Sparse Linear Representation

Halyun Jeong and Young-Han Kim
Dept. of ECE, UCSD
Email: {hajeong, yhk}@ucsd.edu

***Abstract*— This paper studies the question of how well a signal can be reprsented by a sparse linear combination of reference signals from an overcomplete dictionary. When the dictionary size is exponential in the dimension of signal, then the exact characterization of the optimal distortion is given as a function of the dictionary size exponent and the number of reference signals for the linear representation. Roughly speaking, every signal is sparse if the dictionary size is exponentially large, no matter how small the exponent is. Furthermore, an iterative method similar to matching pursuit that successively finds the best reference signal at each stage gives asymptotically optimal representations. This method is essentially equivalent to successive refinement for multiple descriptions and provides a simple alternative proof of the successive refinability of white Gaussian sources.**

I. Introduction and Main Results

Suppose one wishes to represent a signal as a linear combination of reference signals. If the collection $\mathcal{C}$ of reference signals (called *dictionary*) is rich (i.e., the size $M = |\mathcal{C}|$ of dictionary is much larger than the dimension n of the signal) or if one is allowed to take an arbitrarily complex linear combination (i.e., the number k of reference signals forming the linear combination is very large), then one can expect that the linear representation approximates the original signal with very little distortion. As a trivial example, if $\mathcal{C}$ contains n linearly independent reference signals of dimention n, then every signal can be represented faithfully as a linear combination of those n reference signals. On the other extreme point, if $\mathcal{C}$ includes all possible signals, then the original signal can be represented as (a linear combination of) itself without any distortion. More generally, Shannon's rate distortion theory [1] suggests that if the dictionary size $M = 2^{nR}$ is exponential in n with exponent $R > 0$, then the best reference signal (as a singleton) achives the distortion $D(R)$ given as a function of R.

Several interesting questions arise:

1) What will happen if the linear combination is sparse ($k \ll n$)? How well can one represent a signal as a (sparse) linear combination of reference signals?
2) How should one choose the dictionary of reference signals under the size limitation? Is there a dictionary that provides a good representation for all or most signals?
3) How can one find the best linear representation given the dictionary? Is there a low-complexity algorithm with optimal or near-optimal performance?

These questions arise in many applications and naturally have been studied in several different contexts [2]. The current paper provides partial answers to these questions by focusing on asymptotic relationship between the collection size M, the dimension n of the signal, the sparsity k of the representation, and the distortion D of the representation.

More formally, let $\mathcal{C} = \{\boldsymbol{\phi}(1), \boldsymbol{\phi}(2), \dots, \boldsymbol{\phi}(M)\}$ be a collection (dictionary) of M vectors in $\mathbb{R}^n$. For each vector $\mathbf{y} \in \mathbb{R}^n$, we define its best k-linear representation $\hat{\mathbf{y}}_k$ from the dictionary $\mathcal{C}$ as

$$\hat{\mathbf{y}}_k = x_1\boldsymbol{\phi}(m_1) + x_2\boldsymbol{\phi}(m_2) + \cdots + x_k\boldsymbol{\phi}(m_k),$$

where $x_1, \dots, x_k \in \mathbb{R}$ and $m_1, \dots, m_k \in [1 : M] := \{1, 2, \dots, M\}$ are chosen to minimize the squared error

$$d_k(\mathbf{y}, \mathcal{C}) = \|\mathbf{y} - (x_1\boldsymbol{\phi}(m_1) + x_2\boldsymbol{\phi}(m_2) + \cdots + x_k\boldsymbol{\phi}(m_k))\|^2.$$

Here the norm of a vector $\mathbf{z} = (z_1, \dots, z_n) \in \mathbb{R}^n$ is defined as $||\mathbf{z}|| = (\sum_{i=1}^n z_i^2)^{1/2}$.

We further define the worst-case distortion $d_k^*(\mathcal{C})$ of the dictionary $\mathcal{C}$ as

$$d_k^*(\mathcal{C}) := \sup_{\mathbf{y}: \|\mathbf{y}\|^2 \le 1} d_k(\mathbf{y}, \mathcal{C}),$$

where the supremum is taken over all n-vectors $\mathbf{y}$ in the (closed) unit sphere.

