Where is Randomness Needed to Break the Square-Root Bottleneck?

Patrick Kuppinger, Giuseppe Durisi, and Helmut Bölcskei
ETH Zurich, 8092 Zurich, Switzerland
E-mail: {patricku, gdurisi, boelcskei}@nari.ee.ethz.ch

***Abstract*—As shown by Tropp, 2008, for the concatenation of two orthonormal bases (ONBs), breaking the square-root bottleneck in compressed sensing does not require randomization over all the positions of the nonzero entries of the sparse coefficient vector. Rather the positions corresponding to one of the two ONBs can be chosen arbitrarily. The two-ONB structure is, however, restrictive and does not reveal the property that is responsible for allowing to break the bottleneck with reduced randomness. For general dictionaries we show that if a subdictionary with small enough coherence and large enough cardinality can be isolated, the bottleneck can be broken under the same probabilistic model on the sparse coefficient vector as in the two-ONB case.**

I. Introduction

The central idea underlying compressed sensing (CS) is to recover a sparse signal from as few non-adaptive linear measurements as possible [1], [2]. Given the measurement outcome $\mathbf{y} \in \mathbb{C}^M$ and the measurement matrix $\mathbf{D} \in \mathbb{C}^{M \times N}$ ($M \leq N$), often referred to as dictionary,[1] we want to find the sparsest coefficient vector $\mathbf{x} \in \mathbb{C}^N$ that is consistent with the measurement outcome, i.e., that satisfies $\mathbf{y} = \mathbf{D}\mathbf{x}$. This problem can be formalized as follows:

$$\text{(P0)} \quad \text{find } \arg\min \|\mathbf{x}\|_0 \quad \text{subject to } \mathbf{y} = \mathbf{D}\mathbf{x}.$$

Here, $\|\mathbf{x}\|_0$ denotes the number of nonzero entries of the vector $\mathbf{x}$. Unfortunately, solving (P0) for practically relevant problem sizes N, M is infeasible as it requires a combinatorial search. Instead, the CS literature has focused on the convex relaxation of (P0), i.e., on the following ℓ_1-minimization problem:

$$\text{(P1)} \quad \text{find } \arg\min \|\mathbf{x}\|_1 \quad \text{subject to } \mathbf{y} = \mathbf{D}\mathbf{x}$$

commonly referred to as *basis pursuit* (BP) [3]–[8]. Here, $\|\mathbf{x}\|_1 \triangleq \sum_{i=1}^{N} |x_i|$ denotes the ℓ_1-norm of $\mathbf{x}$. Since (P1) can be cast as a linear program (in the real case) or a second-order cone program (in the complex case), it can be solved more efficiently than (P0).

It is now natural to ask under which conditions the solutions of (P0) and (P1) are unique and coincide. A sufficient condition for this to happen[2] [4]–[6] is $\|\mathbf{x}\|_0 < S$, where the *sparsity threshold* $S = (1+1/d)/2$ depends on the dictionary *coherence* $d = \max_{i \neq j} |\mathbf{d}_i^H \mathbf{d}_j|$. Sparsity thresholds S larger than $(1 + 1/d)/2$ can be established if more information on the dictionary is available [6]–[9], e.g., if the dictionary consists of the concatenation of two or more orthonormal bases (ONBs), or—more generally—if a sufficiently large sub-dictionary with coherence much smaller than d can be isolated [9]. We emphasize that the results in [4]–[9] apply to *all* vectors $\mathbf{x}$ with $\|\mathbf{x}\|_0 < S$—irrespective of the positions and the values of the nonzero entries of $\mathbf{x}$.

The line of work presented in [4]–[9] leads to sparsity thresholds S that are on the order of $1/d$. From the Welch lower bound [10]

$$d \geq \sqrt{(N-M)/[M(N-1)]}$$

we can conclude that the thresholds in [4]–[9] are at best on the order of $\sqrt{M}$ (for $N \gg M$). This scaling behavior is sometimes referred to as the *square-root bottleneck*. A better scaling behavior can be obtained by asking for sparsity thresholds that hold for almost all—rather than all (as in [4]–[9])—vectors $\mathbf{x}$, or, more precisely, by asking for sparsity thresholds that hold with high probability, given a probabilistic model on $\mathbf{x}$.[3] Following the terminology used in [11], we refer to sparsity thresholds that hold for almost all S-sparse vectors $\mathbf{x}$ as *robust sparsity thresholds*.

