Democracy in Action: Quantization, Saturation, and Compressive Sensing

Jason N. Laska, Petros T. Boufounos, Mark A. Davenport, and Richard G. Baraniuk

Abstract—**Recent theoretical developments in the area of *compressive sensing* (CS) have the potential to significantly extend the capabilities of digital data acquisition systems such as analog-to-digital converters and digital imagers in certain applications. A key hallmark of CS is that it enables sub-Nyquist sampling for signals, images, and other data. In this paper, we explore and exploit another heretofore relatively unexplored hallmark, the fact that certain CS measurement systems are *democractic*, which means that each measurement carries roughly the same amount of information about the signal being acquired. Using the democracy property, we re-think how to quantize the compressive measurements in practical CS systems. If we were to apply the conventional wisdom gained from conventional Shannon-Nyquist uniform sampling, then we would scale down the analog signal amplitude (and therefore increase the quantization error) to avoid the gross saturation errors that occur when the signal amplitude exceeds the quantizer's dynamic range. In stark contrast, we demonstrate that a CS system achieves the best performance when it operates at a significantly nonzero saturation rate. We develop two methods to recover signals from saturated CS measurements. The first directly exploits the democracy property by simply discarding the saturated measurements. The second integrates saturated measurements as constraints into standard linear programming and greedy recovery techniques. Finally, we develop a simple automatic gain control system that uses the saturation rate to optimize the input gain.**

Index Terms—**compressive sensing, quantization, saturation, inequality constraint, consistent reconstruction**

I. Introduction

Analog-to-digital converters (ADCs) are an essential component in digital sensing and communications systems. They interface the analog physical world, where many signals originate, with the digital world, where they can be efficiently processed and analyzed. As digital processors have become smaller and more powerful, their increased capabilities have inspired applications that require the sampling of ever-higher bandwidth signals. This demand has placed a growing burden on ADCs [1]. As ADC sampling rates push higher, they move toward a physical barrier, beyond which their design becomes increasingly difficult and costly [2].

Fortunately, recent theoretical developments in the area of *compressive sensing* (CS) have the potential to significantly extend the capabilities of current ADCs to keep pace with demand [3, 4]. CS provides a framework for sampling signals at a rate proportional to their *information content* rather than their bandwidth, as in Shannon-Nyquist systems. In CS, the information content of a signal is quantified as the number of nonzero coefficients in a known transform basis over a fixed time interval [5]. Signals that have few nonzero coefficients are called *sparse* signals. More generally, signals with coefficient magnitudes that decay rapidly are called *compressible*, because they can be well-approximated by sparse signals. By exploiting sparse and compressible signal models, CS provides a methodology for simultaneously acquiring and compressing signals. This leads to lower sampling rates and thus simplifies hardware designs. The CS measurements can be used to reconstruct the signal or can be directly processed to extract other kinds of information.

The CS framework employs non-adaptive, linear measurement systems and non-linear reconstruction algorithms. In most cases, CS systems exploit a degree of *randomness* in order to provide theoretical guarantees on the performance of the system. Such systems exhibit additional desirable properties beyond lower sampling rates. In particular, the measurements are *democratic*, meaning that each measurement contributes an equal amount of information to the compressed representation. This is in contrast to both conventional sampling systems and conventional compression algorithms, where the removal of some samples or bits can lead to high distortion, while the removal of others will have negligible effect. One of the contributions of this paper is to formally define and quantify the democracy of the CS measurement process. Although the term was loosely used before, this is the first time, to our knowledge, that such a formulation exists.

Several CS-inspired hardware architectures for acquiring signals, images, and videos have been proposed, analyzed, and in some cases implemented [6–15]. The common element in each of these acquisition systems is that the measurements are ultimately *quantized*, i.e., mapped from real-values to a set of countable values, before they are stored or transmitted. In this paper, we focus on this quantization step.

While the effect of quantization on the CS framework has been previously explored [16–19], prior work has ignored *saturation*. Saturation occurs when measurement values exceed the *saturation level*, i.e., the dynamic range of a quantizer. These measurements take on the value of the saturation level. All practical quantizers have a finite dynamic range for one of two reasons, or both: *(i)* physical limitations allow only a finite range of voltages to be accurately converted to bits and, *(ii)* only a finite number of bits are available to represent each

J. N. Laska, M. A. Davenport, and R. G. Baraniuk with the Department of Electrical and Computer Engineering, Rice University, Houston, TX, 30332 USA e-mail: laska@rice.edu, md@rice.edu, richb@rice.edu.

P. T. Boufounos is with Mitsubishi Electric Research Laboratories (MERL) e-mail: petrosb@merl.com.

This work was supported by the grants NSF CCF-0431150, CCF-0728867, CNS-0435425, and CNS-0520280, DARPA/ONR N66001-08-1-2065, ONR N00014-07-1-0936, N00014-08-1-1067, N00014-08-1-1112, and N00014-08-1-1066, AFOSR FA9550-07-1-0301, ARO MURI W311NF-07-1-0185, ARO MURI W911NF-09-1-0383, and the Texas Instruments Leadership University Program.

value. Quantization with saturation is commonly referred to as *finite-range* quantization.

The challenge in dealing with the errors imposed by finite-range quantization is that, in the absence of an *a priori* upper bound on the measurements, saturation errors are potentially unbounded. Most CS recovery algorithms only provide guarantees for noise that is either bounded or bounded with high probability (for example, Gaussian noise) [20], with the exception of [21, 22] which consider sparse or impulsive noise models, and [23, 24] which consider unbounded noise from particular distributions.

The intuitive approach to dealing with finite-range quantization is to scale the measurements so that saturation never or rarely occurs. However, rescaling the signal comes at a cost. The signal-to-noise ratio (SNR) is decreased on the measurements that do not saturate, and so the SNR of the acquired signal will decrease as well.

In this paper, we present two new approaches for mitigating unbounded quantization errors caused by saturation in CS systems. The first approach simply discards saturated measurements and performs signal reconstruction without them. The second approach is based on a new CS recovery algorithm that treats saturated measurements differently from unsaturated ones. This is achieved by employing a magnitude constraint on the indices of the saturated measurements while maintaining the conventional regularization constraint on the indices of the other measurements. We analyze both approaches and show that both can recover sparse and compressible signals with guarantees similar to those for standard CS recovery algorithms.

Our proposed methods exploit the democratic nature of CS measurements. Because each measurement contributes equally to the compressed representation, we can remove some of them and still maintain a sufficient amount of information about the signal to enable recovery. We prove this fact, which allows us to provide a simple analysis of the two approaches described above and yields concrete bounds on how many measurements are sufficient to ensure that we are robust to the saturation of some specified number of measurements.

When characterizing our methods, we find that in order to maximize the acquisition SNR, the optimal strategy is to allow the quantizer to saturate at some nonzero rate. This is due to the inverse relationship between quantization error and saturation rate: as the saturation rate increases, the distortion of remaining measurements decreases. Our experimental results show that on average, the optimal SNR is achieved at nonzero saturation rates. This demonstrates that just as CS challenges the conventional wisdom of how to sample a signal, it also challenges the conventional wisdom of avoiding saturation events.

Since the optimal signal recovery performance occurs at a nonzero saturation rate, we present a simple *automatic gain control* (AGC) that adjusts the gain of the analog input signal so that the desired saturation rate is achieved. According to one rule of thumb, a conventional AGC will set the gain such that there is an average of 63 clips per million samples [25]. Thus, because the desired saturation rate is close to zero, saturation rate alone cannot be used to design a stable AGC. However, since the optimal CS performance occurs at a significantly non-zero saturation rate, our proposed AGC uses only the saturation rate to determine the gain.

The organization of this paper is as follows. In Section II, we review quantization with saturation and the key concepts of the CS framework. In Section III, we discuss the problem of unbounded saturation error in CS and define our proposed solutions. In Section IV we provide theoretical analysis to show that CS measurements are democratic and that our solutions solve the stated problem. In Section V, we validate our claims experimentally and show that in many scenarios, we achieve improved performance. In Section VI we derive a simple AGC for CS systems and in Section VII we discuss how the democracy property can be useful in other applications.

II. Background

A. Analog-to-digital conversion

ADC consists of two discretization steps: *sampling*, which converts a continuous-time signal to a discrete-time set of measurements, followed by *quantization*, which converts the continuous value of each measurement to a discrete one chosen from a pre-determined, finite set. Both steps are necessary to represent an analog signal in the discrete digital world.

The discretization step can be lossless or lossy. For example, classical results due to Shannon and Nyquist demonstrate that the sampling step induces no loss of information, provided that the signal is bandlimited and a sufficient number of measurements (or samples) are obtained. Similarly, sensing of images assumes that the image is sufficiently smooth such that the integration of light in each pixel of the sensor is sufficient for a good quality representation of the image. Our paper assumes the existence of a discretization that exactly represents the signal, or approximates to sufficient quality. Examples of such discretizations and their implementation in the context of compressive sensing can be found in [7–15]. We briefly discuss aspects of such systems in Sec. II-D.

Instead we focus on the second aspect of digitization, namely quantization. Quantization results in an irreversible loss of information unless the measurement amplitudes belong to the discrete set defined by the quantizer. A central ADC system design goal is to minimize the distortion due to quantization.

B. Scalar quantization

Scalar quantization is the process of converting the continuous value of an individual measurement to one of several discrete values through a non-invertible function $R(\cdot)$. Practical quantizers introduce two kinds of distortion: *bounded* quantization error and *unbounded* saturation error.

