Appl. Comput. Harmon. Anal. 31 (2011) 429–443

Contents lists available at ScienceDirect

Applied and Computational Harmonic Analysis

www.elsevier.com/locate/acha

Democracy in action: Quantization, saturation, and compressive sensing ☆

Jason N. Laska [a,*], Petros T. Boufounos [b], Mark A. Davenport [c], Richard G. Baraniuk [a]

[a] Department of Electrical and Computer Engineering, Rice University, Houston, TX 77005, USA
[b] Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA 02139, USA
[c] Department of Statistics, Stanford University, Stanford, CA 94305, USA

ARTICLE INFO

Article history:
Received 14 September 2010
Revised 3 February 2011
Accepted 14 February 2011
Available online 22 February 2011
Communicated by Charles K. Chui

Keywords:
Compressive sensing
Quantization
Saturation
Consistent reconstruction

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. To date, most of the CS literature has been devoted to studying the recovery of sparse signals from a small number of linear measurements. In this paper, we study more practical CS systems where the measurements are *quantized* to a finite number of bits; in such systems some of the measurements typically *saturate*, causing significant nonlinearity and potentially unbounded errors. We develop two general approaches to sparse signal recovery in the face of saturation error. The first approach merely rejects saturated measurements; the second approach factors them into a conventional CS recovery algorithm via convex consistency constraints. To prove that both approaches are capable of stable signal recovery, we exploit the heretofore relatively unexplored property that many CS measurement systems are *democratic*, in that each measurement carries roughly the same amount of information about the signal being acquired. A series of computational experiments indicate that the signal acquisition error is minimized when a significant fraction of the CS measurements is allowed to saturate (10–30% in our experiments). This challenges the conventional wisdom of both conventional sampling and CS.

1. Introduction

Analog-to-digital converters (ADCs) interface the analog physical world, where many signals originate, with the digital world, where they can be efficiently analyzed and processed. Given a one-dimensional analog input signal $x_a(t)$ that is a function of the continuous time index $t \in \mathbb{R}$, a conventional ADC first periodically *samples* it to create a signal $x(n) = x_a(nT)$ that is a function of the discrete time index $n \in \mathbb{Z}$ and then *quantizes* the samples so that each can be represented by a finite number of digital bits. Two-dimensional ADCs for, say, image data operate analogously with $t \in \mathbb{R} \times \mathbb{R}$ and $n \in \mathbb{Z} \times \mathbb{Z}$. The classical Shannon–Nyquist sampling theorem states that when the sampling frequency $\frac{2\pi}{T}$ is greater than the bandwidth of x_a's Fourier transform then no information is lost in the sampling process, meaning that $x_a(t)$ can be exactly recovered from $x(n)$ via cardinal sine (sinc) interpolation.

As digital computers 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

☆ This work was supported by the grants NSF CCF-0431150, CCF-0728867, CCF-0926127, CNS-0435425, CNS-0520280, and DMS-1004718, 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 and FA9550-09-1-0432, ARO MURI W911NF-07-1-0185 and W911NF-09-1-0383, and the Texas Instruments Leadership University Program.

* Corresponding author.
E-mail addresses: laska@rice.edu (J.N. Laska), petrosb@merl.com (P.T. Boufounos), markad@stanford.edu (M.A. Davenport), richb@rice.edu (R.G. Baraniuk).

1063-5203/$ – see front matter
doi:10.1016/j.acha.2011.02.002

sampling rates are pushed higher, they approach a physical barrier, beyond which their design becomes increasingly difficult and costly [2].

Fortunately, recent mathematical developments in computational harmonic analysis, in particular *compressive sensing* (CS), have enabled a new approach to analog-to-digital conversion that has the potential to keep pace with demand for certain kinds of signals [3,4]. CS provides a framework for sampling signals at a rate proportional to their information content rather than their Fourier bandwidth. In CS, a signal's information content is quantified as the number of nonzero coefficients in a known transform basis over some fixed time interval [5]. Signals sporting few nonzero coefficients are called *sparse*; approximately sparse signals with coefficient magnitudes that decay rapidly are called *compressible*.

A compressive sampler computes more general linear measurements than the periodic samples of a conventional ADC. Over some fixed interval of time (or region of space), we obtain

$$y = F(x_a) = \Phi x \tag{1}$$

where y is a vector of M compressive measurements, F is a sampling operator, and Φ is the equivalent $M \times N$ matrix that operates on the vector of N Nyquist-rate samples, x. The CS theory states that if the signal x is sparse, then under certain conditions on the sampling operator, x (and hence the analog signal x_a) can be recovered exactly from the compressive measurements y. Curiously, CS systems typically exploit a degree of *randomness* in order to provide theoretical performance guarantees [3,4].