The improvements in the scaling behavior that result from the relaxation to robust sparsity thresholds will, of course, depend on the probabilistic model on $\mathbf{x}$ [11]–[13]. A widely used probabilistic model for n-sparse vectors $\mathbf{x}$ is to choose the positions of the n nonzero entries (i.e., the *sparsity pattern*) of $\mathbf{x}$ uniformly at random among all possible $\binom{N}{n}$ support sets of cardinality n. The values of these nonzero entries of $\mathbf{x}$ are drawn from a continuous probability distribution, with the additional constraint that their phases are i.i.d. and uniformly distributed on $[0, 2\pi)$ [11], [12]. For this probabilistic model it

[1] Throughout the paper, we assume that the columns $\mathbf{d}_i$ of $\mathbf{D}$ have unit ℓ_2-norm, i.e., $\|\mathbf{d}_i\|_2 = 1$ for $i = 1, \ldots, N$.

[2] In the remainder of the paper, whenever we speak of a vector $\mathbf{x}$, we implicitly assume that this vector is consistent with the observation $\mathbf{y}$, i.e., it satisfies $\mathbf{y} = \mathbf{D}\mathbf{x}$.

[3] An alternative approach, which we do not pursue in this paper, is to introduce a probabilistic model on the dictionary $\mathbf{D}$ [1], [2].

is shown in [12] that the square-root bottleneck can be broken. More specifically, the main result in [12] states that, assuming a dictionary with coherence on the order of $1/\sqrt{M}$, a robust sparsity threshold on the order of $M/(\log N)$ can be obtained. Put differently, this result shows that to recover almost all vectors $\mathbf{x}$ with S nonzero entries, the required number of non-adaptive linear measurements M is (order-wise) $S \log N$ instead of S^2.

Remarkably, for dictionaries that consist of the concatenation of two ONBs, robust sparsity thresholds on the order of $M/(\log N)$ can be obtained with reduced randomness as compared to the case of general dictionaries. Specifically, it was found in [11], [12] that it suffices to pick the positions of the nonzero entries of $\mathbf{x}$ corresponding to one of the two ONBs uniformly at random, while the positions of the remaining nonzero entries can be chosen arbitrarily. The probabilistic model on the values of the nonzero entries of $\mathbf{x}$ (corresponding to *both* ONBs) remains the same as for the general dictionaries considered in [12].

Contributions: The two-ONB result in [11], [12] is interesting as it shows that one need not choose the locations of all the nonzero entries of the sparse vector randomly to break the square-root bottleneck. However, the two-ONB structure is restrictive and does not reveal which property of the dictionary is responsible for allowing to break the square-root bottleneck with reduced randomness. The two ONBs are on equal footing.

The purpose of this paper is twofold. First, we extend the two-ONB result in [11], [12] to general dictionaries. Second, by virtue of this extension, we show that—for a general dictionary $\mathbf{D}$ with low coherence d—the fundamental property needed to break the square-root bottleneck with reduced randomness is the presence of a sufficiently large sub-dictionary $\mathbf{A}$ with coherence much smaller than d. The positions of the nonzero entries of $\mathbf{x}$ corresponding to $\mathbf{A}$ can be chosen arbitrarily, and the positions of the remaining nonzero entries must be chosen randomly. Naturally, the larger the sub-dictionary $\mathbf{A}$, the more significant the reduction in randomness becomes. Randomization over the remaining part of the dictionary ensures that the sparsity patterns that cannot be recovered through BP occur with small enough probability. More formally, we prove the following result. Consider a general dictionary $\mathbf{D}$ with coherence on the order of $1/\sqrt{M}$ that contains a sub-dictionary $\mathbf{A}$ with coherence on the order of $(\log N)/M$ and cardinality at least on the order of $M/(\log N)$. Then, a robust sparsity threshold on the order of $M/(\log N)$ can be established—and hence the square-root bottleneck is broken—under the same probabilistic model on the vector $\mathbf{x}$ as in the two-ONB case, whenever the spectral norms of $\mathbf{A}$ and of the sub-dictionary containing the remaining columns of $\mathbf{D}$ satisfy certain technical conditions. These technical conditions are trivially satisfied, e.g., for dictionaries that consist of two tight frames.

Our analysis relies heavily on the mathematical tools developed in [12] for the two-ONB setting.