In this paper, we focus on uniform quantizers with quantization interval Δ. Thus, the quantized values become $q_k = q_0 + k\Delta$, for $k \in \mathbb{Z}$, and every measurement g is quantized to the nearest quantization level $R(g) = \text{argmin}_{q_k} |g - q_k| = \Delta/2 + k\Delta$, the midpoint of each quantization interval. This minimizes the expected quantization distortion and implies that the quantization error per measurement, $|g - R(q)|$, is

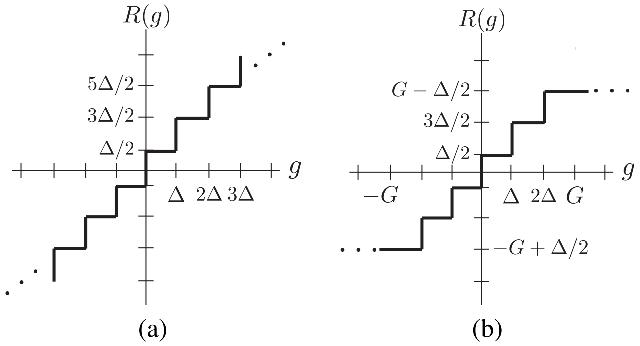


Fig. 1. (a) Midrise scalar quantizer. (b) Finite-range midrise scalar quantizer with saturation level G.

TABLE I
QUANTIZATION PARAMETERS.

G	saturation level
B	number of bits
Δ	bin width
$\Delta/2$	maximum error per (quantized) measurement
unbounded	maximum error per (saturated) measurement

bounded by $\Delta/2$. Figure 1(a) depicts the mapping performed by a midrise quantizer.

In practice, quantizers have a finite dynamic range, dictated by hardware constraints such as the voltage limits of the devices and the finite number of bits per measurement of the quantized representation. Thus, a *finite-range* quantizer represents a symmetric range of values $|g| < G$, where $G > 0$ is known as the saturation level [26]. Values of g between $-G$ and G will not saturate, thus, the quantization interval is defined by these parameters as $\Delta = 2^{-B+1}G$. Without loss of generality we assume a midrise B-bit quantizer, i.e., the quantization levels are $q_k = \Delta/2 + k\Delta$, where $k = -2^{B-1}, \ldots, 2^{B-1}-1$. Any measurement with magnitude greater than G saturates the quantizer, i.e., it quantizes to the quantization level $G - \Delta/2$, implying an unbounded error. Figure 1(b) depicts the mapping performed by a finite range midrise quantizer with saturation level G and Table I summarizes the parameters defined with respect to quantization.

C. Compressive sensing (CS)

In the CS framework, we acquire a signal $\mathbf{x} \in \mathbb{R}^N$ via the linear measurements

$$\mathbf{y} = \mathbf{\Phi}\mathbf{x} + \mathbf{e}, \tag{1}$$

where $\mathbf{\Phi}$ is an $M \times N$ measurement matrix modeling the sampling system, $\mathbf{y} \in \mathbb{R}^M$ is the vector of samples acquired, and $\mathbf{e}$ is an $M \times 1$ vector that represents measurement errors. If $\mathbf{x}$ is K-sparse when represented in the *sparsity basis* $\mathbf{\Psi}$, i.e., $\mathbf{x} = \mathbf{\Psi}\boldsymbol{\alpha}$ with $\|\boldsymbol{\alpha}\|_0 := |\mathrm{supp}(\alpha)| \leq K$, then one can acquire only $M = O(K\log(N/K))$ measurements and still recover the signal $\mathbf{x}$ [3, 4]. A similar guarantee can be obtained for approximately sparse, or *compressible*, signals. Observe that if K is small, then the number of measurements required can be significantly smaller than the Shannon-Nyquist rate.

In [27], Candès and Tao introduced the *restricted isometry property* (RIP) of a matrix $\mathbf{\Phi}$ and established its important role in CS. From [27], we have the definition,

Definition 1. *A matrix $\mathbf{\Phi}$ satisfies the RIP of order K with constant $\delta \in (0, 1)$ if*

$$(1-\delta)\|\mathbf{x}\|_2^2 \leq \|\mathbf{\Phi}\mathbf{x}\|_2^2 \leq (1+\delta)\|\mathbf{x}\|_2^2 \tag{2}$$

holds for all $\mathbf{x}$ such that $\|\mathbf{x}\|_0 \leq K$.

In words, $\mathbf{\Phi}$ acts as an approximate isometry on the set of vectors that are K-sparse in the basis $\mathbf{\Psi}$. An important result is that for any unitary matrix $\mathbf{\Psi}$, if we draw a random matrix $\mathbf{\Phi}$ whose entries ϕ_{ij} are independent realizations from a sub-Gaussian distribution, then $\mathbf{\Phi}\mathbf{\Psi}$ will satisfy the RIP of order K with high probability provided that $M = O(K\log(N/K))$ [28]. In this paper, without the loss of generality, we fix $\mathbf{\Psi} = \mathbf{I}$, the identity matrix, implying that $\mathbf{x} = \boldsymbol{\alpha}$.

The RIP is a necessary condition if we wish to be able to recover all sparse signals $\mathbf{x}$ from the measurements $\mathbf{y}$. Specifically, if $\|\mathbf{x}\|_0 = K$, then $\mathbf{\Phi}$ must satisfy the lower bound of the RIP of order $2K$ with $\delta < 1$ in order to ensure that any algorithm can recover $\mathbf{x}$ from the measurements $\mathbf{y}$. Furthermore, the RIP also suffices to ensure that a variety of practical algorithms can successfully recover any sparse or compressible signal from noisy measurements. In particular, for bounded errors of the form $\|\mathbf{e}\|_2 \leq \epsilon$, the convex program

$$\widehat{\mathbf{x}} = \underset{\theta}{\operatorname{argmin}} \ \|\mathbf{x}\|_1 \ \text{ s.t. } \|\mathbf{\Phi}\mathbf{x} - \mathbf{y}\|_2 \leq \epsilon \tag{3}$$

can recover a sparse or compressible signal $\mathbf{x}$. The following theorem, a slight modification of Theorem 1.2 from [29], makes this precise by bounding the recovery error of $\mathbf{x}$ with respect to the measurement noise norm, denoted by ϵ, and with respect the best approximation of $\mathbf{x}$ by its largest K terms, denoted using $\mathbf{x}_K$.

Theorem 1. *Suppose that $\mathbf{\Phi}\mathbf{\Psi}$ satisfies the RIP of order $2K$ with $\delta < \sqrt{2}-1$. Given measurements of the form $\mathbf{y} = \mathbf{\Phi}\mathbf{\Psi}\mathbf{x} + \mathbf{e}$, where $\|\mathbf{e}\|_2 \leq \epsilon$, then the solution to (3) obeys*

$$\|\widehat{\mathbf{x}} - \mathbf{x}\|_2 \leq C_0\epsilon + C_1\frac{\|\mathbf{x} - \mathbf{x}_K\|_1}{\sqrt{K}},$$

where

$$C_0 = \frac{4(1+\delta)}{1-(\sqrt{2}+1)\delta}, \quad C_1 = \frac{1+(\sqrt{2}-1)\delta}{1-(\sqrt{2}+1)\delta}.$$

While convex optimization techniques like (3) are a powerful method for CS signal recovery, there also exist a variety of alternative algorithms that are commonly used in practice and for which performance guarantees comparable to that of Theorem 1 can be established. In particular, iterative algorithms such as CoSaMP and iterative hard thresholding (IHT) are known to satisfy similar guarantees under slightly stronger assumptions on the RIP constants [30, 31]. Furthermore, alternative recovery strategies based on (3) have been analyzed in [20, 32]. These methods replace the constraint in (3) with an alternative constraint that is motivated by the assumption that the measurement noise is Gaussian in the case of [20] and that is agnostic to the value of ϵ in [32].

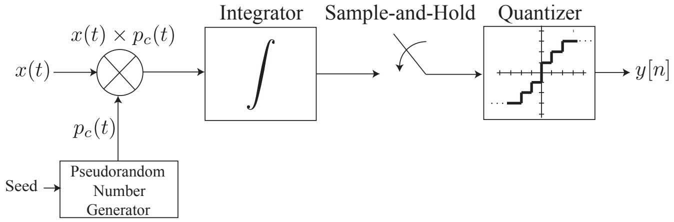


Fig. 2. Random demodulator compressive ADC.

D. *CS in practice*

Several hardware architectures have been proposed and implemented that allow CS to be used in practical settings with analog signals. Examples include the random demodulator, random filtering, and random convolution for signals [7–9, 14, 15], and several compressive imaging architectures [10–12].

We briefly describe the random demodulator as an example of such a system [7]. Figure 2 depicts the block diagram of the random demodulator. The four key components are a pseudo-random ±1 "chipping sequence" $p_c(t)$ operating at the Nyquist rate or higher, a low pass filter, often represented by an ideal integrator with reset, a low-rate ADC, and a quantizer. An input analog signal $x(t)$ is modulated by the chipping sequence and integrated. The output of the integrator is sampled, and the integrator is reset after each sample. The output measurements from the ADC are then quantized.

Systems such as these represent a linear operator mapping the analog input signal to a discrete output vector, followed by a quantizer. It is possible, but beyond the scope of this paper, to relate this operator to a discrete measurement matrix $\mathbf{\Phi}$ which maps, for example, the Nyquist-rate samples of the input signal to the discrete output vector [7, 15, 33]. In this paper we will restrict our focus to settings in which the measurement operator $\mathbf{\Phi}$ can be represented as an $M \times N$ matrix.

III. Signal Recovery from Saturated Measurements

A. *Unbounded saturation error*

A standard CS recovery approach like (3) assumes that the measurement error is bounded. However, when quantizing the measurements $\mathbf{y}$, the error on saturated measurements is unbounded. Thus, conventional wisdom would suggest that the measurements should first be scaled down appropriately so that none saturate.

This approach has two main drawbacks. First, rescaling the measurements reduces the saturation rate at the cost of increasing the quantization error on each measurement that does not saturate. Saturation events may be quite rare, but the additional quantization error will affect every measurement and induce a higher reconstruction error than if the signal had not been scaled and no saturation occurred. Second, in practice, saturation events may be impossible to avoid completely.

