CONCENTRATION OF MEASURE FOR BLOCK DIAGONAL MEASUREMENT MATRICES

Michael B. Wakin,[*] *Jae Young Park,*[♮] *Han Lun Yap,*[†] *and Christopher J. Rozell*[†]

[*] Division of Engineering, Colorado School of Mines
[♮] Department of Electrical Engineering and Computer Science, University of Michigan
[†] School of Electrical and Computer Engineering, Georgia Institute of Technology

ABSTRACT

Concentration of measure inequalities are at the heart of much theoretical analysis of randomized compressive operators. Though commonly studied for dense matrices, in this paper we derive a concentration of measure bound for block diagonal matrices where the nonzero entries along the main diagonal blocks are i.i.d. subgaussian random variables. Our main result states that the concentration exponent, in the best case, scales as that for a fully dense matrix. We also identify the role that the energy distribution of the signal plays in distinguishing the best case from the worst. We illustrate these phenomena with a series of experiments.

***Index Terms*—** Compressive Sensing, concentration of measure, Johnson-Lindenstrauss lemma, block diagonal matrices

1. INTRODUCTION

With many signal processing applications demanding ever greater volumes of higher-resolution data, a significant amount of research in recent years has involved compressive linear measurement operators. Letting $x \in \mathbb{R}^N$ denote a signal, we represent a compressive linear operator by an $M \times N$ matrix Φ ($M < N$). Given a sufficient number of measurements $y = \Phi x$, the goal in the field of compressive sensing (CS) [1, 2] is to recover x from y.

Randomized designs for the compressive operator Φ are particularly appealing because of their universal applicability to different signals. A random $M \times N$ matrix populated with independent and identically distributed (i.i.d.) Gaussian entries, for example, can be used in CS to acquire any signal that has a sparse decomposition in some basis (with high probability, and supposing again that M is sufficiently large [1, 2]). This universality covers not only different signals but different applications: the same randomized constructions used in CS can also be for nearest neighbor search, detection, classification, estimation, and so on [3, 4].

One of the concepts at the heart of theoretical analysis of randomized compressive operators is a phenomenon known as concentration of measure [5]. Put simply, for any fixed signal $x \in \mathbb{R}^N$ a suitable randomized Φ will approximately preserve the norm of x with high probability. More formally, a typical result is the following:[1]

Lemma 1.1. [6] *Let Φ be an $M \times N$ matrix with i.i.d. subgaussian entries having variance $\frac{1}{M}$. Then there exists a constant $C > 0$ such that, for any $\epsilon \in (0, 1)$,*

$$P\left(\left|\|\Phi x\|_2^2 - \|x\|_2^2\right| > \epsilon \|x\|_2^2\right) \leq 2\exp(-CM\epsilon^2). \quad (1)$$

Among the consequences of this powerful result is the Johnson-Lindenstrauss (JL) lemma [3], which states that for a finite cloud of points $Q \subset \mathbb{R}^N$,

$$(1-\epsilon)\|u-v\|_2 \leq \|\Phi(u-v)\|_2 \leq (1+\epsilon)\|u-v\|_2 \quad (2)$$

holds for all $u, v \in Q$ with high probability supposing that $M = O(\log(|Q|)\epsilon^{-2})$. Another consequence is the Restricted Isometry Property (RIP) [7] in CS, which states that (2) holds for all pairs u, v of K-sparse signals in $\mathbb{R}^N$ with high probability supposing that $M = O(K\log(N/K)\epsilon^{-2})$ [6, 8].

Unfortunately, in many signal acquisition settings, one may not have the luxury of using a dense measurement operator Φ as described above. With a long video sequence, for example, it may be desirable for each measurement to come only from a single frame, rather than from the entire video [9]. Similarly, in a network of sensors observing signals relating to a common phenomenon, it may be desirable for each sensor to record measurements of its own incident signal, rather than have each measurement depend on the whole ensemble [10]. In settings such as these, we may envision the signal $x \in \mathbb{R}^{NJ}$ as representing the concatenation $x = [x_1^T\ x_2^T\ \cdots x_J^T]^T$ of J component signals $\{x_j\}_{j=1}^J \subset \mathbb{R}^N$ and the measurement vector $y \in \mathbb{R}^{MJ}$ as representing the concatenation $y = [y_1^T\ y_2^T\ \cdots y_J^T]^T$ of the vectors $y_j = \Phi_j x_j$, where Φ_j is a matrix of size $M \times N$ that measures only signal j. It follows that $y = \Phi x$, where the resulting $MJ \times NJ$ measurement matrix Φ is zero everywhere except for blocks along the main diagonal:

$$\Phi = \begin{pmatrix} \Phi_1 & & & \\ & \Phi_2 & & \\ & & \ddots & \\ & & & \Phi_J \end{pmatrix}. \quad (3)$$

This work was partially supported by NSF grants CCF-0830456 and CCF-0830320, DARPA grant HR0011-08-1-0078, and AFOSR grant FA9550-09-1-0465.