Several CS hardware architectures for acquiring signals, images, and videos have been proposed, analyzed, and in some cases implemented [6–13]. An important aspect of these hardware architectures that is not typically dealt with in the mathematical CS literature is that practical hardware must *quantize* the measurements y from (1). That is, each real-valued entry of the vector y is mapped to a finite set of values that can be encoded using a finite number of digital bits (we will deal only with this so-called finite-range scalar quantization in this paper).

Quantization introduces two kinds of errors into the CS framework. The first kind, *quantization error*, is due to the fact that intervals of real-valued measurement values are mapped to the same digital bit string. The effect and mitigation of quantization error on CS recovery have been explored in several works [14–17]. The second kind, *saturation error*, is due to the fact that a finite number of digital bits can represent only a limited dynamic range of measurement values; arbitrarily larger positive or smaller negative values that are out of the range of the quantizer are unrepresentable and typically assigned to the digital codes corresponding to the upper and lower limits of the quantizer, respectively. The effect and mitigation of saturation error on CS recovery are relatively unexplored and are the primary subject of this paper.

Unlike quantization error, the challenge with saturation error is that, in the absence of an *a priori* upper bound on the measurement magnitudes, it can be unbounded. Most current CS recovery algorithms provide guarantees only for noise that is bounded or bounded with high probability (for example, Gaussian noise) [18]. Several recovery techniques have been developed for sparse or impulsive noise [19,20] and unbounded exponential noise [21,22]; however none of these techniques will tolerate unbounded quantization error due to saturation.

The naïve approach to dealing with saturation is to scale down the amplitude of the signal or its measurements so that saturation never or very rarely occurs. This is the approach pursued in many conventional sensor systems; a typical rule of thumb used with communication system ADCs suggests that one reduce the signal amplitude until only 63 in one million samples saturates [23]. Unfortunately, scaling down the signal amplitude proportionately scales up the amount of quantization noise.

In this paper, we develop and study two approaches for mitigating unbounded saturation errors in CS signal acquisition and recovery. In contrast to previous work, both approaches allow a nontrivial number of measurements to saturate. The first technique, *saturation rejection* simply discards any saturated measurements before performing a conventional CS signal recovery. The second technique, *saturation consistency*, factors the saturation phenomenon into new recovery algorithms via convex inequality constraints. We analyze both approaches below and show that they recover sparse and compressible signals with guarantees similar to standard CS recovery algorithms.

Our analysis of both approaches is based on the heretofore relatively unexplored *democracy* property of CS measurements, namely that each measurement y_i carries roughly the same amount of information about the entirety of the signal x. In saturation rejection, democracy enables us to reject all saturated measurements and yet still achieve accurate signal recovery so long as "enough" unsaturated measurements remain. In saturation consistency, democracy enables us to factor in saturated measurements through weak inequality constraints to attain a more accurate signal recovery.

While democracy has been alluded to in several works related to CS and frame theory [24–27], we are aware of no hard analytical results in CS. Therefore, we carefully define democracy and characterize the number of measurements/rows M required for a CS matrix Φ to have this property. Our results are central to the analysis of the above two recovery algorithms, but are also of independent interest.

In addition to a careful analytical treatment of democracy and the recovery performance of our algorithms, we also conduct a series of computational experiments to validate the basic concepts and test their limits. Broadly speaking, we find that the signal acquisition error is minimized when a significant fraction of the CS measurements is allowed to saturate (10–30% in our experiments). This challenges the conventional wisdom of both conventional sampling and CS.

The organization of this paper is as follows. Section 2 reviews quantization, saturation, and the key concepts of the CS framework. Section 3 develops our two new recovery algorithms for unbounded saturation in CS, introduces a strong

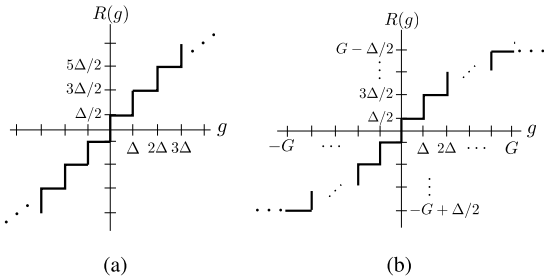


Fig. 1. (a) Midrise scalar quantizer. (b) Finite-range midrise scalar quantizer with saturation level G.

definition of democracy, and validates these algorithms for democratic matrices. Section 4 demonstrates that a large class of matrices satisfy our strong definition of democracy. Section 5 tests and validates our claims experimentally. Section 6 closes the paper with a discussion regarding how democracy can be useful in other applications.

