On The Sample Complexity Of Sparse Dictionary Learning

Matthias Seibert, Martin Kleinsteuber, Rémi Gribonval, Rodolphe Jenatton, Francis Bach

ON THE SAMPLE COMPLEXITY OF SPARSE DICTIONARY LEARNING

M. Seibert[1], M. Kleinsteuber[1], R. Gribonval[2], R. Jenatton[3,4], F. Bach[3,5]

[1]Department of Electrical Engineering and Information Technology, TU München, Munich, Germany.
[2]PANAMA Project-Team (INRIA & CNRS).
[3]SIERRA Project-Team (INRIA Paris).
[4]Centre de Mathématiques Appliquées - Ecole Polytechnique (CMAP).
[5]Laboratoire d'informatique de l'école normale supérieure (LIENS).
{m.seibert,kleinsteuber}@tum.de, gribonval@inria.fr, r.jenatton@criteo.com, francis.bach@ens.fr

ABSTRACT

In the synthesis model signals are represented as a sparse combinations of atoms from a dictionary. Dictionary learning describes the acquisition process of the underlying dictionary for a given set of training samples. While ideally this would be achieved by optimizing the expectation of the factors over the underlying distribution of the training data, in practice the necessary information about the distribution is not available. Therefore, in real world applications it is achieved by minimizing an empirical average over the available samples. The main goal of this paper is to provide a sample complexity estimate that controls to what extent the empirical average deviates from the cost function. This estimate then provides a suitable estimate to the accuracy of the representation of the learned dictionary. The presented approach exemplifies the general results proposed by the authors in [1] and gives more concrete bounds of the sample complexity of dictionary learning. We cover a variety of sparsity measures employed in the learning procedure.

Index Terms— Dictionary learning, sample complexity, sparse coding

1. INTRODUCTION

The sparse synthesis model relies on the assumption that signals $\mathbf{x} \in \mathbb{R}^m$ can be represented as a sparse combination of columns, or atoms, of a dictionary $\mathbf{D} \in \mathbb{R}^{m \times d}$. As an equation this reads as

$$\mathbf{x} = \mathbf{D}\boldsymbol{\alpha} \tag{1}$$

where $\boldsymbol{\alpha} \in \mathbb{R}^d$ is the sparse coefficient vector.

The task of dictionary learning focuses on finding the best dictionary to sparsely represent a set of training samples concatenated in the matrix $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]$. The corresponding sparse representations are stored in the coefficient matrix $\mathbf{A} = [\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_n]$. A common learning approach is to find a solution to the minimization problem

$$\min_{\mathbf{D},\mathbf{A}} \quad \mathcal{L}_{\mathbf{X}}(\mathbf{D}, \mathbf{A}) \tag{2}$$

$$\mathcal{L}_{\mathbf{X}}(\mathbf{D}, \mathbf{A}) \coloneqq \tfrac{1}{2n}\|\mathbf{X} - \mathbf{D}\mathbf{A}\|_F^2 + \tfrac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{\alpha}_i). \tag{3}$$

The function g in (3) serves as a measure of sparsity. Concretely, we consider scaled powers of the ℓ_p-norm, i.e.

$$g(\boldsymbol{\alpha}) \coloneqq \|\boldsymbol{\alpha}/\lambda\|_p^q \tag{4}$$

for any $p, q > 0$ and the weighting parameter $\lambda > 0$.

In order to avoid trivial solutions, the learned dictionary $\mathbf{D}$ is generally forced to be an element of a constraint set $\mathfrak{D}$. In this paper we will focus on dictionaries with unit ℓ_2-norm atoms, which is a commonly used constraint.

The vast amount of dictionary learning algorithms that take different approaches to the topic illustrates the popularity of the synthesis model. A probabilistic method is presented in [2]. Another famous representative is the K-SVD algorithm as proposed in [3] which is based on K-means clustering. Finally, there are learning strategies that aim at learning dictionaries with specific structures that enable fast computations, see e.g. [4, 5].

