On the Observability of Linear Systems from Random, Compressive Measurements

Michael B. Wakin, Borhan M. Sanandaji, and Tyrone L. Vincent

***Abstract*— Recovering or estimating the initial state of a high-dimensional system can require a potentially large number of measurements. In this paper, we explain how this burden can be significantly reduced for certain linear systems when randomized measurement operators are employed. Our work builds upon recent results from the field of Compressive Sensing (CS), in which a high-dimensional signal containing few nonzero entries can be efficiently recovered from a small number of random measurements. In particular, we develop concentration of measure bounds for the observability matrix and explain circumstances under which this matrix can satisfy the Restricted Isometry Property (RIP), which is central to much analysis in CS. We also illustrate our results with a simple case study of a diffusion system. Aside from permitting recovery of sparse initial states, our analysis has potential applications in solving inference problems such as detection and classification of more general initial states.**

I. Introduction

A. Measurement burdens in observability theory

Let us consider a discrete-time linear dynamical system of the form:

$$\begin{aligned} x_k &= Ax_{k-1} \\ y_k &= C_k x_k \end{aligned} \tag{1}$$

where $x_k \in \mathbb{R}^N$ represents the state vector at time $k \in \{0, 1, 2, \dots\}$, $A \in \mathbb{R}^{N \times N}$ represents the state transition matrix, $y_k \in \mathbb{R}^M$ represents a set of measurements (or "observations") of the state at time k, and $C_k \in \mathbb{R}^{M \times N}$ represents the measurement matrix at time k. For any integer $K > 0$, we also define the *observability matrix*

$$\mathcal{O}_K := \begin{bmatrix} C_0 \\ C_1 A \\ \vdots \\ C_{K-1} A^{K-1} \end{bmatrix}. \tag{2}$$

This matrix has size $KM \times N$.

Although we will consider situations in this paper where C_k changes with each k, let us first discuss the classical case where $C_k = C$ for all k. In this setting, the system (1) is said to be *observable* if $\mathcal{O}_K$ has rank N for some value of K. The primary use of observability is in ensuring that a state (say, an initial state x_0) can be recovered from a collection of measurement vectors $\{y_0, y_1, y_2, \dots, y_{K-1}\}$. In particular, defining

$$y^K := \begin{bmatrix} y_0^T & y_1^T & \cdots & y_{K-1}^T \end{bmatrix}^T$$

we have

$$y^K = \mathcal{O}_K x_0. \tag{3}$$

If $\mathcal{O}_K$ has full column rank (i.e., rank N), then it follows from standard linear algebra that we may recover any x_0 from the measurements y^K.

An important and classical result [1] states that the system (1) is observable if and only if $\operatorname{rank}(\mathcal{O}_N) = N$. In other words, if the system is observable, we need consider no more then $K = N$ successive measurement vectors to be able to recover any initial state. One challenge in exploiting this fact is that for some systems, N can be quite large. For example, distributed systems evolving on a spatial domain can have a large state space even after taking a spatially-discretized approximation. In settings such as these, we might therefore require a very large total number MN of measurements to identify an initial state, and moreover, inverting the matrix $\mathcal{O}_N$ could be very computationally demanding.

This raises an interesting question: under what circumstances might we be able to infer the initial state of a system when $K < N$? We might imagine, for example, that the measurement burden could be alleviated in cases where we have a model for the state x_0 that we wish to recover. Alternatively, we may have cases where, rather than needing to recover x_0 from y^K, we desire only to solve a much simpler inference problem such as a binary detection or a classification problem. In this paper, inspired by the emerging theory of Compressive Sensing (CS) [2, 3], we explain how such assumptions can indeed reduce the measurement burden and, in some cases, even allow recovery of the initial state when $KM < N$ and the system of equations (3) is guaranteed to be underdetermined.

B. Compressive Sensing and randomized measurements

The CS theory states that is is possible to solve certain rank-deficient sets of linear equations by imposing some model assumption on the signal to be recovered. In particular, suppose $y = \Phi x$ where Φ is an $m \times n$ matrix with $m < n$. Suppose also that $x \in \mathbb{R}^n$ is S-sparse, meaning that only S out of its n entries are nonzero.[1] Then if Φ satisfies a condition called the Restricted Isometry Property (RIP) of order $2S$, meaning that for a suitably small $\epsilon > 0$

$$(1 - \epsilon)\|u\|_2^2 \leq \|\Phi u\|_2^2 \leq (1 + \epsilon)\|u\|_2^2 \tag{4}$$

holds for all $2S$-sparse vectors $u \in \mathbb{R}^n$, then it is possible to uniquely recover any S-sparse signal x from the measurements $y = \Phi x$ using a tractable convex optimization program

All authors are with Division of Engineering, Colorado School of Mines, Golden, CO 80401, USA. Email: {mwakin, bmolazem, tvincent}@mines.edu. This work was partially supported by AFOSR Grant FA9550-09-1-0465, NSF Grant CCF-0830320, DARPA Grant HR0011-08-1-0078, NSF grant CNS-0931748 and Department of Energy, Office of Energy Efficiency and Renewable Energy Grant DE-FG36-08GO88100.

[1] This is easily extended to the case where x is sparse in some transform basis.

known as ℓ_1-minimization [2–4]. The RIP also ensures that the recovery process is robust to noise and stable in cases where x is not precisely sparse [5].

In order to obtain an RIP matrix Φ with as few rows m as possible, one commonly relies on a randomized construction. Supposing that Φ is populated with independent and identically distributed (i.i.d.) Gaussian random variables having mean zero and variance $\frac{1}{m}$, for example, then Φ satisfies the RIP of order $2S$ with very high probability assuming only that $m = O\left(S \log \frac{n}{S}\right)$. Other random distributions may also be considered, including matrices with random $\pm\frac{1}{\sqrt{m}}$ entries. Consequently, a number of new sensing hardware architectures, from analog-to-digital converters to digital cameras, are being developed to take advantage of the benefits of random measurements [6, 7].

One straightforward way [8] of proving the RIP for a randomized construction of Φ involves first showing that the matrix satisfies a concentration of measure inequality akin to the following.

