Abstract—Sparse representation over redundant dictionaries constitutes a good model for many classes of signals (e.g., patches of natural images, segments of speech signals, etc.). However, despite its popularity, very little is known about the representation capacity of this model. In this paper, we study how redundant a dictionary must be so as to allow any vector to admit a sparse approximation with a prescribed sparsity and a prescribed level of accuracy. We address this problem both in a worst-case setting and in an average-case one. For each scenario we derive lower and upper bounds on the minimal required overcompleteness. Our bounds have simple closed-form expressions that allow to easily deduce the asymptotic behavior in large dimensions. In particular, we find that the required overcompleteness grows exponentially with the sparsity level and polynomially with the allowed representation error. This implies that universal sparse representation is practical only at moderate sparsity levels, but can be achieved with a relatively high accuracy. We illustrate the validity of our results through numerical simulations, which support our findings.

I. INTRODUCTION

Researchers and engineers often use transforms to analyze and process signals. A common desired property from a transform, is that it allow signals to be represented as combinations of a small number of “atoms”. For example, the Fourier transform is commonly used for analyzing audio signals [14] since they tend to be comprised of a small number of harmonic components. Piecewise smooth signals, on the other hand, are much more compactly represented by the wavelet transform, which is thus popular in image processing [6]. The emergence of the field of sparse representations [19], initiated the systematic construction of dictionaries, that allow representing signals as linear combinations of a small number of their atoms [13], [2]. Today, this concept constitutes a key ingredient in numerous areas, ranging from image enhancement to signal recovery and compression [10], [12], [24].

In this paper, we address a fundamental question relating to the expressive power of sparse representations. Specifically, we study conditions under which a redundant dictionary can be used to represent every signal in \( \mathbb{R}^d \) as a linear combination of at most \( k < d \) of its atoms, with an error no larger than \( \varepsilon \). Our goal is to obtain necessary and sufficient conditions on the minimal number of atoms \( n \) allowing this.

This problem has two motivations. First, when the sparse representation model is used as a prior, as in compressed sensing or signal restoration [11], [23], only a small set of signals is meant to be sparsely representable over the dictionary. This is often achieved by learning a dictionary from a set of relevant training examples (e.g., patches from natural images) [13], [2], [16], [18]. In this context, it is of interest to identify when a dictionary has an unnecessarily large overcompleteness (i.e., one which allows sparse representation of every signal, and not only of those from the designated set). A second motivation relates to the use of the sparse representation model as a generic transform, under which all signals are sparse.

It is easy to show that when \( k \) is taken to be a fixed fraction of \( d \), the set of signals in \( \mathbb{R}^d \) that can be approximately represented by a specific choice of \( k \) atoms has a volume which is exponentially small in \( k \) (see Appendix for details). This, is, in fact, the principle underlying the Johnson-Lindenstrauss lemma [15]. This lemma asserts that when projecting points in \( \mathbb{R}^d \) onto a random \( k \)-dimensional space, there is a concentration of measure effect whereby the points’ norms are approximately preserved up to a factor of \( k/d \). Therefore, when \( k \) is much smaller than \( d \), such projections are very far from the original points with high probability. Nevertheless, in our context, if \( k \) is also a fixed fraction of \( n \), then the number of choices of \( k \) atoms from a dictionary of size \( n \) is exponentially large in \( k \). Therefore, it is not a-priori clear whether the overcompleteness should be very large in order to allow universal sparse representation. Interestingly, our results show that for certain regimes of error and sparsity levels, universal sparse representation can be achieved with moderate redundancy. For other regimes, on the other hand, universal sparse representation becomes impractical.

It should be noted that our setting is very different from that of compressed sensing. There, the quantity of interest is the minimal number of linear measurements from which any \( k \)-sparse signal can be uniquely recovered [8], [22], [9], [5]. Merging the measurement matrix into the dictionary, this problem is equivalent to asking what is the minimal number of rows of a dictionary allowing to uniquely recover any signal that is \( k \)-sparse in the standard (non-overcomplete) basis. In contrast, here we analyze the minimal number of atoms (columns of an overcomplete dictionary) with which all signals possess a sparse representation. Furthermore, we do not require uniqueness of the representation.