Notation: Throughout the paper, we use lowercase boldface letters for column vectors, e.g., $\mathbf{x}$, and uppercase boldface letters for matrices, e.g., $\mathbf{D}$. For a given matrix $\mathbf{D}$, we denote its conjugate transpose by $\mathbf{D}^H$ and $\mathbf{d}_i$ stands for its ith column. The spectral norm of a matrix $\mathbf{D}$ is $\|\mathbf{D}\| = \sqrt{\lambda}$, where λ is the maximum eigenvalue of $\mathbf{D}^H\mathbf{D}$. The minimum and maximum singular value of a matrix $\mathbf{D}$ are denoted by $\sigma_{\min}(\mathbf{D})$ and $\sigma_{\max}(\mathbf{D})$, respectively, $\operatorname{rank}(\mathbf{D})$ stands for the rank of $\mathbf{D}$, and $\|\mathbf{D}\|_{1,2} = \max_i\{\|\mathbf{d}_i\|_2\}$. We use $\mathbf{I}_n$ to denote the $n \times n$ identity matrix and $\mathbf{0}$ stands for the all-zero matrix of appropriate size. The natural logarithm is denoted as log. For two functions $f(M)$ and $g(M)$, the notation $f(M) = \mathcal{O}(g(M))$ means that $\lim_{M\to\infty}|f(M)|/|g(M)|$ is bounded above by a finite constant, and $f(M) = \Theta(g(M))$ means that there exist two positive finite constants k_1 and k_2 such that $k_1 \leq \lim_{M\to\infty}|f(M)|/|g(M)| \leq k_2$. Whenever we say that a vector $\mathbf{x} \in \mathbb{C}^N$ has a *randomly* chosen sparsity pattern of cardinality n, we mean that the support set of $\mathbf{x}$ is chosen uniformly at random among all $\binom{N}{n}$ possible support sets of cardinality n.

II. Brief Review of Previous Relevant Results

Robust sparsity thresholds for dictionaries consisting of two ONBs were first obtained in [11] and later improved in [12]. In Theorem 1 below, we restate the result in [12] in a slightly modified form, which is better suited to draw parallels to the more general case. The theorem follows by combining Theorems D, 13, and 14 in [12].

Theorem 1: Assume that[4] $N > 2$. Let $\mathbf{D} \in \mathbb{C}^{M\times N}$ be the concatenation of two ONBs $\mathbf{A}$ and $\mathbf{B}$ for $\mathbb{C}^M$ (i.e., $N = 2M$) and denote the coherence of $\mathbf{D}$ as d. Fix $s \geq 1$. Let the vector $\mathbf{x} \in \mathbb{C}^N$ have an *arbitrarily* chosen sparsity pattern of n_a nonzero entries corresponding to columns of sub-dictionary $\mathbf{A}$ and a *randomly* chosen sparsity pattern of n_b nonzero entries corresponding to columns of sub-dictionary $\mathbf{B}$. Suppose that

$$n_a + n_b < \min\{c\, d^{-2}/(s \log N), d^{-2}/2\} \tag{1}$$

where c is no smaller than 0.004212. If the values of *all* nonzero entries of $\mathbf{x}$ are drawn from a continuous probability distribution, $\mathbf{x}$ is the unique solution of (P0) with probability exceeding $(1 - N^{-s})$. Furthermore, if n_a and n_b, in addition to (1), satisfy

$$n_a + n_b \leq d^{-2}/[8(s+1)\log N] \tag{2}$$

and the phases of *all* nonzero entries of $\mathbf{x}$ are i.i.d. and uniformly distributed on $[0, 2\pi)$, then $\mathbf{x}$ is the unique solution of both (P0) and (P1) with probability exceeding $(1 - 3N^{-s})$.

[4] In [12] $M \geq 3$ (and hence $N \geq 6$) is assumed. However, it can be shown that $N > 2$ is sufficient to establish the result.

Interpretation of Theorem 1: Assume that $\mathbf{D}$ has coherence $d = \mathcal{O}(1/\sqrt{M})$. As a consequence of (1) and (2), Theorem 1 establishes (under certain technical conditions on the values of the nonzero entries of $\mathbf{x}$) the robust sparsity threshold[5] $S > n_a + n_b = \Theta(M/(\log N))$.

This result is interesting as it shows that we do not need the entire sparsity pattern of $\mathbf{x}$ to be chosen at random but rather the positions of the non-zero entries corresponding to one of the two ONBs can be chosen arbitrarily.

In the following section, we first present (in Theorem 2) an extension of the two-ONB result in [11], [12] to general dictionaries. As a consequence of Theorem 2, we then establish that—for a general dictionary $\mathbf{D}$ with low coherence d—the fundamental property that allows to break the square-root bottleneck with reduced randomness is the presence of a sufficiently large sub-dictionary $\mathbf{A}$ with coherence much smaller than d.

III. Main Results

Consider a dictionary $\mathbf{D} = [\mathbf{A}\ \mathbf{B}]$, where the sub-dictionary $\mathbf{A}$ has N_a elements (i.e., columns) and coherence a and the sub-dictionary $\mathbf{B}$ has $N_b = N - N_a$ elements and coherence b. The set of all such dictionaries is denoted as $\mathcal{D}(d, a, b)$. Correspondingly, we view the vector $\mathbf{x}$ as the concatenation of the two vectors $\mathbf{x}_a \in \mathbb{C}^{N_a}$ and $\mathbf{x}_b \in \mathbb{C}^{N_b}$ such that $\mathbf{y} = \mathbf{D}\mathbf{x} = \mathbf{A}\mathbf{x}_a + \mathbf{B}\mathbf{x}_b$. Since $\mathbf{A}$ and $\mathbf{B}$ are sub-dictionaries of $\mathbf{D}$, we have $a, b \leq d$. We now state our main result.