Lemma 1.1: [9] Let $u \in \mathbb{R}^n$ be any fixed signal (not necessarily sparse) and let Φ be a random $m \times n$ Gaussian matrix as described above. Then with probability at least $1 - 2e^{-mc_0(\epsilon)}$, (4) holds for u. In other words,

$$P\left(|\|\Phi u\|_2^2 - \|u\|_2^2| > \|u\|_2^2 \epsilon\right) \leq 2\exp\left\{-mc_0(\epsilon)\right\}.$$

We note that in the above lemma, the failure probability decays exponentially fast in the number of measurements m and also in some function $c_0(\epsilon)$ that depends on the isometry constant ϵ.

Aside from connections to the RIP, concentration inequalities such as the above can also be useful when solving other types of inference problems from compressive measurements. For example, rather than recover a signal x, we may wish only to solve a binary detection problem and determine whether a set of measurements y correspond only to noise (the null hypothesis $y = \Phi(\text{noise})$) or to signal plus noise ($y = \Phi(x + \text{noise})$). When Φ is random, the performance of a compressive detector (and of other multi-signal classifiers) can be well studied using concentration inequalities [10], and in these settings we do not need to assume that x is sparse.

C. Observability from random, compressive measurements

In order to exploit CS concepts in observability analysis, we consider in this paper the case where the measurement matrices C_k are populated with random entries. Physically, such randomized measurements may be taken using the types of CS protocols and hardware mentioned above. Our analysis is therefore appropriate in cases where one has some control over the sensing process.

As is apparent from (2), even with randomness in the matrices C_k, our observability matrices $\mathcal{O}_K$ will contain some structure and cannot simply modeled as being populated with i.i.d. Gaussian random variables. Our main goal in this paper is to show how to account for this structure in deriving a concentration of measure bound for $\mathcal{O}_K$. As we demonstrate, the concentration performance of such a matrix depends on properties of both the state transition matrix A and the initial state x_0. This work builds on two recent papers in which we derive concentration of measure bounds for random, block diagonal measurement matrices [11, 12].

We show that, under certain conditions on A, the observability matrix $\mathcal{O}_K$ will satisfy the RIP of order S with high probability, where the total number of measurements $KM = O(S \log \frac{N}{S})$. Thus, in this best case, the concentration behavior of $\mathcal{O}_K$ can be just as favorable as for an i.i.d. Gaussian matrix of the same size. A major implication of this fact is that for certain N-dimensional systems, we can potentially infer *any* initial state from far fewer than N total measurements as long as that initial state is suitably sparse. Other inference problems concerning x_0 (such as detection or classification) could also be solved from the random, compressive measurements, and the performance of such techniques could be studied using the concentration of measure bound that we provide.

Questions involving observability in compressive measurement settings have also been raised in a recent paper [13] concerned with tracking the state of a system from nonlinear observations. Due to the intrinsic nature of the problems in that paper, however, the observability issues raised are quite different. For example, one argument appears to assume that $M \geq S$, a requirement that we do not have.

D. Paper organization

In Sections II and III, we consider two cases in turn. In both cases, the measurement matrices C_k are populated with i.i.d. Gaussian random variables having mean zero and variance $\sigma^2 = \frac{1}{M}$. In Section II, however, all matrices C_k are generated independently of each other, while in Section III, all matrices C_k are equal. Within each of these two sections, we derive a concentration of measure bound for $\mathcal{O}_K$, discuss the implications of the properties of A and of x_0, and make connections with the RIP.

In Section IV, we illustrate these phenomena with a short case study of a diffusion system. Though simplified to a one-dimensional domain, one may imagine that such problems could arise when sparse contaminants are introduced into particular (i.e., sparse) locations in a water supply or in the air. From the available measurements, we would like to find the source of the contamination.

We conclude in Section V.

II. Independent Random Measurement Matrices

In this section, we consider the case where all matrices C_k are generated independently of each other. Each matrix C_k is populated with i.i.d. Gaussian random variables having mean zero and variance $\sigma^2 = \frac{1}{M}$.

A. Connection with block diagonal matrices

To begin, it will be useful to note that we can write

$$\mathcal{O}_K = \mathcal{C}_K \mathcal{A}_K,$$

where

$$\mathcal{C}_K := \begin{bmatrix} C_0 & & & \\ & C_1 & & \\ & & \ddots & \\ & & & C_{K-1} \end{bmatrix} \tag{5}$$

and

$$\mathcal{A}_K := \begin{bmatrix} I \\ A \\ \vdots \\ A^{K-1} \end{bmatrix}. \tag{6}$$

The matrix $\mathcal{C}_K$ is block diagonal, and focusing just on this matrix for the moment, we have the following bound on its concentration behavior.[2]

Theorem 2.1: [11] Let $v_0, v_1, \ldots, v_{K-1} \in \mathbb{R}^N$ and define

$$v = \begin{bmatrix} v_0^T & v_1^T & \cdots & v_{K-1}^T \end{bmatrix}^T \in \mathbb{R}^{KN}.$$

Suppose $\mathcal{C}_K$ is a block diagonal matrix as in (5) populated with Gaussian random variables having mean zero and variance $\sigma^2 = \frac{1}{M}$. Then

$$P(\left|\|\mathcal{C}_K v\|_2^2 - \|v\|_2^2\right| > \epsilon \|v\|_2^2) \leq \begin{cases} 2\exp\{-\frac{M\epsilon^2\|\gamma\|_1^2}{256\|\gamma\|_2^2}\}, & 0 \leq \epsilon \leq \frac{16\|\gamma\|_2^2}{\|\gamma\|_\infty \|\gamma\|_1} \\ 2\exp\{-\frac{M\epsilon\|\gamma\|_1}{16\|\gamma\|_\infty}\}, & \epsilon \geq \frac{16\|\gamma\|_2^2}{\|\gamma\|_\infty \|\gamma\|_1}, \end{cases}$$

where

$$\gamma = \gamma(v) := \begin{bmatrix} \|v_0\|_2^2 \\ \|v_1\|_2^2 \\ \vdots \\ \|v_{K-1}\|_2^2 \end{bmatrix} \in \mathbb{R}^K.$$

As we will be frequently concerned with applications where ϵ is small, let us consider the first of the cases given in the right hand side of the above bound. (It can be shown [11] that this case always permits any value of ϵ between 0 and $\frac{16}{\sqrt{K}}$.) We define

$$\Gamma = \Gamma(v) := \frac{\|\gamma(v)\|_1^2}{\|\gamma(v)\|_2^2} = \frac{(\|v_0\|_2^2 + \|v_1\|_2^2 + \cdots + \|v_{K-1}\|_2^2)^2}{\|v_0\|_2^4 + \|v_1\|_2^4 + \cdots + \|v_{K-1}\|_2^4} \tag{7}$$

and note that for any $v \in \mathbb{R}^{KN}$, $1 \leq \Gamma(v) \leq K$. (This follows from the standard relation that $\|z\|_2 \leq \|z\|_1 \leq \sqrt{K}\|z\|_2$ for all $z \in \mathbb{R}^K$.)