Mathematically, universal sparse representation can be viewed as a covering problem, a branch of mathematics with many known results [4]. However most works consider ball covering problems, whereas our setting is concerned with covering by dilations of linear subspaces (spanned by subsets of atoms from the dictionary). Moreover, these subspaces share atoms and are thus constrained to intersect. To the best of our knowledge, such settings were not studied in the past.

Only a few attempts were made to characterize the representation ability of overcomplete dictionaries. In [1, Ch. 7] the author provided approximations for the relative volume of signals that admit a sparse representation with an allowable error. However, the expressions depend on properties of the dictionary (minimal and maximal singular values of any subset of columns) and are thus not universal. Furthermore, the accuracy of the approximations in high dimensions is not clear. In [3] the authors analyzed a stochastic setting, for which they provided a lower bound on the achievable mean squared error (MSE) of the representation as a function of the sparsity and the dictionary’s overcompleteness. The analysis is universal in that it holds for all dictionaries. However, it does not provide an upper bound on the error (from which an upper bound on the required overcompleteness could be deduced), and it does not provide deterministic (worst-case) results.

In this paper we study the universal sparse representation problem...
from both a worst case standpoint and an average case one. In the worst case setting we request that the representation error be bounded by \(\epsilon\) for every signal in \(\mathbb{R}^d\). In the average case setting, we assume that the signal is random and require that the probability that it can be sparsely represented with an error less than \(\epsilon\), be high. For each scenario we give lower and upper bounds on the minimal required overcompleteness allowing universal sparse representation. As opposed to previous works, our bounds have simple closed-form expressions, which allow to easily deduce the asymptotic behavior of the required overcompleteness. In particular, our bounds reveal that if \(\epsilon \ll 1\) or \(k \ll d\), then the minimal required overcompleteness behaves like \((1/\epsilon)^{d/k-1}\) up to polynomial factors in \(d/k\). We provide simulations, which show that our bounds correctly predict the threshold at which sparse coding techniques start to succeed in approximating arbitrary signals.

This paper is a short version of [21]. For the proofs, please refer to the extended journal version.

II. MAIN RESULTS

Our goal is to be able to represent every signal \(x \in \mathbb{R}^d\) as a linear combination of a small number \(k \ll d\) of atoms from some dictionary \(\Phi \in \mathbb{R}^{d \times n}\). Note that this is impossible to do without incurring some error, since the set of signals admitting such a \(k\)-sparse representation is a union of \(\binom{n}{k}\) subspaces of dimension at most \(k\), which is strictly contained in \(\mathbb{R}^d\). However, the question we ask is: Under what conditions can we guarantee a \(k\)-sparse representation for every signal in \(\mathbb{R}^d\) with a small error?

**Definition 1 (Normalized \(k\)-sparse representation error):** We define the \(k\)-sparse representation error of a signal \(x \in \mathbb{R}^d\) over a dictionary \(\Phi \in \mathbb{R}^{d \times n}\) as

\[
\epsilon(x, \Phi) \triangleq \min_{\alpha \in \mathbb{R}^n} \frac{\|x - \Phi \alpha\|}{\|x\|} \quad \text{s.t.} \quad \|\alpha\|_0 \leq k, \tag{1}
\]

where the \(\ell_0\) (pseudo) norm \(\|\cdot\|_0\) counts the number of nonzero elements of its vector argument. We shall say that \(x\) has a \(k\)-sparse representation over \(\Phi\) with precision \(\epsilon\) if \(\epsilon(x, \Phi) \leq \epsilon\).

The normalized error is indifferent to scaling of \(x\). Therefore, without loss of generality, we will restrict our analysis to signals lying on the unit sphere (i.e., with \(\|x\| = 1\)). We are interested in the existence of dictionaries \(\Phi\) such that \(\epsilon(x, \Phi)\) is small for all, or at least most, signals \(x\). More specifically, we consider both a worst-case design (Sec. II-A) and an average-case one (Sec. II-B). In the former, we require that the error \(\epsilon(x, \Phi)\) be bounded by \(\epsilon\) for every \(x \in \mathbb{R}^d\). In the latter, we assume that \(x\) is a random vector and require that the probability that \(\epsilon(x, \Phi) \leq \epsilon\) be large.

