Chapter 24
Lower Bounds for Sparse Coding

Andreas Maurer, Massimiliano Pontil and Luca Baldassarre

Abstract We give lower bounds on the reconstruction error for PCA, k-means clustering, and various sparse coding methods. It is shown that the two objectives of good data approximation and sparsity of the solution are incompatible if the data distribution is evasive in the sense that most of its mass lies away from any low dimensional subspace. We give closure properties and examples of evasive distributions and quantify the extent to which distributions of bounded support and bounded density are evasive.

24.1 Introduction

Much recent work in machine learning and signal processing has concentrated on the problem of approximating high dimensional data $x \in \mathbb{R}^N$ by sparse linear combinations of the columns of a dictionary matrix[1] $D = [d_1, ..., d_K] \in \mathbb{R}^{N \times K}$—see, for example, [3, 4, 6–12] and references therein. For a fixed dictionary D every such linear combination has the form

$$Dy = \sum_{i=1}^{K} y_i d_i,$$

[1] Throughout the chapter, with some abuse of notation we use D to denote both the dictionary matrix and the dictionary $D = \{d_1, \ldots, d_K\} \subseteq \mathbb{R}^N$.

A. Maurer
Adalbertstrasse 55, D-80799 Munich, Germany
e-mail: am@andreas-maurer.eu

M. Pontil (✉)
University College London, Malet Place, London WC1E, UK
e-mail: m.pontil@cs.ucl.ac.uk

L. Baldassarre
LIONS, EPFL, 1015 Lausanne, Switzerland
e-mail: l.baldassarre@cs.ucl.ac.uk

V. Vovk et al. (eds.), *Measures of Complexity*,
DOI 10.1007/978-3-319-21852-6_24

where the vector of coefficients y is chosen to be a solution of the optimisation problem

$$\min_{y \in \Lambda} \|x - Dy\|^2 . \tag{24.1}$$

Here $\Lambda \subseteq \mathbb{R}^K$ is a fixed regularizing set, which implements constraints on the complexity of the chosen representations. We denote by $y(x)$ such a solution, and it is one inherent objective that the vectors $y\,(x)$ obtained should be sparse, in that the number of their nonzero coefficients is much smaller than the ambient dimension N. If $y\,(x)$ is not sparse itself, it should at least have a close sparse approximation.

We assume that the data x are distributed at random according to a distribution μ on $\mathbb{R}^N$ corresponding to a random variable X taking values in $\mathbb{R}^N$. The reconstruction error $\|X - Dy\,(X)\|^2$ is then a random variable and its expectation

$$R\,(D) = E\,\|X - Dy\,(X)\|^2$$

measures the failure of D to match the distribution μ. Thus, given Λ, one wishes to choose D so as to minimize $R\,(D)$.

In this chapter we show that these methods are likely to produce poor results for a large class of seemingly well-behaved distributions on $\mathbb{R}^N$, because the two objectives are incompatible: With high probability the representing vector $y\,(X)$ is either not very sparse (or does not have a good sparse approximation), or the reconstruction error is large. Our negative results are independent of any problem of sample-based estimation and still hold if we have complete knowledge of the distribution μ.

The "bad" distributions μ have the following property of *evasiveness*: For any low dimensional subspace M, the overwhelming majority of the mass of μ is bounded away from M. Below we use the notation $d\,(x, M) = \inf_{z \in M} \|x - z\|$.

Definition 24.1 Suppose α_k is a nonincreasing sequence of positive real numbers, $\beta, C > 0$. A random variable X with values in $\mathbb{R}^N$ is said to be (α, β, C)-evasive if for every $k < N$, every k-dimensional subspace M of $\mathbb{R}^N$ and every $t \in (0, \alpha_k)$

$$\Pr\left\{d\,(X, M)^2 \leq \alpha_k - t\right\} \leq Ce^{-N\beta t^2}.$$

A probability measure μ on $\mathbb{R}^N$ is called (α, β, C)-evasive if the corresponding random variable is.