The case $\Gamma(v) = K$ is quite favorable because the failure probability will decay exponentially fast in the total number of measurements KM. In this case, we get the same degree of concentration from the $KM \times KN$ block diagonal matrix $\mathcal{C}_K$ as we would get from a *dense* $KM \times KN$ matrix populated with i.i.d. Gaussian random variables. This event happens if and only if the components v_k have equal energy, i.e., if and only if

$$\|v_0\|_2 = \|v_1\|_2 = \cdots = \|v_{K-1}\|_2.$$

On the other hand, the case $\Gamma(v) = 1$ is quite unfavorable and implies that we get the same degree of concentration from the $KM \times KN$ block diagonal matrix $\mathcal{C}_K$ as we would get from a dense Gaussian matrix having size only $M \times KN$. This event happens if and only if $\|v_k\|_2 = 0$ for all but one k. Thus, more uniformity in the values of the $\|v_k\|_2$ ensures a higher probability of concentration.

B. Relation to the initial state

We now note that, when applying the observability matrix to an initial state, we will have

$$\mathcal{O}_K x_0 = \mathcal{C}_K \mathcal{A}_K x_0.$$

This leads us to the following corollary of Theorem 2.1.

Corollary 2.1: Fix any state $x_0 \in \mathbb{R}^N$. Then for any $\epsilon \in (0, \frac{16}{\sqrt{K}})$,

$$P(\left|\|\mathcal{O}_K x_0\|_2^2 - \|\mathcal{A}_K x_0\|_2^2\right| > \epsilon \|\mathcal{A}_K x_0\|_2^2) \leq 2\exp\left\{-\frac{M\Gamma(\mathcal{A}_K x_0)\epsilon^2}{256}\right\}. \tag{8}$$

There are two important phenomena to consider in this result, and both are impacted by the interaction of A with x_0. First, on the left hand side of (8), we see that the point of concentration of $\|\mathcal{O}_K x_0\|_2^2$ is actually around $\|\mathcal{A}_K x_0\|_2^2$, where

$$\|\mathcal{A}_K x_0\|_2^2 = \|x_0\|_2^2 + \|Ax_0\|_2^2 + \cdots + \|A^{K-1}x_0\|_2^2.$$

For a concentration bound of the same form as Lemma 1.1, however, we might like to ensure that $\|\mathcal{O}_K x_0\|_2^2$ concentrates around some constant multiple of $\|x_0\|_2^2$. In general, for different initial states x_0 and transition matrices A, we may see widely varying ratios $\|\mathcal{A}_K x_0\|_2^2 / \|x_0\|_2^2$. However, in Section II-C, we discuss one scenario where this ratio is predictable and fixed.

Second, on the right hand side of (8), we see that the exponent of the concentration failure probability scales with

$$\Gamma(\mathcal{A}_K x_0) = \frac{(\|x_0\|_2^2 + \|Ax_0\|_2^2 + \cdots + \|A^{K-1}x_0\|_2^2)^2}{\|x_0\|_2^4 + \|Ax_0\|_2^4 + \cdots + \|A^{K-1}x_0\|_2^4}.$$

From the discussion in Section II-A, it follows that $1 \leq \Gamma(\mathcal{A}_K x_0) \leq K$. The case $\Gamma(\mathcal{A}_K x_0) = K$ is quite favorable and happens when $\|x_0\|_2 = \|Ax_0\|_2 = \cdots = \|A^{K-1}x_0\|_2$; in Section II-C, we discuss one scenario where this is guaranteed to occur. The case $\Gamma(\mathcal{A}_K x_0) = 1$ is quite unfavorable and happens if and only if $x_0 \neq 0$ and $x_0 \in \text{null}(A)$.

C. Unitary system matrices

In the special case where A is unitary (i.e., $\|Au\|_2^2 = \|u\|_2^2$ for all $u \in \mathbb{R}^N$), we can draw a particularly strong conclusion. Because a unitary A guarantees both that $\|\mathcal{A}_K x_0\|_2^2 = K\|x_0\|_2^2$ and that $\Gamma(\mathcal{A}_K x_0) = K$, we have the following result.[3]

Corollary 2.2: Fix any state $x_0 \in \mathbb{R}^N$ and suppose that A is a unitary operator. Then for any $\epsilon \in (0, \frac{16}{\sqrt{K}})$,

$$P(\left|\|\frac{1}{\sqrt{K}}\mathcal{O}_K x_0\|_2^2 - \|x_0\|_2^2\right| > \epsilon \|x_0\|_2^2) \leq 2\exp\left\{-\frac{MK\epsilon^2}{256}\right\}. \tag{9}$$

[2] All results in Section II may be extended to the case where the matrices C_k are populated with subgaussian random variables, as in [11].

[3] Corollary 2.2 may be relaxed in a natural way if the singular values of A all cluster around (but may not equal) 1.

What this means is that we get the same degree of concentration from the $KM \times N$ observability matrix $\mathcal{O}_K$ as we would get from a fully random dense $KM \times N$ matrix populated with i.i.d. Gaussian random variables. Consequently, many results from CS carry through directly, including the following.

Corollary 2.3: Suppose that A is a unitary operator and that $KM = O(S \log \frac{N}{S})$. Then with high probability, $\frac{1}{\sqrt{K}}\mathcal{O}_K$ satisfies the RIP of order $2S$, and so any S-sparse initial state x_0 may be uniquely recovered from the measurements (3).

Beyond considering sparse signal families, this concentration result can also be used to prove that finite point clouds [14] and low-dimensional manifolds [15] in $\mathbb{R}^N$ can have stable, approximate distance-preserving embeddings under the matrix $\mathcal{O}_K$. In each of these cases we may be able to solve very powerful signal inference and recovery problems with $KM \ll N$.