[1] Note that subgaussian random variables are defined in Definition 2.1. Also, Lemma 1.1 follows from Theorem 3.1 by setting $J = 1$.

In this paper we derive a concentration of measure bound for such matrices where the nonzero entries of the block diagonal matrix are i.i.d. subgaussian random variables. Our main result, detailed in Theorem 3.1, essentially states that the probability of concentration behaves as

$$P(\big|\|y\|_2^2 - \|x\|_2^2\big| > \epsilon\|x\|_2^2) \leq 2\exp(-CM\Lambda\epsilon^2),$$

where $\Lambda = \Lambda(x)$ is a term that depends on the energy distribution of x. At one extreme, when the energy of x is uniformly spread across its component signals $\{x_j\}$, we have $\Lambda(x) = J$ and the concentration of measure exponent scales exactly as in (1) for a fully dense random matrix; that is, it scales with the total number of measurements, which in this case equals MJ. At the other extreme, however, when the energy of x is concentrated in a single component signal, we have $\Lambda(x) = 1$ and the measurement operator effectiveness is diminished.

2. BACKGROUND

In this section we provide some background on subgaussian random variables [11] and their properties.

Definition 2.1. *A random variable w is* subgaussian *if $\exists a \geq 0$ such that*

$$\mathbf{E}e^{tw} \leq \exp\left(\frac{1}{2}a^2t^2\right) \text{ for all } t \in \mathbb{R}.$$

The quantity

$$\tau(w) := \inf\{a \geq 0 : \mathbf{E}e^{tw} \leq \exp(\frac{1}{2}a^2t^2) \text{ for all } t \in \mathbb{R}\}$$

is known as the Gaussian standard *of w.*

With this definition and Jensen's inequality we infer that subgaussian random variables must always be centered, i.e., $\mathbf{E}w = 0$, and with some simple functional analysis we can also see that the variance $\mathbf{E}w^2 \leq \tau^2(w)$. In [11], a sub-class of subgaussian random variables is introduced as follows.

Definition 2.2. *A subgaussian random variable w is* strictly subgaussian *if $\mathbf{E}w^2 = \tau^2(w)$.*

Examples of strictly subgaussian random variables include Gaussian random variables, ± 1 Bernoulli random variables ($p = \frac{1}{2}$) and uniform random variables on $[-1, 1]$.

As with Gaussian random variables, linear combinations of i.i.d. subgaussian random variables are also subgaussian. We provide a more formal statement in the following lemma.

Lemma 2.1. [11, Theorem 1 and Lemma 3] *Let $\beta \in \mathbb{R}^Z$ be a fixed vector, and suppose $w(1), w(2), \ldots, w(Z)$ are a collection of i.i.d. subgaussian random variables with Gaussian standards all equal to $\tau(w)$. Then the quantity $v := \sum_{i=1}^{Z} \beta(i)w(i)$ is a subgaussian random variable with Gaussian standard $\tau(v) \leq \tau(w)\|\beta\|_2$.*

We also need the following lemma regarding the tail distribution of the square of subgaussian random variables.

Lemma 2.2. [11] *Suppose that w is a subgaussian random variable with Gaussian standard $\tau(w)$. Then*

$$P(|w|^2 > t) \leq 2\exp\left(-\frac{t}{2\tau^2(w)}\right) \text{ for all } t \geq 0.$$

This property of subgaussian random variables allows us to use the following important theorem.