We give two examples:

Example 24.1 (Noisy generative model) If Y is any random variable in $\mathbb{R}^N$ and Z is a centered isotropic Gaussian with variance σ^2 and independent of Y, then the random variable $X = Y + Z$ is evasive with

$$\alpha_k = \sigma^2 \frac{N - k - \pi^2 - 1}{N},\ \beta = \frac{1}{2\sigma^4\pi^2},\ C = 2,$$

as will be shown in Sect. 24.3.2. With Y taking values in a union of low dimensional subspaces generated by some potentially unknown dictionary, the random variable X can be viewed as a generative model contaminated by noise. Here we will prove lower bounds in the order of σ^2.

Example 24.2 (Bounded support and bounded density) While the previous example is of a special form, this example is more generic. If a distribution μ has support in the unit ball B_N of $\mathbb{R}^N$ and a bounded density $d\mu/d\rho$ with respect to the uniform measure ρ on B_N, then μ is evasive with

$$\alpha_k = \left\| \frac{d\mu}{d\rho} \right\|_\infty^{\frac{-2}{N-k}} \frac{N-k}{e^{3/2}N}, \ \beta = 1, \ C = 1,$$

where $\|.\|_\infty$ is the essential supremum norm w.r.t. ρ. This will be shown in Theorem 24.3 below.

We come to the announced negative results. Suppose first that in (24.1) a hard sparsity constraint is implemented by the regularizing set

$$\Lambda_s = \left\{ y \in \mathbb{R}^K : \|y\|_0 \leq s \right\}, \tag{24.2}$$

where $\|y\|_0$ is the number of nonzero components of y and s is any integer $s \leq K$. An easy union bound then gives the following result:

Theorem 24.1 *Let $D = [d_1, ..., d_K] \in \mathbb{R}^{N \times K}$ be any dictionary and suppose that X is (α, β, C)-evasive. Then for any integer $s \leq K$ and $t \in (0, \alpha_s)$*

$$\Pr_{X \sim \mu} \left\{ \min_{y \in \Lambda_s} \|X - Dy\|^2 \leq \alpha_s - t \right\} \leq C \exp\left(-N\beta t^2 + s \ln K \right). \tag{24.3}$$

If $s \ln K \ll N$ the reconstruction error is bounded away from zero with high probability.

We might hope to improve this situation by requiring the encoding vectors y to be sparse only in some approximate sense. The next result holds for all vectors $y \in \mathbb{R}^K$, sparse or not, and exhibits a tradeoff between the quality of two approximations: the approximation of the data by Dy and the ℓ_1-approximation of y by its nearest vector of prescribed sparsity. For $y = (y_1, \ldots, y_K) \in \mathbb{R}^K$ and $s < K$ let y^s denote the s-sparse approximation of y, obtained by setting all components y_i equal to zero except for the s indices where $|y_i|$ is largest.

Theorem 24.2 *Let D be any dictionary with $\|d_i\| = \|De_i\| \leq B$ and suppose that X is (α, β, C)-evasive. Then for every $\delta \in (0, 1)$ we have with probability at least $1 - \delta$ for every $y \in \mathbb{R}^K$ and every $s \in \{1, ..., K\}$ that*

$$\|X - Dy\|^2 \geq \frac{\alpha_s}{2} - \frac{1}{2}\sqrt{\frac{(1+s)\ln K + \ln \frac{C}{\delta}}{N\beta}} - B^2 \left\| y - y^s \right\|_1^2.$$

In many applications we can assume $B = 1$. So if $s \ln K \ll N$ and $\|y - y^s\|_1$ is small (so that y has a close s-sparse approximation) then the reconstruction error is of order α_s.

Below we use these results on PCA, K-means clustering and sparse coding and delimit the class of distributions to which these methods of unsupervised learning can be successfully applied.

24.2 Applications and Examples

The framework described in the introduction is general enough to capture many approaches to unsupervised learning.