III. Identical Random Measurement Matrices

In this section, we consider the case where all matrices C_k are identical and equal to some $M \times N$ matrix C which is populated with i.i.d. Gaussian entries having mean zero and variance $\sigma^2 = \frac{1}{M}$.

A. Connection with block diagonal matrices

We can again write $\mathcal{O}_K = \mathcal{C}_K \mathcal{A}_K$, where this time

$$\mathcal{C}_K := \begin{bmatrix} C_0 & & & \\ & C_1 & & \\ & & \ddots & \\ & & & C_{K-1} \end{bmatrix} = \begin{bmatrix} C & & & \\ & C & & \\ & & \ddots & \\ & & & C \end{bmatrix} \tag{10}$$

and $\mathcal{A}_K$ is as defined in (6). The matrix $\mathcal{C}_K$ is block diagonal with equal blocks on its main diagonal, and we have the following bound on its concentration behavior.

Theorem 3.1: [12] Let $v_0, v_1, \ldots, v_{K-1} \in \mathbb{R}^N$ and define

$$v = \begin{bmatrix} v_0^T & v_1^T & \cdots & v_{K-1}^T \end{bmatrix}^T \in \mathbb{R}^{KN}.$$

Suppose $\mathcal{C}_K$ is a block diagonal matrix as in (10) populated with Gaussian random variables having mean zero and variance $\sigma^2 = \frac{1}{M}$. Then

$$P\left(\left|\|\mathcal{C}_K v\|_2^2 - \|v\|_2^2\right| > \epsilon \|v\|_2^2\right) \leq \begin{cases} 2\exp\{-\frac{M\epsilon^2\|\lambda\|_1^2}{256\|\lambda\|_2^2}\}, & 0 \leq \epsilon \leq \frac{16\|\lambda\|_2^2}{\|\lambda\|_\infty \|\lambda\|_1} \\ 2\exp\{-\frac{M\epsilon\|\lambda\|_1}{16\|\lambda\|_\infty}\}, & \epsilon \geq \frac{16\|\lambda\|_2^2}{\|\lambda\|_\infty \|\lambda\|_1}, \end{cases}$$

where

$$\lambda = \lambda(v) := \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_{\min(K,N)} \end{bmatrix} \in \mathbb{R}^{\min(K,N)}$$

and $\{\lambda_1, \lambda_2, \ldots, \lambda_{\min(K,N)}\}$ are the first (nonzero) eigenvalues of the $K \times K$ matrix $V^T V$, where

$$V = \begin{bmatrix} v_0 & v_1 & \cdots & v_{K-1} \end{bmatrix} \in \mathbb{R}^{N \times K}.$$

Let us again consider the first of the cases given in the right hand side of the above bound. (Once again, this case permits any value of ϵ between 0 and $\frac{16}{\sqrt{\min(K,N)}}$.) We define

$$\Lambda(v) := \frac{\|\lambda(v)\|_1^2}{\|\lambda(v)\|_2^2} \tag{11}$$

and note that for any $v \in \mathbb{R}^{KN}$, $1 \leq \Lambda(v) \leq \min(K, N)$. Moving forward, we will assume for simplicity that $K \leq N$, but this assumption can be removed without much complication.

The case $\Lambda(v) = K$ is quite favorable and implies that we get the same degree of concentration from the $KM \times KN$ block diagonal matrix $\mathcal{C}_K$ as we would get from a dense $KM \times KN$ matrix populated with i.i.d. Gaussian random variables. This event happens if and only if $\lambda_1 = \lambda_2 = \cdots = \lambda_K$, which happens if and only if

$$\|v_0\|_2 = \|v_1\|_2 = \cdots = \|v_{K-1}\|_2$$

and $\langle v_k, v_\ell \rangle = 0$ for all $0 \leq k, \ell \leq K-1$ with $k \neq \ell$. On the other hand, the case $\Lambda(v) = 1$ is quite unfavorable and implies that we get the same degree of concentration from the $KM \times KN$ block diagonal matrix $\mathcal{C}_K$ as we would get from a dense Gaussian matrix having only M rows. This event happens if and only if the dimension of $\operatorname{span}\{v_0, v_1, \ldots, v_{K-1}\}$ equals 1. Thus, comparing to Section II-A, uniformity in the norms of the vectors v_k is no longer sufficient for a high probability of concentration; in addition to this we must have diversity in the directions of the v_k.

B. Relation to the initial state

We again note that, when applying the observability matrix to an initial state, we will have

$$\mathcal{O}_K x_0 = \mathcal{C}_K \mathcal{A}_K x_0.$$

This leads us to the following corollary of Theorem 3.1.

Corollary 3.1: Fix any state $x_0 \in \mathbb{R}^N$. Then for any $\epsilon \in (0, \frac{16}{\sqrt{K}})$,

$$P\left(\left|\|\mathcal{O}_K x_0\|_2^2 - \|\mathcal{A}_K x_0\|_2^2\right| > \epsilon \|\mathcal{A}_K x_0\|_2^2\right) \leq 2\exp\left\{-\frac{M\Lambda(\mathcal{A}_K x_0)\epsilon^2}{256}\right\}. \tag{12}$$

Once again, there are two important phenomena to consider in this result, and both are impacted by the interaction of A with x_0. First, on the left hand side of (12), we see that the point of concentration of $\|\mathcal{O}_K x_0\|_2^2$ is again around $\|\mathcal{A}_K x_0\|_2^2$. Second, on the right hand side of (12), we see that the exponent of the concentration failure probability scales with $\Lambda(\mathcal{A}_K x_0)$, which is determined by the eigenvalues of the $K \times K$ Gram matrix $V^T V$, where

$$V = \begin{bmatrix} x_0 & Ax_0 & \cdots & A^{K-1}x_0 \end{bmatrix} \in \mathbb{R}^{N \times K}.$$

From the discussion in Section III-A, it follows that $1 \leq \Lambda(\mathcal{A}_K x_0) \leq K$. The case $\Gamma(\mathcal{A}_K x_0) = K$ is quite favorable and happens when $\|x_0\|_2 = \|Ax_0\|_2 = \cdots = \|A^{K-1}x_0\|_2$

and $\langle A^k x_0, A^\ell x_0 \rangle = 0$ for all $0 \le k, \ell \le K-1$ with $k \neq \ell$. The case $\Lambda(\mathcal{A}_K x_0) = 1$ is quite unfavorable and happens the dimension of $\text{span}\{x_0, Ax_0, \ldots, A^{K-1}x_0\}$ equals 1.

