Sequential Compressed Sensing

Dmitry M. Malioutov, Sujay R. Sanghavi, and Alan S. Willsky, *Fellow, IEEE*

***Abstract*—Compressed sensing allows perfect recovery of sparse signals (or signals sparse in some basis) using only a small number of random measurements. Existing results in compressed sensing literature have focused on characterizing the achievable performance by bounding the number of samples required for a given level of signal sparsity. However, using these bounds to minimize the number of samples requires *a priori* knowledge of the sparsity of the unknown signal, or the decay structure for near-sparse signals. Furthermore, there are some popular recovery methods for which no such bounds are known. In this paper, we investigate an alternative scenario where observations are available in sequence. For any recovery method, this means that there is now a sequence of candidate reconstructions. We propose a method to estimate the reconstruction error directly from the samples themselves, for every candidate in this sequence. This estimate is universal in the sense that it is based only on the measurement ensemble, and not on the recovery method or any assumed level of sparsity of the unknown signal. With these estimates, one can now stop observations as soon as there is reasonable certainty of either exact or sufficiently accurate reconstruction. They also provide a way to obtain "run-time" guarantees for recovery methods that otherwise lack *a priori* performance bounds. We investigate both continuous (e.g., Gaussian) and discrete (e.g., Bernoulli) random measurement ensembles, both for exactly sparse and general near-sparse signals, and with both noisy and noiseless measurements.**

***Index Terms*—Compressed sensing (CS), sequential measurements, stopping rule.**

I. Introduction

In compressed sensing (CS) [1], [2], a few random linear measurements of a signal are taken, and the signal is recovered using the additional knowledge that either the signal or some linear transform of it is sparse. These ideas have generated a lot of excitement in the signal processing and machine learning communities, and have been applied to a range of applications such as magnetic resonance imaging (MRI) [3], computational photography [4], wireless networks [5], and structure discovery in biological networks [6].

Manuscript received February 20, 2009; revised October 30, 2009. Current version published March 17, 2010. This work was supported in part by the Army Research Office under Grant W911NF-05-1-0207 and in part by the Air Force Office of Scientific Research under Grant FA9550-04-1-0351. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Rick Chartrand.

D. M. Malioutov is with the Microsoft Research, Cambridge CB3 0FB, U.K. (e-mail: dmal@alum.mit.edu).

S. R. Sanghavi is with the Electrical and Computer Engineering Department, University of Texas, Austin, TX 78712-0240 USA (e-mail: sanghavi@mail.utexas.edu).

A. S. Willsky is with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: willsky@mit.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2009.2038211

The applications where compressed sensing is most beneficial (e.g., MRI) have a high cost of acquiring each additional sample. If this cost (in terms of time, power, etc.) is high as compared to the cost of computation, then it is suitable to use sophisticated recovery algorithms which include the ℓ_1-based *basis pursuit* [7], greedy approaches [8], and even non-convex (ℓ_p) or iterative formulations [9]–[11] to enable recovery from fewer measurements.

While some of the recovery methods, especially those based on ℓ_1-regularization, have analytically provable performance guarantees [2], [12], others, such as non-convex ℓ_p, reweighted ℓ_1 [11], and sparse Bayesian learning (SBL) [13] do not, and they have been shown empirically to often require even fewer samples than ℓ_1-based methods. Furthermore, when guarantees do exist, they have been empirically observed to sometimes be highly pessimistic and may require large dimensions to hold with high probability [1], [14]. Another drawback is that much of the existing analysis characterizes how many measurements are needed for a signal with a given sparsity level. However, as the sparsity level is often not known *a priori*, it can be very challenging to use these results in practical settings.

In this paper, we take an alternative approach and we develop estimates and bounds for the reconstruction error using only the observations, without any *a priori* assumptions on signal sparsity, or on the reconstruction method. We consider a scenario where one is able to get observations in sequence, and perform computations in between observations to decide whether enough samples have been obtained—thus allowing to recover the signal either exactly or to a given tolerance from the smallest possible number of random observations. This, however, requires a computationally efficient approach to detect exactly when enough samples have been received. To get an intuition behind our approach—suppose that we first attempt to reconstruct the signal while withholding some available observations, akin to cross-validation. The observations correspond to a known linear function of the true signal, so if the reconstructed signal is quite different from the true signal, then the same linear function applied to our recovered signal will result in a value that is far from the actual observation, with high probability. Our results provide estimates of the reconstruction error based on the statistics of the measurement model. They can thus be used to provide "run-time" guarantees even for decoders that are otherwise not amenable to analysis.

We first consider the case when noiseless measurements are taken using the random Gaussian (or generic continuous) ensemble, and we show that simply checking for one-step agreement provides a way to check exactly when enough samples have been received. Suppose that after receiving M samples $y_i = \mathbf{a}_i'\mathbf{x}$, $i = 1, \ldots, M$, we apply a sparse reconstruction method of our choice, and obtain a solution $\hat{\mathbf{x}}^M$ satisfying all the M measurements. We can use any sparse decoder, including

greedy matching pursuit, SBL, ℓ_p formulations, and even the brute-force decoder, but we require that the solution at each step M satisfies $y_i = \mathbf{a}_i'\hat{\mathbf{x}}^M$, for $i = 1, \ldots, M$. For example, in the case of basis pursuit, we would solve

$$\hat{\mathbf{x}}^M = \arg\min ||\mathbf{x}||_1 \quad \text{s.t.} \quad \mathbf{a}_i'\mathbf{x} = y_i, \quad i = 1, \ldots, M. \quad (1)$$

Next, we receive one more measurement, and check for one step agreement: i.e., if $\hat{\mathbf{x}}^{M+1} = \hat{\mathbf{x}}^M$, then the decoder declares $\hat{\mathbf{x}}^M$ to be the reconstruction and stops requesting new measurements. In Section III, we show in Propositions 1 and 2 that this decoder gives exact reconstruction with probability one.

For some other measurement ensembles, such as random Bernoulli and the ensemble of random rows from a Fourier basis, the one-step agreement stopping rule no longer has zero probability of error. We modify the rule to wait until T subsequent solutions $\hat{\mathbf{x}}^M, \ldots, \hat{\mathbf{x}}^{M+T}$ all agree. In Section IV, we show in Proposition 3 that in the Bernoulli case the probability of making an error using this stopping rule decays exponentially with T, allowing trade-off of error probability and delay.

In Sections V and VI, we show how the error in reconstruction can be estimated from the sequence of recovered solutions. We first present analysis for the Gaussian measurement ensemble in Proposition 4, and then generalize to any sensing matrices with independent and identically distributed (i.i.d.) entries. This enables the decoder to stop once the error is below a required tolerance—even for signals that are not exactly sparse, but in which the energy is largely concentrated in a few components, or for measurements which are corrupted by noise.

Finally, in Section VII we motivate the need for efficient solvers in the sequential setting. We consider the basis pursuit sparse solver and show that rather than resolving the problem from scratch after an additional measurement is received, we could use an augmented linear program that uses the solution at step M to guide its search for the new solution. We show empirically that this approach significantly reduces computational complexity.

During the review process, we learned about a very recent analysis in [15] for the cross-validation setting, using the Johnson–Lindenstrauss lemma. We describe similarities and differences from our work in the discussion in Section V. Our current paper extends our earlier results presented in [16].

II. Brief Overview of Compressed Sensing

As there is no dearth of excellent tutorials on compressed sensing [1], [2], [17], in this section we give only a brief outline mainly to set the stage for the rest of the paper. At the heart of compressed sensing lies the sparse recovery problem,[1] which tries to reconstruct an unknown sparse signal $\mathbf{x}$ from a limited number of measurements $\mathbf{y} = A\mathbf{x}$, where $A \in \mathbb{R}^{M \times N}$, $M \ll N$. Much of excitement in the field stems from the fact that the hard combinatorial problem of searching for sparse solutions in the affine space $\{\mathbf{x} : \mathbf{y} = A\mathbf{x}\}$ under certain suitable conditions can be solved exactly via various tractable methods. The most widely known methods include greedy matching pursuit and its variants [8], and approaches based on convex optimization, using ℓ_1 norms as a proxy for sparsity [7]

$$\min ||\mathbf{x}||_1 \quad \text{subject to} \quad \mathbf{y} = A\mathbf{x}. \quad (2)$$

An early sufficient condition for sparse recovery [18] states that the formulation in (2) recovers the unique sparse solution if A is well-posed and $\mathbf{x}$ is sparse enough, i.e., if $||\mathbf{x}||_0 < (1 + 1/M(A))/2$, where $M(A) = \max_{i \neq j} |\mathbf{a}_i'\mathbf{a}_j|$, and A has columns $\mathbf{a}_i$ normalized to 1. However, this simple condition is very pessimistic. Much tighter conditions are obtained by considering larger subsets of columns of A, e.g., the restricted isometry property (RIP) depends on the maximum and minimum singular values over all $M \times K$ submatrices of A [12]. Namely, a matrix A satisfies the K-RIP with constant δ_K if $(1 - \delta_K)||\mathbf{x}||_2^2 \leq ||A\mathbf{x}||_2^2 \leq (1 + \delta_K)||\mathbf{x}||_2^2$ for every $\mathbf{x}$ which has at most K nonzero entries. While enabling much tighter sufficient conditions for recovery of sparse signals [12], the RIP is very costly (exponential in K) to check for a given matrix.