However, unlike conventional sampling systems that employ linear interpolation-based reconstruction, where each sample contains information for only a localized portion of the signal, CS measurements contain information for a larger portion of the signal. This creates a need for *non-linear* reconstruction algorithms but gives rise to some practical benefits such as robustness to the loss of a small number of measurements.

In this section, we propose two approaches for handling saturated measurements in CS systems:

1) saturation rejection: simply discard saturated measurements and then perform signal recovery on those that remain;
2) constrained optimization: incorporate saturated measurements in the recovery algorithm by enforcing consistency on the saturated measurements.

While both of these approaches are intuitive modifications of standard CS recovery algorithms, it is not obvious that they are guaranteed to work. For instance, in order for the saturation rejection approach to work we must be able to recover the signal using only the measurements that are retained, or equivalently, using only the rows of $\mathbf{\Phi}$ that are retained. An analysis of the properties of this matrix will be essential to understanding the performance of this approach. Similarly, it unclear when the combination of the retained measurements plus the additional information provided by the saturation constraints is sufficient to recover the signal. A main result of this paper, that we prove below, is that there exists a class of matrices $\mathbf{\Phi}$ such that an arbitrary subset of their rows will indeed satisfy the RIP, in which case existing results can provide performance guarantees for both of these approaches.

Before describing our approaches for handling saturated measurements in greater detail, we briefly establish some notation that will prove useful for the remainder of this paper. Let $\Gamma \subset \{1, 2, , \ldots, M\}$. By $\mathbf{\Phi}^{\Gamma}$ we mean the $|\Gamma| \times M$ matrix obtained by selecting the rows of $\mathbf{\Phi}$ indexed by Γ. Alternatively, if $\Lambda \subset \{1, 2, \ldots, N\}$, then we use $\mathbf{\Phi}_{\Lambda}$ to indicate the $M \times |\Lambda|$ matrix obtained by selecting the columns of $\mathbf{\Phi}$ indexed by Λ.

B. *Recovery via saturation rejection*

An intuitive way to handle saturated measurements is to simply discard them [34]. Denote the vector of the measurements that did not saturate as $\widetilde{\mathbf{y}}$ with length $\widetilde{M}$. The matrix $\widetilde{\mathbf{\Phi}}$ is created by selecting the rows of $\mathbf{\Phi}$ that correspond to the elements of $\widetilde{\mathbf{y}}$. Then, as an example, using (3) for reconstruction yields the program:

$$\widehat{\mathbf{x}} = \operatorname*{argmin}_{\mathbf{x}} \|\mathbf{x}\|_1 \quad \text{s.t.} \quad \|\widetilde{\mathbf{\Phi}}\mathbf{x} - \widetilde{\mathbf{y}}\|_2 < \epsilon. \tag{4}$$

There are several advantages to this approach. Any fast or specialized recovery algorithm can be employed without modification. In addition, the speed of most algorithms will be increased since fewer measurements are used.

The saturation rejection approach can also be applied in conjunction with processing and inference techniques such as the *smashed filter* [35] for detection, which utilizes the inner products $\langle \mathbf{\Phi u}, \mathbf{\Phi v} \rangle$ between the measurement of vectors $\mathbf{u}, \mathbf{v}$. Such techniques depend on $\langle \mathbf{\Phi u}, \mathbf{\Phi v} \rangle$ being close to $\langle \mathbf{u}, \mathbf{v} \rangle$. Saturation can induce unbounded errors in $\langle \mathbf{\Phi u}, \mathbf{\Phi v} \rangle$, making it arbitrarily far away from $\langle \mathbf{u}, \mathbf{v} \rangle$. Thus, by discarding saturated measurements, the error between these inner products is bounded.

C. Recovery via convex optimization with consistency constraints

Clearly saturation rejection discards potentially useful information. Thus, in our second approach, we include saturated measurements, but treat them differently from the others by enforcing *consistency*. Consistency means that we constrain the recovered signal $\widehat{\mathbf{x}}$ so that the magnitudes of the values of $\mathbf{\Phi}\widehat{\mathbf{x}}$ corresponding to the saturated measurements are greater than G.

Specifically, let S^+ and S^- correspond be the sets of indices of the positive saturated measurements, and negative saturated measurements, respectively. We define the matrix $\mathring{\mathbf{\Phi}}$ as

$$\mathring{\mathbf{\Phi}} := \left[\begin{array}{c} \mathbf{\Phi}^{S^+} \\ -\mathbf{\Phi}^{S^-} \end{array} \right]. \tag{5}$$

We obtain an estimate $\widehat{\mathbf{x}}$ via the program,

$$\widehat{\mathbf{x}} = \underset{\mathbf{x}}{\operatorname{argmin}} \|\mathbf{x}\|_1 \qquad \text{s.t.} \qquad \|\widetilde{\mathbf{\Phi}}\mathbf{x} - \widetilde{\mathbf{y}}\|_2 < \epsilon \tag{6a}$$

$$\text{and} \qquad \mathring{\mathbf{\Phi}}\mathbf{x} \geq G \cdot \mathbf{1}, \tag{6b}$$

where $\mathbf{1}$ denotes an $(M - \widetilde{M}) \times 1$ vector of ones. In words, we are looking for the $\mathbf{x}$ with the minimum ℓ_1 norm such that the measurements that do not saturate have bounded ℓ_2 error, and the measurements that do saturate are consistent with the saturation constraint. Alternative regularization terms that impose the consistency requirement on the unsaturated quantized measurements can be used on $\widetilde{\mathbf{y}}$, such as those proposed in [16, 17], or alternative techniques for the unsaturated quantized measurements can be used such as those proposed in [18]. In some hardware systems, the measurements that are acquired following a saturation event can have higher distortion than the other unsaturated measurements. This is a physical effect of some quantizers and may happen when the sample rate is high. In this case, an additional ℓ_2 constraint, $\|\widetilde{\mathbf{\Phi}}^\star \mathbf{x} - \widetilde{\mathbf{y}}^\star\|_2 < \epsilon_1$, can be applied where $\star$ denotes the indices of the measurements immediately following a saturation event and where $\epsilon_1 > \epsilon$. The measurements $\widetilde{\mathbf{y}}^\star$ can be determined via measured properties of the physical system.

D. Recovery via greedy algorithms with consistency constraints

Greedy algorithms can also be modified to include a saturation constraint. One example of a greedy algorithm that is typically used for sparse recovery is CoSaMP [30]. In this subsection, we introduce *Saturation Consistent CoSaMP* (SC-CoSaMP), a modified version of CoSaMP that performs consistent reconstruction with saturated measurements.

Algorithm 1 SC-CoSaMP greedy algorithm

```
1: Input: $\mathbf{y}$, $\mathbf{\Phi}$, and $K$
2: Initialize: $\widehat{\mathbf{x}}^{[0]} \leftarrow \mathbf{0}$, $n \leftarrow 0$
3: while not converged do
4:     Compute proxy:
       $\mathbf{p} \leftarrow \widetilde{\mathbf{\Phi}}^T \left(\widetilde{\mathbf{y}} - \widetilde{\mathbf{\Phi}}\widehat{\mathbf{x}}^{[n]}\right) + \mathring{\mathbf{\Phi}}^T \left(G \cdot \mathbf{1} - \mathring{\mathbf{\Phi}}\widehat{\mathbf{x}}^{[n]}\right)_+$
5:     Update coefficient support:
       $\Omega \leftarrow$ union of
           • support of largest $2K$ coefficients from $\mathbf{p}$
           • support of $\widehat{\mathbf{x}}^{[n]}$
6:     Estimate new coefficient values:
       $\widehat{\mathbf{x}}^{[n+1]} \leftarrow \operatorname{argmin}_{\mathbf{x}} \|\widetilde{\mathbf{y}} - \widetilde{\mathbf{\Phi}}_\Omega \mathbf{x}\|_2^2 + \|(G \cdot \mathbf{1} - \mathring{\mathbf{\Phi}}_\Omega \mathbf{x})_+\|_2^2$
7:     Prune:
       $\widehat{\mathbf{x}}^{[n+1]} \leftarrow$ keep largest $K$ coefficients of $\widehat{\mathbf{x}}^{[n+1]}$
8:     $n \leftarrow n + 1$
9: end while
```

CoSaMP estimates the signal $\widehat{\mathbf{x}}$ by finding a coefficient support set Ω and estimating the signal coefficients over that support. The support is found in part by first computing a vector $\mathbf{p} = \mathbf{\Phi}^T(\mathbf{y} - \mathbf{\Phi}\widehat{\mathbf{x}})$, that allows us to infer large signal coefficients, and hence is called the proxy vector [30], and second, by choosing the support of the largest $2K$ elements of $\mathbf{p}$. These $2K$ support locations are merged with the support corresponding to the largest K coefficients of $\widehat{\mathbf{x}}$ to produce Ω. Given Ω, CoSaMP estimates the signal coefficients by solving the least squares problem:

$$\widehat{\mathbf{x}} = \min_{\mathbf{x}} \|\mathbf{\Phi}_\Omega \mathbf{x} - \mathbf{y}\|_2^2. \tag{7}$$

These steps are done successively until the algorithm converges.

We modify two steps of CoSaMP to produce SC-CoSaMP; the proxy step and the coefficient estimate step. When computing the proxy vector, SC-CoSaMP enforces consistency from the contribution of the saturated measurements. When estimating the coefficients, a constraint on the saturated measurements is added to (7).