Theorem 2.1. [12] *Let $X_1, \ldots, X_L$ be independent Banach space valued random variables with $P\{\|X_i\| > t\} \leq a\exp\{-\alpha_i t\}$ for all t and i. Let $d \geq \max_i \alpha_i^{-1}$ and $b \geq a\sum_{i=1}^{L} \alpha_i^{-2}$. Then setting $S = \sum_{i=1}^{L} X_i$ we have*

$$P\{|\|S\| - \mathbf{E}\|S\|| > t\} \leq \begin{cases} 2\exp\{-t^2/32b\}, & 0 \leq t \leq \frac{4b}{d} \\ 2\exp\{-t/8d\}, & t \geq \frac{4b}{d}. \end{cases}$$

3. MAIN RESULT

Let Φ be as in (3), where each Φ_j has size $M \times N$. For $i \in \{1, \ldots, M\}$, let $\phi_{j,i}^T$ denote row i of Φ_j, and let $y_j(i)$ denote the i^{th} component of measurement vector y_j. It follows that $y_j(i) = \langle \phi_{j,i}, x_j \rangle = \sum_{n=1}^{N} \phi_{j,i}(n)x_j(n)$. We will also have use for the following vector describing the energy distribution across x. For any $x \in \mathbb{R}^{NJ}$, we define

$$\lambda = \lambda(x) := [\|x_1\|_2^2 \ \|x_2\|_2^2 \ \cdots \ \|x_J\|_2^2]^T \in \mathbb{R}^J.$$

Equipped with this notation, we now present our main result.

Theorem 3.1. *Suppose $x \in \mathbb{R}^{NJ}$. Let $\{\Phi_j\}_{j=1}^{J}$ be random $M \times N$ matrices populated with i.i.d. subgaussian entries having variance $\sigma^2 = \frac{1}{M}$ and Gaussian standard $\tau^2(\phi) = c\sigma^2$, where $c \geq 1$, and let Φ be an $MJ \times NJ$ block diagonal matrix as defined in (3). Then*

$$P(\big|\|\Phi x\|_2^2 - \|x\|_2^2\big| > \epsilon\|x\|_2^2) \leq \begin{cases} 2\exp\{-\frac{M\epsilon^2\|\lambda\|_1^2}{256c^2\|\lambda\|_2^2}\}, & 0 \leq \epsilon \leq \frac{16c\|\lambda\|_2^2}{\|\lambda\|_\infty\|\lambda\|_1} \\ 2\exp\{-\frac{M\epsilon\|\lambda\|_1}{16c\|\lambda\|_\infty}\}, & \epsilon \geq \frac{16c\|\lambda\|_2^2}{\|\lambda\|_\infty\|\lambda\|_1}. \end{cases} \quad (4)$$

Proof. Let $y = \Phi x$. Each entry of $\phi_{1,1}, \ldots, \phi_{J,m}$ are i.i.d. subgaussian random variables having Gaussian standard $\tau(\phi)$. The expectation of each measurement squared can be written as $\mathbf{E}y_j^2(i) = \mathbf{E}\left(\sum_{n=1}^{N} \phi_{j,i}(n)x_j(n)\right)^2$. Since the $\phi_{j,i}(n)$ are zero mean and independent, all the cross product terms are equal to zero, which then gives us

$$\mathbf{E}y_j^2(i) = \mathbf{E}\sum_{n=1}^{N} \phi_{j,i}^2(n)x_j^2(n) = \sigma^2\|x_j\|_2^2 = \frac{1}{M}\|x_j\|_2^2.$$

Furthermore,

$$\mathbf{E}\|y\|_2^2 = \sum_{j=1}^{J}\sum_{i=1}^{M}\mathbf{E}y_j^2(i) = \sum_{j=1}^{J}\|x_j\|_2^2 = \|x\|_2^2.$$

We would like to find the probability that $\left|\|y\|_2^2 - \|x\|_2^2\right| > \epsilon\|x\|_2^2$. Since $\mathbf{E}\|y\|_2^2 = \|x\|_2^2$, this is equivalent to the condition that $\left|\|y\|_2^2 - \mathbf{E}\|y\|_2^2\right| > \epsilon\mathbf{E}\|y\|_2^2$. By Lemma 2.1, each $y_j(i)$ is a subgaussian random variable with Gaussian standard $\tau(y_j(i)) \leq \tau(\phi)\|x_j\|_2$. By Lemma 2.2 we have $\forall t \geq 0$,

$$\begin{aligned} P(y_j^2(i) > t) &\leq 2\exp\left(-\frac{t}{2\tau^2(y_j(i))}\right) \\ &\leq 2\exp\left(-\frac{t}{2\tau^2(\phi)\|x_j\|_2^2}\right). \end{aligned}$$