The two cardinal parameters in our problem are the sparsity factor \(s\) and the overcompleteness ratio \(o\), defined as

\[
s \triangleq \frac{k}{d}, \quad o \triangleq \frac{n}{d}. \tag{2}
\]

The overcompleteness ratio can be thought of as the aspect ratio of the (wide) dictionary matrix \(\Phi \in \mathbb{R}^{d \times n}\). Similarly, the sparsity factor \(s\) is the aspect ratio of the (tall) sub-matrix of \(\Phi\) containing the \(k\) atoms participating in the decomposition of the signal \(x\). Note that \(s\) is not the percentage of nonzeros in the coefficient vector \(\alpha\) in (1) (which would be \(k/n\)).

Our goal is to characterize the minimal overcompleteness \(o\) such that all/most signals in \(\mathbb{R}^d\) possess a sparse representation with sparsity \(s\), up to some permissible error \(\epsilon\). Before we state our main results, let us first give some intuition into why this problem is not trivial.

As mentioned above, each choice of \(k\) atoms from the dictionary corresponds to a single subspace of dimension at most \(k\). The volume of the set of signals whose normalized distance from this subspace is bounded by \(\epsilon\), can be computed in closed form (see Appendix). The problem is that when the number of atoms \(n\) tends to infinity while keeping \(k/n\) fixed, the relative overlap develops a sharp peak at \(k/n\) and its standard deviation tends to 0. Thus, when \(n\) is large, practically all pairs of groups of \(k\) atoms share \(k\) atoms. Here, \(k/n = 0.25\).

![Fig. 1. Given a set of \(n\) atoms, there exist \(\binom{n}{k}\) distinct subsets of \(k\) atoms. The graph shows the relative amount of pairs of such subsets as a function of their overlap. As \(n\) tends to infinity while keeping \(k/n\) fixed, the relative overlap develops a sharp peak at \(k/n\) and its standard deviation tends to 0. Thus, when \(n\) is large, practically all pairs of groups of \(k\) atoms share \(k\) atoms. Here, \(k/n = 0.25\).](image-url)

To overcome this difficulty, our upper bounds are based on a special type of (sub-optimal) structured dictionaries and also pose a certain (sub-optimal) restriction on the allowed choices of atoms from the dictionary. These assumptions significantly simplify the derivations, and while they may seem to lead to a crude overestimation of the required overcompleteness, we show that the resulting bounds are rather accurate in quite a wide range of settings.

A. Worst-case analysis

We begin by studying the problem from a worst-case standpoint.

**Definition 2 (Universal \(k\)-sparse representation dictionary):** We say that \(\Phi \in \mathbb{R}^{d \times n}\) is a universal \(k\)-sparse representation dictionary with precision \(\epsilon\) if all signals in \(\mathbb{R}^d\) admit a \(k\)-sparse representation with precision \(\epsilon\) over \(\Phi\), namely \(\epsilon(x, \Phi) \leq \epsilon\), \(\forall x \in \mathbb{R}^d\) (equivalently, \(\max_{x \in \mathbb{R}^d} \epsilon(x, \Phi) \leq \epsilon\)).

Let us denote by \(o^\dagger\) the minimal overcompleteness allowing universal sparse representation. In the theorems below we provide upper and lower bounds on \(o^\dagger\). These bounds are expressed only in terms of the sparsity \(s\) and allowable error \(\epsilon\). Particularly, they are independent of the dimension \(d\) (despite the fact that \(o^\dagger\) itself does depend on \(d\)).
Theorem 1 (Worst-case lower bound): If \( \varepsilon \in (0, \sqrt{1 - s}) \), then
\[
o^* \geq c_1(s, \varepsilon) \times s^2 \left( \frac{1}{\varepsilon} \right)^{\frac{1}{s - 1}}, \tag{3}
\]
where \( c_1(s, \varepsilon) = e^{-1}\sqrt{(1 - s)^{1/s - 1}/(1 - \varepsilon^2)} \geq e^{-\frac{2}{s}} \) for all \( s \) and \( \varepsilon \). If \( \varepsilon \in (\sqrt{1 - s}, 1) \), then
\[
o^* = 1. \tag{4}
\]

As expected, when either the sparsity factor \( s \) or the precision \( \varepsilon \) are small, the required overcompleteness is large. However, interestingly, the dependence on \( s \) and \( \varepsilon \) is quite different. While the bound is polynomial in \( \varepsilon^{-1} \), it is exponential in \( s^{-1} \), implying that universal sparse representation is practically impossible at very small sparsity factors.