The steps of SC-CoSaMP are displayed in Algorithm 1. In steps 1 and 2, the algorithm initializes by choosing an estimate $\widehat{\mathbf{x}}^{[0]} = \mathbf{0}$, an N-dimensional vector of zeros, where the superscript $[\cdot]$ denotes iteration. To recover K coefficients, the algorithm loops until a condition in step 3 is met. For each iteration n, the algorithm proceeds as follows:

The proxy vector is computed in step 4. This is accomplished by computing the sum of two proxy vectors; a proxy from $\widetilde{\mathbf{y}}$ and a proxy that uses the supports of the saturated measurements. To compute the proxy from $\widetilde{\mathbf{y}}$, we repeat the same computation as in CoSaMP, $\widetilde{\mathbf{\Phi}}^T(\widetilde{\mathbf{y}} - \widetilde{\mathbf{\Phi}}\widehat{\mathbf{x}}^{[n]})$, where the superscript T denotes the matrix transpose. To compute the proxy from the support of the measurements that saturated, we introduce the saturation residual, denoted as $G \cdot \mathbf{1} - \mathring{\mathbf{\Phi}}\widehat{\mathbf{x}}^{[n]}$. This vector measures how close the elements of $\mathring{\mathbf{\Phi}}\widehat{\mathbf{x}}$ are to G. In consistent reconstruction, the magnitude of the elements of $\mathring{\mathbf{\Phi}}\widehat{\mathbf{x}}$ should be greater than or equal to G, however, once these are greater than G, the magnitude given by the saturation residual cannot be effectively interpreted.

Thus, consistency is achieved by applying a function that selects the positive elements of the saturation residual,

$$(y_i)_+ = \left\{ \begin{array}{ll} 0, & y_i < 0 \\ y_i, & y_i \geq 0, \end{array} \right. \tag{8}$$

where the function is applied element-wise when operating on a vector.

By combining the proxies from $\widetilde{\mathbf{y}}$ and the saturated measurement supports, the proxy vector of step 4 is

$$\mathbf{p} = \widetilde{\mathbf{\Phi}}^T \left(\widetilde{\mathbf{y}} - \widetilde{\mathbf{\Phi}} \widehat{\mathbf{x}}^{[n]} \right) + \ \mathring{\mathbf{\Phi}}^T \left(G \cdot \mathbf{1} - \mathring{\mathbf{\Phi}} \widehat{\mathbf{x}}^{[n]} \right)_+ . \qquad (9)$$

In this arrangement, the elements of $\mathring{\mathbf{\Phi}}\widehat{\mathbf{x}}$ that are below G will contribute new information to $\mathbf{p}$, however, elements that are greater than G will be set to zero, and therefore do not contribute additional information to $\mathbf{p}$. We note that a similar computation can be made in the IHT algorithm [31].

In step 5, the new coefficient support Ω is found by taking the union of the support of the largest $2K$ coefficients of $\mathbf{p}$ and the support of $\widehat{\mathbf{x}}^{[n]}$. This results in a support set Ω with at most $3K$ elements. This step ensures that if coefficients were incorrectly chosen in a previous iteration, they can be replaced.

In step 6 new coefficient values are estimated by finding the $\mathbf{x}$ that minimizes $\|\mathbf{\Phi}_\Omega \mathbf{x} - \mathbf{y}\|_2^2$. Thus in CoSaMP, new coefficient values are estimated via $\mathbf{\Phi}_\Omega^\dagger \mathbf{y}$, where $\dagger$ denotes the Moore-Penrose pseudo-inverse, i.e., $\mathbf{\Phi}_\Omega^\dagger = (\mathbf{\Phi}_\Omega^T \mathbf{\Phi}_\Omega)^{-1} \mathbf{\Phi}_\Omega^T$. We reformulate this step to include the saturation constraint. Specifically, step 6 of SC-CoSaMP finds the solution to

$$\widehat{\mathbf{x}}^{[n+1]} \leftarrow \underset{\mathbf{x}}{\operatorname{argmin}} \|\widetilde{\mathbf{y}} - \widetilde{\mathbf{\Phi}}_\Omega \mathbf{x}\|_2^2 + \|(G \cdot \mathbf{1} - \mathring{\mathbf{\Phi}}_\Omega \mathbf{x})_+\|_2^2 \qquad (10)$$

This can be achieved via gradient descent or other optimization techniques. By employing a one-sided quadratic we ensure a soft application of the constraint and ensure the program is feasible even in the presence of noise [36].

In step 7, we keep the largest K coefficients of the signal estimate. The algorithm repeats until a convergence condition is met.

As demonstrated, SC-CoSaMP is different from CoSaMP in steps 4 and 6. In practice, we have found that applying step 4 of SC-CoSaMP to compute $\mathbf{p}$ provides a significant increase in performance over the equivalent step in CoSaMP, while applying step 6 for coefficient estimation provides only a marginal performance increase.

IV. Random Measurements and Democracy

A. Democracy and recovery

In this section, we demonstrate that the random measurement schemes typically advocated in CS are democratic, i.e., each measurement contributes a similar amount of information about the signal $\mathbf{x}$ to the compressed representation $\mathbf{y}$ [37–39].[1] The fact that random measurements are democratic seems intuitive; when using random measurements, each measurement is a randomly weighted sum of a large fraction (or all) of the coefficients of $\mathbf{x}$, and since the weights are chosen independently at random, no preference is given to any particular set of coefficients. More concretely, suppose that the measurements $y_1, y_2, \ldots, y_M$ are independent and identically distributed (i.i.d.) according to some distribution f_Y, as is the case for the $\mathbf{\Phi}$ considered in this paper. Now suppose that we select $\widetilde{M} < M$ of the y_i at random (or according to some procedure that is *independent* of $\mathbf{y}$). Then we are left with a length-$\widetilde{M}$ measurement vector $\widetilde{\mathbf{y}}$ such that each $\widetilde{y}_i \sim f_Y$. Stated another way, if we set $D = M - \widetilde{M}$, then there is no difference between collecting $\widetilde{M}$ measurements and collecting M measurements and deleting D of them, provided that this deletion is done independently of the actual values of $\mathbf{y}$.

However, following this line of reasoning will ultimately lead to a rather weak definition of democracy. To see this, consider the case where the measurements are deleted by an adversary. Since by adaptively deleting the entries of $\mathbf{y}$ one can change the distribution of $\widetilde{\mathbf{y}}$, the adversary can delete the D largest elements of $\mathbf{y}$, thereby skewing the distribution of $\widetilde{\mathbf{y}}$. In many cases, especially if the same matrix $\mathbf{\Phi}$ will be used repeatedly with different measurements being deleted each time, it would be far better to know that *any* $\widetilde{M}$ measurements will be sufficient to *robustly* reconstruct the signal. This is a significantly stronger requirement.

The RIP also provides us with a way to quantify our notion of democracy.

Definition 2. *Let $\mathbf{\Phi}$ be and $M \times N$ matrix, and let $\widetilde{M} \leq M$ be given. We say that $\mathbf{\Phi}$ is $(\widetilde{M}, K, \delta)$-democratic if for all Γ such that $|\Gamma| \geq \widetilde{M}$ the matrix $\mathbf{\Phi}^\Gamma$ satisfies the RIP of order K with constant δ.*

If Φ is $(\widetilde{M}, 2K, \delta)$-democratic, then both approaches described in Section III will recover sparse and compressible signals. In particular, the democracy property implies that any $\widetilde{M} \times N$ submatrix of Φ has RIP, and in particular that $\widetilde{\mathbf{\Phi}}$ satisfies the RIP. Thus, if $\delta < \sqrt{2} - 1$, it immediately follows from Theorem 1 that the rejection approach (4) yields a recovered signal that satisfies (1) whenever the number of unsaturated measurements exceeds $\widetilde{M}$. Furthermore, under the same conditions, we also have that (6) yields a recovered signal (1). This can be seen by observing that the proof of Theorem 1 in [29] essentially depends on only three facts: *(i)* that the original signal $\mathbf{x}$ is in the feasible set, so that we can conclude *(ii)* that $\|\widehat{\mathbf{x}}\|_1 \leq \|\mathbf{x}\|_1$, and finally *(iii)* that $\|\mathbf{\Phi}\widehat{\mathbf{x}} - \mathbf{\Phi}\mathbf{x}\|_2 \leq \epsilon$, where $\mathbf{\Phi}$ can be any matrix that satisfies the RIP of order $2K$ with constant $\delta < \sqrt{2} - 1$. Since $\mathbf{\Phi}$ is democratic we have that *(iii)* holds for $\widetilde{\mathbf{\Phi}}$ regardless of whether we incorporate the additional constraints. Since the original signal $\mathbf{x}$ will remain feasible in (6), *(i)* and *(ii)* will also hold.

Note that the two approaches will not necessarily produce the same solution. This is because the solution from the rejection approach may not lie in the feasible set of solutions of the consistent approach (6). However, the reverse is true. The solution to the consistent approach does lie in the feasible set of solutions to the rejection approach. While we do not provide a detailed analysis that compares the performance of these two approaches, we expect that the consistent approach will outperform the rejection approach since it incorporates additional information about the signal. We provide experimental confirmation of this in Section V.

We now demonstrate that certain randomly generated ma-

[1] The original introduction of this term was with respect to quantization [37, 38], i.e., a democratic quantizer would ensure that each bit is given "equal weight." As the CS framework developed, it became empirically clear that CS systems exhibited this property with respect to compression [39].

trices are democratic. While the theorem actually holds (with different constants) for the more general class of *sub-Gaussian* matrices, for simplicity we restrict our attention to Gaussian matrices.

Theorem 2. *Let $\mathbf{\Phi}$ by an $M \times N$ matrix with elements ϕ_{ij} drawn according to $\mathcal{N}(0, 1/M)$ and let $\widetilde{M} \le M$, $K < \widetilde{M}$, and $\delta \in (0,1)$ be given. Define $D = M - \widetilde{M}$. If*

$$M = C_1(K+D)\log\left(\frac{N+M}{K+D}\right), \qquad (11)$$

then with probability exceeding $1 - 3e^{-C_2 M}$ we have that $\mathbf{\Phi}$ is $(\widetilde{M}, K, \delta/(1-\delta))$-democratic, where C_1 is arbitrary and $C_2 = (\delta/8)^2 - \log(42e/\delta)/C_1$.