Assuming that these training samples are drawn according to some distribution, the goal of dictionary learning is to find a dictionary $\mathbf{D}^\star$ for which the expected value of the cost function (3) is minimal. In practice the distribution of the available training samples is unknown and therefore only an empirical minimizer $\hat{\mathbf{D}}$ can be obtained. The sample complexity results which we derive in this paper contribute to understand how accurately this empirical minimizer approximates $\mathbf{D}^\star$.

We assume the training signals to be drawn according to a distribution in the ball with unit radius, i.e. the distribution is an element of the set

$$\mathfrak{P} \coloneqq \{\mathbb{P} : \mathbb{P}(\|\mathbf{x}\|_2 \leq 1) = 1\}. \tag{5}$$

This work was partially supported Cluster of Excellence CoTeSys funded by the German DFG.

This work was partially supported by the EU FP7, SMALL project, FET-Open grant number 225913, and by the European Research Council, PLEASE project (ERC-StG-2011-277906)

Our results are based on the work [1] where a more general framework of matrix factorizations has been considered. Our stringent setting here allows for more concrete bounds on the sample complexity.

Previous state of the art sample complexity results are presented in [6, 7]. These results are restricted to the indicator function for ℓ_0 and ℓ_1-norms. These works also cover the case of fast rates which we will not consider here.

2. PROBLEM STATEMENT & NOTATION

Matrices are denoted by boldface capital letters such as $\mathbf{X}$, vectors will be represented as boldface small letters, e.g. $\mathbf{x}$. Scalars will be slanted letters like n, N. The i^{th} entry of a vector $\boldsymbol{\alpha}$ is denoted by α_i. Finally, sets are denoted in blackletter such as $\mathfrak{D}$.

The sparse representation of a given signal $\mathbf{x}$ according to a given dictionary can be found by solving the optimization problem

$$\arg\min_{\boldsymbol{\alpha}\in\mathbb{R}^d} \tfrac{1}{2}\|\mathbf{x}-\mathbf{D}\boldsymbol{\alpha}\|_2^2 + g(\boldsymbol{\alpha}). \tag{6}$$

The quality of how well a signal can be sparsely coded for a dictionary is evaluated via the function

$$f_{\mathbf{x}}(\mathbf{D}) \coloneqq \inf_{\boldsymbol{\alpha}\in\mathbb{R}^d} \tfrac{1}{2}\|\mathbf{x}-\mathbf{D}\boldsymbol{\alpha}\|_2^2 + g(\boldsymbol{\alpha}). \tag{7}$$

For a set of signals $\mathbf{X}$ the overall quality of the sparse representation is measured via

$$F_{\mathbf{X}}(\mathbf{D}) \coloneqq \inf_{\mathbf{A}} \mathcal{L}_{\mathbf{X}}(\mathbf{D},\mathbf{A}) \tag{8}$$

with $\mathcal{L}_{\mathbf{X}}$ as defined in (3). This measure is equal to the mean of the quality of all samples, i.e. $F_{\mathbf{X}}(\mathbf{D}) = \frac{1}{n}\sum_i f_{\mathbf{x}_i}(\mathbf{D})$.

Our goal is to provide a bound for the generalization error of the empirical minimizer, i.e.

$$\sup_{\mathbf{D}\in\mathfrak{D}} |F_{\mathbf{X}}(\mathbf{D}) - \mathbb{E}_{\mathbf{x}\sim\mathbb{P}} f_{\mathbf{x}}(\mathbf{D})| \leq \eta(n,m,d,L) \tag{9}$$

which depends on the number of samples n used in the learning process, the size of the samples m, the number of dictionary atoms d, and a certain Lipschitz constant L for $F_{\mathbf{X}}$.

3. SAMPLE COMPLEXITY RESULTS

The general strategy will be to first find a Lipschitz constant for the functions $F_{\mathbf{X}}$ and $\mathbb{E}f_{\mathbf{x}}$. In combination with the assumed underlying probability distribution and the structure of the dictionary this will allow us to provide an upper bound for the sample complexity.