Theorem 2 (Worst-case upper bound): If \( \varepsilon \in (0, 1) \), \( k \) is a divisor of \( d \) and \( s \leq \frac{1}{2} \), then
\[
o^* \leq c_2(s, \varepsilon) \times \log(s^{-1})s^{-\frac{1}{2}} \left( \frac{1}{\varepsilon} \right)^{\frac{1}{s - 1}}, \tag{5}
\]
where \( c_2(s, \varepsilon) = \sqrt{2\pi}(1 + \frac{2}{\log_2(s^{-1})})e^{\frac{\log\log_2(s^{-1})}{\log_2(s^{-1})}} + \frac{2}{\sqrt{\pi}}(1 - \varepsilon^{2s^{1/2}})\frac{1}{\sqrt{s}} \leq 12 \) for all \( s \) and \( \varepsilon \). If \( k \) is not a divisor of \( d \), then this bound holds true with \( s^{-1} \) replaced by \( [s^{-1}] \).

As can be seen, both bounds are exponentially equivalent \(^1\) to \( (1/\varepsilon)^{1/s - 1} \). This implies that under the conditions of Theorems 1 and 2, the minimal overcompleteness \( o^* \) satisfies
\[
o^* \approx \left( \frac{1}{s} - 1 \right) \text{SNR}_{\text{dB}}, \tag{6}
\]
where \( \approx \) denotes exponential equivalence. One can think of the representation error as noise, in which case the term \( 1/\varepsilon \) can be interpreted as the signal-to-noise ratio (SNR). Therefore, (6) can also be written as
\[
o^*_{\text{dB}} \approx \left( \frac{1}{s} - 1 \right) \text{SNR}_{\text{dB}}, \tag{7}
\]
where \( o^*_{\text{dB}} \) denotes asymptotic equivalence, \( \text{SNR}_{\text{dB}} = 20 \log_{10}(1/\varepsilon) \) and \( o^*_{\text{dB}} = 20 \log_{10}(o^*) \).

Exponential equivalence is agnostic to polynomial dependencies. Thus, to refine our intuition, it is instructive to examine the ratio between the bounds (5) and (3). As can be seen, this ratio is bounded from above as a function of \( \varepsilon \) and is only polynomial in \( s^{-1} \) (behaves as \( \Theta(\log(s^{-1})s^{-2}) \)). This indicates that our bounds are relatively accurate for moderate sparsity factors, even when \( \varepsilon \) is small, but may become inaccurate for very small \( s \).

B. Average-case analysis

The upper bound of Theorem 2 is rather pessimistic as it guarantees that all signals can be sparsely represented with precision \( \varepsilon \), including esoteric and unlikely signals. In many practical situations, it may be enough to loosen this requirement and replace it by a probabilistic one. Specifically, suppose we have prior knowledge in the form of a distribution \( \Omega \) over signals in \( \mathbb{R}^d \). In this case, it may be enough to settle for dictionaries allowing sparse representation only with high probability.

Definition 3 (Optimal success probability): For any given distribution \( \Omega \) over \( \mathbb{R}^d \) and dictionary size \( d \times n \), we define the optimal success probability under \( \Omega \) as
\[
P(\Omega) \triangleq \max_{\Phi \in \mathbb{R}^{d \times n}} P(\epsilon(x, \Phi) \leq \varepsilon), \tag{8}
\]
where \( x \) is a random vector with distribution \( \Omega \).