We apply Theorem 2.1 for the random variables $y_j^2(i)$, $\forall i, j$, with $a = 2$ and $\alpha_j^{-1}(i) = 2\tau^2(\phi)\|x_j\|_2^2 = \frac{2c}{M}\|x_j\|_2^2$, to compute the concentration result for $\|y\|_2^2$. Note that $\alpha_j^{-1}(i)$ is constant for a fixed j. Hence, for $d \geq \max_{i,j}\alpha_j^{-1}(i) = \frac{2c}{M}\max_j\|x_j\|_2^2$ and $b \geq a\sum_{j,i}\alpha_j^{-2}(i) = \frac{8c^2}{M}\sum_j\|x_j\|_2^4$,

$$P(\left|\|y\|_2^2 - \|x\|_2^2\right| > \epsilon\|x\|_2^2) \leq \begin{cases} 2\exp\{-\frac{\epsilon^2\|x\|_2^4}{32b}\}, & 0 \leq \epsilon \leq \frac{4b}{d\|x\|_2^2} \\ 2\exp\{-\frac{\epsilon\|x\|_2^2}{8d}\}, & \epsilon \geq \frac{4b}{d\|x\|_2^2}. \end{cases} \quad (5)$$

Note that $\|x\|_2^2 = \|\lambda\|_1$ and $\|x\|_2^4 = \|\lambda\|_1^2$. Substituting $d = \frac{2c}{M}\max_j\|x_j\|_2^2 = \frac{2c}{M}\|\lambda\|_\infty$ and $b = \frac{8c^2}{M}\sum_j\|x_j\|_2^4 = \frac{8c^2}{M}\|\lambda\|_2^2$ into (5) completes the proof. □

As we will be frequently concerned with applications where ϵ is small, let us consider the first of the cases given in (4). We define $\Lambda(x) = \frac{\|\lambda\|_1^2}{\|\lambda\|_2^2}$ and note that for any $x \in \mathbb{R}^N$, $1 \leq \Lambda(x) \leq J$, where equality on the left is obtained when $\|x_j\|_2^2 = 0$ for all but one j, and equality on the right is obtained when all $\|x_j\|_2^2$ are equal. (This follows from the standard relation that $\|z\|_2 \leq \|z\|_1 \leq \sqrt{J}\|z\|_2$ for all $z \in \mathbb{R}^J$.) The first case ($\Lambda = 1$) is unfavorable, and implies that compared to a full dense Φ of size $MJ \times NJ$ (for which the concentration exponent would scale with MJ), we diminish the effectiveness of the measurements by a factor of J; this is to be expected since only M measurements will be nonzero. The second case ($\Lambda = J$) is favorable as—remarkably—the concentration exponent scales at the same rate as for a full dense Φ of size $MJ \times NJ$. Incidentally, for the second case in (4), it is worth noting that $1 \leq \frac{\|\lambda\|_1}{\|\lambda\|_\infty} \leq J$, with the extreme values attained in the same two cases mentioned above. So, in both cases the critical scaling of the exponent is between $M \cdot f(\epsilon)$ and $MJ \cdot f(\epsilon)$ where $f(\epsilon)$ is some function of ϵ.

Two final comments about Theorem 3.1 are in order. First, the bounds in (4) are most favorable for strictly subgaussian random variables, for which $c = 1$. Second, regarding the demarcation between the two cases in (4), one can show that $\frac{2(\sqrt{J}-1)}{J-1} \leq \frac{\|\lambda\|_2^2}{\|\lambda\|_\infty\|\lambda\|_1} \leq 1$. For $J \geq 2$, the left hand term obeys $\frac{2(\sqrt{J}-1)}{J-1} \geq \frac{1}{\sqrt{J}}$ implying that the upper case is guaranteed to include at least $0 \leq \epsilon \leq \frac{16c}{\sqrt{J}}$.

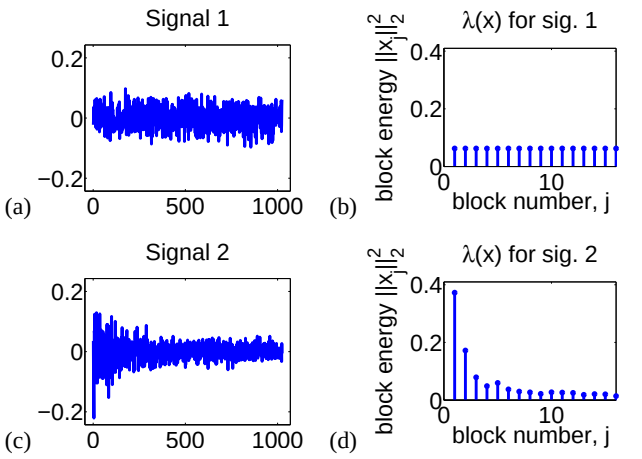

Fig. 1. *(a) Signal with uniform energy across $J = 16$ blocks. (b) $\lambda(x)$ for signal 1. (c) Signal with nonuniform energy across $J = 16$ blocks. (d) $\lambda(x)$ for signal 2.*

4. EXPERIMENTS

Let us illustrate the concentration phenomena suggested in Theorem 3.1 by considering two specific signals. The plots we show are typical of our experiments with other signals.