Results in compressed sensing take advantage of RIP by bringing in the theory of random matrices into the picture. In compressed sensing, we receive random measurements $\mathbf{y} = \Psi\mathbf{s}$, where the unknown signal of interest $\mathbf{s}$ is itself sparse in some basis, i.e., $\mathbf{s} = \Phi\mathbf{x}$, with $\mathbf{x}$ sparse. Hence, the problem reduces to finding sparse solutions satisfying $\mathbf{y} = \Psi\Phi\mathbf{x} = A\mathbf{x}$, where $A = \Psi\Phi$ is a random matrix.

A collection of results have been established that RIP holds for random matrices of certain size from given ensembles: Gaussian, Bernoulli, random Fourier rows [2], [12], [14]. The general conclusion of these results is that the convex ℓ_1 formulation can recover (with high probability) a signal $\mathbf{x} \in \mathbb{R}^N$ with K nonzeros from only $CK\log(N)$ measurements, where C is a constant depending on the random measurement ensemble. This is indeed remarkable—as it only requires a logarithmic dependence of the number of measurements on N.

However, when each additional measurement is very costly there are several problems with these bounds—first, since they are high-probability results independent of $\mathbf{y}$, they tend to be conservative, and also the constants C are typically generous upper-bounds. Second, the number of measurements depends on the number of nonzero components of $\mathbf{x}$ which may not be known *a priori*. Finally, there are successful approaches which we mentioned in Section I for which no such results are available.

In Fig. 1, we illustrate the drawbacks of using upper bounds on the number of measurements. We find the minimum number M of random samples which were needed to recover a sparse signal $\mathbf{x}$ with $N = 100$, and $K = 10$ from random Gaussian and Bernoulli measurements using the ℓ_1-formulation in (2), over 500 random trials. We plot a histogram of these numbers, and we see that they exhibit high variance, and so relying on conditions that guarantee recovery with high probability often means taking many unnecessary samples. This motivates the need for sequential compressed sensing scenario that can adaptively minimize the number of samples for each observed $\mathbf{y}$, which we describe next.

[1]The ground-breaking results [18] predating compressed sensing were in context of sparse signal representation where one seeks to represent a vector $\mathbf{y}$ in an overcomplete dictionary $A \in \mathbb{R}^{M \times N}$, $M \ll N$, with coefficients $\mathbf{x}$, i.e., $\mathbf{y} = A\mathbf{x}$.

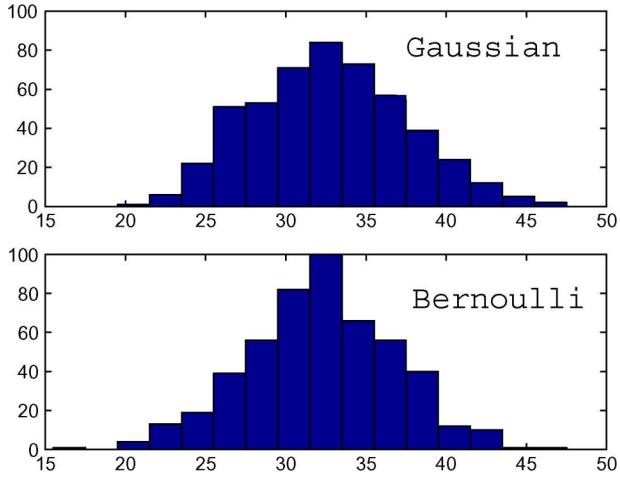

Fig. 1. Histogram of the stopping times distribution for Gaussian and Bernoulli measurement ensembles: $N = 100$, and $K = 10$, and ℓ_1 decoding.

III. Stopping Rule in the Noiseless Continuous Case

We now analyze the sequential CS approach for the case when the measurements vectors $\mathbf{a}_i$ come from a continuous ensemble (e.g., the i.i.d. Gaussian ensemble), having the property that with probability 1 a new vector $\mathbf{a}_{M+1}$ will not be in any lower-dimensional subspace determined by previous vectors $\{\mathbf{a}_i\}_{i=1}^{M}$. Suppose that the underlying sparse signal $\mathbf{x}^* \in \mathbb{R}^N$ has K nonzero components (we denote the number of nonzero entries in $\mathbf{x}$ by $||\mathbf{x}||_0$). We sequentially receive random measurements $y_i = \mathbf{a}_i'\mathbf{x}^*$, where for concreteness $\mathbf{a}_i \sim \mathcal{N}(0, I)$ is a N-vector of i.i.d. Gaussian samples, but the analysis also holds if entries of $\mathbf{a}_i$ are i.i.d. samples of an arbitrary continuous random variable. At step M, we use a sparse solver of our choice to obtain a *feasible*[2] solution $\hat{\mathbf{x}}^M$ using all the received data. Results in compressed sensing [1], [14] indicate that if we use basis pursuit or matching pursuit methods, then after receiving around $M \propto K\log(N)$ measurements we can recover the signal $\mathbf{x}^*$ with high probability. This requires the knowledge of K, which may not be available, and only rough bounds on the scaling constants are known. Our approach is different—we compare the solutions at step M and $M+1$, and if they agree, we declare correct recovery.

Proposition 1: If in the Gaussian (generic continuous) measurement ensemble it holds that $\hat{\mathbf{x}}^{M+1} = \hat{\mathbf{x}}^M$, then $\hat{\mathbf{x}}^M = \mathbf{x}^*$, with probability 1.

Proof: Let $\mathbf{y}_{1:M} \triangleq [y_1, \ldots, y_M]'$, and $A^M \triangleq [\mathbf{a}_1, \ldots, \mathbf{a}_M]'$. Suppose that $\hat{\mathbf{x}}^M \neq \mathbf{x}^*$. We have that $\mathbf{y}_{1:M} = A^M\hat{\mathbf{x}}^M$ and $\mathbf{y}_{1:M} = A^M\mathbf{x}^*$: both $\mathbf{x}^*$ and $\hat{\mathbf{x}}^M$ belong to the $(N-M)$-dimensional affine space $\{\mathbf{x}|\mathbf{y}_{1:M} = A^M\mathbf{x}\}$. The next measurement passes a random hyperplane $y_{M+1} = \mathbf{a}_{M+1}'\mathbf{x}^*$ through $\mathbf{x}^*$ and reduces the dimension of the affine subspace of feasible solutions by 1. In order for $\hat{\mathbf{x}}^M$ to remain feasible at step $M+1$, it must hold that $\mathbf{y}_{M+1} = \mathbf{a}_{M+1}'\hat{\mathbf{x}}^M$. Since we also have $\mathbf{y}_{M+1} = \mathbf{a}_{M+1}'\mathbf{x}^*$, then $\hat{\mathbf{x}}^M$ remains feasible only if $(\hat{\mathbf{x}}^M - \mathbf{x}^*)'\mathbf{a}_{M+1} = 0$, i.e., if $\mathbf{a}_{M+1}$ falls in the $N-1$ dimensional subspace of $\mathbb{R}^N$ corresponding to $Null((\hat{\mathbf{x}}^M - \mathbf{x}^*)')$. As $\mathbf{a}_{M+1}$ is random and independent of $\hat{\mathbf{x}}^M$ and of the previous samples $\mathbf{a}_1, \ldots, \mathbf{a}_M$, the probability that this happens is 0 (event with measure zero). See Fig. 2 for illustration. □

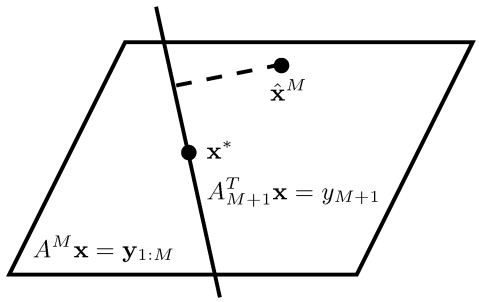

Fig. 2. A new constraint is added: $\mathbf{a}_{M+1}'\mathbf{x} = y_{M+1}$. Probability that this hyperplane passing through $\mathbf{x}^*$ also passes through $\hat{\mathbf{x}}^M$ is zero.