Theorem 2: Assume that $N > 2$. Let $\mathbf{D} = [\mathbf{A}\ \mathbf{B}]$ be a dictionary in $\mathcal{D}(d, a, b)$. Fix $s \geq 1$ and $\gamma \in [0, 1]$. Consider a random vector $\mathbf{x} = \left[\mathbf{x}_a^T\ \mathbf{x}_b^T\right]^T$ where $\mathbf{x}_a$ has an *arbitrarily* chosen sparsity pattern of cardinality n_a such that

$$6\sqrt{2}\sqrt{n_a d^2 s \log N} + 2(n_a - 1)a \leq (1-\gamma)e^{-1/4} \qquad (3)$$

and $\mathbf{x}_b$ has a *randomly* chosen sparsity pattern of cardinality n_b such that

$$24\sqrt{n_b b^2 s \log N} + \frac{4n_b}{N_b}\|\mathbf{B}\|^2 + 2\sqrt{\frac{n_b}{N_b}}\|\mathbf{A}\|\|\mathbf{B}\| \leq \gamma e^{-1/4}. \qquad (4)$$

Furthermore, assume that the total number of nonzero entries of $\mathbf{x}$ satisfies

$$n_a + n_b < d^{-2}/2. \qquad (5)$$

Then, if the values of *all* nonzero entries of $\mathbf{x}$ are drawn from a continuous probability distribution, $\mathbf{x}$ is the unique solution of (P0) with probability exceeding $(1 - N^{-s})$. Furthermore, if n_a and n_b, in addition to (3)–(5), satisfy

$$n_a + n_b \leq d^{-2}/[8(s+1)\log N] \qquad (6)$$

and the phases of *all* nonzero entries of $\mathbf{x}$ are i.i.d. and uniformly distributed on $[0, 2\pi)$, then, $\mathbf{x}$ is the unique solution of both (P0) and (P1) with probability exceeding $(1 - 3N^{-s})$.

Proof: The proof is based on the following lemma, which is the main technical result of this paper and whose proof can be found in Appendix A.

Lemma 1: Fix $s \geq 1$ and $\gamma \in [0, 1]$. Let $\mathbf{S}$ be a sub-dictionary of $\mathbf{D} = [\mathbf{A}\ \mathbf{B}] \in \mathcal{D}(d, a, b)$ that contains n_a *arbitrarily* chosen columns of $\mathbf{A}$ and n_b columns of $\mathbf{B}$ chosen uniformly at *random*. If n_a and n_b satisfy conditions (3) and (4), then, the minimum singular value $\sigma_{\min}(\mathbf{S})$ of the sub-dictionary $\mathbf{S}$ obeys

$$\mathbb{P}\left\{\sigma_{\min}(\mathbf{S}) \leq 1/\sqrt{2}\right\} \leq N^{-s}.$$

The proof of Theorem 2 is then obtained from Lemma 1 and the results in [12] as follows. The sparsity pattern of $\mathbf{x}$ assumed in the statement of Theorem 2 induces a sub-dictionary $\mathbf{S}$ of $\mathbf{D}$ containing n_a arbitrarily chosen columns of $\mathbf{A}$ and n_b randomly chosen columns of $\mathbf{B}$. As a consequence of Lemma 1, the smallest singular value of $\mathbf{S}$ exceeds $1/\sqrt{2}$ with probability at least $(1 - N^{-s})$. This property of the sub-dictionary $\mathbf{S}$, together with condition (5) and the requirement that the values of *all* nonzero entries of $\mathbf{x}$ are drawn from a continuous probability distribution, implies, as a consequence of [12, Thm. 13], that $\mathbf{x}$ is the unique solution of (P0) with probability at least $(1 - N^{-s})$. If, in addition, condition (6) is satisfied and the phases of *all* nonzero entries of $\mathbf{x}$ are i.i.d. and uniformly distributed on $[0, 2\pi)$, we can apply [12, Thm. 14] (with $\delta = N^{-s}$) to infer that $\mathbf{x}$ is the unique solution of both (P0) and (P1) with probability at least $(1 - N^{-s})(1 - 2N^{-s}) \geq (1 - 3N^{-s})$. ■