24.2.1 PCA

In problem (24.1), if Λ is all of $\mathbb{R}^K$ with $K = s \ll N$, then an optimal D is found to be an isometry from $\mathbb{R}^K$ to a maximal K-dimensional subspace of the covariance of X. The resulting method is PCA, and trivially every representing vector is s-sparse, namely $y(x)$ has at most $s = K$ nonzero components.

We could apply Theorem 24.1, but this would incur a superfluous term $K \ln K$ in the exponent of the bound. Instead, by directly appealing to the definition, we find that for (α, β, C)-evasive X and any dictionary D

$$\Pr\left\{\min_{y\in\mathbb{R}^K} \|X - Dy\|^2 < \alpha_K - t\right\} \leq Ce^{-\beta N t^2}.$$

An illustration of the evasiveness of bounded densities (Example 24.2 above) is the following: Suppose we do PCA in one thousand dimensions, and we know that the data distribution is contained in the unit ball. If we find a 100-dimensional subspace which achieves an expected reconstruction error of ≈ 0.1, then the supremum of the distribution density (if such exists, and relative to the uniform measure on the ball) must be at least in the order of 10^{45}. The supremum relative to the Lebesgue measure must be at least $10^{45}/V_{1000} \approx 10^{1800}$, where V_N is the volume of the unit ball in $\mathbb{R}^N$. To derive this we use $(\alpha_K - t)\left(1 - C\exp\left(-\beta N t^2\right)\right)$ as a simple lower bound on the expected reconstruction error with $t = 0.05$, $N = 1000$, $K = 100$, $\beta = 1$, $C = 1$, equate the bound to 0.1, and solve for the bound on the density.

24.2.2 K-means Clustering

At the other extreme from PCA, if $\Lambda = \{e_1, ..., e_K\}$ is a basis for $\mathbb{R}^K$, then the method becomes K-means clustering or vector quantization, and the optimal dictionary atoms $d_1, ..., d_K$ are just the optimal centers. In this case the complexity constraint can be seen as a maximal sparsity requirement, as every member y of Λ satisfies $\|y\|_0 = 1$, but we may now allow $K > N$.

With Λ_s defined as in (24.2) we find $\{e_1, ..., e_K\} \subseteq \Lambda_1$, so appealing to Theorem 24.1 we find for (α, β, C)-evasive X and any dictionary D

$$\begin{aligned}\Pr\left\{\min_{y\in\{e_1,...,e_K\}} \|X - Dy\|^2 < \alpha_1 - t\right\} &\leq \Pr\left\{\min_{y\in\Lambda_1} \|X - Dy\|^2 < \alpha_1 - t\right\} \\ &\leq C\exp\left(-N\beta t^2 + \ln K\right).\end{aligned}$$

Of course there is a slight giveaway here, because Λ_1 is somewhat more expressive than $\{e_1, ..., e_K\}$.

24.2.3 Sparse Coding Methods

To make Λ more expressive we can relax the extreme sparsity constraint, setting $\Lambda = \Lambda_s$ with $1 \leq s \ll N$. This is the situation directly addressed by Theorem 24.1, which immediately gives a lower error bound. The corresponding method is not very practical, however, because of the unwieldy nature of the ℓ_0-function.

The alternative is to replace (24.1) with the following optimization problem

$$\min_{y\in\mathbb{R}^K} \|x - Dy\|^2 + \gamma \|y\|_1, \tag{24.4}$$

where γ is some positive constant, thus encouraging sparsity through the use of the ℓ_1-norm regularizer. A large body of work has been dedicated to the study of this and related methods, the success of which depends on different coherence properties of D, see [1–3] and references therein. The search for an optimal D in this case corresponds to the sparse coding method proposed by Olshausen and Field [10], which was originally motivated by neurophysiological studies of the mammalian visual system.

A similar approach uses the initial formulation (24.1) and takes Λ to be a multiple of the ℓ_1-unit ball. It relates to (24.4) as Ivanov regularization relates to Tychonov regularization.