To obtain bounds that do not depend on the prior \( \Omega \), we will examine the worst-case optimal success probability over all possible distributions \( \Omega \). Mathematically, let \( D(\mathbb{R}^d) \) be the collection of all distributions over \( \mathbb{R}^d \). Then the worst-case optimal success probability is defined as
\[
P^* = \min_{\Omega \in \mathbb{R}^d} P(\Omega). \tag{9}
\]

Studying the behavior of \( P^* \) is particularly interesting in high dimensions. In this setting there is a sharp transition between the regime of overcompleteness factors at which \( P^* \) tends to 1 and the regime at which it tends to 0. We would therefore like to study the limit of \( P^* \) as \( d \) tends to infinity, while keeping the sparsity factor \( s \) fixed. To this end, we denote by \( o^* \) the minimal overcompleteness such that \( \lim_{d \to \infty} P^* = 1 \).

There are several important distinctions between \( o^* \) of the average case scenario and \( o^* \) of the worst case setting. First, \( o^* \) is only affected by typical signals, whereas \( o^* \) takes into account all signals. Therefore, we necessarily have that \( o^* \leq o^* \). Second, as can be seen from (8), it may be that for each distribution \( \Omega \) the optimal dictionary is different. Thus, as opposed to the worst-case analysis, in the average case setting we do not guarantee the existence of a single dictionary that is good for all signals. Finally, \( o^* \) is defined only for \( d \to \infty \), whereas \( o^* \) is defined for all \( d \). This is particularly important when bounding these quantities from above, since the minimal required overcompleteness becomes smaller as \( d \) increases.

This further contributes to our ability to obtain an upper bound on \( o^* \), which is lower than the upper bound on \( o^* \) in Theorem 2.

The next two theorems are analogous to Theorems 1 and 2. The first statement in each theorem bounds \( o^* \), and thus provides an asymptotic analysis. The second statement characterizes the convergence to the asymptotic behavior, and is relevant for any finite dimension.

Theorem 3 (Average-case lower bound): If \( \varepsilon \in (0, \sqrt{1 - s}) \), then
\[
o^* \geq c_1(s, \varepsilon) \times s^2 \left( \frac{1}{\varepsilon} \right)^{\frac{1}{s - 1}}, \tag{10}
\]
where \( c_1(s, \varepsilon) \) is as in (3). Furthermore, for any finite dimension \( d \),
\[
P^* \leq \frac{1}{\sqrt{2\pi s}} \exp\left[-c_3(s, \varepsilon, \alpha) d\right]. \tag{11}
\]

Here, \( c_3(s, \varepsilon, \alpha) = \frac{1}{2} H(1 - s, \varepsilon^2) - \alpha S_\alpha(z) \), where \( S(\alpha) = -\alpha \log(\alpha) - (1 - \alpha) \log(1 - \alpha) \) is the entropy of a Bernoulli(\( \alpha \)) random variable, and \( H(\alpha, \beta) = \alpha \log(\frac{\alpha}{\beta}) + (1 - \alpha) \log(\frac{1 - \alpha}{1 - \beta}) \) is the Kullback-Leibler divergence between the Bernoulli(\( \alpha \)) and Bernoulli(\( \beta \)) distributions.

In fact, as shown in [21, Sec. 3], when the overcompleteness \( o \) is smaller than the right hand side of (10), not only that \( P^* \) does not tend to 1, it actually tends to 0. This implies that below this bound, universal sparse representation is practically impossible in high dimensions (for the worst case distribution).

The next theorem is stated in terms of the incomplete regularized beta function \( I_\beta(\alpha, \beta) \), which is the probability that a Beta(\( \alpha, \beta \)) random variable is smaller than \( x \).

Theorem 4 (Average-case upper bound): If \( s = \frac{1}{m} \) for some \( m \in \mathbb{N} \), then
\[
o^* \leq c_4(s, \varepsilon) \times s^2 \left( \frac{1}{\varepsilon} \right)^{\frac{1}{s - 1}}, \tag{12}
\]
overcompleteness ratio satisfies

$$o \geq \frac{s}{I \cdot \frac{1}{2} \left( \frac{1}{2} - \frac{1}{s} \right)} = \Theta \left( s^2 \left( \frac{1}{\delta} \right)^{\frac{1}{2} - 1} \right)$$  \hspace{1cm} (13)

for some $\delta \in (0, \varepsilon)$, then

$$\mathcal{P}^* \geq 1 - \frac{c_5(s, \varepsilon, o, \delta)}{c_5(s, \varepsilon, o, \delta) + d^{1/5}}$$  \hspace{1cm} (14)

where $c_5(s, \varepsilon, o, \delta) = \frac{(1 - 2\varepsilon^2)(1 - 2\varepsilon^2)(\sqrt{2} - 1 + \varepsilon^2)}{8(1 + 2\varepsilon^2)}$ is a constant independent of the dimension $d$.