Note that the proof implies that we can simplify the decoder to checking whether $\mathbf{a}_{M+1}'\hat{\mathbf{x}}^M = y_{M+1}$, avoiding the need to solve for $\hat{\mathbf{x}}^{M+1}$ at the last step.[3] Moreover, if using any sparse solver in the continuous ensemble case the solution $\hat{\mathbf{x}}^M$ has fewer than M nonzero entries, then $\hat{\mathbf{x}}^M = \mathbf{x}^*$ with probability 1.

Proposition 2: For a Gaussian (continuous) measurement ensemble, if $||\hat{\mathbf{x}}^M||_0 < M$, then $\hat{\mathbf{x}}^M = \mathbf{x}^*$ with probability 1.[4]

Proof: Denote the support of our unknown sparse vector $\mathbf{x}^*$ by $\mathcal{I}$, i.e., $\mathcal{I} = \{i|x_i^* \neq 0\}$. We next generate a random measurement matrix A^M. Let $A = A^M$ to simplify notation. We receive the corresponding measurements $\mathbf{y} = A\mathbf{x}^*$. Now A is $M \times N$, with $M < N$. The key fact about random matrices with i.i.d. entries from a continuous distribution is that any $M \times M$ submatrix of A is nonsingular with probability 1.[5] We now argue that with probability 1 after receiving $\mathbf{y}$ there will not exist another sparse feasible solution $\hat{\mathbf{x}} \neq \mathbf{x}^*$, i.e., $\hat{\mathbf{x}}$ with fewer than M nonzero entries satisfying $\mathbf{y} = A\hat{\mathbf{x}}$. We consider all possible sparse supports $\mathcal{J} \subset \{1, \ldots, N\}$, with $|\mathcal{J}| < M$, and show that a feasible solution $\hat{\mathbf{x}} \neq \mathbf{x}^*$ can have this support only with probability 0. There are two cases: $\mathcal{I} \subset \mathcal{J}$ and $\mathcal{I} \neq \mathcal{I} \cap \mathcal{J}$.

First suppose $\mathcal{I} \subset \mathcal{J}$, $|\mathcal{J}| < M$, and suppose there exists some feasible $\hat{\mathbf{x}}$ supported on $\mathcal{J}$. Then $\hat{\mathbf{x}} - \mathbf{x}^* \in Null(A)$, and support of $\hat{\mathbf{x}} - \mathbf{x}^*$ is a subset of $\mathcal{J}$; hence, it is smaller than M, but that means that there is a subset of fewer than M columns of A that are linearly dependent, which can only happen with probability zero.

Now consider the case $\mathcal{I} \neq \mathcal{I} \cap \mathcal{J}$. For a fixed $\mathcal{I}$, we consider all such possible sets $\mathcal{J}$, with $|\mathcal{J}| < M$. First fix one such set $\mathcal{J}$. We use the notation $\mathcal{I} \setminus \mathcal{J} = \{i \in I | i \notin J\}$. Note that we have $\mathbf{y} = A\mathbf{x}^* = A_{\mathcal{J}}\mathbf{x}_{\mathcal{J}}^* + A_{\mathcal{I}\setminus\mathcal{J}}\mathbf{x}_{\mathcal{I}\setminus\mathcal{J}}^*$. Let $\tilde{\mathbf{y}} = A_{\mathcal{I}\setminus\mathcal{J}}\mathbf{x}_{\mathcal{I}\setminus\mathcal{J}}^*$. Now since we require $\hat{\mathbf{x}}$ to be feasible, we also need $\mathbf{y} = A\hat{\mathbf{x}} =$

[2]This requirement is essential for the noiseless case (it is relaxed in later sections). For greedy methods such as matching pursuit this means that we allow enough iterations until all the measurements received so far are satisfied perfectly. Noiseless basis pursuit formulations satisfy it by construction.

[3]We thank the anonymous reviewer for this simplification.

[4]Note that a random measurement model is essential: for a fixed matrix A if $2K > M$ then there exist $\mathbf{x}_1$ and $\mathbf{x}_2$ such that $A\mathbf{x}_1 = A\mathbf{x}_2$ and $||\mathbf{x}_i||_0 \leq K$. However, for a fixed $\mathbf{x}^*$ with $||\mathbf{x}^*||_0 < M$ the probability that it will have ambiguous sparse solutions for a random choice of A is zero.

[5]This is easy to see: fix $T \subset \{1, \ldots, N\}$ with $|T| = M$. Then probability that $A_{T_M} \in span(A_{T_1}, \ldots, A_{T_{M-1}})$ is zero, as A_{T_M} is a random vector in $\mathbb{R}^M$ and the remaining columns span a lower dimensional subspace.

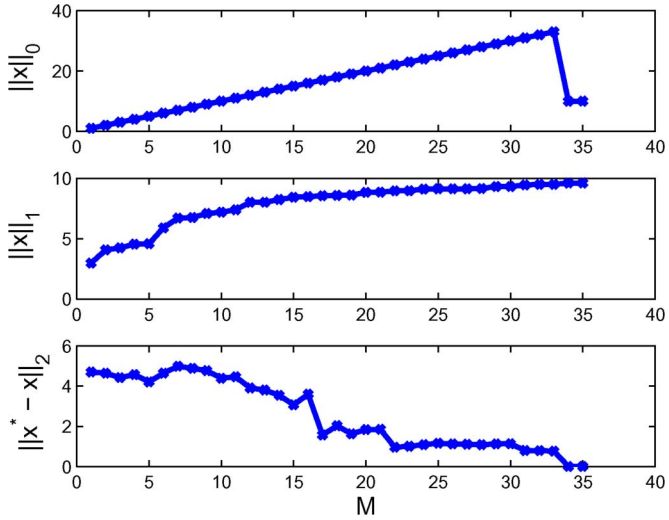

Fig. 3. Gaussian ensemble example: $N = 100$, and $K = 10$. (Top): $\|\hat{\mathbf{x}}^M\|_0$. (Middle): $\|\hat{\mathbf{x}}^M\|_1$. (Bottom): $\|\mathbf{x}^* - \hat{\mathbf{x}}^M\|_2$.

$A_{\mathcal{J}}\hat{\mathbf{x}}_{\mathcal{J}}$ which would imply that $\tilde{\mathbf{y}} = A_{\mathcal{J}}(\hat{\mathbf{x}}_{\mathcal{J}} - \mathbf{x}^*_{\mathcal{J}})$. This means that the vector $\tilde{\mathbf{y}}$ would also have to be in the span of $A_{\mathcal{J}}$. However, $\tilde{\mathbf{y}}$ is a random vector in $\mathbb{R}^M$ (determined by $\mathbf{x}^*$ and $A_{\mathcal{I}\setminus\mathcal{J}}$), and span of $A_{\mathcal{J}}$ is an independent random subspace of dimension strictly less than M. Hence, the event that $\tilde{\mathbf{y}}$ also falls in the span of $A_{\mathcal{J}}$ has measure zero. This means that for a fixed $\mathcal{J}$ a distinct sparse solution can only exist with probability 0. Now the number of possible subsets $\mathcal{J}$ is finite (albeit large), so even when we take all such supports $\mathcal{J}$, a distinct sparse solution supported on $\mathcal{J}$ can only exist with probability 0. Hence, with probability 1 there is only one solution with $\|\mathbf{x}\|_0 < M$, namely $\mathbf{x}^*$. □

This proposition allows to stop making measurements when a feasible solution has less than M nonzero entries—avoiding the need to make the last $(M+1)$st measurement.

Consider an example in Fig. 3 with $N = 100$, and $K = 10$. We keep receiving additional measurements and solving (1) until we reach one-step agreement, $\hat{\mathbf{x}}^M = \hat{\mathbf{x}}^{M+1}$. The top plot shows that $\|\hat{\mathbf{x}}^M\|_0$ increases linearly with M until one step agreement occurs at $M = 35$, at which point it drops to $K = 10$ and a and we recover the correct sparse solution, $\hat{\mathbf{x}}^M = \mathbf{x}^*$. The middle plot shows the monotonic increase in $\|\hat{\mathbf{x}}^M\|_1$ (as the feasible set is shrinking with M). The bottom plot shows the error-norm of the solution, $\|\hat{\mathbf{x}}^M - \mathbf{x}^*\|_2$. On average it tends to go down with more observations, but non-monotonically. After $M = 35$, the error becomes zero. We see that in the ideal conditions of no measurement noise, sparse unknown signals and Gaussian measurement ensembles, the number of measurements can be indeed minimized by a simple stopping rule.