Another example in this suite of methods is nonnegative matrix factorization, as proposed by Lee and Seung [6], where the d_i are required to be in the positive orthant of $\mathbb{R}^N$.

Theorem 24.2 immediately applies to all these cases and shows that for evasive distributions the requirements of good data approximation and approximate sparsity are incompatible.

24.3 Proofs

We review our notation and then prove the announced results.

For every vector $y \in \mathbb{R}^K$, we let $\|y\|_0$ denote the number of nonzero components of y. We say that y is s-sparse if $\|y\|_0 = s$. We denote by y^s an s-sparse vector which is nearest to y according to the ℓ_1 norm. For every linear subspace M of $\mathbb{R}^N$, we let P_M be the corresponding projection matrix and define $d(x, M) = \inf_{z \in M} \|x - z\|$, namely the distance of x to the linear subspace M. Note that $d(x, M) = \|P_{M^\perp} x\|$, where $M^\perp$ is the orthogonal complement of M. We denote by $\|\cdot\|$ the ℓ_2 norm of a vector and by $|||\cdot|||$ the operator norm of a matrix. For every $n \in \mathbb{N}$, we let B_n be the unit ball in $\mathbb{R}^n$ and let V_n be its volume.

If ν and ρ are measures on the same space and $\nu(A) = 0$ for every A satisfying $\rho(A) = 0$, then ν is called absolutely continuous w.r.t. ρ and there exists a measurable density function $d\mu/d\rho$, called the Radon-Nykodym derivative, such that $d\nu = (d\nu/d\rho)\, d\rho$.

24.3.1 Limitations of Sparse Coding

We prove Theorems 24.1 and 24.2.

Proof (Proof of Theorem 24.1) For $S \subseteq \{1, ..., K\}$ let M_S denote the subspace spanned by the d_i with $i \in S$. In the event on the left-hand side of (24.3) there is some subset $S \subseteq \{1, ..., K\}$ with cardinality $|S| \leq s$ such that $d(X, M_S)^2 \leq \alpha_s - t$. The dimension of M_S is at most s, so we get from a union bound that

$$\begin{aligned} &\Pr\left\{\min_{y:\|y\|_0 \leq s} \|X - Dy\|^2 \leq \alpha_s - t\right\} \\ &\leq \Pr\left\{\exists S \subseteq \{1, ..., K\}, |S| \leq s, d(X, M_S)^2 \leq \alpha_s - t\right\} \\ &\leq \binom{K}{s} C e^{-N\beta t^2} \leq C \exp\left(-N\beta t^2 + s \ln K\right). \end{aligned}$$

□

Proof (Proof of Theorem 24.2) Denote

$$t_s(s) = \min\left\{\alpha_s, \sqrt{((1+s)\ln K + \ln(C/\delta))/(N\beta)}\right\}.$$

For any $s \in \{1, ..., K\}$, $x \in \mathbb{R}^N$ and $y \in \mathbb{R}^K$ we have the following sequence of implications:

$$\|x - Dy\|^2 < \frac{\alpha_s - t(s)}{2} - B^2 \left\|y - y^s\right\|_1^2 \quad (24.5)$$

$$\text{implies } \|x - Dy\| + B\left\|y - y^s\right\|_1 < \sqrt{\alpha_s - t(s)} \quad (24.6)$$

$$\text{implies } \|x - Dy\| + \left\|D\left(y - y^s\right)\right\|_1 < \sqrt{\alpha_s - t(s)} \quad (24.7)$$

$$\text{implies } \left\|x - Dy^s\right\|^2 < \alpha_s - t(s). \quad (24.8)$$

(24.5) $\Longrightarrow$ (24.6) follows from $(a+b)^2 \leq 2a^2 + 2b^2$, (24.6) $\Longrightarrow$ (24.7) from $\|Dy\| \leq B\|y\|_1$, because of the bound on $\|d_i\|$, and (24.7) $\Longrightarrow$ (24.8) from the triangle inequality. Thus

