\leq 2\exp\left(\frac{-n\epsilon^2}{2(V+\frac{V}{3})}\right)\\
&=2\exp\left(-\beta n\epsilon^{2-\frac{\ln\left(\frac{8\beta V}{3}\right)}{\ln\epsilon}}\right).
\end{aligned}$$

For any $\delta\in(0,1)$, let

$$2\exp\left(-\beta n\epsilon^{2-\frac{\ln\left(\frac{8\beta V}{3}\right)}{\ln\epsilon}}\right)=\delta. \tag{16}$$

Then, with probability at least $1-\delta$, we have

$$|R(T)-R_n(T)| \leq \epsilon. \tag{17}$$

Combining (16) and (17), with probability at least $1-\delta$, we have

$$\frac{\ln 2/\delta}{\beta n} = \epsilon^{2-\frac{\ln\left(\frac{8\beta V}{3}\right)}{\ln\epsilon}} \geq \epsilon^{2-\frac{\ln\left(\frac{8\beta V}{3}\right)}{\ln|R(T)-R_n(T)|}}$$

and

$$\epsilon \leq \left(\frac{\ln 2/\delta}{\beta n}\right)^{\frac{1}{2-\frac{\ln\left(\frac{8\beta V}{3}\right)}{\ln|R(T)-R_n(T)|}}}. \tag{18}$$

Combining (17) and (18), with probability at least $1-\delta$, we have

$$|R(T)-R_n(T)| \leq \left(\frac{\ln 2/\delta}{\beta n}\right)^{\frac{1}{2-\frac{\ln\left(\frac{8\beta V}{3}\right)}{\ln|R(T)-R_n(T)|}}}.$$

Thus, the Second part of Lemma 6 holds. ■

Similar to the proof of Theorem 14, Theorem 16 can be proven by using Lemma 6 and a union bound of probability.

Theorem 5 can be proven by combining Theorem 16 and Lemma 1.

5.6 Proof of Lemma 2

The proof method is the same as that of Lemma 2 in (Maurer & Pontil, 2010).

Proof. Let

$$h(y) = \left\| x - \sum_{i=1}^{k} T_i y_i \right\|^2.$$

Assume that y is a minimizer of h and $\|y\| > r$. Because T is normalized, $\|T_i\| = 1, i, \ldots, k$. Then

$$\left\| \sum_{i=1}^{k} T_i y_i \right\|^2 = \|y\|^2 + \sum_{i \neq j} y_i y_j \langle T_i, T_j \rangle > r^2.$$

Let the real-valued function f be defined as

$$f(t) = h(ty).$$

Then

$$\begin{aligned} f'(1) &= 2\left(\left\|\sum_{i=1}^{k} T_i y_i\right\|^2 - \left\langle x, \sum_{i=1}^{k} T_i y_i \right\rangle\right) \\ &\text{(Using Cauchy-Schwarz inequality)} \\ &\geq 2\left(\left\|\sum_{i=1}^{k} T_i y_i\right\|^2 - r\left\|\sum_{i=1}^{k} T_i y_i\right\|\right) \\ &= 2\left(\left\|\sum_{i=1}^{k} T_i y_i\right\| - r\right)\left\|\sum_{i=1}^{k} T_i y_i\right\| > 0. \end{aligned}$$

So f cannot have a minimum at 1, whence y cannot be a minimizer of h. Thus, the minimizer y must be contained in the ball with radius r in the m-dimensional space. ■

5.7 Proof of Lemma 4

Proof. As in the proof of Lemma 1, we can pick out a set $\mathcal{S}$, where $|\mathcal{S}| \leq \left(\frac{4c}{\xi}\right)^{mk}$, having the property that for every T, there exists a $T' \in \mathcal{S}$ such that $\sup_x |f_T(x) - f_{T'}(x)| \leq \xi'$ with $\xi' = (rs + cs^2 k^{1-1/p})\sqrt{m}\xi k^{1-1/p}$. The detail is as follows.