C. Unitary system matrices

In the special case where A is unitary, we know that $\|\mathcal{A}_K x_0\|_2^2 = K\|x_0\|_2^2$. However, a unitary system matrix does not guarantee a favorable value for $\Lambda(\mathcal{A}_K x_0)$. Indeed, if $A = I_{N\times N}$ we obtain the worse case value $\Lambda(\mathcal{A}_K x_0) = 1$. If, on the other hand, A acts as a rotation that takes a state into an orthogonal subspace, we will have a stronger result.

Corollary 3.2: Fix any state $x_0 \in \mathbb{R}^N$ and suppose that A is a unitary operator. Suppose also that $\langle A^k x_0, A^\ell x_0 \rangle = 0$ for all $0 \le k, \ell \le K-1$ with $k \neq \ell$. Then for any $\epsilon \in (0, \frac{16}{\sqrt{K}})$,

$$P\left(\left|\|\frac{1}{\sqrt{K}}\mathcal{O}_K x_0\|_2^2 - \|x_0\|_2^2\right| > \epsilon\|x_0\|_2^2\right) \le 2\exp\left\{-\frac{MK\epsilon^2}{256}\right\}. \quad (13)$$

This result requires a particular relationship between A and x_0, namely that $\langle A^k x_0, A^\ell x_0 \rangle = 0$ for all $0 \le k, \ell \le K-1$ with $k \neq \ell$. Thus, given a particular system matrix A, it is possible that it might hold for some x_0 and not others. One must therefore be cautious in using this concentration result for CS applications (such as proving the RIP) that involve applying the concentration bound to a prescribed collection of vectors [8]; one must ensure that the "orthogonal rotation" property holds for each vector in the prescribed set. We defer a deeper discussion of this topic to a subsequent paper.

IV. Example: Estimating the Initial State in a Diffusion Process

We now use a simple case study to illustrate some of the phenomena raised in the previous sections.

A. System model

We consider the problem of estimating the initial state in a system governed by the diffusion equation

$$\frac{\partial x}{\partial t} = \nabla \cdot (D(p)\nabla x(p,t))$$

where $x(p,t)$ is the concentration, or density, at position p at time t, and $D(p)$ is the diffusion coefficient at position p. If D is independent of position, then this simplifies to

$$\frac{\partial x}{\partial t} = D\nabla^2 x(p,t).$$

The boundary conditions can vary according to the surroundings of the domain Ω. If Ω is bounded by an impermeable surface (e.g., a lake surrounded by the shore), then the boundary conditions are $n(p) \cdot \frac{\partial x}{\partial p}\Big|_{p\in\partial\Omega} = 0$, where $n(p)$ is normal to $\partial\Omega$ at p.

We will work with an approximate model discretized in time and in space and having one spatial dimension. We let $p = \begin{bmatrix} p(1) & p(2) & \cdots & p(N) \end{bmatrix}^T$ be a vector of equally spaced locations with spacing Δ_s, and $x(p,t) = \begin{bmatrix} x(p(1),t) & x(p(2),t) & \cdots & x(p(N),t) \end{bmatrix}^T$. Then a first difference approximation in space gives the model

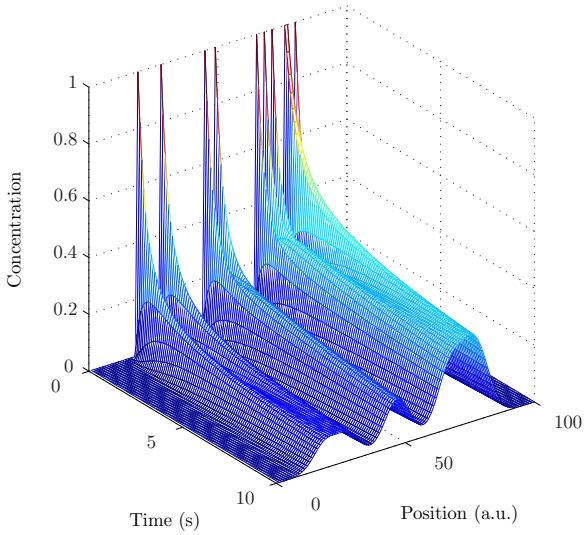

Fig. 1. Simulation of diffusion equation with sparse initial state.

$$\dot{x}(p,t) = Gx(p,t) \quad (14)$$

where G represents the discrete Laplacian:

$$G = \frac{D}{\Delta_s^2}\begin{bmatrix} -1 & 1 & 0 & 0 & \cdots & 0 \\ 1 & -2 & 1 & 0 & \cdots & 0 \\ 0 & 1 & -2 & 1 & \cdots & 0 \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & & & 1 & -1 \end{bmatrix}.$$

To obtain a discrete time model, we choose sampling time T_s, and let the vector $x_k = x(p, kT_s)$ be the concentration at positions $p(1), p(2), \ldots, p(N)$ at sampling time k. Using a first difference approximation in time, we have

$$x_k = Ax_{k-1}$$

where $A = I + GT_s$.

For all experiments in this section we take $D = 1$, $\Delta_s = 1$, $N = 100$, and $T_s = 0.1$. An example simulation of this system is shown in Figure 1, where we have initialized the system with a sparse initial state x_0 containing unit impulses at $S = 10$ randomly chosen locations.

From compressive measurements of this system, it is sometimes possible to recover the initial state. In Section IV-C, we provide several demonstrations of this fact, and we discuss the effects of choosing different times at which to measure the system. Before dealing with the problem of recovery, however, we start in Section IV-B by examining the concentration behavior of this system. Because this system is not unitary and we cannot directly invoke Corollaries 2.2 or 2.3, we explore the connection between concentration and recovery numerically.