In [21] we also present tighter bounds on $o^*$ and $\mathcal{P}^*$, which do not have close-form expressions, but their value can be easily approximated numerically (see Eq. (74) and Eq. (78) in [21]).

Note that Theorem 3 provides a lower bound on the minimal required overcompleteness (see (10)) and an upper bound on the probability of success (see (11)). Similarly, Theorem 4 provides an upper bound on the minimal required overcompleteness (see (12)), which is further refined via a lower bound on the probability of success (see (13),(14)).

Comparing Theorems 2 and 4, it can be seen that the average-case analysis provides an improvement of $\Theta(s^{-1} \log s^{-1})$ over the worst-case analysis (compare (5) with (12)). Another interesting comparison is between the lower and upper bounds in the average-case scenario. In general, the ratio between (10) and (12) is a complicated function of $s$ and $\varepsilon$ which behaves as $\Theta(s^{-1})$. Yet, for certain regimes of $\varepsilon$ and $s$ we can obtain simple expressions. When $\varepsilon \ll 1$, the ratio becomes approximately $\sqrt{\frac{s^2}{2(1 - s^2)}} \cdot s^{-1}$. This expression is independent of $\varepsilon$, implying that the bounds are relatively tight for moderate values of $s$ when $\varepsilon$ is small. Similarly, when $s \ll 1$, the ratio becomes approximately $\sqrt{\frac{s^2}{2}(1 - s^2)} \cdot s^{-1}$.

III. NUMERICAL SIMULATIONS

In this section, we present simulations that demonstrate the bounds from Section II. Figure 2 depicts the values of the worst-case and average-case lower-bounds (3),(10), the worst-case upper bound (5) and the average-case upper bound (12), as functions of the allowed error $\varepsilon$ and the sparsity $s$. Note that the color scale in all plots is logarithmic, thus highlighting the fact that the minimal required overcompleteness becomes extremely large for small values of $\varepsilon$ and $s$. We can see the asymmetrical dependency of the overcompleteness on $s$ and $\varepsilon$, which is exponential in $s^{-1}$ and only polynomial in $\varepsilon^{-1}$. This illustrates that for small values of $s$, it is practically impossible to achieve universal sparse representation with any reasonable error $\varepsilon$ (the required overcompleteness is extremely large). However, for small values of $\varepsilon$, universal sparse representation may still be practical if the sparsity is not too small (e.g., $s \approx 0.3$).

To better visualize the differences between the bounds, Figs. 3 and 4 show slices from the two-dimensional surfaces. Specifically, Fig. 3
Fig. 5. Probability of succeeding to sparsely represent a white Gaussian signal, as a function of the dictionary’s overcompleteness. Here, the dimension is $d = 1600$, the sparsity factor is $s = 0.2$, and the permissible error is $\text{SNR} = 10\text{[dB]}$.

depicts the bounds as functions of $s^{-1}$ at a constant representation error $\varepsilon$ corresponding to $\text{SNR}_{\text{min}} = 20 \log_{10}(1/\varepsilon) = 10\text{dB}$. Figure 4 shows the bounds as functions of the SNR at a constant sparsity factor of $s = 0.2$. Here we can see that the worst-case upper bound is quite pessimistic with respect to the average case one.