IV. Stopping Rule in the Bernoulli Case

In this section, we study a simple but popular measurement ensemble that is not one of the generic continuous ensembles described in the previous section. Suppose that the measurement vectors $\mathbf{a}_i$ have equiprobable i.i.d. Bernoulli entries ± 1. A difference emerges from the Gaussian case: the probability that all $M \times M$ submatrices of A^M are nonsingular is no longer 0. This makes it possible (with nonzero probability) for $\hat{\mathbf{x}}^{M+1}$ to agree with $\hat{\mathbf{x}}^M$ even though $\hat{\mathbf{x}}^M \neq \mathbf{x}^*$, and for erroneous solutions $\hat{\mathbf{x}}^M$ to have cardinality less than M. We modify the stopping rule to require agreement for several steps—success is declared only when last T solutions all agree. We show in proposition 3 that the probability of error decays exponentially with T. We use the following Lemma from [19]:

Lemma 1 (Tao and Vu): Let $\mathbf{a}$ be an i.i.d. equiprobable Bernoulli vector with $\mathbf{a} \in \{-1, 1\}^N$. Let W be a deterministic d-dimensional subspace of $\mathbb{R}^N$, $0 \leq d < N$. Then $P(\mathbf{a} \in W) \leq 2^{d-N}$.

We are now ready to establish the following claim.

Proposition 3: Consider the Bernoulli measurement case. If $\hat{\mathbf{x}}^M = \hat{\mathbf{x}}^{M+1} = \ldots = \hat{\mathbf{x}}^{M+T}$, then $\hat{\mathbf{x}}^M = \mathbf{x}^*$ with probability greater than or equal to $1 - 2^{-T}$.

Proof: Suppose that $\hat{\mathbf{x}}^M \neq \mathbf{x}^*$. Denote the support of $\mathbf{x}^*$ by $\mathcal{I}$ and the support of $\hat{\mathbf{x}}^M$ by $\mathcal{J}$. At step M we have $A^M\mathbf{x}^* = A^M\hat{\mathbf{x}}^M$. Let $W = \{\mathbf{a} | (\hat{\mathbf{x}}^M - \mathbf{x}^*)'\mathbf{a} = 0\}$, i.e., the nullspace of $(\hat{\mathbf{x}}^M - \mathbf{x}^*)'$. Then W is an $(N-1)$-dimensional subspace of $\mathbb{R}^N$.

Given a new random Bernoulli sample $\mathbf{a}_{M+1}$, the vector $\hat{\mathbf{x}}^M$ can remain feasible at step $M+1$ only if $(\hat{\mathbf{x}}^M - \mathbf{x}^*)'\mathbf{a}_{M+1} = 0$, i.e., if $\mathbf{a}_{M+1}$ falls into W. By Lemma 1, the probability that $\mathbf{a}_{M+1} \in W$ is a most 1/2. The same argument applies to all subsequent samples of $\mathbf{a}_{M+i}$ for $i = 1, \ldots, T$, so the probability of having T-step agreement with an incorrect solution is bounded above by 2^{-T}. □

Note that as in the discussion for the continuous case, we can simply check that $\mathbf{a}'_{M+i}\hat{\mathbf{x}}^M = y_{M+i}$ for $i = 1, \ldots, T$, avoiding the need to solve for $\hat{\mathbf{x}}^{M+T}$.

We now pursue an alternative heuristic analysis, more akin to Proposition 2. For the Bernoulli case, $\|\hat{\mathbf{x}}^M\|_0 < M$ does not imply $\hat{\mathbf{x}}^M = \mathbf{x}^*$. However, we believe that once we obtain enough samples so that $N^2 2^{1-M} \ll 1$ then $\|\hat{\mathbf{x}}^M\|_0 < M$ will imply that $\hat{\mathbf{x}}^M = \mathbf{x}^*$ with high probability. Since the elements of $\mathbf{a}_i$ belong to finite set $\{-1, 1\}$, an $M \times M$ submatrix of A^M can be singular with nonzero probability. Surprisingly, characterizing this probability is a very hard question. It is conjectured [19] that the dominant source of singularity is the event that two columns or two rows are equal or opposite in sign. This leads to the following estimate (here X_M is $M \times M$):[6]

$$P(\det X_M = 0) = (1 + o(1))\, M^2 2^{1-M}. \tag{3}$$

However, the very recent best provable bound on this probability is still rather far: $P(\det X_M = 0) = (((3/4) + o(1))^M)$ [19]. If we assume that the simple estimate based on pairs of columns is accurate, similar analysis shows that the probability that a random ± 1 $M \times N$ matrix with $M \ll N$ having all $M \times M$ submatrices nonsingular is $(1 + o(1))N^2 2^{1-M}$.

V. Near-Sparse Signals

In practical settings, e.g., when taking Fourier and wavelet transforms of smooth signals, we may only have approximate sparseness: a few values are large, and most are very small. In this section, we extend our approach to this case; again, and in contrast to existing work, we do not need to assume a specific

[6]Probability that two columns are equal or opposite in sign is 2^{1-M}, and there are $O(M^2)$ pairs of columns.

near-sparse structure, like power-law decay, but instead provide bounds that hold for any signal.

The exact one-step agreement stopping rule from Section III is vacuous for near-sparse signals, as $\|\mathbf{x}^*\|_0 = N$, and all samples are needed for perfect recovery. We start by considering Gaussian measurements, and show that we can gather information about the current reconstruction error by obtaining a small number of additional measurements, and computing the distance between the current reconstruction and the affine space determined by these new measurements. The reconstruction error is then equal to an unknown constant times this distance

$$\|\mathbf{x}^* - \hat{\mathbf{x}}^M\|_2 = C_T \, d(\hat{\mathbf{x}}^M, H_{M+T}) \tag{4}$$

where $H_{M+T} \triangleq \{x | y_i = \mathbf{a}_i' x, 1 \leq i \leq M+T\}$ is the affine space determined by all $M+T$ measurements, C_T is a random variable that we will bound, and $d(\hat{\mathbf{x}}^M, H_{M+T})$ denotes the distance from $\hat{\mathbf{x}}^M$ to H_{M+T}. We characterize $E[C_T]$ and $Var[C_T]$—this gives us a confidence interval on the reconstruction error using the observed distance $d(\hat{\mathbf{x}}^M, H_{M+T})$. We can now stop taking new measurements once the error falls below a desired tolerance. Note that our analysis does not assume a model of decay, and bounds the reconstruction error by obtaining a small number of additional measurements, and computing the prediction error. In contrast, some related results in CS literature assume a power-law decay of entries of $\mathbf{x}^*$ (upon sorting) and show that with roughly $O(K \log N)$ samples, $\hat{\mathbf{x}}^M$ in (1) will have similar error to that of keeping the K largest entries in $\mathbf{x}^*$ [1].

We now outline the analysis leading to a bound based on (4). Consider Fig. 4. Let $H_M = \{\mathbf{x} : A^M \mathbf{x} = \mathbf{y}_{1:M}\}$ be the subspace of feasible solutions after M measurements. Both $\mathbf{x}^*$ and $\hat{\mathbf{x}}^M$ lie in H_M. The affine space H_{M+T} is contained in H_M. Let $L = N - M$, and θ_T be the angle between the vector $\hat{\mathbf{x}}^M - \mathbf{x}^*$ and the affine space H_{M+T}. Both are contained in the L-dimensional space H_M. Centering around $\mathbf{x}^*$, we see that θ_T is the angle between a fixed vector in R^L and a random $L-T$ dimensional subspace of R^L, and the constant C_T in (4) is equal to $1/\sin(\theta_T)$

$$\|\mathbf{x}^* - \hat{\mathbf{x}}^M\|_2 = \frac{d(\hat{\mathbf{x}}^M, H_{M+T})}{\sin(\theta_T)}. \tag{5}$$

We next analyze the distribution of θ_T and hence of C_T. In distribution, θ_T is equivalent to the angle between a fixed $L-T$ dimensional subspace, say the one spanned by the last $L-T$ coordinates, and an i.i.d. Gaussian vector (whose direction falls uniformly on a unit sphere in $\mathbb{R}^L$). This holds because the distribution of an i.i.d. Gaussian sample does not get changed after applying an arbitrary orthogonal transformation. Let H be the span of the last $L-T$ coordinate vectors, and $\mathbf{h}$ be i.i.d. Gaussian. Then

$$C_T = \frac{1}{\sin(\theta)} = \sqrt{\sum_{i=1}^{L} h_i^2} \Big/ \sqrt{\sum_{i=1}^{T} h_i^2}. \tag{6}$$

Using the properties of χ_L, χ_L^2, and inverse-χ_L^2 distributions [20] and Jensen's inequality, we have an estimate of the mean

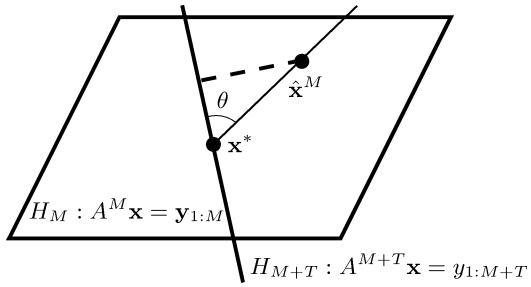

Fig. 4. Geometry of the analysis for near-sparse signals. The unknown reconstruction error is related to $d(\hat{\mathbf{x}}^M, H_{M+T})$ and the angle θ between the line from $\mathbf{x}^*$ to $\hat{\mathbf{x}}^M$ and the affine space H_{M+T} defined by the new measurements.