Interpretation of Theorem 2: We next present an interpretation of our result and reveal the fundamental property that allows to break the square-root bottleneck with reduced randomness. In particular, we determine conditions on the dictionary such that both $n_a = \Theta(M/(\log N))$ and $n_b = \Theta(M/(\log N))$. As a consequence, a robust sparsity threshold $S > n_a + n_b = \Theta(M/(\log N))$ is established. In the following, for clarity of exposition, we only consider the dependency of n_a and n_b on the dictionary parameters d, a, b, N_a, N_b, and the spectral norms of $\mathbf{A}$ and $\mathbf{B}$, and absorb all constants that are independent of these quantities in $c(\gamma, s)$, where γ and s are defined in Theorem 2. Note that $c(\gamma, s)$ can change its value at each appearance. Condition (3) together with $n_a \leq N_a$ yields the following constraint on n_a:

$$n_a \leq c(\gamma, s)\min\left\{d^{-2}/(\log N), a^{-1}, N_a\right\}.$$

This constraint is compatible with $n_a = \Theta(M/(\log N))$, if the following three requirements are fulfilled:

i) the coherence of $\mathbf{D}$ satisfies $d = \mathcal{O}(1/\sqrt{M})$

[5] Whenever for some function $g(M, N)$ we write $\Theta(g(M, N))$ or $\mathcal{O}(g(M, N))$, we mean that the ratio N/M remains fixed while $M \to \infty$.

ii) the coherence of $\mathbf{A}$ satisfies $a = \mathcal{O}((\log N)/M)$

iii) the cardinality of $\mathbf{A}$ satisfies $N_a \geq c\,M/(\log N)$

where c is a constant that can change at each appearance. Condition (4), which can be rewritten as

$$n_b \leq c(\gamma, s) \min\left\{ \frac{b^{-2}}{\log N}, \frac{N_b}{\|\mathbf{B}\|^2}, \frac{N_b}{\|\mathbf{A}\|^2 \|\mathbf{B}\|^2} \right\} \tag{7}$$

is more laborious to interpret. For the constraint (7) to be compatible with $n_b = \Theta(M/(\log N))$, we need requirement i) above to be fulfilled (recall that $b \leq d$), together with the following two requirements on the spectral norms of $\mathbf{B}$ and $\mathbf{A}$, namely

iv) $\|\mathbf{B}\|^2 \leq c\,N_b(\log N)/M$

v) $\|\mathbf{A}\|^2 \leq c\,N_b(\log N)/(\|\mathbf{B}\|^2 M)$.

We finally note that when the requirements i) – v) are met, conditions (5) and (6), which can then be rewritten as $n_a + n_b \leq c\,M$ and $n_a + n_b \leq c\,M/(\log N)$, respectively, are compatible with both $n_a = \Theta(M/(\log N))$ and $n_b = \Theta(M/(\log N))$.

Hence, a robust sparsity threshold $S > n_a + n_b = \Theta(M/(\log N))$ can be established under the same probabilistic model on $\mathbf{x}$ as in the two-ONB case; namely, the positions of the nonzero entries of $\mathbf{x}$ corresponding to $\mathbf{B}$ have to be chosen randomly, while the positions of the nonzero entries of $\mathbf{x}$ corresponding to $\mathbf{A}$ can be chosen arbitrarily.

The requirements iv) and v) are difficult to interpret because they depend on the spectral norms of the sub-dictionaries $\mathbf{A}$ and $\mathbf{B}$. To get more insight into these two requirements, we consider the special case of $\mathbf{A}$ and $\mathbf{B}$ being tight frames for $\mathbb{C}^M$ [14] (with the frame elements ℓ_2-normalized to one). Then, $\|\mathbf{A}\|^2 = N_a/M$ and $\|\mathbf{B}\|^2 = N_b/M$, so that iv) is trivially satisfied and v) reduces to $N_a \leq c\,M \log N$. However, because of the Welch lower bound [10] condition ii) puts a more stringent restriction on the cardinality of N_a for large M. Hence, a robust sparsity threshold of $\Theta(M/(\log N))$ is obtained, under the same probabilistic model on the vector $\mathbf{x}$ as in the two-ONB case, if the coherence of sub-dictionary $\mathbf{A}$ satisfies $a = \mathcal{O}((\log N)/M)$.

A simple dictionary that satisfies i) - v): For $M = p^k$, with p prime and $k \in \mathbb{N}^+$, a dictionary $\mathbf{D}$ with coherence equal to $1/\sqrt{M}$ can be obtained by concatenating $M+1$ ONBs for $\mathbb{C}^M$ [6]. Since $\mathbf{D}$ constitutes a tight frame for $\mathbb{C}^M$, by [12] a robust sparsity threshold of $\Theta(M/(\log N))$ is obtained by randomizing over all positions of the nonzero entries of $\mathbf{x}$. Note, however, that we can write $\mathbf{D} = [\mathbf{A}\ \mathbf{B}]$, where $\mathbf{A}$ is an ONB ($a = 0$) and $\mathbf{B}$ is the concatenation of the remaining M ONBs and hence a tight frame for $\mathbb{C}^M$. As $N_a = M$ the requirements iii) and v) are satisfied. Therefore, by the results of the previous paragraph, a robust sparsity threshold of $\Theta(M/(\log N))$ is obtained by randomizing only over the positions of the nonzero entries of $\mathbf{x}$ corresponding to $\mathbf{B}$.