2. Background

2.1. Analog-to-digital conversion

An analog-to-digital converter (ADC) performs two discretization steps: *sampling*, which converts a function $x_a(t)$ of a continuous variable $t \in \mathbb{R}$ to a function $x(n) = x_a(nT)$ of a discrete variable $n \in \mathbb{Z}$, followed by *quantization*, which converts the real-numbered value of each measurement $x(n)$ to a discrete value $R(x(n))$ chosen from a pre-determined, finite set. Classical results due to Shannon, Nyquist, and others have demonstrated that when the sampling frequency $\frac{2\pi}{T}$ is greater than the bandwidth of x_a's Fourier transform no information is lost in the sampling process, meaning that $x_a(t)$ can be exactly recovered from $x(n)$ via cardinal sine (sinc) interpolation. Quantization, on the other hand, generally results in an irreversible loss of information.

2.2. Scalar quantization

In this paper, without loss of generality, we focus on scalar uniform symmetrical quantizers with quantization bin width Δ. Such a quantizer first takes a real-valued input value g (typically $g = x(n)$) and assigns it to the nearest quantized value $R(g)$ from the discrete set $\{q_k = q_0 + k\Delta,\ k \in \mathbb{Z}\}$. The choice $q_0 = \Delta/2$ corresponds to a so-called "midrise" quantizer, while $q_0 = 0$ corresponds to a "midtread" quantizer. We will emphasize midrise quantizers in this paper. Note that the maximum quantization error can be bounded as $|g - R(g)| \leqslant \Delta/2$. See Fig. 1(a) for a graphical depiction of a midrise quantizer.

In practice, quantizers have a finite *dynamic range* that is dictated by the voltage limits of hardware devices and the desire to use a finite number of bits B to represent the set of quantized values. A *finite-range quantizer* of B bits can uniquely represent up to 2^B quantization levels q_k. Inputs g of magnitude larger than $G - \Delta$ where $G := \Delta 2^{B-1}$ *saturate* and incur a potentially unbounded error $|g - R(g)|$ [28]. See Fig. 1(b) for a graphical depiction of a midrise quantizer with finite dynamic range.

2.3. Compressive sensing (CS)

CS is a new approach to signal sampling that aims to reproduce Shannon–Nyquist performance using a smaller number of samples. Over some fixed interval of time (or region of space), we obtain the CS measurements as follows

$$y = F(x_a) + e = \Phi x + e \tag{2}$$

where y again represents the vector of M compressive measurements, F is the sampling operator, and Φ is the equivalent $M \times N$ matrix that operates on the vector of N Nyquist-rate samples, x. In contrast to the over-simplified model in (1), (2) includes an $M \times 1$ vector e that represents measurement errors. We now briefly review the CS theory; readers interested in a more detailed treatment can see [3,4].

If x is K-sparse when represented in a *sparsity basis* Ψ, i.e., $x = \Psi\alpha$ with $\|\boldsymbol{\alpha}\|_0 := |\mathrm{supp}(\alpha)| \leqslant K$, then one can acquire just $M = O(K \log(N/K))$ measurements and still stably recover the signal x [3,4]. A similar guarantee holds for approximately sparse, or *compressible* signals whose coefficients α decay with a power law when sorted; that is $|\alpha_n| \propto n^{-\frac{1}{p}}$, $p \leqslant 1$. Observe that when K is small, the number of measurements M required can be significantly smaller than the number of samples N at the Shannon–Nyquist rate.

In order to establish these results, it is necessary to make some assumptions concerning the measurement matrix Φ. While there are many possibilities, one of the most common and powerful assumptions is the *restricted isometry property* [29].

Definition 1. A matrix Φ satisfies the RIP of order K with constant $\delta \in (0,1)$ if

$$(1-\delta)\|x\|_2^2 \leqslant \|\Phi x\|_2^2 \leqslant (1+\delta)\|x\|_2^2 \tag{3}$$

holds for all x such that $\|x\|_0 \leqslant K$.

In words, Φ acts as an approximate isometry on the set of vectors that are K-sparse in the basis Ψ. An important result is that for any unitary matrix Ψ, if we draw a random matrix Φ whose entries ϕ_{ij} are independent realizations from a sub-Gaussian distribution, then $\Phi\Psi$ will satisfy the RIP of order K with high probability provided that $M = O(K\log(N/K))$ [30]. In this paper, without loss of generality, we fix $\Psi = \mathbf{I}$, the identity matrix, implying that $x = \alpha$.