$$\begin{aligned}
&\Pr\left\{\exists s \in \mathbb{N}, \exists y \in \mathbb{R}^n, \ \|X - Dy\|^2 \leq \frac{\alpha_s - t(s)}{2} - B^2\left\|y - y^s\right\|_1^2\right\} \\
&\leq \Pr\left\{\exists s \in \mathbb{N}, \exists y \in \mathbb{R}^n, \|y\|_0 \leq s, \|X - Dy\|^2 \leq \alpha_s - t(s)\right\} \\
&\leq \sum_{s=1}^{K} \Pr\left\{\exists y \in \mathbb{R}^n, \|y\|_0 \leq s, \|X - Dy\|^2 \leq \alpha_s - t(s)\right\} \\
&\leq \sum_{s=1}^{K} C\exp\left(-N\beta t_s^2 + s\ln K\right) \text{ by Theorem 24.1} \\
&\leq \delta \text{ by definition of } t(s).
\end{aligned}$$

□

24.3.2 Evasive Distributions

The parameters (α, β, C) of an evasive distribution transform under the operations of scaling, translation and convolution.

Proposition 24.1 *Let X be (α, β, C)-evasive with values in $\mathbb{R}^N$. Then*

(i) AX is $\left(|||A^{-1}|||^{-2}\alpha, |||A^{-1}|||^4\beta, C\right)$-evasive for every nonsingular $N \times N$ matrix A;
(ii) cX is $\left(c^2\alpha, c^{-4}\beta, C\right)$-evasive for every $c \in \mathbb{R}$;
(iii) $X + z$ is $\left(\alpha', \beta, C\right)$-evasive with $\alpha'_k = \alpha_{k+1}$, for every $z \in \mathbb{R}^N$;
(iv) $X + Z$ is $\left(\alpha', \beta, C\right)$-evasive with $\alpha'_k = \alpha_{k+1}$, for every random variable Z independent of X.

Proof If A is nonsingular and M is any k-dimensional subspace of $\mathbb{R}^N$ then for $z \in M$

$$\left\| X - A^{-1}z \right\| = \left\| A^{-1}\left(AX - z\right) \right\| \leq |||A^{-1}|||\, \|AX - z\|\,,$$

which shows that $d\left(AX, M\right) \geq |||A^{-1}|||^{-1} d\left(X, A^{-1}M\right)$. We therefore have for $t \in \left(0, |||A^{-1}|||^{-2}\alpha\right)$ that

$$\begin{aligned}
&\Pr\left\{ d\left(AX, M\right)^2 < |||A^{-1}|||^{-2}\alpha_k - t \right\} \\
&\quad \leq \Pr\left\{ d\left(X, A^{-1}M\right)^2 < \alpha_k - |||A^{-1}|||^2 t \right\} \\
&\quad \leq \exp\left(-N\beta |||A^{-1}|||^4 t^2 \right),
\end{aligned}$$

since $A^{-1}M$ is also k-dimensional. (ii) is just (i) applied to $A = cI$. (iii) follows from

$$d\left(X + z, M\right) = d\left(X, M - z\right) \geq d\left(X, \mathrm{Span}\left(M, z\right)\right)$$

and the observation that the dimension of $\mathrm{Span}(M, z)$ is at most $\dim M + 1$. Finally (iv) follows from (iii) by first conditioning on Z:

$$\begin{aligned}
\Pr\left\{ d\left(X + Z, M\right)^2 < \alpha_{k+1} - t \right\} &= E\left[E\left[1_{\{X : d(X+Z,M)^2 < \alpha_{k+1} - t\}} | Z \right] \right] \\
&\leq E\left[Ce^{-N\beta t^2} \right].
\end{aligned}$$

□

Next we show that the normalized isotropic Gaussian in $\mathbb{R}^N$ is evasive.