Proof: Our proof consists of two main steps. We begin by defining the $M \times (N+M)$ matrix $\boldsymbol{A} = [\mathbf{I}\ \mathbf{\Phi}]$ formed by appending $\mathbf{\Phi}$ to the $M \times M$ identity matrix. Theorem 1 from [22] demonstrates that under the assumptions in the theorem statement, with probability exceeding $1 - 3e^{-C_2 M}$ we have that $\boldsymbol{A}$ satisfies the RIP of order $K + D$ with constant δ. The second step is to use this fact to show that all possible $\widetilde{M} \times N$ submatrices of $\mathbf{\Phi}$ satisfy the RIP of order K with constant $\delta/(1-\delta)$.

Towards this end, we let $\Gamma \subset \{1, 2, \ldots, M\}$ be an arbitrary subset of rows such that $|\Gamma| \ge \widetilde{M}$. Define $\Lambda = \{1, 2, \ldots, M\} \setminus \Gamma$ and note that $|\Lambda| = D$. Additionally, let

$$\mathbf{P}_\Lambda \triangleq \boldsymbol{A}_\Lambda \boldsymbol{A}_\Lambda^\dagger, \qquad (12)$$

be the orthogonal projector onto $\mathcal{R}(\boldsymbol{A}_\Lambda)$, i.e., the range, or column space, of $\boldsymbol{A}_\Lambda$. Furthermore, we define

$$\mathbf{P}_\Lambda^\perp \triangleq \mathbf{I} - \mathbf{P}_\Lambda, \qquad (13)$$

as the orthogonal projector onto the orthogonal complement of $\mathcal{R}(\boldsymbol{A}_\Lambda)$. In words, this projector nulls the columns of $\boldsymbol{A}$ corresponding to the index set Λ. Now, note that $\Lambda \subset \{1, 2, \ldots, M\}$, so $\boldsymbol{A}_\Lambda = \mathbf{I}_\Lambda$. Thus,

$$\mathbf{P}_\Lambda = \mathbf{I}_\Lambda \mathbf{I}_\Lambda^\dagger = \mathbf{I}_\Lambda (\mathbf{I}_\Lambda^T \mathbf{I}_\Lambda)^{-1} \mathbf{I}_\Lambda^T = \mathbf{I}_\Lambda \mathbf{I}_\Lambda^T = \mathbf{I}(\Lambda),$$

where we use $\mathbf{I}(\Lambda)$ to denote the $M \times M$ matrix with all zeros except for ones on the diagonal entries corresponding to the columns indexed by Λ. (We distinguish the $M \times M$ matrix $\mathbf{I}(\Lambda)$ from the $M \times D$ matrix $\mathbf{I}_\Lambda$ — in the former case we replace columns not indexed by Λ with zero columns, while in the latter we remove these columns to form a smaller matrix.) Similarly, we have

$$\mathbf{P}_\Lambda^\perp = \mathbf{I} - \mathbf{P}_\Lambda = \mathbf{I}(\Gamma).$$

Thus, we observe that the matrix $\mathbf{P}_\Lambda^\perp \boldsymbol{A} = \mathbf{I}(\Gamma)\boldsymbol{A}$ is simply the matrix $\boldsymbol{A}$ with zeros replacing all entries on any row i such that $i \notin \Gamma$, i.e., $(\mathbf{P}_\Lambda^\perp \boldsymbol{A})^\Gamma = \boldsymbol{A}^\Gamma$ and $(\mathbf{P}_\Lambda^\perp \boldsymbol{A})^\Lambda = \mathbf{0}$. Furthermore, Theorem X from [40] states that for $\boldsymbol{A}$ satisfying the RIP of order $K + D$ with constant δ, we have that

$$\left(1 - \frac{\delta}{1-\delta}\right)\|\mathbf{u}\|_2^2 \le \|\mathbf{P}_\Lambda^\perp \boldsymbol{A}\mathbf{u}\|_2^2 \le (1+\delta)\|\mathbf{u}\|_2^2, \qquad (14)$$

holds for all $\mathbf{u} \in \mathbb{R}^{N+M}$ such that $\|\mathbf{u}\|_0 = K + D - |\Lambda| = K$ and $\mathrm{supp}(\mathbf{u}) \cap \Lambda = \emptyset$. Equivalently, letting $\Lambda^c = \{1, 2, \ldots, N+M\} \setminus \Lambda$, this result states that $(\mathbf{I}(\Gamma)\boldsymbol{A})_{\Lambda^c}$ satisfies the RIP of order K with constant $\delta/(1-\delta)$. To complete the proof, we note that if $(\mathbf{I}(\Gamma)\boldsymbol{A})_{\Lambda^c}$ satisfies the RIP of order K with constant $\delta/(1-\delta)$, then we trivially have that $\mathbf{I}(\Gamma)\mathbf{\Phi}$ also has the RIP of order at least K with constant $\delta/(1-\delta)$, since $\mathbf{I}(\Gamma)\mathbf{\Phi}$ is just a submatrix of $(\mathbf{I}(\Gamma)\boldsymbol{A})_{\Lambda^c}$. Since $\|\mathbf{I}(\Gamma)\mathbf{\Phi}\mathbf{x}\|_2 = \|\mathbf{\Phi}^\Gamma \mathbf{x}\|_2$, this establishes the theorem. ■

B. Robustness and stability

Observe that we require roughly $O(D\log(N))$ additional measurements to ensure that $\mathbf{\Phi}$ is $(\widetilde{M}, K, \delta)$-democratic compared to the number of measurements required to simply ensure that $\mathbf{\Phi}$ satisfies the RIP of order K. This seems intuitive; if we wish to be robust to the loss of any D measurements while retaining the RIP of order K, then we should expect to take *at least* D additional measurements. This is not unique to the CS framework. For instance, by *oversampling*, i.e., sampling faster than the minimum required Nyquist rate, uniform sampling systems can also improve robustness with respect to the loss of measurements. Reconstruction can be performed in principle on the remaining non-uniform grid, as long as the remaining samples satisfy the Nyquist range on average [41].

However, linear reconstruction in such cases is known to be unstable. Furthermore the linear reconstruction kernels are difficult to compute. Under certain conditions stable non-linear reconstruction is possible, although this poses further requirements on the subset set of samples that can be lost and the computation can be expensive [42]. For example, dropping contiguous groups of measurements can be a challenge for the stability of the reconstruction algorithms. Instead, the democratic principle of CS allows dropping of an arbitrary subset D of the measurements without compromising the reconstruction stability, independent of the way these measurements are chosen.

In some applications, this difference may have significant impact. For example, in finite dynamic range quantizers, the measurements saturate when their magnitude exceeds some level. Thus, when uniformly sampling with a low saturation level, if one sample saturates, then the likelihood that any of the neighboring samples will saturate is high, and significant oversampling may be required to ensure any benefit. However, in CS, if many adjacent measurements were to saturate, then for only a slight increase in the number of measurements we can mitigate this kind of error by simply rejecting the saturated measurements; the fact that $\mathbf{\Phi}$ is democratic ensures that this strategy will be effective.

Theorem 2 further guarantees graceful degradation due to loss of samples. Specifically, the theorem implies that reconstruction from any subset of CS measurements is stable to the loss of a potentially larger number of measurements than anticipated. To see this, suppose that and $M \times N$ matrix $\mathbf{\Phi}$ is $(M - D, K, \delta)$-democratic, but consider the situation where $D + \widetilde{D}$ measurements are dropped. It is clear from the proof of Theorem 2 that if $\widetilde{D} < K$, then the resulting matrix $\mathbf{\Phi}^\Gamma$ will satisfy the RIP of order $K - \widetilde{D}$ with constant δ. Thus, from [43], if we define $\widetilde{K} = (K - \widetilde{D})/2$, then the reconstruction error is then bounded by

$$\|\mathbf{x} - \widehat{\mathbf{x}}\|_2 \leq C_3 \frac{\|\mathbf{x} - \mathbf{x}_{\widetilde{K}}\|_1}{\sqrt{\widetilde{K}}}, \tag{15}$$

where $\mathbf{x}_{\widetilde{K}}$ denotes the best $\widetilde{K}$-term approximation of $\mathbf{x}$ and C_3 is an absolute constant depending on $\mathbf{\Phi}$ that can be bounded using the constants derived in Theorem 2. Thus, if $\widetilde{D}$ is small then the additional error caused by dropping too many measurements will also be relatively small. To our knowledge, there is simply no analog to this kind of graceful degradation result for uniform sampling with linear reconstruction. When the number of dropped samples exceeds D, there is are no guarantees as to the accuracy of the reconstruction.

V. Experimental Validation

In the previous sections, we discussed three approaches for recovering sparse signals from finite-range, quantized CS measurements;

1) the *conventional approach*, scaling the signal so that the saturation rate is zero and reconstructing with the program (3);
2) the *rejection approach*, discarding saturated measurements before reconstruction with (4); and
3) the *consistent approach*, incorporating saturated measurements as a constraint in the program (6).

In this section we compare these approaches via a suite of simulations to demonstrate that, on average, using the saturation constraint outperforms the other approaches for a given saturation level G. Our main findings include:

- In many cases the optimal performance for the consistent and rejection approaches is superior to the optimal performance for the conventional approach and occurs when the saturation rate is nonzero.
- The difference in optimal performance between the consistent and rejection approaches is small for a given ratio of M/N.
- The consistent reconstruction approach is more robust to saturation than the rejection approach. Also, for a large range of saturation rates, consistent reconstruction outperforms the conventional approach even if the latter is evaluated under optimal conditions.

We find these behaviors for both sparse and compressible signals and for both optimization and greedy recovery algorithms.

A. Experimental setup

Signal model: We study the performance of our approaches using two signal classes:

- K-sparse: in each trial, K nonzero elements x_n are drawn from an i.i.d. Gaussian distribution and where the locations n are randomly chosen;
- weak ℓ_p-compressible: in each trial, elements x_n are first generated according to

$$x_n = v_n n^{-1/p}, \tag{16}$$

for $p \leq 1$ where v_n is a ±1 Rademacher random variable. The positions n are then permuted randomly.