Note that $d_k^*(\mathcal{C}) \le 1$ for all $\mathcal{C}$ and all n, with $d_k^*(\mathcal{C}) = 1$ attained by a singleton dictionary $\mathcal{C} = \{\mathbf{0}\}$. Conversely, if $M < n$, then $d_k^*(\mathcal{C}) = 1$ for any dictionary $\mathcal{C}$ of size M. Hence, we consider the case $M \ge n$ only, that is, the case in which the dictionary is *overcomplete*.

Similarly, we define the average-case distortion $\bar{d}_k(\mathcal{C})$ of the dictionary $\mathcal{C}$ as

$$\bar{d}_k(\mathcal{C}) = E\left(d_k(\mathbf{Y}, \mathcal{C})\right),$$

where the expectation is taken with respect to a random signal $\mathbf{Y}$ uniformly drawn from the unit sphere $\{\mathbf{y} \in \mathbb{R}^n : \|\mathbf{y}\| \le 1\}$.

Now we are ready to state our main results. The first result concerns the existence of an asymptotically good dictionary.

Theorem 1: Suppose $M = M_n$ satisfies

$$\liminf_{n\to\infty} \frac{\log M}{n} > 0.$$

Then there exists a sequence of dictionaries $\mathcal{C}_n$ of respective sizes M_n such that

$$\limsup_{n\to\infty} \left[\log d_k^*(\mathcal{C}_n) + \frac{2k \log M}{n}\right] \le 0. \qquad (1)$$

In particular, if $k \to \infty$ as $n \to \infty$, then $d_k^*(\mathcal{C}_n) \to 0$.

An interesting implication of Theorem 1 is that if we choose a good dictionary of exponentially large size, no matter how small the exponent is, every signal is essentially sparse (say, $k = \log\log n$) with respect to that dictionary in the asymptotics.

The proof of Theorem 1 will be given in Section II. The major ingredients of the proof include Wyner's uniform sphere covering lemma [3] and its application in successive linear representation. Simply put, given a good dictionary for singleton representations ($k = 1$), we iteratively represent the signal, the error, the error of the error, etc. by scaling the same dictionary.

This representation method is intimately related to successive refinement coding [4]. Indeed, Theorem 1, specialized to $k = 1$, is essentially equivalent to Shannon's rate distortion theorem for white Gaussian sources [1]. At the same time, the representation method gives a very simple proof of successive refinability [4] and additive successive refinability [5] of white Gaussian sources under the mean squared error distortion.

It turns out that the asymptotic distortion in Theorem 1, which is achieved by the simple successive representation method, is in fact optimal. The following result, essentially due to Fletcher *et al.* [6], provides the performance bound for the optimal dictionary.

Theorem 2 ([6, Theorem 2]): For any sequence of dictionaries $\mathcal{C}_n$ of size $M = M_n$ and any nondecreasing sequence $k = k_n$,

$$\liminf_{n\to\infty} \left[\log \bar{d}_k(\mathcal{C}_n) + \frac{2\log\binom{M}{k}}{n-k} + c_n \right] \geq 0.$$

where

$$c_n = \log\frac{n}{n-k} + \frac{k}{n-k}\log\frac{n}{k}.$$

In particular, if k is bounded, then for any sequence of dictionaries $\mathcal{C}_n$ of size $M = M_n$,

$$\liminf_{n\to\infty} \left[\log \bar{d}_k(\mathcal{C}_n) + \frac{2k\log M}{n-k} \right] \geq 0.$$

Note that if $M = 2^{nR}$ for some $R > 0$ and k is a constant, then Theorem 2 implies that the average distortion is lower bounded by

$$\liminf_{n\to\infty} \left[\log \bar{d}_k(\mathcal{C}_n) + \frac{2k\log M}{n} \right] \geq 0.$$

(Therefore so is the worst-case distortion.) Thus the distortion bound in (1) Theorem 1 is tight when the dictionary size grows exponentially in n.

The asymptotic optimality of successive representation method provides a theoretical justification for matching pursuit [7] or similar greedy algorithms in signal processing. This conclusion is especially appealing since these iterative methods have linear complexity in dictionary size M (or even lower complexity if the dictionary has further structures), while finding the optimal representation in a single shot, even when tractable, can have much higher complexity. However, there are two caveats. First, the dictionary size here is exponential in the signal dimension. Second, the dictionary should represent all signals with singletons uniformly well.