B. Concentration behavior with compressive measurements

As we have discussed in Sections II-A and III-A, a favorable situation occurs when repeated applications of the

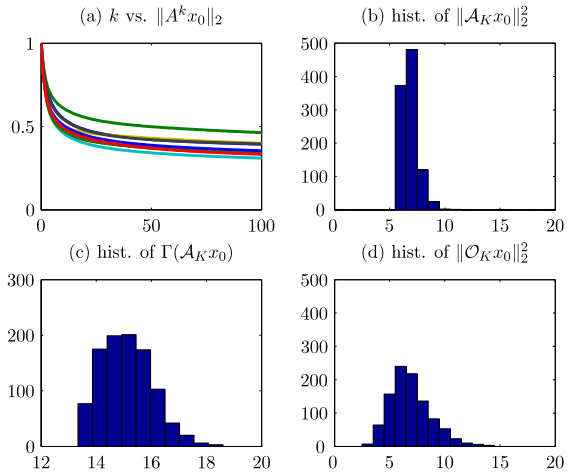

Fig. 2. Concentration behavior for various sparse initial states x_0 having $S = 10$ random positive entries in random locations. (a) Decay of $\|A^k x_0\|_2$. (b) Concentration of $\|\mathcal{A}_K x_0\|_2^2$ with $K = 20$ and $M = 2$. (c) Histogram of $\Gamma(\mathcal{A}_K x_0)$ values. (d) Concentration of $\|\mathcal{O}_K x_0\|_2^2$ when independent measurement matrices are used.

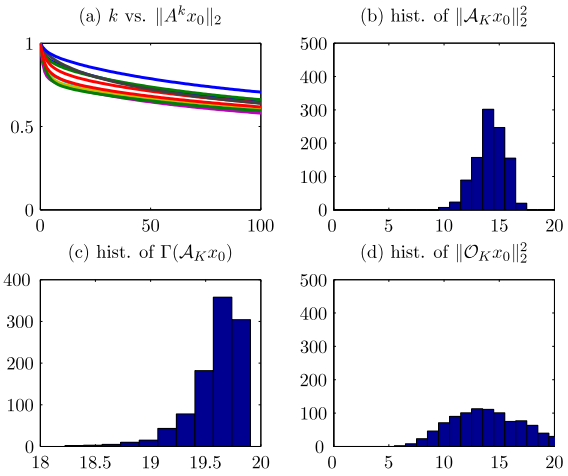

Fig. 3. Concentration behavior for various sparse initial states x_0 having $S = 10$ random positive entries in one contiguous block at the center of the vector. (a) Decay of $\|A^k x_0\|_2$. (b) Concentration of $\|\mathcal{A}_K x_0\|_2^2$ with $K = 20$ and $M = 2$. (c) Histogram of $\Gamma(\mathcal{A}_K x_0)$ values. (d) Concentration of $\|\mathcal{O}_K x_0\|_2^2$ when independent measurement matrices are used.

transition matrix A do not significantly change the energy of the state. Because A is not unitary for this system, it will preserve the norm of some vectors more than others.[4] Figure 2(a) plots the norm $\|A^k x_0\|_2$ as a function of k for various random sparse vectors $x_0 \in \mathbb{R}^N$ having $S = 10$ random positive entries in random locations. Initially, each vector x_0 is normalized to have unit energy. For values of k in this range, we see that the energy of the initial vector is mostly preserved. This leads to a reasonable concentration for values of $\|\mathcal{A}_K x_0\|_2^2$; we plot a histogram of this quantity in Figure 2(b), where we have set $K = 20$ and $M = 2$ and generated 1000 random sparse signals. This also leads to favorable values of $\Gamma(\mathcal{A}_K x_0)$, which are relatively close to the best possible value of $K = 20$ as shown in the histogram of Figure 2(c). In contrast, the lack of diversity in $A^k x_0$ over time leads to poor values for $\Lambda(\mathcal{A}_K x_0)$, which tend to cluster around 1.2. This suggests that such a diffusion system should be measured using matrices C_k that are generated randomly and independently of each other. Figure 2(d) shows the resulting concentration of $\|\mathcal{O}_K x_0\|_2^2$ when independent measurement matrices are used.

While we have considered generic sparse vectors above, the behavior of A on any particular initial state can depend very much on that particular state. In Figure 3, we repeat all of the above experiments, again with random sparse vectors $x_0 \in \mathbb{R}^N$ having $S = 10$ random positive entries, but where the nonzero entries all occur in one "block" in the middle of the vector (centered around position $N/2$). We see that the values of $\Gamma(\mathcal{A}_K x_0)$ tend to be even higher, but the point of concentration for $\|\mathcal{O}_K x_0\|_2^2$ is markedly different.

Finally, we recall that properties such as the RIP require favorable concentration behavior for *any* sparse initial state.

[4]This can be understood more formally by considering the eigendecomposition of A. In this case, since A is nearly a circulant matrix, it is approximately diagonalized by the Discrete Fourier Transform. Its eigenvalues decay from 1 as the frequency of the eigenvector increases.

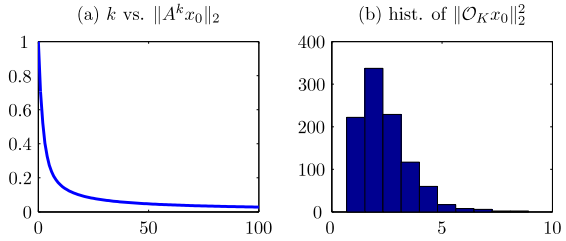

Fig. 4. Concentration behavior for various sparse initial state x_0 having ± 1 entries. (a) Decay of $\|A^k x_0\|_2$. (b) Concentration of $\|\mathcal{O}_K x_0\|_2^2$ when independent measurement matrices are used.

We create a "high frequency" sparse signal by setting $x_0 = 0$ everywhere except position $N/2$, where it equals $1/\sqrt{2}$, and position $(N/2)+1$, where it equals $-1/\sqrt{2}$. (Although this vector has a negative entry and is not itself a plausible initial state, it is relevant for applications such as proving the RIP, where we must consider *differences* between plausible initial states.) As shown in Figure 4, this vector has a much faster decay of $\|A^k x_0\|_2$ and the point of concentration for $\|\mathcal{O}_K x_0\|_2^2$ is therefore quite small.