Once a signal is drawn, it is normalized to have unit ℓ_2 norm. Aside from quantization we do not add any additional noise sources.

Measurement matrix: For each trial a measurement matrix is generated using an i.i.d. Gaussian distribution with variance $1/M$. Our extended experimentation, not shown here in the interest of space, demonstrates consistent results across a variety of measurement matrix classes including i.i.d. ±1 Rademacher matrices and other sub-Gaussian matrices, as well as the random demodulator and random time-sampling.

Reconstruction metric: We report the reconstruction *signal-to-noise ratio* (SNR) in decibels (dB):

$$\text{SNR} \triangleq 10 \log_{10} \left(\frac{\|\mathbf{x}\|_2^2}{\|\mathbf{x} - \widehat{\mathbf{x}}\|_2^2} \right), \tag{17}$$

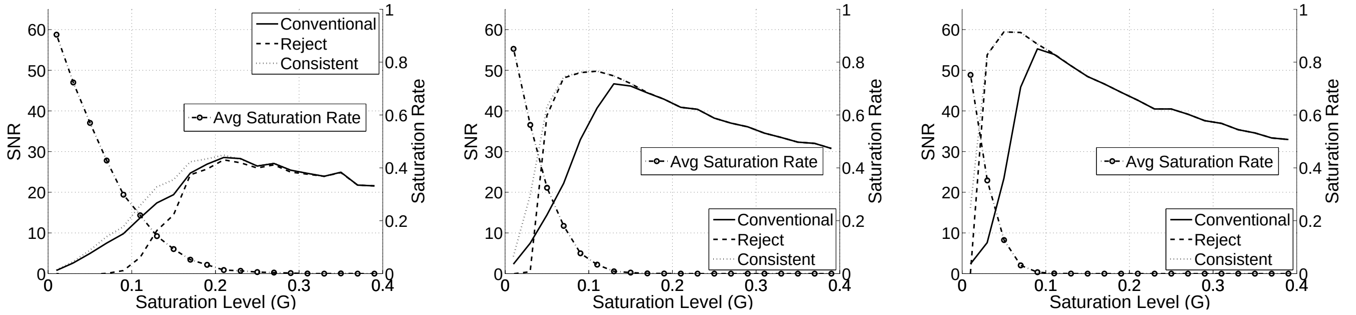
where $\widehat{\mathbf{x}}$ denotes the reconstructed signal.

B. Reconstruction SNR: K-sparse signals

We compare the reconstruction performance of the three approaches by applying each to the same set of measurements. We fix the parameters, $N = 1024$, $K = 20$, and $B = 4$ and vary the saturation level parameter G over the range $[0, 0.4]$. We varied the ratio M/N in the range $[1/16, 1]$ but plot results for only the three ratios $M/N = 2/16$, $6/16$, and $15/16$ that exhibit typical behavior for their regime. For each parameter combination, we performed 100 trials, and computed the average performance. The results were similar for other parameters, thus those experiments are not displayed here.

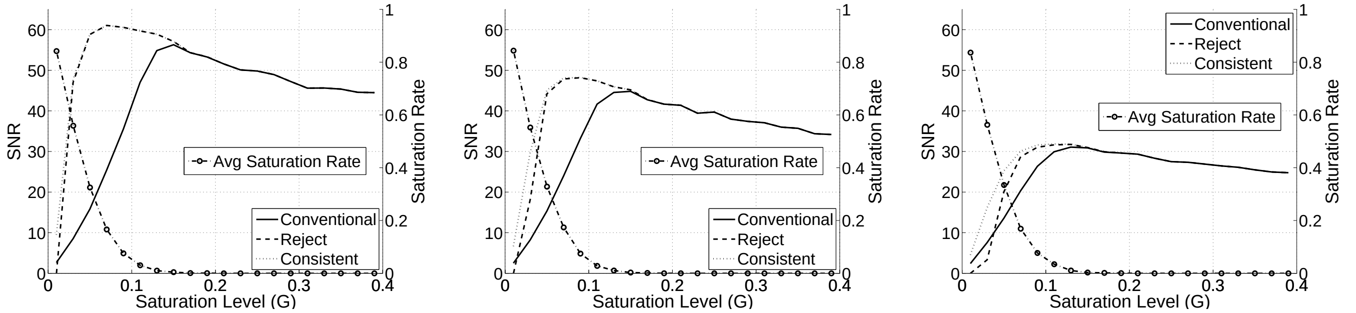
The experiments were performed as follows. For each trial we draw a new sparse signal $\mathbf{x}$ and a new matrix $\mathbf{\Phi}$ according to the details in Section V-A and compute $\mathbf{y} = \mathbf{\Phi x}$. We quantize the measurements using a quantizer with saturation level G and then use them to reconstruct the signal using the three approaches described above. The reconstructions were performed using CVX [44, 45], a general purpose optimization package.

Figures 3(a), 3(b), and 3(c) display the reconstruction SNR performance of the three approaches in dB for $M/N = 2/16$, $M/N = 6/16$, $M/N = 15/16$, respectively. The solid line depicts the conventional approach, the dashed line depicts the rejection approach, and the dotted line depicts the consistent approach. Each of these lines follow the scale on the left y-axis. The dashed-circled line denotes the average saturation rate, $(M - \widetilde{M})/M$, and correspond to the right y-axis. In Figure 3(a), the three lines meet at $G = 0.25$, as expected, because the saturation rate is effectively zero at this point. This is the operating point for the conventional approach and is the largest SNR value for the solid line. In this case, only the consistent approach obtains SNRs greater than the conventional approach. In Figure 3(b), the three lines meet at $G = 0.15$. Both the consistent and the rejection approaches achieve their optimal performance at around $G = 0.1$, where the saturation rate is 0.09. In Figure 3(c), the three lines meet at $G = 0.1$ and both the consistent and rejection approaches achieve their optimal performance at $G = 0.06$.



(a) $M/N = 2/16$ (b) $M/N = 6/16$ (c) $M/N = 15/16$

Fig. 3. Comparison of reconstruction approaches using CVX for K-sparse signals with $N = 1024$, $K = 20$, and $B = 4$. Solid line depicts reconstruction for the conventional approach. Dotted line depicts reconstruction for the consistent approach. Dashed line depicts reconstruction for the rejection approach. The left y-axis corresponds to each of these lines. The dashed-circled line represents the average saturation rate and corresponds to the right y-axis. Each plot represents a different measurement regime: (a) low $M/N = 2/16$, (b) medium $M/N = 6/16$, and (c) high $M/N = 15/16$.



(a) $p = 0.4$ (b) $p = 0.8$ (c) $p = 1$

Fig. 4. Comparison of reconstruction approaches using CVX for weak ℓ_p compressible signals with $N = 1024$, $M/N = 6/16$, and $B = 4$. Solid line depicts reconstruction for the conventional approach. Dotted line depicts reconstruction for the consistent approach. Dashed line depicts reconstruction for the rejection approach. The left y-axis corresponds to each of these lines. The dashed-circled line represents the average saturation rate and corresponds to the right y-axis. Each plot represents different rate of decay for the coefficients: (a) fast decay $p = 0.4$, (b) medium decay $p = 0.8$, and (c) slow decay $p = 1$.

The implications of this experiment are threefold: First, the saturation constraint offers the best approach for reconstruction. Second, if the signal is very sparse or there is an excess of measurements, then saturated measurements can be rejected with negligible loss in performance. Third, if given control over the parameter G, then the quantizer should be tuned to operate with a positive saturation rate.

C. Reconstruction SNR: Compressible signals

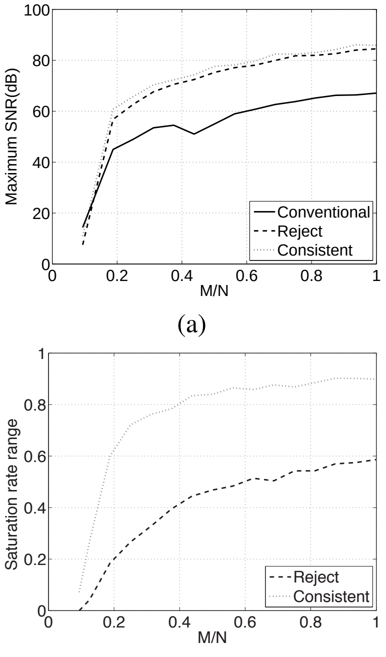
In addition to sparse signals, we also compare the reconstruction performance of the three approaches with compressible signals. As in the strictly sparse experiments, we use CVX for reconstruction. Similar to the sparse reconstruction experiments, we choose the parameters, $N = 1024$, $M/N = 6/16$, and $B = 4$ and vary the saturation level parameter G over the range $[0, 0.4]$. The decay parameter p is varied in the range $[0.4, 1]$, but we will discuss only three decays $p = 0.4$, 0.8, and 1. Some signals are known to exhibit p in (16) in this range, for instance, it has been shown that the wavelet coefficients of natural images have decay rates between $p = 0.3$ and $p = 0.7$ [46]. For each parameter combination, we perform 100 trials, and compute the average performance. The experiments are performed in the same fashion as with the sparse signals.

For signals with smaller p, fewer coefficients are needed to approximate the signals with low error. This also implies that fewer measurements are needed for these signals. The plots in Figure 4 reflect this intuition. Figures 4(a), 4(b), and 4(c) depict the results for $p = 0.4$, $p = 0.8$, and $p = 1$, respectively. The highest SNR for $p = 0.4$ is achieved at a saturation rate of 17%, while for $p = 0.8$ the saturation rate can only be 13%, and for $p = 1$ the highest SNR occurs at a saturation rate of 5%. This means that the smaller the p, the more the measurements should be allowed to saturate.