3.1. Lipschitz Property

In this section we provide Lipschitz constants for the function $F_{\mathbf{X}}(\mathbf{D})$. For the ℓ_p-penalty function the Hölder inequality yields

$$\begin{aligned}\|\boldsymbol{\alpha}\|_1 = \sum_{i=1}^d |\alpha_i| &\leq \left(\sum_{i=1}^d |\alpha_i|^p\right)^{1/p}\left(\sum_{i=1}^d 1^{1/(1-1/p)}\right)^{1-1/p}\\ &= d^{1-1/p}\cdot\|\boldsymbol{\alpha}\|_p\end{aligned}$$

for $1 \leq p < +\infty$. To cover quasi-norms with $0 < p < 1$, we employ the estimate $\|\boldsymbol{\alpha}\|_1 \leq \|\boldsymbol{\alpha}\|_p$ which provides the overall inequality

$$\|\boldsymbol{\alpha}\|_1 \leq d^{(1-1/p)_+}\cdot\|\boldsymbol{\alpha}\|_p, \tag{10}$$

where the function $(\cdot)_+\colon \mathbb{R}\to\mathbb{R}_0^+$ is defined as $(t)_+ = \max\{t,0\}$. Thus, we get the two estimates

$$\|\boldsymbol{\alpha}\|_p \leq t \quad\Rightarrow\quad \|\boldsymbol{\alpha}\|_1 \leq d^{(1-1/p)_+}\cdot t, \tag{11}$$

$$\|\boldsymbol{\alpha}/\lambda\|_p^q \leq t \quad\Rightarrow\quad \|\boldsymbol{\alpha}\|_1 \leq \lambda\cdot d^{(1-1/p)_+}\cdot t^{1/q} \tag{12}$$

which will become of use in the following discussion. The matrix norm $\|\mathbf{A}\|_{1\to 2} \coloneqq \max_i \|\boldsymbol{\alpha}_i\|_2$ is used for the rest of this paper while the subscript is omitted to improve readability. We also make use of the corresponding dual norm which is defined as $\|\mathbf{A}\|_\star \coloneqq \sup_{\mathbf{U},\|\mathbf{U}\|\leq 1}\langle\mathbf{A},\mathbf{U}\rangle_F$ for an appropriately sized matrix $\mathbf{U}$ and the Frobenius inner product $\langle\cdot,\cdot\rangle_F$.

For $\epsilon > 0$ we define the nonempty set of matrices $\mathbf{A}$ that are "ϵ-near solutions" as

$$\mathfrak{A}_\epsilon(\mathbf{X},\mathbf{D}) \coloneqq \{\mathbf{A} : \boldsymbol{\alpha}_i\in\mathbb{R}^d, \mathcal{L}_{\mathbf{x}_i}(\mathbf{D},\boldsymbol{\alpha}_i) \leq f_{\mathbf{x}_i}(\mathbf{D}) + \epsilon\}.$$

Proposition 1. *The set $\mathfrak{A}_0$ is not empty and is bounded.*

Proof. The function $\|\cdot\|_p^q$ is non-negative and coercive. Thus, $\mathcal{L}_{\mathbf{X}}(\mathbf{D},\mathbf{A})$ is non-negative and $\lim_{k\to\infty}\mathcal{L}_{\mathbf{X}}(\mathbf{D},\mathbf{A}_k) = \infty$ whenever $\lim_{k\to\infty}\|\mathbf{A}_k\| = \infty$. Therefore, the function $\mathbf{A}\mapsto\mathcal{L}_{\mathbf{X}}(\mathbf{D},\mathbf{A})$ has bounded sublevel sets. Moreover, since powers of the ℓ_p-norm are continuous, then so is $\mathbf{A}\mapsto\mathcal{L}_{\mathbf{X}}(\mathbf{D},\mathbf{A})$ and thus attains its infimum value. □