C. State recovery from compressive measurements

To address the problem of recovering the initial state x_0, let us consider the situation where we collect measurements only of x_0 itself. We set $M = 32$ and construct measurement matrices C_0 of size 32×100 that are populated with i.i.d. Gaussian entries having variance $\frac{1}{32}$. We then generate random sparse vectors x_0 with varying sparsity levels S, and for each of these we collect the measurements $y_0 = C_0 x_0$. From these measurements, we attempt to recover the initial state using the canonical ℓ_1 minimization problem from CS:

$$\widehat{x}_0 = \arg\min_{x \in \mathbb{R}^N} \|x\|_1 \quad \text{subject to} \quad y_k = C_k A^k x \qquad (15)$$

with $k = 0$. (In the next paragraph, we repeat this experiment for different k.) In Figures 5(a) and 5(b) we plot, as a

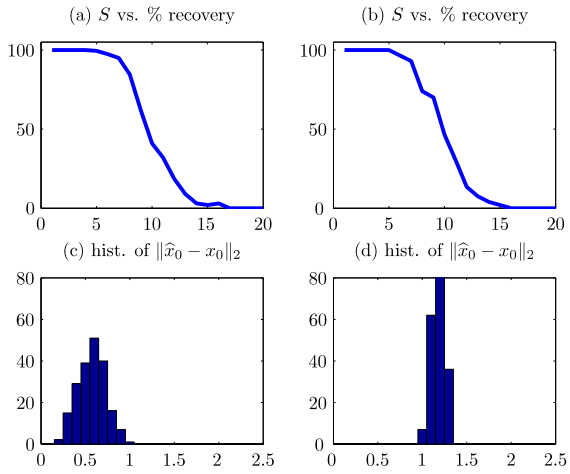

Fig. 5. Signal recovery from $M = 32$ compressive measurements at time $k = 0$. (a),(b) Percent of trials with perfect recovery of S-sparse signals, where the signals have random entries in (a) random and (b) clustered locations. (c),(d) Recovery error of signals with sparsity $S = 10$, where the signals have random entries in (c) random and (d) clustered locations.

	k	0	10	100	10,11,12,13†
rand	mean($\|C_k A^k x_0\|_2^2$)	1.00	0.26	0.13	0.25
	std($\|C_k A^k x_0\|_2^2$)	0.25	0.08	0.04	0.07
block	mean($\|C_k A^k x_0\|_2^2$)	1.00	0.68	0.41	0.66
	std($\|C_k A^k x_0\|_2^2$)	0.25	0.19	0.11	0.18
± 1	mean($\|C_k A^k x_0\|_2^2$)	1.00	0.025	0.0008	0.021
	std($\|C_k A^k x_0\|_2^2$)	0.24	0.007	0.0002	0.005

TABLE I

MEASUREMENT VECTOR ENERGIES FOR THREE TYPES OF SPARSE SIGNALS x_0.†IN THE FINAL COLUMN, WE LIST CONCENTRATION VALUES FOR THE CONCATENATION $[(C_{10}A^{10}x_0)^T \ \cdots \ (C_{13}A^{13}x_0)^T]^T$.

function of S, the percent of trials (with x_0 and C_0 randomly chosen in each trial) in which the initial state is recovered perfectly, i.e., $\widehat{x}_0 = x_0$. The first plot corresponds to sparse vectors generated according to the same model used in Figure 2 (i.e., with random positions) and the second plot corresponds to sparse vectors generated according to the same model used in Figure 3 (i.e., with a cluster of nonzeros). Naturally, we see that states x_0 with higher sparsity levels S are more difficult to recover. In Figures 5(c) and 5(d), we consider only states with sparsity $S = 10$, introduce white noise in the measurements with standard deviation 0.05, use a noise-aware version of the ℓ_1 recovery algorithm [5], and plot a histogram of the recovery errors $\|\widehat{x}_0 - x_0\|_2$. Finally, in Table I, we provide a small collection of concentration results for this measurement operator, listing the mean and standard deviation of $\|C_0 x_0\|_2^2$ for the same three types of signals x_0 considered in Section IV-B: sparse signals with $S = 10$ having random values and positions, sparse signals with $S = 10$ having random values in a cluster, and a fixed $\pm\frac{1}{\sqrt{2}}$ signal with $S = 2$.

As can be seen in Figure 1, the diffusion process causes a profound "spreading" of the spikes that should make them difficult to distinguish as time evolves. What seems intuitively clear is that measurements should be taken as

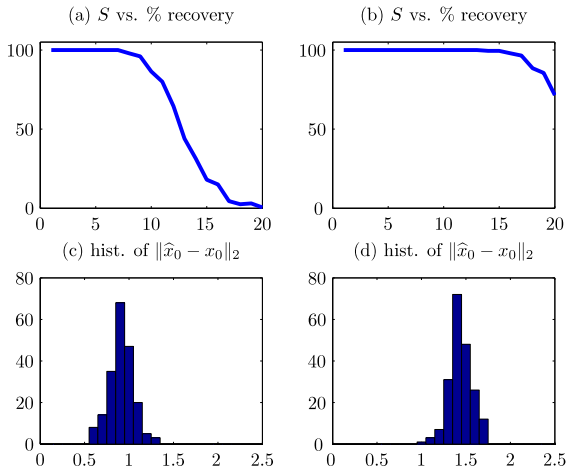

Fig. 6. Signal recovery from $M = 32$ compressive measurements at time $k = 10$.

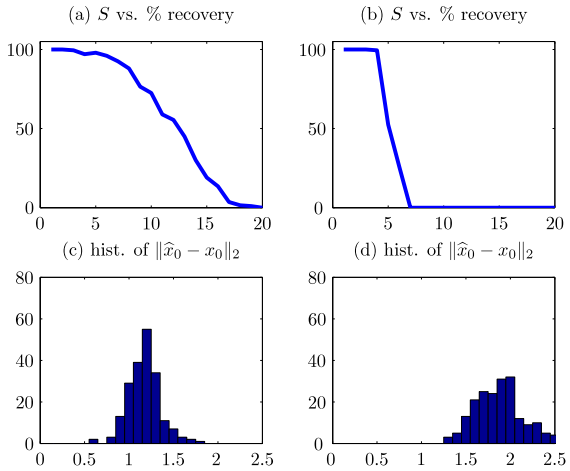

Fig. 7. Signal recovery from $M = 32$ compressive measurements at time $k = 100$. In each of panels (c) and (d), a small number of trials are omitted from the histogram in which recovery error is as large as 5.5.

soon as possible after the diffusion begins. For example, we repeat the recovery experiments described in the previous paragraph, but with $M = 32$ compressive measurements only of state x_{10} or only of state x_{100}; that is, for $k = 10$ or $k = 100$, we collect only the measurements $y_k = C_k x_k$ and solve the recovery program (15) (or the noise-aware version) to recover x_0. The results are plotted in Figures 6 and 7, respectively. When measuring x_{10} we see a surprising improvement in noise-free recovery of the clustered sparse signals, but this improvement vanishes when measuring x_{100}. More significantly, however, we see that recovery of x_{10} and then x_{100} are progressively less and less robust to measurement noise. While we cannot efficiently verify whether our sensing matrices meet the RIP, this increased noise sensitivity is likely due directly to the poor concentration the system exhibits for certain high-frequency vectors, as demonstrated in Figure 4. This is also evident in Table I, where the three signal types exhibit markedly different concentration behavior in $\|C_k A^k x_0\|_2^2$ as k grows.