$$\begin{aligned} |f_T - f_{T'}| &= \left|\min_y \|x - Ty\|^2 - \min_y \|x - T'y\|^2\right| \\ &\leq \left|\max_y \left(\|x - Ty\|^2 - \|x - T'y\|^2\right)\right| \\ &\leq \left|\max_y 2x^\top Ty - 2x^\top T'y\right| + \left|\max_y \|Ty\|^2 - \|T'y\|^2\right| \qquad (19) \\ &= \left|\max_y \sum_{i=1}^{k} y_i \langle 2x, (T - T')e_i \rangle\right| + \left|\max_y \sum_{i,j}^{k} y_i y_j \langle (T + T')e_i, (T - T')e_j \rangle\right|. \end{aligned}$$

Using Hölder's inequality, we have

$$\begin{aligned}
&\left|\max_y \sum_{i=1}^{k} y_i \left\langle 2x, (T-T')e_i \right\rangle\right| \\
&\leq \left|\max_y \|y\|_p \left(\sum_{i=1}^{k} |\langle 2x, (T-T')e_i\rangle|^q\right)^{1/q}\right| \qquad (20) \\
&\leq \left|\max_y \|y\|_p \left(\sum_{i=1}^{k} |\|2x\|\|(T-T')e_i\||^q\right)^{1/q}\right| \\
&\leq \sqrt{m}sr\xi k^{1/q} \\
&\leq \sqrt{m}sr\xi k^{1-1/p}.
\end{aligned}$$

Using Hölder's inequality again, we have inequalities (21) and (22):

$$\begin{aligned}
&\left|\max_y \sum_{i,j}^{k} y_i y_j \left\langle (T+T')e_i, (T-T')e_j \right\rangle\right| \\
&\leq \left|\max_y \|y\|_p \left(\sum_{i}^{k} \left|\sum_{j}^{k} \left\langle (T+T')e_i, (T-T')e_j \right\rangle y_j\right|^q\right)^{1/q}\right|, \qquad (21)
\end{aligned}$$

and

$$\begin{aligned}
&\left|\sum_{j}^{k} \left\langle (T+T')e_i, (T-T')e_j \right\rangle y_j\right| \\
&\leq \left|\left(\sum_{j}^{k} \left\langle (T+T')e_i, (T-T')e_j \right\rangle^q\right)^{1/q} \left(\sum_{j}^{k} |y_j|^p\right)^{1/p}\right| \\
&\leq \left|\left(\sum_{j}^{k} (\|(T+T')e_i\|\|(T-T')e_j\|)^q\right)^{1/q} \left(\sum_{j}^{k} |y_j|^p\right)^{1/p}\right| \\
&\leq \left|\left(\sum_{j}^{k} ((\|Te_i\| + \|T'e_i\|)\|(T-T')e_j\|)^q\right)^{1/q} \left(\sum_{j}^{k} |y_j|^p\right)^{1/p}\right| \qquad (22) \\
&\leq \sqrt{m}sc\xi k^{1/q} = \sqrt{m}sc\xi k^{1-1/p}.
\end{aligned}$$

Combining inequalities (21) and (22), it gives

$$\begin{aligned}
&\left|\max_y \sum_{i,j}^{k} y_i y_j \left\langle (T+T')e_i, (T-T')e_j \right\rangle \right| \\
&\leq \left|\max_y \|y\|_p \left(\sum_{i}^{k} \left|\sqrt{m} s c \xi k^{1-1/p}\right|^q \right)^{1/q} \right| \\
&\leq \sqrt{m} s^2 c \xi k^{2-2/p}.
\end{aligned} \tag{23}$$