Next, note that for any $\mathbf{D}'$ the inequality

$$\begin{aligned}&F_{\mathbf{X}}(\mathbf{D}') - F_{\mathbf{X}}(\mathbf{D})\\ &\qquad\leq L_{\mathbf{X}}(\mathbf{D})\|\mathbf{D}'-\mathbf{D}\| + C_{\mathbf{X}}(\mathbf{D})\|\mathbf{D}'-\mathbf{D}\|^2\end{aligned} \tag{13}$$

holds with

$$L_{\mathbf{X}}(\mathbf{D}) \coloneqq \inf_{\epsilon>0}\sup_{\mathbf{A}\in\mathfrak{A}_\epsilon} \tfrac{1}{n}\|(\mathbf{X}-\mathbf{D}\mathbf{A})\mathbf{A}^\top\|_\star, \tag{14}$$

$$C_{\mathbf{X}}(\mathbf{D}) \coloneqq \inf_{\epsilon>0}\sup_{\mathbf{A}\in\mathfrak{A}_\epsilon} \tfrac{1}{2n}\sum_{i=1}^n \|\boldsymbol{\alpha}_i\|_1^2. \tag{15}$$

A detailed derivation of these parameters can be found in [1].

Proposition 2. *There exist upper bounds for $L_{\mathbf{X}}(\mathbf{D})$ and $C_{\mathbf{X}}(\mathbf{D})$ which are independent of the used dictionary.*

Proof. For $\mathbf{A} = [\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_n] \in \mathfrak{A}_0$ we have

$$\tfrac{1}{2}\|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2^2 + \|\boldsymbol{\alpha}_i/\lambda\|_p^q \leq f_{\mathbf{x}_i}(\mathbf{D})$$

which immediately results in the estimates

$$0 \leq \|\boldsymbol{\alpha}_i/\lambda\|_p^q \leq f_{\mathbf{x}_i}(\mathbf{D}) \leq \mathcal{L}(\mathbf{D}, \mathbf{0}) = \tfrac{1}{2}\|\mathbf{x}_i\|_2^2, \quad (16)$$

$$\|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2 \leq \sqrt{2 f_{\mathbf{x}_i}(\mathbf{D})} \leq \|\mathbf{x}_i\|_2. \quad (17)$$

Furthermore, the above in combination with (12) lets us bound the ℓ_1-norm of $\boldsymbol{\alpha}$ via

$$\|\boldsymbol{\alpha}\|_1 \leq \lambda \cdot d^{(1-1/p)_+} \left(\tfrac{1}{2}\|\mathbf{x}\|_2^2\right)^{1/q}. \quad (18)$$

This yields the upper bound $C_{\mathbf{X}}(\mathbf{D}) \leq C_{\mathbf{X}}$ with

$$C_{\mathbf{X}} \coloneqq \tfrac{1}{2n} \sum_{i=1}^{n} \lambda \cdot d^{(1-1/p)_+} \left(\tfrac{1}{2}\|\mathbf{x}_i\|_2^2\right)^{1/q} \quad (19)$$

for (15). In order to provide an upper bound for $L_{\mathbf{X}}(\mathbf{D})$ which is independent of the dictionary $\mathbf{D}$ we first note that

$$\langle (\mathbf{X} - \mathbf{D}\mathbf{A})\mathbf{A}^\top, \mathbf{U} \rangle \leq \sum \|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2 \cdot \|\boldsymbol{\alpha}_i\|_1 \cdot \|\mathbf{U}\|.$$

By using the definition of the dual norm and the estimate developed above, we obtain the upper bound $L_{\mathbf{X}}(\mathbf{D}) \leq L_{\mathbf{X}}$ with

$$L_{\mathbf{X}} \coloneqq \tfrac{1}{n} \sum_{i=1}^{n} \|\mathbf{x}_i\|_2 \cdot \lambda \cdot d^{(1-1/p)_+} \left(\tfrac{1}{2}\|\mathbf{x}_i\|_2^2\right)^{1/q} \quad (20)$$

which concludes the proof. □

Proposition 2 allows us to rewrite (13) as

$$\frac{|F_{\mathbf{X}}(\mathbf{D}') - F_{\mathbf{X}}(\mathbf{D})|}{\|\mathbf{D}' - \mathbf{D}\|} \leq L_{\mathbf{X}} \cdot \left(1 + \tfrac{C_{\mathbf{X}}}{L_{\mathbf{X}}}\|\mathbf{D}' - \mathbf{D}\|\right), \quad (21)$$

which implies that the function $F_{\mathbf{X}}$ is uniformly locally Lipschitz with constant $L_{\mathbf{X}}$.