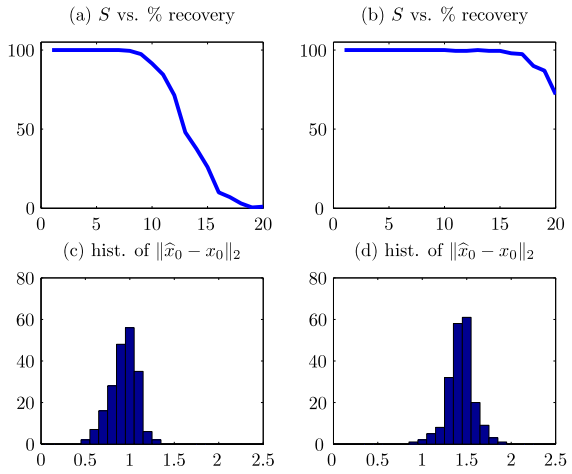

Fig. 8. Signal recovery from $M = 8$ compressive measurements at each of times $k = 10, 11, 12, 13$.

Of course, it is not necessary to take all measurements of this system at one time instant. What may not obvious a priori is how spreading the measurements in time may impact the signal recovery. In Figure 8 we repeat the signal recovery experiments when taking $M = 8$ measurements of 4 successive states x_k for $k = 10, 11, 12, 13$. (The random elements of the measurement matrices C_{10}, C_{11}, C_{12}, C_{13} were generated with variance $\frac{1}{32}$ for a fair comparison with Figures 5-7 regarding noise tolerance, and the recovery program (15) and its noise-aware version were adapted to incorporate constraints from all four measurement vectors $y_{10}, y_{11}, y_{12}, y_{13}$.) Based on the results of Section IV-B, we do not expect dramatic changes in $\|A^k x_0\|_2$ over this range of k, and so the overall recovery performance should not be much different compared to, say, taking $M = 32$ measurements at the single instant $k = 10$. By comparing the second and the fourth columns of Table I, and by comparing Figure 6 to Figure 8, we see that the overall concentration and recovery performance is indeed quite similar, and so there is no significant penalty that one pays by slightly spreading out the measurement collection process in time, as long as a different random measurement matrix is used at each time instant.[5] Though we omit the results, the same similarities have been observed when measuring states x_k for $k = 100, 101, 102, 103$ and comparing recovery performance to Figure 7.

V. Conclusions

In this paper, we have built upon CS principles to demonstrate that the certain initial states of certain high-dimensional linear systems can be recovered from very small numbers of randomized measurements. Our analysis centered around the principle of concentration of measure, and we studied the signal and system properties that are most desirable in ensuring favorable concentration behavior. In particular, when using different measurement matrices at each time instant, it is most favorable when the energy of the state is slowly changing. Moreover, when using identical measurement matrices at each time instant, it is desirable that the states traverse various subspaces of $\mathbb{R}^N$. As discussed in Section I, other inference problems aside from signal recovery (such as detection or classification) could also be solved from the random, compressive measurements, and following [10], the performance of such techniques could be studied using the concentration of measure bounds that we provide.

In ongoing work, we are studying the implications of these results in systems beyond the diffusion example we have given here. One potential future application may be in helping optimize the sensor placements for seismic imaging, in which sparse representations are currently employed to improve reconstruction [16].

[5] In fact, in all trials of this experiment with 4 observation times, the observed value of Γ was always between 3.9 and 4.0 (which is the best possible).

Acknowledgments

The authors gratefully acknowledge Chris Rozell, Han Lun Yap, Jae Young Park, and Armin Eftekhari for helpful conversations during the development of this work.

References

[1] C. T. Chen, *Linear System Theory and Design*, Oxford University Press, 3rd edition, 1999.
[2] D.L. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, vol. 52, no. 4, pp. 1289–1306, April 2006.
[3] E. Candès, "Compressive sampling," in *Proc. Int. Congress Math.*, Madrid, Spain, August 2006, vol. 3, pp. 1433–1452.
[4] E. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. Inform. Theory*, vol. 12, no. 51, pp. 4203–4215, Dec 2005.
[5] E. Candès, "The restricted isometry property and its implications for compressed sensing," in *Compte Rendus de l'Academie des Sciences, Paris, Series I*, 2008, vol. 346, pp. 589–592.
[6] M.F. Duarte, M.A. Davenport, D. Takhar, J.N. Laska, T. Sun, K.F. Kelly, and R.G. Baraniuk, "Single-pixel imaging via compressive sampling," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 83–91, 2008.
[7] D. Healy and D. J. Brady, "Compression at the physical interface," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 67–71, 2008.
[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] 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.*
[10] M.A. Davenport, P.T. Boufounos, M.B. Wakin, and R.G. Baraniuk, "Signal processing with compressive measurements," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 2, pp. 445–460, 2010.
[11] M. B. Wakin, J. Y. Park, H. L. Yap, and C. J. Rozell, "Concentration of measure for block diagonal measurement matrices," in *Proc. Int. Conf. Acoustics, Speech, Signal Proc. (ICASSP)*, March 2010.
[12] C. J. Rozell, H. L. Yap, J. Y. Park, and M. B. Wakin, "Concentration of measure for block diagonal matrices eith repeated blocks," in *Proc. Conf. Inform. Sci. Sys. (CISS)*, March 2010.
[13] E. Wang, J. Silva, and L. Carin, "Compressive particle filtering for target tracking," in *Proc. Stat. Signal Process. Workshop (SSP)*, 2009.
[14] 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.
[15] R. G. Baraniuk and M. B. Wakin, "Random projections of smooth manifolds," *Foundations of Computational Mathematics*, vol. 9, no. 1, pp. 51–77, Feb 2009.
[16] W. Tang, J. Ma, and F. J. Herrmann, "Optimized compressed sensing for curvelet-based seismic data reconstruction," 2009, Preprint.