Combining inequalities (19), (20) and (23), we have

$$\begin{aligned}
|f_T - f_{T'}| &\leq \left|\max_y \sum_{i=1}^{k} y_i \left\langle 2x, (T-T')e_i \right\rangle \right| \\
&\quad + \left|\max_y \sum_{i,j}^{k} y_i y_j \left\langle (T+T')e_i, (T-T')e_j \right\rangle \right| \\
&\leq \sqrt{m} s r \xi k^{1-1/p} + \sqrt{m} s^2 c \xi k^{2-2/p} \\
&= (rs + cs^2 k^{1-1/p}) \sqrt{m} \xi k^{1-1/p} = \xi'.
\end{aligned}$$

According to Definition 1, for $\forall f_T \in F_{\mathcal{T}}$, there is a $T' \in \mathcal{S}$ such that

$$\|d(f_T(X), f_{T'}(X))\|_1 = \left[\sum_{i=1}^{2} d(f_T(x_i), f_{T'}(x_i))\right] \leq 2\xi'.$$

Thus,

$$\mathcal{N}_1(F_{\mathcal{T}}, \xi', n) \leq |\mathcal{S}| \leq \left(\frac{4c}{\xi}\right)^{mk} = \left(\frac{4(rs + cs^2 k^{1-1/p})\sqrt{m} c k^{1-1/p}}{\xi'}\right)^{mk}.$$

Taking log on both sides, we have

$$\ln \mathcal{N}_1(F_{\mathcal{T}}, \xi', n) \leq mk \ln \left(\frac{4(rs + cs^2 k^{1-1/p})\sqrt{m} c k^{1-1/p}}{\xi'}\right),$$

which concludes the proof. ■

5.8 Proof of Lemma 5

The proof method of Lemma 5 is similar to that of Lemma 1.

Proof. For k-means clustering and vector quantization, we can easily prove that $\|Te_i\| \leq r, i = 1, \ldots, k$. As in the proof of Lemma 1 and Lemma 4, we can pick out a set $\mathcal{S}$, where $|\mathcal{S}| \leq \left(\frac{4r}{\xi}\right)^{mk}$, having the property that for every T there exists a $T' \in \mathcal{S}$ such that $\sup_x |f_T(x) - f_{T'}(x)| \leq \xi'$ with $\xi' = 2r\sqrt{m}\xi$. The proof is as follows:

$$
\begin{aligned}
&|f_T - f_{T'}| \\
&\leq \left| \max_{i \in \{1,\ldots,k\}} \left(\|x - Te_i\|^2 - \|x - T'e_i\|^2\right) \right| \\
&\leq \left| \max_{i \in \{1,\ldots,k\}} 2x^\top Te_i - 2x^\top T'e_i \right| + \left| \max_{i \in \{1,\ldots,k\}} \|Te_i\|^2 - \|T'e_i\|^2 \right| \\
&= \left| \max_{i \in \{1,\ldots,k\}} \langle 2x, (T - T')e_i \rangle \right| + \left| \max_{i \in \{1,\ldots,k\}} \langle (T + T')e_i, (T - T')e_i \rangle \right| \\
&\quad \text{(Using Cauchy-Schwarz inequality)} \\
&\leq \left| \max_{i \in \{1,\ldots,k\}} \|2x\| \|(T - T')e_i\| \right| + \left| \max_{i \in \{1,\ldots,k\}} \left(\|Te_i\| + \|T'e_i\|\right) \|(T - T')e_i\| \right| \\
&\leq \sqrt{m} r\xi + \sqrt{m} r\xi \\
&= 2r\sqrt{m}\xi = \xi'.
\end{aligned}
$$

Thus,

$$
\mathcal{N}_1(F_\mathcal{T}, \xi', n) \leq |\mathcal{S}| \leq \left(\frac{4r}{\xi}\right)^{mk} = \left(\frac{8r^2\sqrt{m}}{\xi'}\right)^{mk}.
$$

Taking log on both sides, we have

$$
\ln \mathcal{N}_1(F_\mathcal{T}, \xi', n) \leq mk \ln\left(\frac{8r^2\sqrt{m}}{\xi'}\right),
$$

which concludes the proof. ■

6 Conclusion

Here we propose a method to analyze the dimensionality-dependent generalization bounds for k-dimensional coding schemes, which are the abstract and general descriptions of a set of methods that encode random vectors in Hilbert space $\mathcal{H}$. There are several specific forms of k-dimensional coding schemes, including NMF, dictionary learning, sparse coding, k-means clustering and vector quantization, which have achieved great successes in pattern recognition and machine learning.