We next compare our bounds to the actual performance of a sparse coding algorithm. To the best of our knowledge, there exists no practical method for calculating the worst-case error $\max_{x \in \mathbb{R}^d} \epsilon(x, \Phi)$ for a given dictionary $\Phi$. This means that we cannot verify whether a given $\Phi$ is a universal $k$-sparse representation dictionary. Consequently, we focus on examining only the average case scenario. In this setting, we take $x \in \mathbb{R}^d$ to be a Gaussian vector with i.i.d coordinates, which, according to [21, Lemma 4], is a worst case distribution. For any given $\Phi$, this allows us to easily approximate the probability of success $\mathbb{P}(\epsilon(x, \Phi) \leq \varepsilon)$, simply by applying the OMP algorithm on many draws of $x$ and counting the relative number of times the resulting sparse approximation satisfies our error constraint. This still leaves us with the problem of choosing the optimal $\Phi$. Since there is no closed form expression for the optimal $\Phi$, here we make a suboptimal choice, which is to take the dictionary to be a random matrix with Gaussian i.i.d entries. We note that according to [21, Lemma 6], the best possible probability of success is the same whether we restrict the search for deterministic dictionaries or also allow random dictionaries.

Figure 5 compares the probability of success of the OMP algorithm over a Gaussian dictionary to the lower and upper bounds on the probability of success (11) and (14) as well as to the lighter bound [21, Eq. (78)], which we calculated numerically. This simulation was carried out for a fixed dimension of $d = 1600$. As can be seen, the success probability is indeed between the upper and lower bounds. Moreover, it exhibits a sharp phase-transition at some overcompleteness. Below this critical overcompleteness, the probability of success is nearly 0, while above the threshold it climbs very steeply towards 1. Obviously, both our choice of sparse-coding algorithm and our choice of dictionary are suboptimal. Therefore, we must keep in mind that the simulation gives us an underestimation of the optimal performance (the true best achievable probability of success is actually higher than the black dash-dotted curve).

Figure 6 demonstrates the asymptotic behavior in Theorems 3 and 4 (i.e., the bounds in (10) and (12)). Here, we compare the bounds to the minimal overcompleteness that allows to obtain a sparse representation with overwhelming probability. Specifically, we set a threshold of 0.99 on the success probability, and numerically found the minimal overcompleteness that allowed to surpass this success rate. This was done by gradually increasing the overcompleteness ratio, for each dimension $d$, until we hit the 0.99 success probability threshold for the first time. As can be seen, at high dimensions $d$, the overcompleteness required for overwhelming success probability is indeed between the two bounds, as Theorems 3 and 4 predict.

IV. CONCLUSION

In this paper, we presented and studied the universal sparse representation problem, which relates to the ability of constructing sparse approximations to all signals in the space, up to a predefined error. We analyzed the problem in a deterministic setting as well as in a stochastic one. In both cases, we derived necessary and sufficient conditions on the minimal required overcompleteness. Our conditions have simple explicit forms, and, as we illustrated through simulations, accurately capture the behavior of sparse coding algorithms in practice.

APPENDIX

When analyzing the expressive power of sparse representation over redundant dictionaries, it is instructive to examine the representation capacity of a single choice of $k$ atoms from the dictionary. Therefore, our objective in this appendix is to evaluate\(^2\) the volume of signals form the unit sphere $S_{d-1}^d$, which can be approximately represented by a specific choice of $k$ atoms, up to an error of $\varepsilon$. This volume of signals corresponds to the intersection between $S_{d-1}^d$ and the $\varepsilon$-dilation of the subspace spanned by those atoms, as illustrated in Fig. 7. Let $\psi$ denote the subspace spanned by $k$ atoms. Then the relative area covered by an $\varepsilon$-dilation of $\psi$, can be interpreted as the probability that a point $x$ chosen uniformly at random from the unit sphere $S_{d-1}^d$ be $\varepsilon$-close to $\psi$. The square distance between $x$ and $\psi$ is $\|x - P_{\psi} x\|^2 = \|x\|^2 - \|P_{\psi} x\|^2$, where $P_{\psi}$ is the orthogonal projection matrix onto $\psi$. Thus, since $\|x\| = 1$, $x$ is $\varepsilon$-close to $\psi$ if and only if $\|P_{\psi} x\|^2 \geq 1 - \varepsilon^2$. This implies that we have the relation

$$A(d, k, \varepsilon) = \mathbb{P}(\|P_{\psi} x\|^2 \geq 1 - \varepsilon^2),$$

where $A(d, k, \varepsilon)$ denotes the relative covered area. Our problem thus boils down to determining the distribution of the length of a random vector with uniform distribution on the unit sphere, projected down