Proposition 24.2 *Let X be an isotropic Gaussian random variable with values in $\mathbb{R}^N$ and $E\,\|X\|^2 = 1$. Then for any k-dimensional subspace M we have*

$$\Pr\left\{ d\left(X, M\right)^2 < \frac{N - k - \pi^2}{N} - t \right\} \leq 2e^{-Nt^2/(2\pi^2)}.$$

Proof For any k-dimensional subspace M we consider the Gaussian random variable $P_M X$ and find

$$E\left[\|P_M X\|^2 \right] = k/N. \tag{24.9}$$

We also note the Gaussian concentration property for the norm [5]

$$\Pr\left\{ \left| E\,\|P_M X\| - \|P_M X\| \right| > t \right\} \leq 2\exp\left(-2Nt^2/\pi^2 \right), \tag{24.10}$$

which we will use repeatedly. For a bound on the variance of the norm we first use it together with integration by parts to get

$$E\left[\left(\|P_M X\| - E\|P_M X\|\right)^2\right] \leq 4\int_0^\infty t\exp\left(-2Nt^2/\pi^2\right)dt = \frac{\pi^2}{N}.$$

This implies that $E\left[\|P_M X\|^2\right] \leq (E\|P_M X\|)^2 + \pi^2/N$, and hence

$$\begin{aligned}&\Pr\left\{E\left[\|P_M X\|^2\right] - \|P_M X\|^2 > t + \pi^2/N\right\}\\ &\leq \Pr\left\{(E\|P_M X\|)^2 - \|P_M X\|^2 > t\right\}\\ &= \Pr\left\{\left(E\|P_M X\| - \|P_M X\|\right)\left(E\|P_M X\| + \|P_M X\|\right) > t\right\}.\end{aligned}$$

Observe that the latter probability is nonzero only if $\|P_M X\| \leq E\|P_M X\|$, and that by Jensen's inequality and (24.9) $E\|P_M X\| \leq \sqrt{k/N} \leq 1$. Using (24.10) again we therefore obtain

$$\Pr\left\{E\left[\|P_M X\|^2\right] - \|P_M X\|^2 > t\right\} \leq \Pr\left\{E\|P_M X\| - \|P_M X\| > \frac{t}{2}\right\}.$$

From (24.9) and (24.10)

$$\Pr\left\{\|P_M X\|^2 < \frac{k-\pi^2}{N} - t\right\} \leq 2e^{-Nt^2/(2\pi^2)},$$

and applying this inequality to the orthogonal complement $M^\perp$ instead of M gives the conclusion. □

The isotropic Gaussian is thus evasive with $\alpha_k = \frac{N-k-\pi^2}{N}$, $\beta = 1/2\pi^2$, $C = 2$. Using Proposition 24.1 (ii) and (iv) with $c = \sigma$ and addition of an appropriate independent random variable Y proves the claim about noisy generative models made in the introduction.

We now show that evasiveness is a generic behavior in high dimensions, if the distribution in question has a bounded support and a bounded density.

Theorem 24.3 *Let the random variable X be distributed as μ in $\mathbb{R}^N$, where μ is absolutely continuous w.r.t. the uniform measure ρ on the unit ball B_N of $\mathbb{R}^N$. For every k let*

$$\alpha_k = \left\|\frac{d\mu}{d\rho}\right\|_\infty^{\frac{-2}{N-k}} \frac{N-k}{e^{3/2}N}.$$