Lemma 3. *The function $F_{\mathbf{X}}$ is globally Lipschitz with constant $L_{\mathbf{X}}$, i.e.*

$$|F_{\mathbf{X}}(\mathbf{D}') - F_{\mathbf{X}}(\mathbf{D})| \leq L_{\mathbf{X}}\|\mathbf{D}' - \mathbf{D}\| \quad (22)$$

for any $\mathbf{X}$ and any $\mathbf{D}, \mathbf{D}' \in \mathfrak{D}$.

Proof. Let $\epsilon > 0$ be arbitrary but fixed. For $\|\mathbf{D}' - \mathbf{D}\| \leq \epsilon L_{\mathbf{X}}/C_{\mathbf{X}}$ we have

$$|F_{\mathbf{X}}(\mathbf{D}') - F_{\mathbf{X}}(\mathbf{D})| \leq (1+\epsilon) L_{\mathbf{X}}\|\mathbf{D}' - \mathbf{D}\|. \quad (23)$$

If the bound on the distance between $\mathbf{D}$ and $\mathbf{D}'$ does not hold, we construct the sequence $\mathbf{D}_i \coloneqq \mathbf{D} + \frac{i}{k}(\mathbf{D}' - \mathbf{D})$, $i = 0, \ldots, k$ such that $\|\mathbf{D}_{i+1} - \mathbf{D}_i\| \leq \epsilon L_{\mathbf{X}}/C_{\mathbf{X}}$. This enables us to show that the bound (23) holds for any $\mathbf{D}, \mathbf{D}'$. Note that there are no restrictions on $\mathbf{D}, \mathbf{D}'$, as the derived Lipschitz constant $L_{\mathbf{X}}$ is independent of the constraint set $\mathfrak{D}$. Since $\epsilon > 0$ is chosen arbitrarily, (22) follows. □

3.2. Probability Distribution

As mentioned in the introduction we consider probability distributions within the unit ball. In order to obtain meaningful results the distribution according to which the samples are drawn has to fulfill two properties. First, we need to control the Lipschitz constant $L_{\mathbf{X}}$ for signals drawn according to the distribution when the number of samples n is large. This quantity will be measured by the function

$$\Lambda_n(L) \coloneqq \mathbb{P}\left(L_{\mathbf{X}} > L\right). \quad (24)$$

Furthermore, for a given $\mathbf{D}$ we need to control the concentration of the empirical average $F_{\mathbf{X}}(\mathbf{D})$ around its expectation. This is measured via

$$\Gamma_n(\gamma) \coloneqq \sup_{\mathbf{D} \in \mathfrak{D}} \mathbb{P}\left(|F_{\mathbf{X}}(\mathbf{D}) - \mathbb{E}_{\mathbf{x}\sim\mathbb{P}} f_{\mathbf{x}}(\mathbf{D})| > \gamma\right). \quad (25)$$

For the considered distribution these quantities are well controlled, as can be seen in the following.

Proposition 4. *For $\mathbb{P} \in \mathfrak{P}$ as defined in* (5) *we have $\Lambda_n(L) = 0$ for $L \geq d^{(1-1/p)_+}(1/2)^{1/q}$, and*

$$\Gamma_n\left(\tau/\sqrt{8}\right) \leq 2\exp(-n\tau^2),\ \forall\, 0 \leq \tau < +\infty. \quad (26)$$