In a broad context, these results are intimately related to *recovery* of sparse signals via linear measurements. Indeed, the sparse linear representation can be expressed as

$$\mathbf{y} = \Phi\mathbf{x} + \mathbf{z}, \tag{2}$$

where Φ is an $n \times M$ matrix with columns in $\mathcal{C}$, $\mathbf{x} \in \mathbb{R}^M$ is a sparse vector with k nonzero elements $x_1, \ldots, x_k$, and $\mathbf{z}$ is the representation error. The award-winning papers by Candes and Tao [8] and Donoho [9] showed that a sparse signal $\mathbf{x}$ can be reconstructed *exactly* and *efficiently* from the measurement $\mathbf{y}$ given by the underdetermined system (2) of linear equations (when the measurement noise $\mathbf{z} = 0$), opening up the exciting field of compressed sensing. There have been several follow-up discussions that connect compressed sensing and information theory; we refer the reader to [10], [11], [12], [13], [14], [15], [16] for various aspects of the connection between two fields.

While it is quite unfair to summarize in a single sentence a variety of problems studied by compressed sensing and more generally sparse signal recovery, the central focus therein is to recover the *true* sparse signal $\mathbf{x}$ from the measurement $\mathbf{y}$. In particular, when the measurement process is corrupted by noise, the main goal becomes mapping a noise-corrupted measurement output $\mathbf{y}$ to its corresponding cause $\mathbf{x}$ in an efficient manner.

The sparse signal representation problem is in a sense dual to the sparse signal recovery problem (just like source coding is dual to channel coding). Here the focus is on $\mathbf{y}$ and its representation (approximation). There is no *true* representation, and a good dictionary should have several alternative representations of similar distortions. As mentioned above, the problem of general—not necessarily linear—sparse representation (also called the sparse approximation problem) has a history of longer than a century [17], [18] and has been studied in several different contexts. Along with the parallel development in compressed sensing, the recent focus has been efficient algorithms and their theoretical properties; see, for example, [19], [20], [21].

In comparison, studies in [6], [22], and this paper focus on finding asymptotically optimal dictionaries, regardless of computational complexity,[1] and study the tradeoff among the sparsity of the representation, the size of the dictionary, and the fidelity of the approximation. For example, Fletcher *et al.* [6] found a lower bound on the approximation error using rate distortion theory for Gaussian sources with mean squared distortion. A similar lower bound is obtained by Akcakaya and Tarokh [22] based on careful calculation of volumes of spherical caps. Thus, the main contribution of this paper is twofold. First, our Theorem 1 shows that these lower bounds (in particular the one in [6, Theorem 2]) are tight in asymptotic when the dictionary size is exponential

[1] Fortuitously, the associated representation method is highly efficient.

in signal length. Second, we show that a simple successive representation method achieves the lower bound, revealing an intimate connection between sparse signal representation and multiple description coding.

The rest of the paper is organized as follows. We give the proof of Theorem 1 in Section II. In Section III, we digress a little to discuss the implication of Theorem 1 on successive refinement for multiple descriptions of white Gaussian sources and its dual—successive cancelation for additive white Gaussian noise multiple access channels. Finally, the proof of Theorem 2 is presented in Section IV.

II. Successive Linear Representation

In this section, we prove that there exists a codebook of exponential size that is asymptotically good for all signals and all sparsity level. More constructively, we demonstrate that a simple iterative representation method finds a good representation.

More precisely, we show that if $R_0' > R_0 > 0$, there exists a sequence of dictionaries $\mathcal{C}_n$ with sizes $M = 2^{nR_0'}$ such that for $n = n(R_0', R_0)$ sufficiently large,

$$d_k^*(\mathcal{C}_n) \leq 2^{-2kR_0}$$

for every k (independent of n). Since the above inequality holds for all $R_0 \in (0, R_0')$, we have

$$\limsup_{n\to\infty} \left[\log d_k^*(\mathcal{C}_n) + \frac{2k \log M}{n} \right] \leq 0.$$

The following result by Wyner [3] (rephrased for our application) is crucial in proving the above claim:

Lemma 1 (Uniform covering lemma): Given $D \in (0,1)$, let $R' > R(D) = (1/2)\log(1/D)$. Then, for $n = n(R', D)$ sufficiently large, there exists a dictionary $\mathcal{C}_n = \{\hat{\mathbf{y}}(m) : m \in [1 : 2^{nR'}]\}$ such that for all $\mathbf{y}$ in the sphere of radius r,

$$\min_{m\in[1:2^{nR'}]} ||\mathbf{y} - \hat{\mathbf{y}}(m)||^2 \leq r^2 D.$$

In particular, for all $\mathbf{y} \in \mathbb{R}^n$,

$$\min_{x\in\mathbb{R}} \min_{m\in[1:2^{nR'}]} ||\mathbf{y} - x\hat{\mathbf{y}}(m)||^2 \leq \|\mathbf{y}\|^2 D.$$

Note that Wyner's uniform covering lemma shows the existence of a dictionary sequence $\mathcal{C}_n$ satisfying

$$\limsup_{n\to\infty} d_1^*(\mathcal{C}_n) \leq D = 2^{-2R},$$

which is simply a restatment of the claim for $k = 1$.