Our proof approach is based on an upper bound for the covering number of the loss function class induced by the reconstruction error. We explained that the covering number is more suitable for deriving dimensionality-dependent generalization bounds for k-dimensional coding schemes, because it avoids the worst case dependency *w.r.t.* the number k of the columns of the linear implementation. If k is larger than the dimensionality m, our bound could be much tighter than the dimensionality-independent generalization bound. Moreover, according to Bennett's inequality, we derived a dimensionality-dependent generalization bound of order $\mathcal{O}\left(mk\ln(mkn)/n\right)^{\lambda_n}$, where $\lambda_n > 0.5$ when the sample size n is finite, for k-dimensional coding schemes. Our method therefore provides state-of-the-art dimensionality-dependent generalization bounds for NMF, dictionary learning, sparse coding, k-means clustering and vector quantization.

References

Abbott, L., & Dayan, P. (1999). The effect of correlated variability on the accuracy of a population code. *Neural computation*, *11*(1), 91–101.

Alexander, K. (1984). Probability inequalities for empirical processes and a law of the iterated logarithm. *Annals of Probability*, *12*(4), 1041–1067.

Amiri, A., & Haykin, S. (2014). Improved sparse coding under the influence of perceptual attention. *Neural computation*, *26*(2), 377–420.

Anderberg, M. R. (1973). *Cluster analysis for applications*. Academic Press.

Antos, A. (2005). Improved minimax bounds on the test and training distortion of empirically designed vector quantizers. *IEEE Transactions on Information Theory*, *51*(11), 4022-4032.

Antos, A., Györfi, L., & György, A. (2005). Improved convergence rates in empirical vector quantizer design. *IEEE Transactions on Information Theory*, *51*(11), 4013-4022.

Bartlett, P. L., Linder, T., & Lugosi, G. (1998). The minimax distortion redundancy in empirical quantizer design. *IEEE Transactions on Information Theory*, *44*(5), 1802-1813.

Bartlett, P. L., & Mendelson, S. (2003). Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, *3*, 463–482.

Biau, G., Devroye, L., & Lugosi, G. (2008). On the performance of clustering in hilbert spaces. *IEEE Transactions on Information Theory*, *54*(2), 781–790.

Boucheron, S., Lugosi, G., & Massart, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press.

Chen, S. S., Donoho, D. L., & Saunders, M. A. (1999). Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, *20*(1), 33–61.

Chou, P. A. (1994). The distortion of vector quantizers trained on n vectors decreases to the optimum at $\mathcal{O}_p(1/n)$. In *Proceedings of ISIT*.

Cucker, F., & Smale, S. (2002). On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, *39*(1), 1–49.

Devroye, L., Györfi, L., & Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer.

Dhillon, I. S., Guan, Y., & Kulis, B. (2007). Weighted graph cuts without eigenvectors a multilevel approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(11), 1944–1957.

Ding, C., He, X., & Simon, H. D. (2005). On the equivalence of nonnegative matrix factorization and spectral clustering. In *Proceedings of ICDM*.

Févotte, C., Bertin, N., & Durrieu, J.-L. (2009). Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis. *Neural computation*, *21*(3), 793–830.

Gillis, N., & Vavasis, S. A. (2014). Fast and robust recursive algorithms for separable nonnegative matrix factorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(4), 698–714.

Gray, R. (1984). Vector quantization. *IEEE Acoustics, Speech and Signal Processing Magazine*, *1*(2), 4–29.