Then for every k-dimensional subspace M we have, for $t \in (0, \alpha_k)$,

$$\Pr\left\{d(X, M)^2 \leq \alpha_k - t\right\} \leq e^{-Nt^2}. \tag{24.11}$$

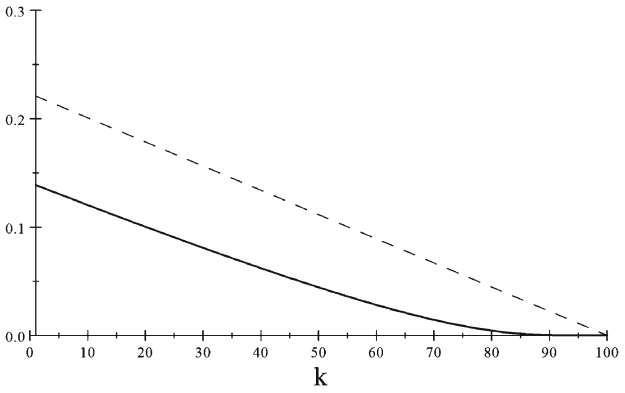

Fig. 24.1 The sequence α_k for $\|d\rho/d\mu\|_\infty = 1$ (*dashed line*) and $\|d\rho/d\mu\|_\infty = 10^{10}$ (*solid line*) with $N = 100$

For applications of the properties of evasive distributions it is crucial that the numbers α_k be reasonably large and decrease slowly. Figure 24.1 plots the decay of the α_k when $N = 100$ and $\|d\mu/d\rho\|_\infty = 1$ (corresponding to the uniform distribution on the ball B_{100}) or $\|d\mu/d\rho\|_\infty = 10^{10}$ respectively.

To prove Theorem 24.3 we need the following technical lemma. Recall that we use V_n to denote the volume of the unit ball in $\mathbb{R}^n$.

Lemma 24.1 *For every $N, k \in \mathbb{N}$ and $1 \leq k < N$ we have*

$$\frac{N-k}{Ne} \leq \left(\frac{V_N}{V_k V_{N-k}}\right)^{\frac{2}{N-k}}.$$

Proof For simplicity we only prove the case where k and N are even. The formula

$$V_n = \frac{\pi^{n/2}}{\Gamma(n/2+1)}$$

shows that

$$\begin{aligned}\frac{V_k V_{N-k}}{V_N} &= \frac{\Gamma(N/2+1)}{\Gamma(k/2+1)\,\Gamma(N/2-k/2+1)} \\ &= \binom{N/2}{k/2} = \binom{N/2}{(N-k)/2} \\ &\leq \left(\frac{Ne}{N-k}\right)^{\frac{N-k}{2}},\end{aligned}$$

where the last inequality is a well-known bound on the binomial coefficients. The result follows. □

Proof (Proof of Theorem 24.3) First we work relative to Lebesgue measure λ. Let

$$a = \left\| \frac{d\mu}{d\lambda} \right\|_\infty = \left\| \frac{d\mu}{d\rho} \right\|_\infty V_N^{-1}.$$

Fix a k-dimensional subspace M. We prove the bound by considering the worst possible density which maximizes the probability in (24.11), subject to the constraint that $\|d\mu/d\lambda\|_\infty = a$ and that μ be supported in the unit ball. Relaxing the constraint on the support of μ from the unit ball to a larger set X will only increase the probability. We can therefore compute a bound on the probability by considering a distribution μ' which maximizes it, subject to the constraint that $\left\|d\mu'/d\lambda\right\|_\infty = a$ and that μ' is supported in the cylinder $(M \cap B_N) \times M^\perp$, which contains the unit ball B_N. Clearly a solution to this optimization problem is given by the density

$$\frac{d\mu'}{d\lambda}(x) = \begin{cases} a \text{ if } d(x, M) \leq r_{\max} \\ 0 \text{ if } d(x, M) > r_{\max} \end{cases}, x \in (M \cap B_N) \times M^\perp,$$

where $r_{\max}$ is determined from the normalization requirement on μ'. This density $d\mu'/d\lambda$ has the maximal value a on a slab of thickness $2r_{\max}$, parallel and symmetric to $M \cap B_N$ and it is zero elsewhere. If V_n denotes the volume of the unit ball in $\mathbb{R}^n$ the volume of this slab is $V_k V_{N-k} r_{\max}^{N-k}$, from which we find

$$r_{\max} = (a V_k V_{N-k})^{-1/(N-k)} = \left(\left\| \frac{d\mu}{d\lambda} \right\|_\infty V_k V_{N-k} \right)^{-1/(N-k)}.$$

A similar computation for the volume of an analogous slab of thickness $2\sqrt{a_k - t}$ gives

$$\Pr\left\{ d(X, M)^2 \leq \alpha_k - t \right\} = \Pr\left\{ d(X, M) \leq \sqrt{\alpha_k - t} \right\} = \left(\frac{\sqrt{\alpha_k - t}}{r_{\max}} \right)^{N-k}, \tag{24.12}$$

where the probability is computed according to μ'. Now we have to show that this is bounded by e^{-Nt^2} for $t \in (0, \alpha_k)$.