Equipped with the lemma, it is straightforward to prove the desired claim for $k > 1$. Given an arbitrary $\mathbf{y}$ in the unit sphere, let $\hat{\mathbf{y}}(m)$ be the best singleton representation of $\mathbf{y}$ and $\mathbf{z}_1 = \mathbf{y} - \hat{\mathbf{y}}(m_1)$ be the resulting error. Then we find the best singleton representation $\hat{\mathbf{z}}_1 = x_2\hat{\mathbf{y}}(m_2)$ of $\mathbf{z}_1$ from the dictionary, resulting in the error $\mathbf{z}_2 = \mathbf{z}_1 - \hat{\mathbf{z}}_1$. In general, at the k-th iteration, the error $\mathbf{z}_{k-1}$ from the previous stage is represented by $\hat{\mathbf{z}}_{k-1} = x_k\hat{\mathbf{y}}(m_k)$, resulting in the error $\mathbf{z}_k$. Thus this process gives a k-linear representation of $\mathbf{y}$ as

$$\begin{aligned}
\mathbf{y} &= \hat{\mathbf{y}}(m_1) + \mathbf{z}_1 \\
&= \hat{\mathbf{y}}(m_1) + x_2\hat{\mathbf{y}}(m_2) + \mathbf{z}_2 \\
&= \cdots \\
&= \hat{\mathbf{y}}(m_1) + x_2\hat{\mathbf{y}}(m_2) + \cdots + x_k\hat{\mathbf{y}}(m_k) + \mathbf{z}_k.
\end{aligned}$$

But by simple induction and the uniform covering lemma, we have

$$\|\mathbf{z}_k\|^2 \leq D\|\mathbf{z}_{k-1}\|^2 \leq D^2\|\mathbf{z}_{k-2}\|^2 \leq D^{k-1}\|\mathbf{z}_1\|^2 \leq D^k,$$

which completes the proof of the claim. Note that each of k reprsentations attains mean square error 2^{-2jR_0} for its sparsity level $j = 1, \ldots, k$.

III. Successive Refinement for Gaussian Sources

The proof in the previous section leads to a deceptively simple proof of successive refinability of white Gaussian sources [23]. First note that in the successive linear representation method we can take $x_k = D^{(k-1)/2}$ for each k. Moreover, if $\mathbf{U} = (U_1, \ldots, U_n)$ is drawn independently and identically according to the standard normal distribution, then it can be shown that

$$E\left(\frac{1}{\sqrt{n}}\|\mathbf{U}\| \;\middle|\; \frac{1}{\sqrt{n}}\|\mathbf{U}\| > 1+\epsilon\right) \cdot \mathrm{P}\left(\frac{1}{\sqrt{n}}\|\mathbf{U}\| > 1+\epsilon\right) \to 0$$

as $n \to \infty$ for any $\epsilon > 0$. Hence, a good representation of a random vector $\mathbf{U}$ when the vector is inside the sphere of radius $(1+\epsilon)\sqrt{n}$ is sufficient to a good description of $\mathbf{U}$ in general.

Now our successive representation method achieves the (expected) mean square distortion $(1+\epsilon)D^k$ after k iterations with a dictionary of size $2^{nR'}$, where $R' > R(D) = (1/2)\log(1/D)$, which is nothing but the Gaussian rate distortion function. Hence, by describing the index of the sigleton reprsentation at each iteration using nR' bits, we can achieve distortion levels $D, D^2, \ldots, D^k$ and trace the Gaussian rate distortion function for $R', 2R', \ldots, kR'$. (Recall that we don't need to describe the scaling factors $x_k = D^{(k-1)/2}$, since these are constants independent of n.)

More generally, the same argument easily extends to the case in which incremental rates $R_1, R_2, \ldots, R_k$ are not necessarily identical; one can even prove the existence of *nested* codebooks (up to scaling) that uniformly cover the unit sphere.