Appendix A
Proof of Lemma 1

Since the minimum singular value $\sigma_{\min}(\mathbf{S})$ of the sub-dictionary $\mathbf{S}$ can be lower-bounded as $\sigma^2_{\min}(\mathbf{S}) \geq 1 - \left\|\mathbf{S}^H\mathbf{S} - \mathbf{I}_{n_a+n_b}\right\|$, we have

$$\begin{aligned}\mathbb{P}\Big\{\sigma_{\min}(\mathbf{S}) \leq 1/\sqrt{2}\Big\} &= \mathbb{P}\big\{\sigma^2_{\min}(\mathbf{S}) \leq 1/2\big\} \\ &\leq \mathbb{P}\big\{1 - \left\|\mathbf{S}^H\mathbf{S} - \mathbf{I}_{n_a+n_b}\right\| \leq 1/2\big\} \\ &= \mathbb{P}\big\{\left\|\mathbf{S}^H\mathbf{S} - \mathbf{I}_{n_a+n_b}\right\| \geq 1/2\big\}\,. \end{aligned} \tag{8}$$

Next, we quantify the tail behavior of the random variable $H = \left\|\mathbf{S}^H\mathbf{S} - \mathbf{I}_{n_a+n_b}\right\|$, which will then lead to an upper bound on the probability of $\sigma_{\min}(\mathbf{S})$ falling below $1/\sqrt{2}$. To this end the following lemma will be useful.

Lemma 2 ([12, Prop. 10]): If the moments of the non-negative random variable R can be upper-bounded as $[\mathbb{E}(R^q)]^{1/q} \leq \alpha\sqrt{q} + \beta$ for all $q \geq Q \in \mathbb{Z}_0^+$, where $\alpha, \beta \in \mathbb{R}_0^+$, then,

$$\mathbb{P}\{R \geq e^{1/4}(\alpha u + \beta)\} \leq e^{-u^2/4}$$

for all $u \geq \sqrt{Q}$.

To be able to apply Lemma 2 to $H = \left\|\mathbf{S}^H\mathbf{S} - \mathbf{I}_{n_a+n_b}\right\|$, we first need an upper bound on $[\mathbb{E}(H^q)]^{1/q}$ that is of the form $\alpha\sqrt{q} + \beta$. We start by writing the sub-dictionary $\mathbf{S}$ as $\mathbf{S} = [\mathbf{S}_a\ \mathbf{S}_b]$, where $\mathbf{S}_a$ and $\mathbf{S}_b$ denote the matrices containing the columns chosen arbitrarily from $\mathbf{A}$ and randomly from $\mathbf{B}$, respectively. We then obtain

$$\mathbf{S}^H\mathbf{S} - \mathbf{I}_{n_a+n_b} = \begin{bmatrix} \mathbf{S}_a^H\mathbf{S}_a - \mathbf{I}_{n_a} & \mathbf{S}_a^H\mathbf{S}_b \\ \mathbf{S}_b^H\mathbf{S}_a & \mathbf{S}_b^H\mathbf{S}_b - \mathbf{I}_{n_b} \end{bmatrix}.$$

Applying the triangle inequality for operator norms, we can now upper-bound H according to

$$\begin{aligned} H &= \left\|\begin{bmatrix} \mathbf{S}_a^H\mathbf{S}_a - \mathbf{I}_{n_a} & \mathbf{S}_a^H\mathbf{S}_b \\ \mathbf{S}_b^H\mathbf{S}_a & \mathbf{S}_b^H\mathbf{S}_b - \mathbf{I}_{n_b} \end{bmatrix}\right\| \\ &\leq \left\|\begin{bmatrix} \mathbf{S}_a^H\mathbf{S}_a - \mathbf{I}_{n_a} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_b^H\mathbf{S}_b - \mathbf{I}_{n_b} \end{bmatrix}\right\| + \left\|\begin{bmatrix} \mathbf{0} & \mathbf{S}_a^H\mathbf{S}_b \\ \mathbf{S}_b^H\mathbf{S}_a & \mathbf{0} \end{bmatrix}\right\| \\ &\leq \max\big\{\left\|\mathbf{S}_a^H\mathbf{S}_a - \mathbf{I}_{n_a}\right\|, \left\|\mathbf{S}_b^H\mathbf{S}_b - \mathbf{I}_{n_b}\right\|\big\} + \left\|\mathbf{S}_a^H\mathbf{S}_b\right\| \\ &\leq \left\|\mathbf{S}_a^H\mathbf{S}_a - \mathbf{I}_{n_a}\right\| + \left\|\mathbf{S}_b^H\mathbf{S}_b - \mathbf{I}_{n_b}\right\| + \left\|\mathbf{S}_a^H\mathbf{S}_b\right\| \end{aligned} \tag{9}$$