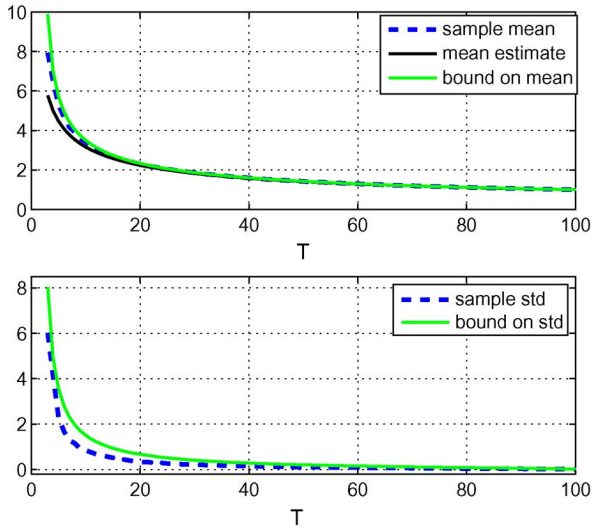

Fig. 5. (Top) sample mean, estimate of the mean, and a bound on the mean of C_T. (Bottom) sample standard deviation, and a bound on the standard deviation of C_T. Sample mean is based on 1000 samples. $L = 100$.

$E[C_T] \approx \sqrt{L/T}$ and an upper bound on both the mean and the variance

$$E[C_T] \leq \sqrt{\frac{L-2}{T-2}}, \tag{7}$$

$$Var[C_T] \leq \frac{L-2}{T-2} - \frac{L}{T}. \tag{8}$$

We describe the analysis in Appendix A. Using these bounds in conjunction with the Chebyshev inequality,[7] ($|a - E[a]| \geq k\sigma_a) \leq 1/k^2$, we have the following result.

Proposition 4: In the Gaussian measurement ensemble we have: $\|\mathbf{x}^* - \hat{\mathbf{x}}^M\|_2 \leq \bar{C}_T^k d(\hat{\mathbf{x}}^M, H_{M+T})$ with probability at least $1 - (1/k^2)$, where $\bar{C}_T^k = \sqrt{(L-2)/(T-2)} + k\sqrt{((L-2)/(T-2)) - (L/T)}$, for any $k > 0$.

In Fig. 5 (top), we plot the mean estimate, and our bound in (7) for C_T and (bottom) the standard deviation bound for $L = 100$ and a range of T. We compare them to sample mean and standard deviation of C_T based on 5000 samples. The figure shows that both bounds provide very good approximation for most of the range of $T > 2$, and also that the standard deviation quickly

[7]To improve upon Chebyshev bounds we could directly characterize the cumulative density function of C_T—either analytically, or by simple Monte Carlo estimates.

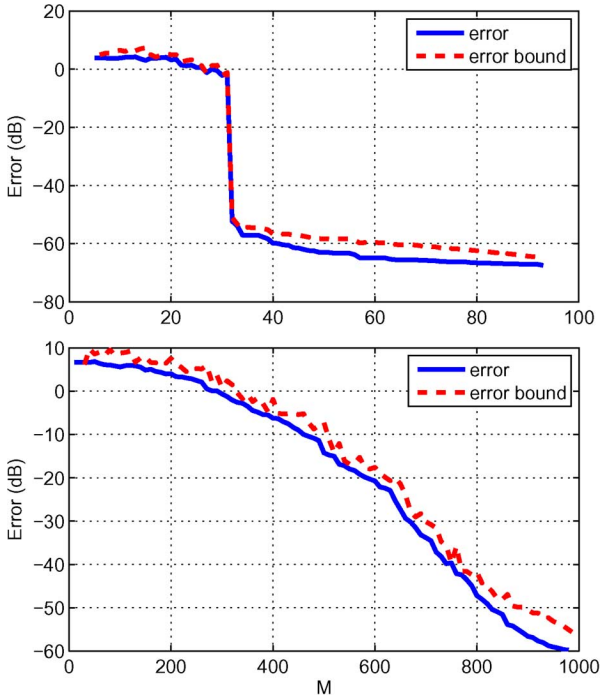

Fig. 6. (Top) Error confidence bounds and actual errors for a sparse signal, $N = 100, T = 5, K = 10$. (Bottom): Error confidence bound and actual errors for a signal with power-law decay, $N = 1000, T = 10$.

falls off with T, giving tight confidence intervals. In Fig. 6, we perform numerical experiments with two example signals, a sparse signal, $N = 100, K = 10, T = 5$ (top) and a near-sparse signal with power-law decay, $N = 1000, T = 10$ (bottom). We use basis pursuit to recover the signals as we obtain progressively more measurements, and we compare our error bounds (via Chebyshev inequality) to the actual errors. We see that the bounds reliably indicate the reconstruction error—after a small delay of T additional measurements. We have used basis pursuit in the experiments, but we could substitute any sparse solver instead, for example we could have also computed error estimates for matching pursuit.

A. *Analysis for More General Ensembles*

To get the bound in (4) we characterized the distribution of $1/\sin(\theta_T)$ and used the properties of the Gaussian measurement ensemble. Analysis of θ_T for general ensembles is challenging. We now consider a simpler analysis which provides useful estimates when $T \ll L$, i.e., the case of main interest for compressed sensing, and when the measurement coefficients a_{ij} are from an i.i.d. zero-mean ensemble. The previous bound for the Gaussian case depended on both M, the number of samples used for the current reconstruction, and T, the number of extra samples. Now, in the following we give estimates and bounds that depend only on T, and in that sense could be weaker for the Gaussian case when M is large; they are however more generally applicable—in particular we no longer require $\hat{\mathbf{x}}^M$ to satisfy the measurements exactly.

Suppose we have a current reconstruction $\hat{\mathbf{x}}$, and suppose $\mathbf{x}^*$ is the (unknown) true signal. We now take T new samples $y_i = \mathbf{a}_i'\mathbf{x}^*$, for $1 \le i \le T$. For each of these samples we compute $\hat{y}_i = \mathbf{a}_i'\hat{\mathbf{x}}^M$ to be the *same* vector $\mathbf{a}_i$ applied to the current reconstruction. Denote the current error vector by $\boldsymbol{\delta} = \hat{\mathbf{x}}^M - \mathbf{x}^*$, and compute $z_i = \hat{y}_i - y_i$, the deviations from the actual measurements. Then

$$z_i = \mathbf{a}_i'\boldsymbol{\delta}, \quad 1 \le i \le T. \tag{9}$$

The new measurements $\mathbf{a}_i$ are independent of $\hat{\mathbf{x}}$ and of $\mathbf{x}^*$, hence of $\boldsymbol{\delta}$. The z_i's are i.i.d. from some (unknown) distribution, which has zero mean and variance $\|\boldsymbol{\delta}\|_2^2 Var(a_{ij})$. We can estimate $\|\boldsymbol{\delta}\|_2^2$ by estimating the variance of the z_i's from the T samples. The quality of the estimate will depend on the exact distribution of $\mathbf{a}_{ij}$.

Consider the case where $\mathbf{a}_i$ are i.i.d. Gaussian. Then z_i is Gaussian as well. For simplicity suppose that $Var(a_{ij}) = 1$, then the distribution of z_i is i.i.d. Gaussian with zero-mean and variance $\|\boldsymbol{\delta}\|_2^2$. Let $Z_T = \sum_{i=M+1}^{M+T} z_i^2$. Then $\tilde{Z}_T \triangleq Z_T/\|\delta\|^2 \sim \chi_T^2$, i.e., χ^2 random variable with T degrees of freedom. Now to obtain a confidence interval for $\|\boldsymbol{\delta}\|_2^2$ we use the cumulative χ_T^2 distribution. We pick a confidence level $1 - \alpha$ (for some small $\alpha > 0$), and we use the χ_T^2 cumulative distribution to find the largest z^* such that $p(\tilde{Z}_T \le z^*) \le \alpha$.[8]

During the review process, a related analysis in [15] was brought to our attention: the paper considers compressed sensing in a cross-validation scenario, and it proposes to estimate the errors in the reconstruction from a few additional (cross-validation) measurements. The paper cleverly uses the Johnson–Lindenstrauss (JL) lemma to find out how many random measurements are needed for predicting the error to a desired accuracy. For Gaussian measurements ensembles our χ^2-based analysis can be seen as a special case (where all the constants are computed explicitly since we use the exact sampling distribution of Z_T), but JL lemma also generalizes to other ensembles satisfying certain requirements on the decay of the tails [15], [21].