Operationally, the recursive coding scheme for successive refinement (i.e., describing the error, the error of the error, and so on) can be viewed as a dual procedure to *successive cancelation* [24], [25] for the Gaussian multiple access channels, in which the messages for each user is peeled off iteratively. In both cases, one strives to best solve the single-user source [channel] coding problem at each stage and progresses recursively by subtracting off the encoded [decoded] part of the source [channel output] $\mathbf{y}$. This duality can be complemented by an interesting connection between the orthogonal matching pursuit and the sucessive cancelation [26] and the duality between signal recovery and signal representation. Note, however, that the duality here is mostly conceptual and cannot be made more precise. For example, while we can use a single codebook (dictionary) for each of k successive descriptions (again up to scaling) as shown above, one cannot use the same codebook for all k users in the Gaussian multiple access channel. If the channel gains are identical among users, it is impossible to distinguish who sent which message (from the same codebook), even without any additive noise! There is no uniform packing lemma that matches the Gaussian capacity function, to begin with.

IV. Lower Bound on the Distortion

We show that for any sequence of dictionaries $\mathcal{C}_n$ of size $M = M_n$ and any nondecreasing sequence $k = k_n$,

$$\log \bar{d}_k(\mathcal{C}_n) + \frac{2\log\binom{M}{k}}{n-k} + \log\frac{n}{n-k} + \frac{k}{n-k}\log\frac{n}{k} \geq o(1).$$

While a similar proof is given in [6, Theorem 2], we present our version for completeness, which slightly generalizes the proof in [6].

The basic idea of the proof between is to bound the mean square error between the random vector $\mathbf{Y}$ and its representation vector $\hat{\mathbf{Y}}$ by computing the mean square error between $\mathbf{Y}$ and $\hat{\mathbf{Y}}'$ (a quantized version of $\hat{\mathbf{Y}}$) and the quantization error (the mean square error between $\hat{\mathbf{Y}}$ and $\hat{\mathbf{Y}}'$). Then, the tradeoff between the error and the complexity of the representation is analyzed via rate distortion theory.Details are as follows.

Without loss of generality, assume that

$$\liminf_{n\to\infty} \bar{d}_k(\mathcal{C}_n) \leq D < 1.$$

Let $\hat{\mathbf{y}} = \hat{\mathbf{y}}(\mathbf{y}) = \sum_{i=1}^{k} x_i \boldsymbol{\phi}(m_i)$ be the best k-sparse linear representation of a given vector $\mathbf{y}$ in the unit sphere. Then $\hat{\mathbf{y}}$ can be rewritten as

$$\hat{\mathbf{y}} = \sum_{i=1}^{k} \lambda_i(\mathbf{y})\boldsymbol{\psi}_i(\mathbf{y}), \tag{3}$$

where $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_k$ form an orthonormal basis of the subspace spanned by $\boldsymbol{\phi}(m_1), \ldots, \boldsymbol{\phi}(m_k)$, uniquely obtained from the Gram–Schmidt orthogonalization. Since $\|\hat{\mathbf{y}}\|^2 = \sum_{i=1}^{k} \lambda_i^2 \leq 1$ from the orthogonality of the vectors $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_k$, $\lambda_i \in [-1, 1]$ for all i.

We consider two cases:

(a) *Bounded* k: Suppose the sequence $k = k_n$ is bounded. Since $c_n \to 0$ as $n \to \infty$ in Theorem 2 for any bounded sequence k, it is suffices to show that the following inequality holds for any sequence $\mathcal{C}_n$ of dictionaries for a bounded sequence k.

$$\frac{2\log\binom{M}{k}}{n-k} + \log \bar{d}_k(\mathcal{C}_n) \geq o(1).$$

Next, we approximate $\hat{\mathbf{y}}$ by quantizing $\lambda_1, \ldots, \lambda_k$ into $\lambda_1', \ldots, \lambda_k' \in \left\{-1, -\frac{l_n-1}{l_n}, \ldots, -\frac{1}{l_n}, 0, \frac{1}{l_n}, \frac{2}{l_n}, \ldots, \frac{l_n-1}{l_n}, 1\right\}$ with quantization step size $1/l_n$. Let

$$\hat{\mathbf{y}}'(\mathbf{y}) = \sum_{i=1}^{k} \lambda_i'(\mathbf{y})\boldsymbol{\psi}_i(\mathbf{y}). \tag{4}$$