D. Robustness to saturation

We also compare the optimal performance of the rejection and consistent reconstruction approaches. First, we find the maximum SNR versus M/N for these approaches and demonstrate that their difference is small. Second, we determine the robustness to saturation of each approach. Because these experiments require many more trials than in the previous experiments, we use SC-CoSaMP from Section III-D for the consistent approach and CoSaMP for the rejection and conventional approaches.

We experimentally measure, by tuning G, the best SNR achieved on average for the three strategies. The experiment is performed as follows. Using the same parameters as in the K-sparse experiments, for each value of M and for each approach, we search for the saturation level G that yields the highest average SNR and report this SNR. This is equivalent to finding the maximum point on each of the curves of each plot in Figure 3 but over a larger range of M.



(a)

(b)

Fig. 5. SNR performance using SC-CoSaMP for $N = 1024$, $K = 20$, and $B = 4$. (a) Best-achieved average SNR vs. M/N. (b) Maximum saturation rate such that average SNR performance is as good or better than the best average performance of the conventional approach. For best-case saturation-level parameters, the rejection and constraint approaches can achieve SNRs exceeding the conventional SNR performance by 20dB. The best performance between the reject and consistent approaches is similar, differing only by 3dB, but the range of saturation rates for which they achieve high performance is much larger for the consistent approach. Thus, the consistent approach is more robust to saturation.

Figure 5(a) depicts the results of this experiment. The solid curve denotes the best performance for the conventional approach; the dashed curve denotes the performance with saturation rejection; and the dotted curve denotes the performance with the constraint. For these parameters, in the best case, saturation rejection can improve performance by 20dB, and the saturation constraint can improve performance over the conventional case by 23dB.

There are two important implications from this experiment. First, when the number of measurements exceeds the minimum required number of measurements, then intentionally saturating measurements can greatly improve performance. Second, in terms of the maximum SNR, the consistent approach performs only marginally better than the rejection approach, assuming that the quantizer operates under the optimal saturation conditions for each approach.

In practice it may be difficult to efficiently determine or maintain the saturation level that achieves the maximum SNR. In those cases, it is beneficial to know the robustness of each approach to changes in the saturation rate. Specifically, we compare the range of saturation rates for which the two approaches outperform the conventional approach when the latter is operating under optimal conditions.

This experiment first determines the maximum SNR achieved by the conventional approach (i.e., the solid curve in Figure 5(a)). Then, for the other approaches, we increase the saturation rate by tuning the saturation level. We continue to increase the saturation rate until the SNR is lower than the best SNR of the conventional approach.

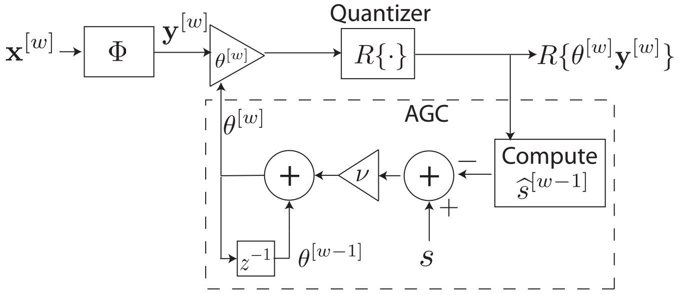


Fig. 6. Automatic gain control (AGC) for tuning to nonzero saturation rates in CS systems.

The results of this experiment are depicted in Figure 5(b). The dashed line denotes the range of saturation rates for the rejection approach and the dotted line denotes the range of saturation rates for the consistent approach. At best, the rejection approach achieves a range of $[0, 0.6]$ while the consistent approach achieves a range of $[0, 0.9]$. Thus, these experiments show that the consistent approach is more robust to saturation rate.

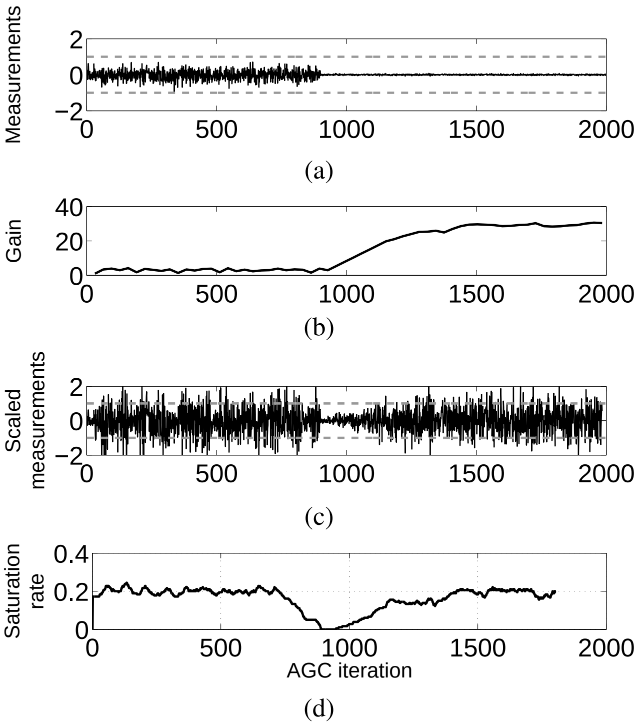
VI. Extensions

A. Automatic gain control (AGC) for CS

Most CS reconstruction approaches (with the exception of [47]) consider finite-length signals $\mathbf{x}$. However, in many applications of CS the measured signal is a time-varying, streaming signal of length unknown in advance. To apply CS methods to such applications, a blocking approach is usually pursued. The signal is split into blocks and each block is compressively sampled and reconstructed separately from the other blocks. In such streaming applications, the signal power does not remain constant but changes throughout the operation of the system and from block to block. Such changes affect the performance, especially in terms of Signal-to-Quantization noise level and saturation rate.

To adapt to changes in signal power and to avoid saturation events, modern sampling systems employ automatic gain control (AGC). These AGC's typically target saturation rates that are close to zero. In this case, saturation events can be used to detect high signal strength; however detecting low signal strength is more difficult. Thus, in conventional systems, saturation rate alone does not provide sufficient feedback to perform automatic gain control. Other measures, such as measured signal power are used in addition to saturation rate to ensure that the signal gain is sufficiently low but not too low.

In this section we demonstrate that in a CS system, where a positive saturation rate is desirable, the saturation rate can by itself provide sufficient feedback to the AGC circuit. Since the desired rate is significantly greater than zero, deviation from the desired rate can be used to both increase and decrease the gain in an AGC circuit to maintain a target saturation rate. Saturation events can be detected more easily and at earlier stages of the signal acquisition systems compared to measures



(a)

(b)

(c)

(d)

Fig. 7. CS AGC in practice. (a) CS measurements with no saturation. Signal strength drops by 90% at measurement 900. (b) Output gain from AGC. (c) Measurements scaled by gain from AGC. (d) Saturation rate of scaled measurements. This figure demonstrates that the CS AGC is sensitive to decreases in signal strength.

such as the signal variance. Thus the effectiveness of AGC increases and the cost decreases.

Our setup is as follows. The signal $\mathbf{x}$ is split into consecutive blocks of length N, and $\mathbf{\Phi}$ is applied to each block separately such that there are M measurements per block. We index each successive block of measurements by w and denote this with the superscript $[\cdot]$. In this example we apply a boxcar window to each block of $\mathbf{x}$, but in general any window can be applied. For each block, a gain $\theta^{[w]}$ is applied to the measurements and then quantized, resulting in a set of M output measurements $R\{\theta^{[w]}\mathbf{y}^{[w]}\}$. Note that in different hardware implementations, the gain might be applied before, after, or within the measurement matrix $\mathbf{\Phi}$; this change does not fundamentally affect our design. Our goal is to tune the gain so that it produces a desired measurement saturation rate s. We also assume that the signal energy does not deviate significantly between consecutive blocks.

A simple AGC that uses saturation rate to tune the gain is depicted in Figure 6 and operates as follows. We compute the saturation rate of the previous block of measurements, $\widehat{s}^{[w-1]}$, after quantization. The new gain is then computed by adding the error between s and $\widehat{s}^{[w-1]}$ to the previous gain, i.e.,

$$\theta^{[w]} = \theta^{[w-1]} + \nu(s - \widehat{s}^{[w-1]}), \tag{18}$$

where $\nu > 0$ is constant. This negative feedback system is BIBO[2] stable for any finite positive ν with $0 < s < 1$ [48].

To demonstrate that this AGC is sensitive to both increases in signal strength as well as decreases, we perform an experiment where the signal strength drops suddenly and significantly. The experiment is depicted in Figure 7 and was performed as follows. We generated a signal such that the parameters per block were $N = 512$, $K = 5$, and $M = 32$. We generated 63 blocks resulting in approximately 2000 measurements in total. The example measurements before the AGC is applied are depicted in Figure 7(a). The dashed lines represent the quantizer range $[-1, 1]$. We have generated the measurements so that the saturation rate is zero, and starting at measurement 900, the signal strength drops by 90%. These measurements are input into the AGC previously described with $\nu = 12$ and we set a desired saturation rate of $s = 0.2$.

Figure 7(b) shows the gain that the AGC applies as it receives each measurement. Figure 7(c) shows the resulting output signal with quantizer range, and Figure 7(d) shows the estimated output saturation rate. Initially, we achieve the desired saturation rate of 0.2 within approximately 10 iterations. The system adapts to the sudden change in signal strength after measurement 900 within approximately 500 iterations. This experiment demonstrates that the saturation rate is by itself sufficient to tune the gain of CS systems.

Of course more elaborate gain update loops can be considered to provide better adaptability and more rapid updates to the gain from block to block. Such methods are beyond the scope of this paper.

VII. Discussion

In this paper, we have presented two new approaches for handling unbounded saturation errors on compressive measurements; rejecting saturated measurements and applying consistency constraints to saturated measurements. We also proposed a greedy algorithm for the latter approach. Both approaches exploit the *democracy* property of measurements from randomized measurement systems. These approaches are not limited to time-varying signals, for example, they can be used with the single-pixel camera [10].