To compare our analysis in (5), based on C_T, to the one in (9) we note that the latter simply estimates the error $\|\boldsymbol{\delta}\|$ as $\|(1/\sqrt{T})\tilde{A}\boldsymbol{\delta}\|$, where $\tilde{A}$ are the new measurements.[9] Now unlike the analysis in (9), in (5) we require that the solution at step M is feasible (matches all the measurements) and instead we compute the error of projecting $\boldsymbol{\delta}$ onto the null-space of A and adjust it by the expected value of $1/\sin(\theta)$, i.e., we estimate $\|\boldsymbol{\delta}\|$ as $\sqrt{L/T}\|A'(AA')^{-1}A\boldsymbol{\delta}\|$, where A includes all $M + T$ measurements. To compare the quality of the two estimates we conducted a simulation with $N = 250$ and $T = 25$, and computed the estimates for random unit-norm vectors $\boldsymbol{\delta}$. We plot the histograms for $M = 0$ and $M = 200$ over 5000 trials in Fig. 7. In the first case with $M = 0$, we see that both estimates have about the same accuracy (similar error distributions); however, as M becomes appreciable the approach in (5) becomes more accurate.

[8]We have that $\sigma_z^2 = Z_T/z^*$ gives the smallest value of σ_z^2 such that probability of observing Z_T is at least α. That is to say, the bound $\|\boldsymbol{\delta}\|^2 < Z_T/z^*$ will hold for at least $1 - \alpha$ fraction of realizations of Z_T.

[9]This is essentially the same estimate as the one based on JL lemma in [15], as the expected value of χ_T^2 is T, hence $E[Z_T] = T\|\boldsymbol{\delta}\|_2^2$.

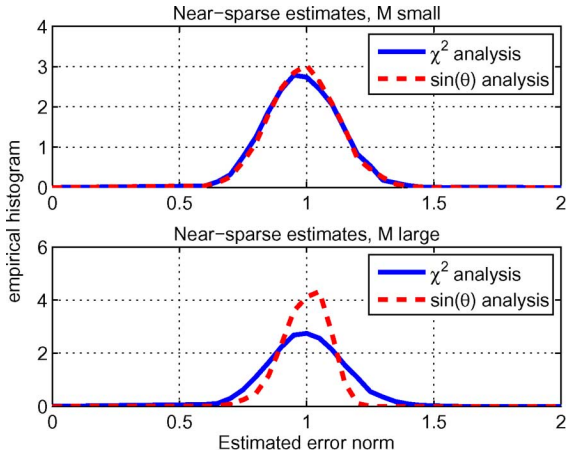

Fig. 7. Comparison of χ^2 and $\sin(\theta)$ analysis. Given a unit-norm vector $\boldsymbol{\delta}$, we obtain T additional measurements, and compute our two estimates of $\|\boldsymbol{\delta}\|$. We plot the histogram of the estimates over 5000 trials with $N = 250$, $T = 25$, and (a) $M = 0$, (b) $M = 200$.

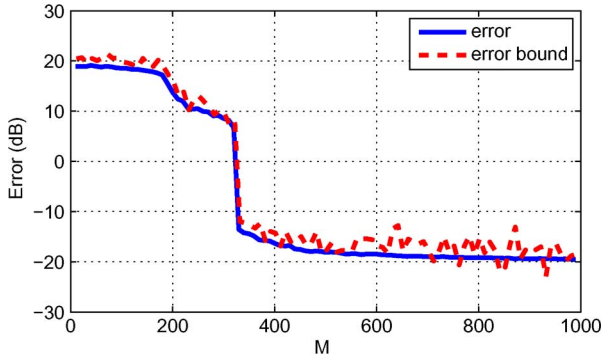

Fig. 8. Error estimate in the noisy case: true error and a 90% confidence bound (dB scale): $N = 1000$, $T = 10$, $K = 100$.

VI. Noisy Case

Next we consider the sequential version of the noisy measurement setting, where the observations are corrupted by additive uncorrelated i.i.d. Gaussian noise with variance σ_n^2

$$y_i = \mathbf{a}_i'\mathbf{x} + n_i, \quad i \in \{1, \ldots, M\}. \tag{10}$$

To solve this problem, one can adapt a variety of sparse solvers which allow inexact solutions $\hat{\mathbf{x}}^M$ in the sequential setting—for example matching pursuit methods with a fixed number of steps, or the noisy versions of basis pursuit. All of these methods have a tradeoff between sparsity of the desired solution and the accuracy in representing the measurements. In the case of basis pursuit denoising, a regularization parameter λ balances these two costs

$$\hat{\mathbf{x}}^M = \arg\min \frac{1}{2}\|\mathbf{y}_{1:M} - A^M\mathbf{x}\|_2^2 + \lambda_M\|\mathbf{x}\|_1. \tag{11}$$

For greedy sparse solvers such as matching pursuit and its variants the tradeoff is controlled directly by deciding how many columns of A to use to represent $\mathbf{y}$. We are interested in a stopping rule which tells us that $\hat{\mathbf{x}}$ is reasonably close to $\mathbf{x}^*$ for any sparse solver and for any user defined choice of the tradeoff between sparsity and measurement likelihood. We do not discuss the question of selecting a choice for the tradeoff—we refer the readers to [22], [23], and also to [15] for a discussion of how this can be done in a cross-validation setting. Now, due to the presence of noise, exact agreement will not occur no matter how many samples are taken. We consider a stopping rule similar to the one in Section V. In principle, the analysis in (4) can be extended to the noisy case, but we instead follow the simplified analysis in Section V-A.

We establish that the reconstruction error can be bounded with high probability by obtaining a small number of additional samples, and seeing how far the measurements deviate from $\hat{y}_i = \mathbf{a}_i'\hat{\mathbf{x}}^M$. With such a bound one can stop receiving additional measurements once the change in the solution reaches levels that can be explained due to noise. The deviations z_i now include contribution due to noise

$$z_i = \hat{y}_i - y_i = \mathbf{a}_i'(\hat{\mathbf{x}}^M - \mathbf{x}^*) - n_i. \tag{12}$$

Let $Z_T = \sum z_i^2$. Consider the Gaussian measurement ensemble. Then $z_i = \mathbf{a}_i'\boldsymbol{\delta} + n_i$, and $\tilde{Z}_T \triangleq Z_T/(\|\delta\|^2 + \sigma_n^2) \sim \chi_T^2$. The distribution of z_i is Gaussian with mean zero and variance $\|\boldsymbol{\delta}\|_2^2 + \sigma_n^2$. Now following a similar analysis as in previous section we can obtain an estimate of $\|\boldsymbol{\delta}\|_2^2 + \sigma_n^2$ from a sample of Z_T, and subtracting σ_n^2 we get an estimate of $\|\boldsymbol{\delta}\|_2^2$.

We show an example in Fig. 8 where the true error appears along with a 90% confidence bound. We have $N = 1000$, $K = 100$, $T = 10$ and $\sigma_n = 0.01$. We use basis pursuit denoising (12) as our choice for sparse solver, and we set $\lambda_M \propto \sqrt{M\log(N)}$ motivated by the universal rule for wavelet denoising [22] to account for noise added with additional measurements. The bound clearly shows where the sparse signal has been recovered up to the noise floor (the signal is sparse with $K = 100$ nonzero elements).

VII. Efficient Sequential Solution

The main motivation for the sequential approach is to reduce the number of measurements to as few as possible. Yet, we would also like to keep the computational complexity of the sequential approach low. We focus on the ℓ_1-based formulations here, and show that there is some potential of using "memory" in the sequential setting for reducing the computational complexity. For the static setting there exists a great variety of approaches to solve both noiseless and noisy basis pursuit (i.e., basis pursuit denoising) in various forms, e.g., [23]–[25]. However, instead of resolving the linear program (1) after each new sample, we would like to use the solution to the previous problem to guide the current problem. It is known that interior point methods are not well-suited to take advantage of such "warm-starts" [23]. Some methods are able to use warm-starts in the context of following the solution path in (11) as a function of λ [23], [26], [27]. In that context the solution path $\hat{\mathbf{x}}(\lambda)$ is continuous (nearby values of λ give nearby solutions) enabling warm-starts. However, once a new measurement $\mathbf{a}_i$ is received, this in general makes the previous solution infeasible, and can dramatically change the optimal solution, making warm-starts more challenging.[10]

We now investigate a linear programming approach for warm-starts using the simplex method to accomplish this in

[10]In related work, [28] proposed to use Row-action methods for compressed sensing, which rely on a quadratic programming formulation equivalent to (1) and can take advantage of sequential measurements.