Then, $\|\hat{\mathbf{y}} - \hat{\mathbf{y}}'\|^2 \leq k(1/l_n)^2 = k/l_n^2$ Since $\hat{\mathbf{y}}$ is obtained by orthogonal projection of $\mathbf{y}$ to the subspace spanned by $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_k$ and $\hat{\mathbf{y}}'$ is a vector in the subspace, $\mathbf{y} - \hat{\mathbf{y}}$ and $\hat{\mathbf{y}} - \hat{\mathbf{y}}'$ are orthogonal. Thus, we have

$$\begin{aligned}\|\mathbf{y} - \hat{\mathbf{y}}'\|^2 &= \|\mathbf{y} - \hat{\mathbf{y}}\|^2 + \|\hat{\mathbf{y}} - \hat{\mathbf{y}}'\|^2 \\ &\leq \|\mathbf{y} - \hat{\mathbf{y}}\|^2 + k/l_n^2 =: d_k(\mathbf{y}, \mathcal{C}_n) + \epsilon_n.\end{aligned}$$

Now consider a random signal $\mathbf{Y}$ drawn uniformly from the unit sphere and its quantized representation

$$\hat{\mathbf{Y}}' = \sum_{i=1}^{n} \lambda_i'(\mathbf{Y})\boldsymbol{\psi}_i(\mathbf{Y}). \tag{5}$$

Then, we have $\|\mathbf{Y} - \hat{\mathbf{Y}}'\|^2 \leq \bar{d}_k(\mathcal{C}_n) + \epsilon_n$.

We have the following chain of inequalities:

$$\begin{aligned}\log\binom{M}{k} &+ k\log(2l_n+1) \\ &\geq H(m_1(\mathbf{Y}), \ldots, m_k(\mathbf{Y}), \lambda_1'(\mathbf{Y}), \ldots, \lambda_k'(\mathbf{Y})) \\ &\geq H(\hat{\mathbf{Y}}') \\ &\geq R(\bar{d}_k(\mathcal{C}_n) + \epsilon_n),\end{aligned} \tag{6}$$

where

$$R(D) = \min_{p(\hat{\mathbf{y}}'|\mathbf{y}): E[\|\mathbf{Y} - \hat{\mathbf{Y}}'\|^2] \leq \bar{d}_k(\mathcal{C}_n) + \epsilon_n} I(\mathbf{Y}; \hat{\mathbf{Y}}')$$

is the rate distortion function for $\mathbf{Y}$ under the mean square distortion $\bar{d}_k(\mathcal{C}_n) + \epsilon_n$.

Here are justification for above steps. The first inequality follows from the ranges of the number of k-dimensional subspaces and λ_j'. The second inequality follows from the fact that $\hat{\mathbf{Y}}'$ is a function of $(m_1(\mathbf{Y}), \ldots, m_k(\mathbf{Y}), \lambda_1'(\mathbf{Y}), \ldots, \lambda_k'(\mathbf{Y}))$. The last inequality follows from the rate distortion theorem.

By the Shannon lower bound on rate distortion function and the (Euclidean) volume of the unit sphere,

$$\begin{aligned}R(D) &\geq h(\mathbf{Y}) - \frac{n}{2}\log(2\pi e(\bar{d}_k(\mathcal{C}_n) + \epsilon_n)) \\ &\geq \frac{n}{2}\log\left(\frac{1}{\bar{d}_k(\mathcal{C}_n) + \epsilon_n}\right) - \log(\pi n) - \frac{1}{6n}.\end{aligned}$$

Combined together with (6), this yields

$$\frac{1}{n}\left(\log\binom{M}{k} + k\log(2l_n+1)\right) \tag{7}$$

$$\geq \frac{1}{2}\log\left(\frac{1}{\bar{d}_k(\mathcal{C}_n) + \epsilon_n}\right) - \frac{\log(\pi n)}{n} - \frac{1}{6n^2} \tag{8}$$

$$= \frac{1}{2}\log\left(\frac{1}{\bar{d}_k(\mathcal{C}_n)}\right) + \frac{1}{2}\log\left(\frac{1}{1 + \epsilon_n/\bar{d}_k(\mathcal{C}_n)}\right) - o(1). \tag{9}$$