Gribonval, R., Jenatton, R., Bach, F., Kleinsteuber, M., & Seibert, M. (2015). Sample complexity of dictionary learning and other matrix factorizations. *IEEE Transactions on Information Theory*, *61*(6), 3469–3486.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, *58*(301), 13–30.

Hunt, J. J., Ibbotson, M., & Goodhill, G. J. (2012). Sparse coding on the spot: Spontaneous retinal waves suffice for orientation selectivity. *Neural computation*, *24*(9), 2422–2433.

Ivana, T., & Pascal, F. (2011). Dictionary learning: What is the right representation for my signal? *IEEE Signal Processing Magazine*, *4*(2), 27–38.

Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y. (2002). An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(7), 881–892.

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by nonnegative matrix factorization. *Nature*, *401*(6755), 788–791.

Levrard, C., et al. (2013). Fast rates for empirical vector quantization. *Electronic Journal of Statistics*, *7*, 1716–1746.

Levrard, C., et al. (2015). Nonasymptotic bounds for vector quantization in hilbert spaces. *The Annals of Statistics*, *43*(2), 592–619.

Linder, T. (2000). On the training distortion of vector quantizers. *IEEE Transactions on Information Theory*, *46*(4), 1617-1623.

Linder, T., Lugosi, G., & Zeger, K. (1994). Rates of convergence in the source coding theorem, in empirical quantizer design, and in universal lossy source coding. *IEEE Transactions on Information Theory*, *40*(6), 1728-1740.

MacQueen, J., et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth berkeley symposium on mathematical statistics and probability.*

Mairal, J., Bach, F., & Ponce, J. (2012). Task-driven dictionary learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*(4), 791–804.

Maurer, A., & Pontil, M. (2010). K-dimensional coding schemes in Hilbert spaces. *IEEE Transactions on Information Theory*, *56*(11), 5839-5846.

Maurer, A., Pontil, M., & Romera-Paredes, B. (2013). Sparse coding for multitask and transfer learning. In *Proceedings of ICML.*

Mehta, N., & Gray, A. G. (2013). Sparsity-based generalization bounds for predictive sparse coding. In *Proceedings of ICML.*

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, *381*(6583), 607–609.

Pehlevan, C., Hu, T., & Chklovskii, D. B. (2015). A hebbian/anti-hebbian neural network for linear subspace learning: A derivation from multidimensional scaling of streaming data. *Neural computation*, *27*(1), 1461–1495.

Pollard, D. (1982). A central limit theorem for k-means clustering. *IEEE Transactions on Information Theory*, *10*(4), 912-926.

Quiroga, R. Q., Nadasdy, Z., & Ben-Shaul, Y. (2004). Unsupervised spike detection and sorting with wavelets and superparamagnetic clustering. *Neural computation*, *16*(8), 1661–1687.

Schneider, P., Biehl, M., & Hammer, B. (2009a). Adaptive relevance matrices in learning vector quantization. *Neural Computation*, *21*(12), 3532–3561.

Schneider, P., Biehl, M., & Hammer, B. (2009b). Distance learning in discriminative vector quantization. *Neural Computation*, *21*(10), 2942–2969.

Talagrand, M. (1994). Sharper bounds for gaussian and empirical processes. *Annals of Probability*, *22*(1), 28–76.

Vainsencher, D., Mannor, S., & Bruckstein, A. M. (2011). The sample complexity of dictionary learning. *Journal of Machine Learning Research*, *12*, 3259–3281.

Vapnik, V. (2000). *The nature of statistical learning theory*. Springer.

Wright, J., Yang, A. Y., Ganesh, A., Sastry, S. S., & Ma, Y. (2009). Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(2), 210–227.

Xu, M., & Lafferty, J. D. (2012). Conditional sparse coding and grouped multivariate regression. In *Proceedings of ICML.*

Zhang, C. (2013). Bennett type generalization bounds: large deviation case and faster rate of convergence. In *Proceedings of UAI.*

Zhang, T. (2002). Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, *2*, 527–550.