where the second inequality follows because the spectral norm of both a block-diagonal matrix and an anti-block-diagonal matrix is given by the largest among the spectral norms of the individual nonzero blocks. Next, we define $H_a = \left\|\mathbf{S}_a^H\mathbf{S}_a - \mathbf{I}_{n_a}\right\|$, $H_b = \left\|\mathbf{S}_b^H\mathbf{S}_b - \mathbf{I}_{n_b}\right\|$, and $Z = \left\|\mathbf{S}_a^H\mathbf{S}_b\right\|$. It then follows from (9) that for all $q \geq 1$

$$\begin{aligned} [\mathbb{E}(H^q)]^{1/q} &\leq [\mathbb{E}((H_a + H_b + Z)^q)]^{1/q} \\ &\leq [\mathbb{E}(H_a^q)]^{1/q} + [\mathbb{E}(H_b^q)]^{1/q} + [\mathbb{E}(Z^q)]^{1/q} \\ &\leq H_a + [\mathbb{E}(H_b^q)]^{1/q} + [\mathbb{E}(Z^q)]^{1/q} \end{aligned} \tag{10}$$

where the second inequality is a consequence of the triangle inequality for the norm $[\mathbb{E}(|\cdot|^q)]^{1/q}$ (recall that $q \geq 1$), and in the last step we used the fact that H_a is a deterministic quantity. All expectations in (10) are with respect to the random choice of columns from sub-dictionary $\mathbf{B}$.

We next upper-bound the three terms on the right-hand side (RHS) of (10) individually. Applying Geršgorin's disc theorem [15, Th. 6.1.1] to the first term, we obtain

$$H_a = \left\|\mathbf{S}_a^H\mathbf{S}_a - \mathbf{I}_{n_a}\right\| \leq (n_a - 1)a. \tag{11}$$

For the second term on the RHS of (10) we can use [12, Eq. 6.1] to get

$$\begin{aligned}[\mathbb{E}(H_b^q)]^{1/q} &= \left[\mathbb{E}\left(\left\|\mathbf{S}_b^H\mathbf{S}_b - \mathbf{I}_{n_b}\right\|^q\right)\right]^{1/q} \\ &\leq \sqrt{144 b^2 n_b r_1} + 2n_b\|\mathbf{B}\|^2/N_b \end{aligned} \tag{12}$$

where $r_1 = \max\{1, \log(n_b/2+1), q/4\}$. Assuming that $q \geq \max\{4\log(n_b/2+1), 4\}$ and hence $r_1 = q/4$, we can simplify (12) to

$$[\mathbb{E}(H_b^q)]^{1/q} \leq 6\sqrt{b^2 n_b}\sqrt{q} + \frac{2n_b}{N_b}\|\mathbf{B}\|^2. \tag{13}$$

To bound the third term on the RHS of (10), we use the upper bound on the spectral norm of a random compression [12, Thm. 8] combined with $\operatorname{rank}(\mathbf{S}_a^H\mathbf{S}_b) \leq n_b$. This yields

$$\begin{aligned}[\mathbb{E}(Z^q)]^{1/q} &= \left[\mathbb{E}\left(\left\|\mathbf{S}_a^H\mathbf{S}_b\right\|^q\right)\right]^{1/q} \\ &\leq 3\sqrt{r_2}\left\|\mathbf{S}_a^H\mathbf{B}\right\|_{1,2} + \sqrt{\frac{n_b}{N_b}}\left\|\mathbf{S}_a^H\mathbf{B}\right\| \end{aligned} \tag{14}$$

where $r_2 = \max\{2, 2\log n_b, q/2\}$. Assuming that $q \geq \max\{4\log n_b, 4\}$, we can further bound the RHS of (14) to get

$$\begin{aligned}[\mathbb{E}(Z^q)]^{1/q} &\leq \frac{3}{\sqrt{2}}\sqrt{q}\left\|\mathbf{S}_a^H\mathbf{B}\right\|_{1,2} + \sqrt{\frac{n_b}{N_b}}\left\|\mathbf{S}_a^H\mathbf{B}\right\| \\ &\leq \frac{3}{\sqrt{2}}\sqrt{d^2 n_a}\sqrt{q} + \sqrt{\frac{n_b}{N_b}}\left\|\mathbf{S}_a^H\mathbf{B}\right\| \qquad (15) \\ &\leq \frac{3}{\sqrt{2}}\sqrt{d^2 n_a}\sqrt{q} + \sqrt{\frac{n_b}{N_b}}\|\mathbf{A}\|\|\mathbf{B}\| \qquad (16) \end{aligned}$$