Now, let f_n be an increasing sequence satisfying

$$\lim_{n\to\infty} f_n = \infty \quad \text{and} \quad \lim_{n\to\infty} \frac{\log f_n}{n} = 0, \qquad (10)$$

and take $l_n = f_n(\bar{d}_k(\mathcal{C}_n))^{-\frac{1}{2}}$. By plugging l_n to (8), we have

$$\frac{1}{n}\left(\log\binom{M}{k} + k\log\left(2f_n\Big/\sqrt{\bar{d}_k(\mathcal{C}_n)}+1\right)\right)$$
$$\geq \frac{1}{2}\log\left(\frac{1}{\bar{d}_k(\mathcal{C}_n)}\right) + \frac{1}{2}\log\left(\frac{1}{1+\epsilon_n/\bar{d}_k(\mathcal{C}_n)}\right) - o(1).$$

Arranging the terms in the above inequality yields

$$\frac{1}{n}\left(\log\binom{M}{k} + k\log\left(2f_n+\sqrt{\bar{d}_k(\mathcal{C}_n)}\right)\right) + o(1)$$
$$\geq \frac{n-k}{2n}\log\left(\frac{1}{\bar{d}_k(\mathcal{C}_n)}\right) + \frac{1}{2}\log\left(\frac{1}{1+\epsilon_n/\bar{d}_k(\mathcal{C}_n)}\right).$$

Then, we can note that $\epsilon_n/\bar{d}_k(\mathcal{C}_n) = (k/l_n^2)\bar{d}_k(\mathcal{C}_n) = k/f_n^2$ and $e_n/\bar{d}_k(\mathcal{C}_n) \to 0$ as $n \to \infty$. Also, from (9) $\left(k\log(2f_n+\sqrt{\bar{d}_k(\mathcal{C}_n)})\right)\Big/n \leq (k\log(2f_n+1))/n \to 0$ as $n \to \infty$. Hence, taking the limit $n \to \infty$ to the last inequality, we get

$$\liminf_{n\to\infty}\left[\log\bar{d}_k(\mathcal{C}_n) + \frac{2k\log M}{n-k}\right] \geq 0.$$

Finally, it is easy to show that the inequality in Theorem 2 reduces to the above inequality for the case when k is bounded.

(b) *Unounded* k: In this case, the scalar quantization in part (a) gives a loose bound. Wyner's uniform covering lemma, however, can be applied to provide a sharper tradeoff between the description complexity and the quantization error.

We continue the proof from the orthogonal representation of $\hat{\mathbf{y}}$ in (3). Since $\hat{\mathbf{y}}$ is a vector with length ≤ 1 in the k-dimensional subspace spanned by $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_k$ and k_n is an increasing sequence, we can invoke the unform covering lemma. Therefore, there must exist a dictionary $\mathcal{C}'_k$ of size 2^b and $\hat{\mathbf{y}}' \in \mathcal{C}'_k$ satisfying

$$\|\hat{\mathbf{y}} - \hat{\mathbf{y}}'\|^2 \leq 2^{-2b/k}.$$

Following the same arguments as in (5)–(9), we have

$$\frac{1}{n}\left(\log\binom{M}{k} + b\right) \geq \frac{1}{2}\log\left(\frac{1}{\bar{d}_k(\mathcal{C}_n)+2^{-2b/k}}\right) - o(1).$$

Finally, optimizing over b yields

$$\bar{d}_k(\mathcal{C}_n) \geq 2^{-2\log\binom{M}{k}/(n-k)} \cdot \left(\frac{n-k}{n}\right) \cdot \left(\frac{k}{n}\right)^{k/(n-k)}.$$

Taking the logarithm and letting $n \to \infty$ on both sides, we have the desired inequality.

Acknowledgments

The authors wish to thank Yuzhe Jin and Bhaskar Rao for stimulating discussions on their formulation of the sparse signal position recovery problem, which motivated the current work.

References

[1] C. E. Shannon, "Coding theorems for a discrete source with a fidelity criterion," in *IRE Int. Conv. Rec., part 4*, 1959, vol. 7, pp. 142–163, reprinted with changes in *Information and decision processes*, R. E. Machol, Ed. New York: McGraw-Hill, 1960, pp. 93-126.

[2] S. Mallat, *A Wavelet Tour of Signal Processing: The Sparse Way*. London: AP Professional, 2009.

[3] A. D. Wyner, "Random packings and coverings of the unit n-sphere," *Bell System Tech. J.*, vol. 46, pp. 2111–2118, 1967.