Proof. The function evaluations $y_i = f_{\mathbf{x}_i}(\mathbf{D})$ are independent random variables. For samples drawn according to a distribution within the unit sphere it immediately follows that they are bounded by $0 \leq y_i \leq \|\mathbf{x}_i\|_2^2/2 \leq 1/2$. Using Hoeffding's Inequality we get

$$\mathbb{P}\left[\tfrac{1}{n}\left(\sum_{i=1}^{n} y_i - \mathbb{E}[y_i]\right) \geq c\tau\right] \leq \exp\left(-8c^2 n\tau^2\right)$$

and thus $\Gamma_n(\tau/\sqrt{8}) \leq 2\exp(-n\tau^2)$. Furthermore, due to the chosen set of viable probability distributions $\mathfrak{P}$, we have $L_{\mathbf{X}} \leq \lambda \cdot d^{(1-1/p)_+}(1/2)^{1/q}$ and hence $\Lambda_n(L) = 0$ for $L \geq \lambda \cdot d^{(1-1/p)_+}(1/2)^{1/q}$. □

3.3. Role of the Constraint Set

In order to provide a sample complexity bound, it is necessary to take the structure of the set to which the learned dictionary is constrained into account. Of particular interest is the covering number of the concerning set. For more information on covering numbers, the interested reader is referred to [8].

We will confine the discussion to the set of dictionaries with unit norm atoms, i.e. each dictionary column is an element of the sphere. It is well known that the covering number of the sphere is upper bounded by $\mathcal{N}_\epsilon(\mathfrak{S}^{m-1}) \leq \left(1 + \frac{2}{\epsilon}\right)^m$. By using the metric $\|\cdot\|_{1\to 2}$ this is readily extended to the product of unit spheres

$$\mathcal{N}_\epsilon\left(\mathfrak{D}(m,d)\right) \leq \left(1 + \tfrac{2}{\epsilon}\right)^{md} \leq \left(\tfrac{3}{\epsilon}\right)^{md}. \quad (27)$$

3.4. Main Result

Theorem 5. *For a given $0 \leq t < \infty$ and the Lipschitz constant $L > \lambda \cdot d^{(1-1/p)_+}(1/2)^{1/q}$, we have*

$$\sup_{\mathbf{D}\in\mathfrak{D}} |F_{\mathbf{X}}(\mathbf{D}) - \mathbb{E}_{\mathbf{x}\sim\mathbb{P}} f_{\mathbf{x}}(\mathbf{D})| \leq \eta(n, m, d, L) \quad (28)$$

with probability at least $1-2e^{-t}$*. The bound is defined as*

$$\eta(n,m,d,L) \coloneqq 2\sqrt{\frac{\beta \log n}{n}} + \sqrt{\frac{\beta + t/\sqrt{8}}{n}} \tag{29}$$

with the driving constant

$$\beta \coloneqq \tfrac{md}{8} \cdot \max\left\{\log\left(6\sqrt{8}L\right), 1\right\}. \tag{30}$$

The parameter controlling the sample complexity is dependent on the size of the dictionary, the determined Lipschitz constant, and the number of samples. It is also dependent on the underlying distribution of the samples, which is an arbitrary distribution in the unit ball in the examined case. Better estimates may hold for fixed probability distributions.

Proof. First, note that $\mathbb{E}f_{\mathbf{x}}$ is Lipschitz with constant $L > L_{\mathbf{X}}$ for the considered case. Let $\epsilon, \gamma > 0$ be given. The constraint set $\mathfrak{D}$ can be covered by and ϵ-net with $\mathcal{N}_\epsilon$ elements $\mathbf{D}_j$. For a fixed dictionary $\mathbf{D}$ there exists an index j for which $\|\mathbf{D} - \mathbf{D}_j\| \leq \epsilon$ and we can write

$$\begin{aligned}|F_{\mathbf{X}}(\mathbf{D}) - \mathbb{E}f_{\mathbf{x}}(\mathbf{D})| \leq & |F_{\mathbf{X}}(\mathbf{D}) - F_{\mathbf{X}}(\mathbf{D}_j)| \\ & + \sup_j |F_{\mathbf{X}}(\mathbf{D}_j) - \mathbb{E}f_{\mathbf{x}}(\mathbf{D}_j)| \\ & + |\mathbb{E}f_{\mathbf{x}}(\mathbf{D}_j) - \mathbb{E}f_{\mathbf{x}}(\mathbf{D})|.\end{aligned}$$