where (15) follows from the fact that the magnitude of each entry of $\mathbf{S}_a^H\mathbf{B}$ is upper-bounded by d and, thus, $\left\|\mathbf{S}_a^H\mathbf{B}\right\|_{1,2} \leq \sqrt{d^2 n_a}$. To arrive at (16) we used $\left\|\mathbf{S}_a^H\mathbf{B}\right\| \leq \left\|\mathbf{S}_a^H\right\|\|\mathbf{B}\| \leq \|\mathbf{A}\|\|\mathbf{B}\|$, which follows from the sub-multiplicativity of the spectral norm and the fact that the spectral norm of the submatrix $\mathbf{S}_a$ of $\mathbf{A}$ cannot exceed that of $\mathbf{A}$. We can now combine the upper bounds (11), (13), and (16) to obtain

$$\begin{aligned}[\mathbb{E}(H^q)]^{1/q} &\leq (n_a-1)a + 6\sqrt{b^2 n_b}\sqrt{q} + \frac{2n_b}{N_b}\|\mathbf{B}\|^2 + \\ &\quad + \frac{3}{\sqrt{2}}\sqrt{d^2 n_a}\sqrt{q} + \sqrt{\frac{n_b}{N_b}}\|\mathbf{A}\|\|\mathbf{B}\| \\ &= \underbrace{\left(6\sqrt{b^2 n_b} + 3\sqrt{d^2 n_a/2}\right)}_{\alpha}\sqrt{q} + \\ &\quad \underbrace{+ (n_a-1)a + \frac{2n_b}{N_b}\|\mathbf{B}\|^2 + \sqrt{\frac{n_b}{N_b}}\|\mathbf{A}\|\|\mathbf{B}\|}_{\beta} \\ &= \alpha\sqrt{q} + \beta \end{aligned}$$

for all $q \geq Q_1 = \max\{4\log(n_b/2+1), 4\log n_b, 4\}$. Hence, Lemma 2 yields

$$\mathbb{P}\{H \geq e^{1/4}(\alpha u + \beta)\} \leq e^{-u^2/4}$$

for all $u \geq \sqrt{Q_1}$. In particular, under the assumption $N \geq e \approx 2.7$, it follows that the choice $u = \sqrt{4s\log N}$ satisfies $u \geq \sqrt{Q_1}$ for any $s \geq 1$. Straightforward calculations reveal that conditions (3) and (4) ensure that $e^{1/4}(\alpha u + \beta) \leq 1/2$, which together with (8) then leads to

$$\begin{aligned}\mathbb{P}\left\{\sigma_{\text{min}}(\mathbf{S}) \leq 1/\sqrt{2}\right\} &\leq \mathbb{P}\{H \geq 1/2\} \\ &\leq \mathbb{P}\{H \geq e^{1/4}(\alpha u + \beta)\} \\ &\leq e^{-u^2/4} = N^{-s}. \end{aligned}$$

REFERENCES

[1] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

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

[3] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 33–61, 1998.

[4] D. L. Donoho and X. Huo, "Uncertainty principles and ideal atomic decomposition," *IEEE Trans. Inf. Theory*, vol. 47, no. 7, pp. 2845–2862, Nov. 2001.

[5] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization," *Proc. Natl. Acad. Sci. USA*, vol. 100, no. 5, pp. 2197–2202, Mar. 2003.

[6] R. Gribonval and M. Nielsen, "Sparse representations in unions of bases," *IEEE Trans. Inf. Theory*, vol. 49, no. 12, pp. 3320–3325, Dec. 2003.

[7] M. Elad and A. M. Bruckstein, "A generalized uncertainty principle and sparse representation in pairs of bases," *IEEE Trans. Inf. Theory*, vol. 48, no. 9, pp. 2558–2567, Sep. 2002.

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

[9] P. Kuppinger, G. Durisi, and H. Bölcskei, "Improved sparsity thresholds through dictionary splitting," *Proc. IEEE Inf. Theory Workshop (ITW), Taormina, Italy*, pp. 338–342, Oct. 2009.

[10] L. Welch, "Lower bounds on the maximum cross correlation of signals," *IEEE Trans. Inf. Theory*, vol. 20, no. 3, pp. 397–399, 1974.

[11] E. J. Candès and J. Romberg, "Quantitative robust uncertainty principles and optimally sparse decompositions," *Foundations of Comput. Math.*, vol. 6, no. 2, pp. 227–254, Apr. 2006.

[12] J. A. Tropp, "On the conditioning of random subdictionaries," *Appl. Comp. Harmonic Anal.*, vol. 25, pp. 1–24, 2008.

[13] R. Calderbank, S. Howard, and S. Jafarpour, "Construction of a large class of deterministic sensing matrices that satisfy a statistical isometry property," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 2, pp. 358–374, Apr. 2010.

[14] O. Christensen, *An Introduction to Frames and Riesz Bases*. Boston, MA. U.S.A.: Birkhäuser, 2003.

[15] R. A. Horn and C. R. Johnson, *Matrix Analysis*. New York, NY: Cambridge Press, 1985.