\(^2\)A similar analysis was provided in [3] for spaces over the field of complex numbers.
Thus, we conclude that for any \( \alpha \) in [20], which states onto a fixed \( k \)-dimensional subspace. This result can be found e.g., in [20], which states

\[
\|P_\varepsilon x\|^2 \sim \text{Beta} \left( \frac{k}{2}, \frac{d-k}{2} \right) \tag{16}
\]

From (16) we have that \( P(\|P_\varepsilon x\|^2 \geq 1 - \varepsilon^2) = 1 - I_{\varepsilon}^2 \left( \frac{k}{2}, \frac{d-k}{2} \right) \), where \( I_{\varepsilon}(\alpha, \beta) \) is the cumulative distribution function of the \( \text{Beta}(\alpha, \beta) \) distribution, also known as the regularized incomplete beta function. From the properties of the beta distribution, we can also write \( 1 - I_{\varepsilon}^2 \left( \frac{k}{2}, \frac{d-k}{2} \right) = I_{2\varepsilon} \left( \frac{d-k}{4}, \frac{k}{4} \right) \). Thus, from (15) and (16) we reach the following conclusion.

**Corollary 1 (Subspace coverage):** Let \( \psi \) be a \( d \)-dimensional linear subspace in \( \mathbb{R}^d \). Then the relative area covered by an \( \varepsilon \)-dilation of \( \psi \) from the unit sphere is precisely \( I_{\varepsilon} \left( \frac{d-k}{2}, \frac{k}{2} \right) \).

We remark that Corollary 1 can be seen as a generalization of [17], which proved it for one dimensional subspaces.

To obtain a better understanding of how large this volume is, let us upper bound it using the following tail bound from [7].

**Lemma 1 ([7]):** Let \( L \sim \text{Beta} \left( \frac{k}{2}, \frac{d-k}{2} \right) \), where \( k < d \) are natural numbers. Then for any \( \beta > 1 \),

\[
\Pr \left( L \geq \beta k \right) \leq \beta^\frac{k}{2} \left( 1 + \frac{(1 - \beta)k}{d - k} \right)^\frac{d-k}{2} \tag{17}
\]

To match our setting in (15), we choose \( \beta = 1 - \varepsilon^2 \). This relation implies that \( \beta = 1 - \varepsilon^2 \) and \( 1 + \frac{(1 - \beta)k}{d - k} = \varepsilon^2 \). Note that the lemma applies to \( \beta > 1 \), which translates to the requirement that \( \varepsilon^2 < 1 - s \). Applying the lemma, gives us the bound

\[
I_{\varepsilon} \left( \frac{d-k}{2}, \frac{k}{2} \right) \leq \left( \frac{1 - \varepsilon^2}{s} \right)^\frac{k}{2} \frac{d-k}{2} \tag{18}
\]

To present this bound more compactly, we use the function \( H(\alpha, \beta) = \frac{\log \left( \frac{\alpha}{\beta} \right) + (1 - \alpha) \log \left( 1 - \frac{\alpha}{\beta} \right)}{\log \left( \frac{\alpha}{\beta} \right)} \), which is the Kullback Leibler divergence between the Bernoulli(\( \alpha \)) and Bernoulli(\( \beta \)) distributions. Then, the right-hand-side of (18) can be equivalently expressed as

\[
\left( \frac{1 - \varepsilon^2}{s} \right)^\frac{k}{2} \frac{d-k}{2} = \exp \left\{ -\frac{d}{2} H(1-s, \varepsilon^2) \right\} \tag{19}
\]

Thus, we conclude that for any \( 0 \leq \varepsilon \leq \sqrt{1 - \frac{k}{d}} \),

\[
A(d, k, \varepsilon) \leq \exp \left\{ -\frac{d}{2} H(1 - \frac{k}{d}, \varepsilon^2) \right\} \tag{20}
\]

From this inequality we deduce that when \( k \) is taken to be a fixed fraction of \( d \), the set of signals in \( \mathbb{R}^d \) that can be approximately represented by a specific choice of \( k \) atoms has a volume which is exponentially small with \( d \) (and equivalently with \( k \)).

**References**