To begin, we randomly construct a signal of length 1024 having uniform energy over each of $J = 16$ blocks. The signal x is plotted in Figure 1(a) and the uniform energy distribution is shown in the plot of $\lambda(x)$ in Figure 1(b). Fixing this signal, we first generate a series of random Φ matrices having size 64×1024. Each matrix is dense: the entries are chosen as i.i.d. zero-mean Gaussian random variables with variance $1/64$. Over the course of 10000 randomly generated Φ, we plot in Figure 2(a) a histogram of $\|\Phi x\|_2/\|x\|_2$. (Ideally, we desire this quantity to be sharply concentrated around 1.)

Using the same fixed signal x, we now consider block diagonal matrices Φ of the form (3), where each of the $J = 16$ blocks on the main diagonal has size $M \times N = 4 \times 64$ and is populated with i.i.d. zero-mean Gaussian random variables with variance $1/4$. The resulting Φ matrices have size $MJ \times NJ = 64 \times 1024$, and because our test signal has $\Lambda(x) = J$, Theorem 3.1 implies that concentration exponent scales at the same rate as for a full dense Φ of size 64×1024. We plot in Figure 2(b) a histogram of $\|\Phi x\|_2/\|x\|_2$. Despite the dramatically different structure of the sensing matrices, the concentration of measure behavior for this signal is virtually identical to the case of fully dense matrices. For each type of matrix, we plot in Figure 3 as a function of ϵ the percent of trials for which $(1-\epsilon) \leq \|\Phi x\|_2/\|x\|_2 \leq (1+\epsilon)$. The curves

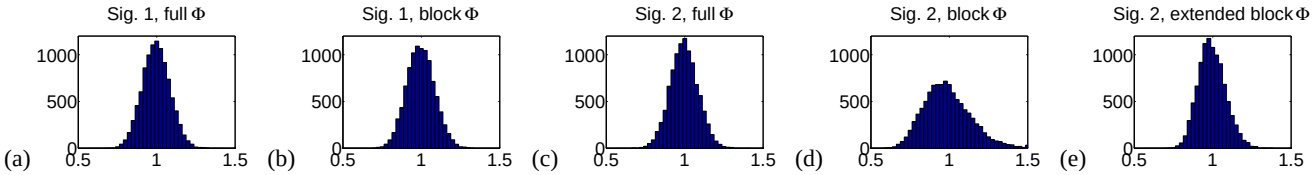

Fig. 2. *Histogram of $\|\Phi x\|_2/\|x\|_2$ for fixed x across 10000 randomly generated matrices Φ. (a) Uniform energy signal, fully dense 64×1024 Φ. (b) Uniform energy signal, block diagonal 64×1024 Φ. (c) Nonuniform energy signal, fully dense 64×1024 Φ. (d) Nonuniform energy signal, block diagonal 64×1024 Φ. (e) Nonuniform energy signal, block diagonal 192×1024 Φ.*

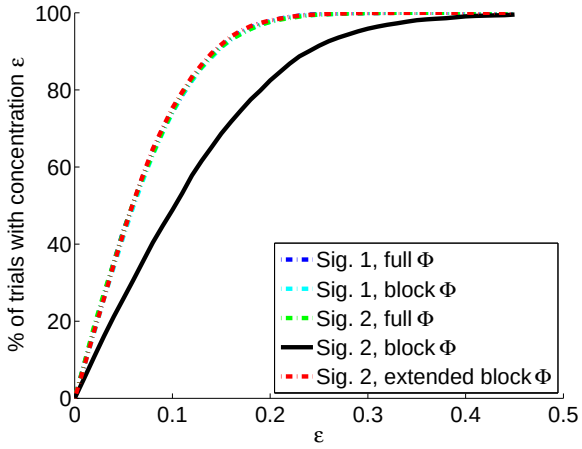

Fig. 3. *The percent of trials for which $(1-\epsilon)\leq\|\Phi x\|_2/\|x\|_2\leq(1+\epsilon)$. Note that all curves overlap except for the nonuniform energy signal with block diagonal Φ.*

for the two types of matrices are indistinguishable.