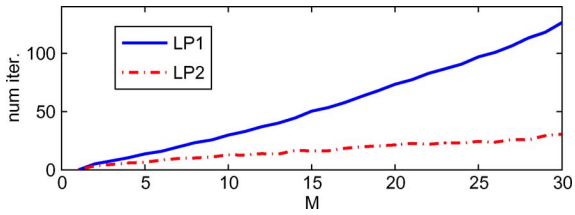

Fig. 9. Comparison of the number of simplex iterations when solving (1) from scratch (LP1) and using the solution at step $M-1$ (LP2). We plot the average number of iterations vs. M, over 100 trials.

the noiseless case (a similar strategy can be used with the Dantzig decoder [1] for the noisy case). We cannot use the solution $\hat{\mathbf{x}}^M$ directly as a starting point for the new problem at step $M+1$, because in general it will not be feasible. In the Gaussian measurement case, unless $\hat{\mathbf{x}}^M = \mathbf{x}^*$, the new constraint $\mathbf{a}'_{M+1}\hat{\mathbf{x}}^M = y_{M+1}$ will be violated. One way to handle this is through a dual formulation,[11] but we instead use an augmented primal formulation [29].

First, to model (1) as a linear program we use the standard trick: define $x_i^+ = \max(x_i, 0)$, $x_i^- = \max(-x_i, 0)$, and $\mathbf{x} = \mathbf{x}^+ - \mathbf{x}^-$. This gives a linear program in standard form:

$$\begin{aligned} &\min \mathbf{1}'\mathbf{x}^+ + \mathbf{1}'\mathbf{x}^- \\ &\mathbf{y}_{1:M} = [A^M \quad -A^M]\begin{bmatrix}\mathbf{x}^+\\ \mathbf{x}^-\end{bmatrix}, \quad \text{and} \quad \mathbf{x}^+, \mathbf{x}^- \geq 0. \end{aligned} \tag{13}$$

Next we need to add an extra constraint $y_{M+1} = \mathbf{a}'_{M+1}\mathbf{x}^+ - \mathbf{a}'_{M+1}\mathbf{x}^-$. Suppose that $\mathbf{a}'_{M+1}\hat{\mathbf{x}}^M > y_{M+1}$. We add an extra slack variable z to the linear program, and a high positive cost Q on z. This gives the following linear program:

$$\begin{aligned} &\min \mathbf{1}'\mathbf{x}^+ + \mathbf{1}'\mathbf{x}^- + Qz \\ &\mathbf{y}_{1:M} = [A^M \quad -A^M]\begin{bmatrix}\mathbf{x}^+\\ \mathbf{x}^-\end{bmatrix}, \text{ and } \mathbf{x}^+, \mathbf{x}^- \geq 0 \\ &y_{M+1} = \mathbf{a}'_{M+1}\mathbf{x}^+ - \mathbf{a}'_{M+1}\mathbf{x}^- - z, \text{ and } z \geq 0. \end{aligned} \tag{14}$$

Now using $\hat{\mathbf{x}}^M$ and $z = \mathbf{a}'_{M+1}(\hat{\mathbf{x}}^M)^+ - \mathbf{a}'_{M+1}(\hat{\mathbf{x}}^M)^- - y_{M+1}$ yields a basic feasible solution to this augmented problem. By selecting Q large enough,[12] z will be removed from the optimal basis (i.e., z is set to 0), and the solutions to this problem and the $(M+1)$th sequential problem are the same.

We test the approach on an example with $N = 200$, $K = 10$, and 100 trials. In Fig. 9, we plot the number of iterations of the simplex method required to solve the problem (1) at step M from scratch (LP1) and using the formulation in (14) (LP2). To solve (13), we first have to find a basic feasible solution, BFS, (phase 1) and then move from it to the optimal BFS. An important advantage of (14) is that we start right away with a BFS, so phase 1 is not required. The figure illustrates that for large M the approach LP2 is significantly faster.

We note that recently a very appealing approach for sequential solution in the noisy setting has been proposed based on the homotopy continuation idea [30], [31], where a homotopy (a continuous transition) is constructed from the problem at step M to the problem at step $M+1$ and the piecewise-smooth path is followed. The efficiency of the approach depends on the number of break-points in this piecewise-smooth path, but the simulations results in the papers are very promising. We also note that [30] proposes an approach to select the trade-off in the noisy case, using cross-validation ideas.

[11]If at step M the optimal dual solution is $\mathbf{p}$, then a feasible solution at step $M+1$ is $[\mathbf{p}; 0]$. However, it may not be a basic feasible solution.

[12]E.g. the big-M approach [29] suggests treating Q as an undetermined value, and assumes that Q dominates when compared to any other value.

VIII. Conclusion and Discussion

This paper presents a formulation for compressed sensing in which the decoder receives samples sequentially, and can perform computations in between samples. We showed how the decoder can estimate the error in the current reconstruction; this enables stopping once the error is within a required tolerance. Our results hold for any decoding algorithm, since they only depend on the distribution of the measurement vectors. This enables "run-time" performance guarantees in situations where *a priori* guarantees may not be available, e.g., if the sparsity level of the signal is not known, or for recovery methods for which such guarantees have not been established.

We have studied a number of scenarios including noiseless, noisy, sparse and near sparse, and involving Gaussian and Bernoulli measurements, and demonstrated that the sequential approach is practical, flexible and has wide applicability. A very interesting problem is to both extend the results to other measurement ensembles, e.g., for sparse ensembles, and moreover, to go beyond results for particular ensembles and develop a general theory of sequential compressed sensing. Furthermore, in many important applications the sparse signal of interest may also be evolving with time during the measurement process. Sequential CS with a notion of "time of a measurement" is a natural candidate setting in which to explore this important extension to the CS literature.

We also remark that there is a closely related problem of recovering low-rank matrices from a small number of random measurements [32], [33], where instead of searching for sparse signals one looks for matrices with low-rank. This problem admits a convex "nuclear-norm" relaxation (much akin to ℓ_1 relaxation of sparsity). Some of our results can be directly extended to this setting—for example if in the Gaussian measurement case with no noise there is one-step agreement, then the recovered low-rank matrix is the true low-rank solution with probability one.

Finally, we comment on an important question [6], [34] of whether it is possible to do better than simply using random measurements—using, e.g., experiment design or active learning techniques. In [6], the authors propose to find a multivariate Gaussian approximation to the posterior $p(\mathbf{x}|\mathbf{y})$ where $p(\mathbf{y}|\mathbf{x}) \propto \exp((1/\sigma^2)\|\mathbf{y} - A\mathbf{x}\|^2)$, and $p(\mathbf{x}) \propto \exp(-\lambda\|\mathbf{x}\|_1)$. Note that MAP estimation in this model $\hat{\mathbf{x}} = \arg\max_{\mathbf{x}} p(\mathbf{x}|\mathbf{y})$ is equivalent to the formulation in (11), but does not provide uncertainties. Using the Bayesian formalism it is possible to do experiment design, i.e., to select the next measurement to maximally reduce the expected uncertainty. This is a very exciting development, and although much more complex than the sequential approach presented here, may reduce the number of required samples even further.

APPENDIX A
DERIVATION OF THE DISTRIBUTION FOR $1/\sin\theta$

Consider $E[\sin^2(\theta)] = E[(\sum_{i=1}^{T} h_i^2)/\|\mathbf{h}\|_2^2]$. Since $\sum_i E[h_i^2/\|\mathbf{h}\|_2^2] = 1$, and each h_i is i.i.d., we have $E[h_i^2/\|\mathbf{h}\|_2^2] = 1/L$. In fact, $E[h_i^2/\|\mathbf{h}\|_2^2]$ follows a Dirichlet distribution. Therefore, $E[\sin^2(\theta)] = T/L$.

Using Jensen's inequality with the convex function $\sqrt{1/x}$, $x > 0$, we have $E[1/\sin(\theta)] \geq \sqrt{L/T}$.