We get from the lemma that

$$\begin{aligned} \alpha_k &= \left\| \frac{d\mu}{d\rho} \right\|_\infty^{\frac{-2}{N-k}} \left(\frac{N-k}{e^{3/2} N} \right) \\ &\leq e^{-1/2} \left\| \frac{d\mu}{d\rho} \right\|_\infty^{\frac{-2}{N-k}} \left(\frac{V_N}{V_k V_{N-k}} \right)^{\frac{2}{N-k}} \\ &= e^{-1/2} \left(\left\| \frac{d\mu}{d\lambda} \right\|_\infty V_k V_{N-k} \right)^{\frac{-2}{N-k}} \end{aligned}$$

$$= r_{\max}^2 e^{-1/2}$$
$$\leq t + r_{\max}^2 \exp\left(-\frac{2N}{N-k}t^2\right).$$

The last step follows from

$$0 \leq t \leq \alpha_k = (N-k)\,/\left(e^{3/2}N\right) \leq \sqrt{(N-k)\,/\,(4N)}.$$

Thus

$$0 \leq \alpha_k - t \leq r_{\max}^2 \exp\left(-\frac{2N}{N-k}t^2\right),$$

and substitution in (24.12) gives the conclusion. □

References

1. Baraniuk, R., Davenport, M., DeVore, R., Wakin, M.: A simple proof of the restricted isometry property for random matrices. Constr. Approx. **28**(3), 253–263 (2008)
2. Candès, E.J.: The restricted isometry property and its implications for compressed sensing. Comptes Rendus Mathematique **346**(9), 589–592 (2008)
3. Elad, M., Aharon, M., Bruckstein, A.M.: On the uniqueness of overcomplete dictionaries and a practical way to retrieve them. Linear Algebra Appl. **416**(1), 48–67 (2006)
4. Gribonval, R., Schnass, K.: Dictionary identification: sparse matrix-factorization via ℓ_1-minimization. IEEE Trans. Inf. Theory **56**(7), 3523–3539 (2010)
5. Ledoux, M., Talagrand, M.: Probability in Banach Spaces. Springer, Berlin (1991)
6. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature **401**, 788–791 (1999)
7. Mairal, J., Bach, F., Ponce, J., Sapiro, G.: Online learning for matrix factorization and sparse coding. J. Mach. Learn. Res. **11**, 19–60 (2010)
8. Maurer, A., Pontil, M.: K-dimensional coding schemes in Hilbert spaces. IEEE Trans. Inf. Theory **56**(11), 5839–5846 (2010)
9. Maurer, A., Pontil, M., Romera-Paredes, B.: Sparse coding for multitask and transfer learning. In: Proceedings of the 30th International Conference on Machine Learning, pp. 343–351 (2013)
10. Olshausen, B.A., Field, D.A.: Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature **381**(6583), 607–609 (1996)
11. Olshausen, B.A., Field, D.J.: Sparse coding with an overcomplete basis set: a strategy employed by V1? Vis. Res. **37**(23), 3311–3325 (1997)
12. Ranzato, M.A., Poultney, C., Chopra, S., LeCun, Y.: Efficient learning of sparse representations with an energy-based model. In: Scholkopf, B., Platt, J.C., Hoffman, T. (eds.) Advances in Neural Information Processing Systems, vol. 19, pp. 1137–1144. MIT Press, Cambridge (2007)