Now, we consider instead a signal which has nonuniform energy distribution. The signal x is plotted in Figure 1(c) and $\lambda(x)$ is plotted in Figure 1(d). Fixing this signal, we plot in Figure 2(c) a histogram of $\|\Phi x\|_2/\|x\|_2$ where the sensing matrices Φ are of size 64×1024 and are fully dense. From this histogram and the plot in Figure 3 we see that the concentration of measure behavior for this signal is virtually identical to the case of the signal with uniform energy.

Next, using this same signal x with nonuniform energy, we plot in Figure 2(d) a histogram of $\|\Phi x\|_2/\|x\|_2$ where the sensing matrices Φ are of size $MJ\times NJ$ and have block diagonal structure with $J=16$, $M=4$, and $N=64$. In this case, because our signal has $\Lambda(x)=5.335$, Theorem 3.1 suggests that the effectiveness of our matrix will be diminished. Based on the histogram and on Figure 3, we observe that the concentration of $\|\Phi x\|_2/\|x\|_2$ is decidedly less sharp.

Finally, using this same signal x, we consider constructing block diagonal matrices that contain extra rows to compensate for the fact that $\Lambda(x)<J$. In particular, we consider matrices Φ of size $M'J\times NJ$ with $J=16$, $M'=12$, and $N=64$. Applying Theorem 3.1, we see that $M'\Lambda(x)=64$, and so we obtain the same concentration exponent as for a full dense Φ of size 64×1024. Remarkably, based on Figure 2(e) and Figure 3, we see that the concentration of measure behavior for this nonuniform energy signal with the 192×1024 block diagonal Φ matrix is indistinguishable from the case of the uniform energy signal with the 64×1024 dense Gaussian Φ matrix. This suggests that the factor $\Lambda(x)$ is playing an important and precise role in dictating the concentration of measure phenomena for block diagonal matrices.

5. CONCLUSION

We have derived a concentration of measure bound for block diagonal matrices composed of i.i.d. subgaussian random variables. Our main result, Theorem 3.1, shows that the concentration exponent may scale as that for a fully dense matrix. We have also identified the role that the signal energy distribution (i.e., $\Lambda(x)$) plays in this concentration exponent.

Using this bound to prove results such as the JL lemma or the RIP will require applying Theorem 3.1 to differences between signals in a set Q or to differences between sparse signals. Depending on the application, these difference vectors can be expected to take either large or small values of $\Lambda(x)$. Understanding the role of signal energy distribution for these settings will be critical for successful applications, and this remains a topic of ongoing research.

6. REFERENCES

[1] D.L. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, vol. 52, no. 4, pp. 1289–1306, April 2006.

[2] E. Candès, "Compressive sampling," in *Proc. Int. Congress Math.*, Madrid, Spain, August 2006, vol. 3, pp. 1433–1452.

[3] P. Indyk and R. Motwani, "Approximate nearest neighbors: towards removing the curse of dimensionality," in *Proc. ACM Symposium on Theory of Computing*, 1998, pp. 604–613.

[4] M. A. Davenport, P. T. Boufounos, M. B. Wakin, and R. G. Baraniuk, "Signal processing with compressive measurements," 2009, Preprint.

[5] M. Ledoux, "The concentration of measure phenomenon," *Mathematical Surveys and Monographs, American Math. Society*, 2001.

[6] R. DeVore, G. Petrova, and P. Wojtaszczyk, "Instance-optimality in probability with an ℓ_1-minimization decoder," 2008, to appear in *Appl. Comp. Harmonic Anal.*

[7] E. J. Candes and T. Tao, "Decoding by linear programming," *IEEE Trans. Inform. Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[8] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, "A simple proof of the restricted isometry property for random matrices," *Const. Approx.*, vol. 28, no. 3, pp. 253–263, 2008.

[9] J. Y. Park and M. B. Wakin, "A multiscale framework for compressive sensing of video," in *Proc. Picture Coding Symposium*, 2009.

[10] D. Baron, M. B. Wakin, M. Duarte, S. Sarvotham, and R. G. Baraniuk, "Distributed compressed sensing," Rice University Technical Report TREE-0612, Nov 2006.

[11] V. V. Buldygin and Yu. V. Kozachenko, "Sub-gaussian random variables," *Ukrainian Mathematical Journal*, vol. 32, Nov. 1980.

[12] R. Vershynin, "On large random almost Euclidean bases," *Acta Math. Univ. Comenianae*, vol. 69, no. 2, pp. 137–144, 2000.