In our experimental results, we find that the given enough initial measurements, the rejection and consistent approaches outperform the conventional approach for quantization with saturation. We also find that best performance in these new methods occurs when the saturation rate is nonzero, implying that the gain for CS systems should be tuned to have a positive saturation rate, even when the distribution of the input and the sampling matrix ensures that the measurements are bounded and saturation can be avoided.

Our reconstruction approaches are not limited to quantization with saturation. Any application where highly corrupted measurements can be easily detected can employ similar techniques to those described in this paper. For instance, some sensors such as the photo-diode used in the single-pixel camera [10], have a linear regime that produces low distortion measurements and a non-linear regime that produces high distortion measurements.

Beyond proposing and demonstrating the benefits of our approaches, we also proved the claim that CS measurements are $\widetilde{M}$-democratic for a large class of random matrices. This means that once a $M \times N$ matrix is drawn, every $\widetilde{M} \times N$ submatrix has the RIP.

[2] BIBO = Bounded Input Bounded Output

The democracy property has applications that extend beyond the scope of this paper. For instance, it can be used to show that CS measurements are robust to erasure channels when using a similar transmission methodology as fountain codes [49] or when applying CS as an multiple description coding (MDC) [50] code.

References

[1] D. Healy. (2005) Analog-to-information. BAA #05-35. [Online]. Available: http://www.darpa.mil/mto/solicitations/baa05-35/s/index.html

[2] R. Walden, "Analog-to-digital converter survey and analysis," *IEEE J. Selected Areas in Comm.*, vol. 17, no. 4, pp. 539–550, 1999.

[3] D. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, vol. 6, no. 4, pp. 1289–1306, 2006.

[4] E. Candès, "Compressive sampling," in *Proc. Int. Congress Math.*, Madrid, Spain, Aug. 2006.

[5] M. Vetterli, P. Marziliano, and T. Blu, "Sampling signals with finite rate of innovation," *IEEE Trans. Signal Processing*, vol. 50, no. 6, pp. 1417–1428, 2002.

[6] J. Laska, S. Kirolos, M. Duarte, T. Ragheb, R. Baraniuk, and Y. Massoud, "Theory and implementation of an analog-to-information converter using random demodulation," in *Proc. IEEE Int. Symp. Circuits and Systems (ISCAS)*, New Orleans, LA, May 2007.

[7] J. Tropp, J. Laska, M. Duarte, J. Romberg, and R. Baraniuk, "Beyond Nyquist: Efficient sampling of sparse, bandlimited signals," *to appear in IEEE Trans. Inform. Theory*, 2009.

[8] J. Romberg, "Compressive sensing by random convolution," *to appear in SIAM J. Imaging Sciences*, 2009.

[9] J. Tropp, M. Wakin, M. Duarte, D. Baron, and R. Baraniuk, "Random filters for compressive sampling and reconstruction," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Toulouse, France, May 2006.

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

[11] R. Robucci, L. Chiu, J. Gray, J. Romberg, P. Hasler, and D. Anderson, "Compressive sensing on a CMOS separable transform image sensor," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Las Vegas, NV, Apr. 2008.

[12] R. Marcia, Z. Harmany, and R. Willett, "Compressive coded aperture imaging," in *Proc. SPIE Symp. Elec. Imaging: Comput. Imaging*, San Jose, CA, Jan. 2009.

[13] Y. Eldar and M. Mishali, "Robust recovery of signals from a structured union of subspaces," *to appear in IEEE Trans. Inform. Theory*, 2009.

[14] M. Mishali, Y. Eldar, and J. Tropp, "Efficient sampling of sparse wideband analog signals," in *Proc. Conv. IEEE in Israel (IEEEI)*, Eilat, Israel, Dec. 2008.

[15] M. Mishali and Y. Eldar, "From theory to practice: Sub-Nyquist sampling of sparse wideband analog signals," *Preprint*, 2009.

[16] L. Jacques, D. Hammond, and M. Fadili, "Dequantizing compressed sensing: When oversampling and non-gaussian contraints combine," *Preprint*, 2009.

[17] W. Dai, H. Pham, and O. Milenkovic, "Distortion-rate functions for quantized compressive sensing," *Preprint*, 2009.

[18] A. Zymnis, S. Boyd, and E. Candès, "Compressed sensing with quantized measurements," *Preprint*, 2009.

[19] J. Sun and V. Goyal, "Quantization for compressed sensing reconstruction," in *Proc. Sampling Theory and Applications (SampTA)*, Marseille, France, May 2009.

[20] E. Candès and T. Tao, "The Dantzig selector: Statistical estimation when p is much larger than n," *Annals of Statistics*, vol. 35, no. 6, pp. 2313–2351, 2007.

[21] R. Carrillo, K. Barner, and T. Aysal, "Robust sampling and reconstruction methods for compressed sensing," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Taipei, Taiwan, Apr. 2009.

[22] J. Laska, M. Davenport, and R. Baraniuk, "Exact signal recovery from corrupted measurements through the pursuit of justice," in *Proc. Asilomar Conf. on Signals Systems and Computers*, Asilomar, CA, Nov. 2009.

[23] Z. Harmany, R. Marcia, and R. Willett, "Sparse poisson intensity reconstruction algorithms," in *Proc. IEEE Work. Stat. Signal Processing (SSP)*, Cardiff, Wales, Aug. 2009.

[24] I. Rish and G. Grabarnik, "Sparse signal recovery with exponential-family noise," in *Proc. Allerton Conf. Comm., Control, and Comput.*, Monticello, IL, Sept. 2009.

[25] J. Triechler, *Personal Communication*, Oct. 2009.

[26] G. Gray and G. Zeoli, "Quantization and saturation noise due to analog-to-digital conversion," *IEEE Trans. Aerospace and Elec. Systems*, vol. 7, no. 1, pp. 222–223, 1971.

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

[28] 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.

[29] E. Candès, "The restricted isometry property and its implications for compressed sensing," *Comptes rendus de l'Académie des Sciences, Série I*, vol. 346, no. 9-10, pp. 589–592, 2008.

[30] D. Needell and J. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Appl. Comput. Harmon. Anal.*, vol. 26, no. 3, pp. 301–321, 2009.

[31] T. Blumensath and M. Davies, "Iterative hard thresholding for compressive sensing," *Appl. Comput. Harmon. Anal.*, vol. 27, no. 3, pp. 265–274, 2009.

[32] P. Wojtaszczyk, "Stability and instance optimality for Gaussian measurements in compressed sensing," *to appear in Found. Comput. Math.*, 2009.

[33] J. Treichler, M. Davenport, and R. Baraniuk, "Application of compressive sensing to the design of wideband signal acquisition receivers," in *U.S./Australia Joint Work. Defense Apps. of Signal Processing (DASP)*, Lihue, Hawaii, Sept. 2009.

[34] J. Laska, P. Boufounos, and R. Baraniuk, "Finite-range scalar quantization for compressive sensing," in *Proc. Sampling Theory and Applications (SampTA)*, Marseille, France, May 2009.

[35] M. Davenport, M. Duarte, M. Wakin, J. Laska, D. Takhar, K. Kelly, and R. Baraniuk, "The smashed filter for compressive classification and target recognition," in *Proc. SPIE Elec. Imaging: Comput. Imaging*, San Jose, CA, Jan. 2007.

[36] P. Boufounos and R. Baraniuk, "1-bit compressive sensing," in *Proc. Conf. Inform. Science and Systems (CISS)*, Princeton, NJ, Mar. 2008.

[37] A. Calderbank and I. Daubechies, "The pros and cons of democracy," *IEEE Trans. Inform. Theory*, vol. 48, no. 6, 2002.

[38] S. Güntürk, "Harmonic analysis of two problems in signal compression," Ph.D. dissertation, Program in Applied and Computation Mathematics, Princeton University, Princeton, NJ, Sept. 2000.

[39] E. Candès, "Integration of sensing and processing," *IMA Annual Program Year Work.*, Dec. 2005.

[40] M. Davenport, P. Boufounos, and R. Baraniuk, "Compressive domain interference cancellation," in *Structure et parcimonie pour la représentation adaptative de signaux (SPARS)*, Saint-Malo, France, Apr. 2009.

[41] F. Beutler, "Error-free recovery of signals from irregularly spaced samples," *SIAM Rev.*, vol. 8, pp. 328–335, July 1966.

[42] A. Aldroubi and K. Gröchenig, "Nonuniform sampling and reconstruction in shift-invariant spaces," *SIAM Rev.*, vol. 43, no. 4, pp. 585–620, 2001.

[43] E. Candès, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Comm. Pure and Appl. Math.*, vol. 59, no. 8, pp. 1207–1223, 2006.

[44] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming," Feb. 2009, available online at http://stanford.edu/ boyd/cvx.

[45] ——, "Graph implementations for nonsmooth convex programs," in *Recent Advances in Learning and Control*, ser. Lecture Notes in Control and Information Sciences, V. Blondel, S. Boyd, and H. Kimura, Eds. Springer, 2008, pp. 95–110.

[46] R. DeVore, B. Jawerth, and B. Lucier, "Image compression through wavelet transform coding," *IEEE Trans. Inform. Theory*, vol. 38, no. 2, 1992.

[47] M. Mishali and Y. Eldar, "Blind multi-band signal reconstruction: compressed sensing for analog signals," *IEEE Trans. Signal Processing*, vol. 57, no. 3, pp. 993–1009, 2009.

[48] A. Oppenheim and A. Willsky, *Signals and systems*. Prentice-Hall, 1996.

[49] D. MacKay, "Fountain codes," *IEE Proc. Comm.*, vol. 152, no. 6, pp. 1062–1068, 2005.

[50] V. Goyal, "Multiple description coding: Compression meets the network," *IEEE Signal Processing Mag.*, vol. 18, no. 5, 2001.