[4] W. H. R. Equitz and T. M. Cover, "Successive refinement of information," *IEEE Trans. Inf. Theory*, vol. IT-37, no. 2, pp. 269–275, 1991.

[5] E. Tuncel and K. Rose, "Additive successive refinement," *IEEE Trans. Inf. Theory*, vol. IT-49, no. 8, pp. 1983–1991, 2003.

[6] A. K. Fletcher, S. Rangan, V. K. Goyal, and K. Ramchandran, "Denoising by sparse approximation: Error bounds based on rate–distortion theory," in *EURASIP J. Applied Signal Processing, Special Issue on Frames and Overcomplete Representations*, vol. 2006, 2006.

[7] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Processing*, vol. 41, no. 12, pp. 3397–3415, Dec 1993.

[8] E. J. Candes and T. Tao, "Near-optimal signal recovery from random projections: universal encoding strategies?" *IEEE Trans. Inf. Theory*, vol. IT-52, no. 12, pp. 5406–5425, 2006.

[9] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. IT-52, no. 4, pp. 1289–1306, 2006.

[10] S. Sarvotham, D. Baron, and R. G. Baraniuk, "Measurements vs. bits: Compressed sensing meets information theory," in *Proceedings of 44th Allerton Conf. Comm., Ctrl., Computing*, Monticello, IL, 2006.

[11] S. Aeron, M. Zhao, and V. Saligrama, "Fundamental tradeoffs between sparsity, sensing diversity and sensing capacity," in *Proceedings of the 40th Asilomar Conf. Signals, Systems, and Computers*, Pacific Grove, CA, 2006.

[12] A. K. Fletcher, S. Rangan, and V. K. Goyal, "On the rate-distortion performance of compressed sensing," in *Proc. ICASSP*, Honolulu, HI, 2007.

[13] M. Wainwright, "Information-theoretic bounds on sparsity recovery in the high-dimensional and noisy setting," in *Proc. IEEE International Symposium on Information Theory*, Nice, France, June 2007.

[14] F. Zhang and H. D. Pfister, "Compressed sensing and linear codes over real numbers," in *Proceedings of the UCSD Information Theory and Applications Workshop*, La Jolla, CA, 2008.

[15] W. Dai, H. V. Pham, and O. Milenkovic, "Distortion-rate functions for quantized compressive sensing," *CoRR*, vol. abs/0901.0749, 2009.

[16] Y. Jin and B. Rao, "Insights into the stable recovery of sparse solutions in overcomplete representations using network information theory," in *Proc. ICASSP*, Las Vegas, NV, 2008.

[17] V. N. Temlyakov, "Nonlinear methods of approximation," *Found. Comput. Math.*, vol. 3, no. 1, pp. 33–107, 2003.

[18] E. Schmidt, "Zur Theorie der linearen und nichtlinearen Integralgleichungen," *Math. Ann.*, vol. 63, no. 4, pp. 433–476, 1907.

[19] J. A. Tropp, "Greed is good: algorithmic results for sparse approximation," *IEEE Trans. Inf. Theory*, vol. IT-50, no. 10, pp. 2231–2242, 2004.

[20] ——, "Just relax: convex programming methods for identifying sparse signals in noise," *IEEE Trans. Inf. Theory*, vol. IT-52, no. 3, pp. 1030–1051, 2006.

[21] D. L. Donoho, M. Elad, and V. N. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Trans. Inf. Theory*, vol. IT-52, no. 1, pp. 6–18, 2006.

[22] M. Akcakaya and V. Tarokh, "A frame construction and a universal distortion bound for sparse representations," *IEEE Trans. Signal Processing*, vol. 56, no. 6, pp. 2443–2450, June 2008.

[23] Y.-H. Kim, "Multiple descriptions with codebook reuse," in *Proceedings of the 42th Asilomar Conf. Signals, Systems, and Computers*, Pacific Grove, CA, 2008.

[24] T. M. Cover, "Some advances in broadcast channels," in *Advances in Communication Systems*, A. J. Viterbi, Ed. San Francisco: Academic Press, 1975, vol. 4, pp. 229–260.

[25] A. D. Wyner, "Recent results in the Shannon theory," *IEEE Trans. Inf. Theory*, vol. IT-20, pp. 2–10, 1974.

[26] Y. Jin and B. Rao, "Performance limits of matching pursuit algorithms," in *Proc. IEEE International Symposium on Information Theory*, Toronto, ON, July 2008.