By using the concentration assumption (25) and the Lipschitz continuity of $F_{\mathbf{X}}$ and $\mathbb{E}f_{\mathbf{x}}$ this implies

$$\sup_{\mathbf{D}} |F_{\mathbf{X}}(\mathbf{D}) - \mathbb{E}_{\mathbf{x}\sim\mathbb{P}} f_{\mathbf{x}}(\mathbf{D})| \leq 2L\epsilon + \gamma \tag{31}$$

except for probability at most $\mathcal{N}_\epsilon \cdot \Gamma_n(\gamma)$. Since the above holds for and $\epsilon, \gamma > 0$, we specify the constants

$$\begin{aligned}\epsilon &\coloneqq \tfrac{1}{2L}\sqrt{\tfrac{\beta \log n}{n}}, \\ \tau &\coloneqq \tfrac{1}{\sqrt{n}}\sqrt{md\log\left(\tfrac{3}{\epsilon}\right) + t}\end{aligned}$$

with $\gamma \coloneqq \tau/\sqrt{8}$ which fulfill the conditions $0 < \epsilon < 1$ and $0 \leq \tau < \infty$. Given these parameters we get

$$\mathcal{N}_\epsilon \cdot \Gamma_n(\tau/\sqrt{8}) = 2e^{-t}. \tag{32}$$

For the final estimate, recall that due to the definition of β the inequalities

$$\log\left(\tfrac{6L}{\sqrt{\beta}}\right) \leq \log\left(6\sqrt{8}L\right) \leq 8\beta/(md)$$

and $\log n \geq 1$ hold. This allows us to provide the estimate

$$2L\epsilon + \tau/\sqrt{8} \leq 2\sqrt{\tfrac{\beta \log n}{n}} + \sqrt{\tfrac{\beta + t/\sqrt{8}}{n}} \tag{33}$$

which concludes the proof of Theorem 5. □

In order to illustrate the results, we will discuss a short example.

Example 1. The general assumption is that the learned dictionary is an element of $\mathfrak{D}(m,d)$ and the training samples are drawn according to a distribution in the unit ball. Let the sparsity promoting function be defined as the ℓ_p-norm with $0 < p < 1$. Then Theorem 5 holds with the sample complexity driving constant

$$\beta = \tfrac{md}{8} \cdot \log(3\sqrt{8}). \tag{34}$$

4. CONCLUSION

Based on the general framework introduced in [1] we provide a sample complexity result for learning dictionaries with unit ℓ_2-norm atoms. Powers of the ℓ_p-norm as a penalty in the learning process while the samples are drawn according to a distribution within the unit ball. In general, we can say that the sample complexity bound η exhibits the behavior $\eta \propto \sqrt{\frac{\log n}{n}}$ with high probability. The sample complexity results achieved in this paper recover those of [6, 7] for a different choice of sparsity measure. We suspect that the achieved results can be further improved by utilizing Rademacher's complexity. This approach will be discussed in future work.

5. REFERENCES

[1] R. Gribonval, R. Jenatton, F. Bach, M. Kleinsteuber, and M. Seibert, "Sample complexity of dictionary learning and other matrix factorizations," *pre-print*, 2014, http://arxiv.org/abs/1312.3790.

[2] B.A. Olshausen and D.J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by VI?," *Vision Research*, vol. 37, no. 23, pp. 3311–3326, 1997.

[3] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, 2006.

[4] R. Rubinstein, M. Zibulevsky, and M. Elad, "Double sparsity: Learning sparse dictionaries for sparse signal approximation," *IEEE Transactions on Signal Processing*, vol. 58, no. 3, pp. 1553–1564, 2010.

[5] S. Hawe, M. Seibert, and M. Kleinsteuber, "Separable dictionary learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

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

[7] D. Vainsencher, S. Mannor, and A.M. Bruckstein, "The sample complexity of dictionary learning," *Journal of Machine Learning Research*, vol. 12, pp. 3259–3281, 2011.

[8] M. Anthony and P.L. Bartlett, *Neural network learning: Theoretical foundations*, Cambridge University Press, 2009.