Now, $E[1/\sin^2(\theta)] = (L-2)/(T-2)$ (for $T > 2$). This is true because $E[1/\sin^2(\theta)] = E(\sum_{i=1}^{L} h_i^2)/(\sum_{i=1}^{T} h_i^2) = 1 + E(\sum_{i=T+1}^{L} h_i^2/\sum_{i=1}^{T} h_i^2) = 1 + (L-T)(1/(T-2))$. The second term is a product of a χ^2 random variable with $(L-T)$ degrees of freedom and an independent inverse-χ^2 distribution with T degrees of freedom: $E[\sum_{i=T+1}^{L} h_i^2] = L-T$, and $E[1/(\sum_{i=1}^{T} h_i^2)] = 1/(T-2)$, see [20]. Now $1 + (L-T)/(T-2) = (L-2)/(T-2)$.

Finally, using Jensen's inequality with the concave function $\sqrt{x}$, $E[1/\sin(\theta)] \leq \sqrt{(T-2)/(L-2)}$.

REFERENCES

[1] E. J. Candes, "Compressive sampling," in *Proc. Int. Congr. Math.*, Madrid, Spain, 2006.

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

[3] M. Lustig, D. L. Donoho, and J. M. Pauly, "Sparse MRI: The application of compressed sensing for rapid MR imaging," *Magn. Resonance Med.*, vol. 58, no. 6, pp. 1182–1195, Dec. 2007.

[4] M. F. Duarte, M. A. Davenport, D. Takhar, J. N. Laska, T. Sun, K. F. Kelly, and R. Baraniuk, "Single pixel imaging via compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 83–91, Mar. 2008.

[5] W. Bajwa, J. Haupt, A. Sayeed, and R. Nowak, "Compressive wireless sensing," in *Proc. Int. Conf. Inf. Process. Sensor Netw. (IPSN)*, Apr. 2006.

[6] F. Steinke, M. Seeger, and K. Tsuda, "Experimental design for efficient identification of gene regulatory networks using sparse Bayesian models," *BMC Syst. Biol.*, vol. 1, no. 51, 2007.

[7] 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.

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

[9] R. Chartrand and W. Yin, "Iteratively reweighted algorithms for compressive sensing," in *Proc. ICASSP*, 2008, pp. 3869–3872.

[10] M. Cetin, D. M. Malioutov, and A. S. Willsky, "A variational technique for source localization based on a sparse signal reconstruction perspective," in *Proc. ICASSP*, 2002, pp. 2965–2968.

[11] E. J. Candes, M. Wakin, and S. Boyd, "Enhancing sparsity by reweighted l1 minimization," California Inst. of Technol., Pasadena, 2007, Tech. Rep..

[12] E. 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.

[13] D. Wipf and B. D. Rao, "Sparse Bayesian learning for basis selection," *IEEE Trans. Signal Process.*, vol. 52, no. 8, pp. 2153–2164, 2004.

[14] M. Rudelson and R. Vershynin, "Sparse reconstruction by convex relaxation: Fourier and Gaussian measurements," in *Proc. CISS'06*, 2006.

[15] R. Ward, "Compressed sensing with cross validation," *IEEE Trans. Inf. Theory*, vol. 55, no. 12, pp. 5773–5782, Dec. 2009.

[16] D. M. Malioutov, S. R. Sanghavi, and A. S. Willsky, "Compressed sensing with sequential observations," in *Proc. ICASSP*, 2008, pp. 3357–3760.

[17] R. Baraniuk, "Compressive sensing," *IEEE Signal Process. Mag.*, vol. 24, no. 4, pp. 118–121, Jul. 2007.

[18] 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.

[19] T. Tao and V. Vu, "On the singularity probability of random Bernoulli matrices," *J. Amer. Math. Soc.*, vol. 20, pp. 603–628, 2007.

[20] S. Kotz, N. Balakrishnan, and N. L. Johnson, *Continuous Multivariate Distributions*. New York: Wiley, 2000.

[21] S. Dasgupta and A. Gupta, "An elementary proof of a theorem of Johnson and Lindenstrauss," *Random Structures and Algorithms*, vol. 22, no. 1, pp. 60–65, 2003.

[22] I. M. Johnstone and D. L. Donoho, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, no. 3, pp. 425–455, 1994.

[23] M. Figueiredo, R. Nowak, and S. Wright, "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," *IEEE J. Sel. Topics Signal Process.*, vol. 1, no. 4, pp. 586–598, Aug. 2007.

[24] S. J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, "A method for large-scale l1-regularized least squares," *IEEE J. Sel. Topics Signal Process.*, vol. 1, no. 6, pp. 606–617, Dec. 2007.

[25] I. Daubechies, M. De Friese, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Comm. in Pure Appl. Math.*, vol. 57, pp. 1413–1457, 2004.

[26] M. R. Osborne, B. Presnell, and B. A. Turlach, "A new approach to variable selection in least squares problems," *IMA J. Numer. Anal.*, vol. 20, no. 3, pp. 389–403, 2000.

[27] D. M. Malioutov, M. Cetin, and A. S. Willsky, "Homotopy continuation for sparse signal representation," in *Proc. ICASSP*, 2005, pp. 733–736.

[28] S. Sra and J. A. Tropp, "Row-action methods for compressed sensing," in *Proc. ICASSP*, 2006, vol. 3, pp. 868–871.

[29] D. Bertsimas and J. N. Tsitsiklis, *Introduction to Linear Optimization*. Nashua, NH: Athena Scientific, 1997.

[30] P. J. Garrigues and L. El Ghaoui, "An homotopy algorithm for the Lasso with online observations," in *Neural Information Processing Systems (NIPS)*, Dec. 2008.

[31] M. S. Asif and J. Romberg, "Streaming measurements in compressive sensing: L1 filtering," in *Proc. 42nd Asilomar Conf. Signals, Syst., Comput.*, Oct. 2008.

[32] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum rank solutions to linear matrix equations via nuclear norm minimization," *SIAM Rev.*, 2007, accepted for publication.

[33] R. Keshavan, A. Montanari, and S. Oh, "Learning low rank matrices from O(n) entries," in *Proc. Allerton Conf.*, Oct. 2008.

[34] Y. Weiss, H. S. Chang, and W. T. Freeman, "Learning compressed sensing," in *Proc. Allerton Conf.*, Sep. 2007.

Dmitry M. Malioutov received the B.S. degree in electrical and computer engineering from Northeastern University, Boston, MA, in 2001, and the M.S. and Ph.D. degrees in electrical engineering and computer science from the Massachusetts Institute of Technology (MIT), Cambridge, in 2003 and 2008, respectively.

After a postdoctoral position in the Machine Learning and Perception Group, Microsoft Research, Cambridge, UK, he is currently a Researcher in the Algorithmic Trading Group at DRW, Chicago, IL. His research interests include statistical signal and image processing, machine learning, and convex optimization with emphasis on graphical models, message passing algorithms, and sparse signal representation.

Sujay R. Sanghavi received the Ph.D. degree in electrical engineering, with a focus on communication networks from the University of Illinois, Urbana-Champaign, in 2006.

From 2006 to 2008, he was a Postdoctoral Researcher at the Massachusetts Institute of Technology, Cambridge, with a focus on Statistical Learning and Inference. He joined Purdue University, West Lafayette, IN, in 2008 and the University of Texas, Austin, in 2009, both as an Assistant Professor in electrical and computer engineering. His research interests are the design of algorithms for large-scale data networks and high-dimensional statistical problems.

Alan S. Willsky (S'70–M'73–SM'82–F'86) joined the Massachusetts Institute of Technology, Cambridge, in 1973 and is the Edwin Sibley Webster Professor of Electrical Engineering and Director of the Laboratory for Information and Decision Systems. He was a founder of Alphatech, Inc., and Chief Scientific Consultant, a role in which he continues at BAE Systems Advanced Information Technologies. From 1998 to 2002, he served on the U.S. Air Force Scientific Advisory Board. He has delivered numerous keynote addresses and is coauthor of the text *Signals and Systems* (Prentice-Hall, 1996). His research interests are in the development and application of advanced methods of estimation, machine learning, and statistical signal and image processing.

Prof. Willsky received several awards including the 1975 American Automatic Control Council Donald P. Eckman Award, the 1979 ASCE Alfred Noble Prize, the 1980 IEEE Browder J. Thompson Memorial Award, the IEEE Control Systems Society Distinguished Member Award in 1988, the 2004 IEEE Donald G. Fink Prize Paper Award, and Doctorat Honoris Causa from Université de Rennes in 2005. Dr. Willsky and his students, colleagues, and postdoctoral associates have also received a variety of Best Paper Awards at various conferences and for papers in journals, including the 2001 IEEE Conference on Computer Vision and Pattern Recognition, the 2003 Spring Meeting of the American Geophysical Union, the 2004 Neural Information Processing Symposium, Fusion 2005, and the 2008 award from the journal *Signal Processing* for the outstanding paper in the year 2007.