The lower bound in the RIP is a necessary condition if we wish to be able to recover all sparse signals x from the measurements y. Specifically, if $\|x\|_0 = K$, then Φ must satisfy the lower bound of the RIP of order $2K$ with $\delta < 1$ in order to ensure that any algorithm can recover x from the measurements 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 $\|e\|_2 \leqslant \epsilon$, the convex program

$$\hat{x} = \underset{x}{\operatorname{argmin}} \|x\|_1 \text{ s.t. } \|\Phi x - y\|_2 \leqslant \epsilon \tag{4}$$

can recover a sparse or compressible signal x. The following theorem makes this notion precise by bounding the recovery error of x with respect to the measurement noise norm, denoted by ϵ, and with respect the best approximation of x by its largest K terms, denoted by x_K.

Theorem 1. *(See Theorem 1.2 of [31].) Suppose that Φ satisfies the RIP of order $2K$ with $\delta < \sqrt{2}-1$. Given measurements of the form $y = \Phi x + e$, where $\|e\|_2 \leqslant \epsilon$, the solution to* (4) *obeys*

$$\|\hat{x} - x\|_2 \leqslant C_0\epsilon + C_1\frac{\|x - x_K\|_1}{\sqrt{K}}, \tag{5}$$

where

$$C_0 = \frac{4\sqrt{1+\delta}}{1-(\sqrt{2}+1)\delta}, \qquad C_1 = 2\frac{1+(\sqrt{2}-1)\delta}{1-(\sqrt{2}+1)\delta}. \tag{6}$$

While convex optimization programs such as (4) are powerful methods 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 [32] and iterative hard thresholding (IHT) [33,34] are known to satisfy similar guarantees under slightly stronger assumptions on the RIP constants. Furthermore, alternative recovery strategies based on (4) have been analyzed in [18,35]. These methods replace the constraint in (4) with an alternative constraint motivated by the assumption that the measurement noise is Gaussian in the case of [18] and that it is agnostic to the value of ϵ in the case of [35].

Finally we note that several hardware architectures have been proposed and implemented to perform CS in practical settings with analog signals. Examples include the random demodulator, compressive multiplexer, random filtering, and random convolution for one-dimensional time signals [7–9,13,36] and several compressive imaging architectures for two-dimensional images [10–12].

3. Signal recovery from saturated measurements

3.1. Unbounded saturation error

In a CS system with finite-range quantization, our task is to recover the signal x from the quantized measurements

$$y = R(\Phi x + e) = \Phi x + w, \tag{7}$$

where w represents the combination of the CS measurement error, quantization error, and saturation error. Standard CS recovery approaches such as (4) assume that the error w is bounded, which ensures that the original signal x is a feasible solution to the optimization program. However, as we saw above in Section 2.2, without restrictions on x, certain entries of w can be unbounded. Scaling down the measurements to avoid saturation is problematic, since rescaling increases the quantization error on every measurement that does not saturate. And, indeed, in many applications, saturation events are impossible to avoid completely. In other words, like the conventional wisdom on saturated fats, saturated measurements are bad for you.

Therefore, in this section, we propose two new CS recovery algorithms that are resilient to measurement saturation:

- **saturation rejection:** simply discard saturated measurements and then perform signal recovery on those that remain;

- **saturation consistency:** incorporate saturated measurements into standard recovery algorithms through convex inequality constraints that enforce consistency on the saturated measurements.

While both of these approaches involve intuitive modifications of standard CS recovery algorithms, it is *not* obvious that they are guaranteed to work. For instance, in order for saturation rejection to work we must be able to recover the signal using only the unsaturated measurements that are retained, or equivalently, using only the rows of Φ that are retained. An analysis of the properties of this submatrix will be essential to understanding the performance of this approach. Similarly, it is 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, which we prove below in Section 4, is that there exists a class of matrices Φ such that an arbitrary subset of their rows will indeed satisfy the RIP. These *democratic* matrices hold the key to the performance guarantees for our two approaches.

We briefly establish some notation that will prove useful for the remainder of this paper. Let $\Gamma \subset \{1, 2, \ldots, M\}$. By Φ^Γ we mean the $|\Gamma| \times M$ matrix obtained by selecting the rows of Φ indexed by Γ. Alternatively, if $\Lambda \subset \{1, 2, \ldots, N\}$, then we use Φ_Λ to indicate the $M \times |\Lambda|$ matrix obtained by selecting the columns of Φ indexed by Λ. Denote the vector of unsaturated measurements as $\tilde{y}$ of length $\widetilde{M}$. The matrix $\widetilde{\Phi}$ is created by selecting the rows of Φ corresponding to the elements of $\tilde{y}$.

3.2. Saturation rejection signal recovery

A simple and intuitive way to handle saturated measurements is to simply discard them and then run a standard CS recovery algorithm [37]. Using, for instance, (4) for reconstruction yields the program:

$$\hat{x} = \underset{x}{\operatorname{argmin}} \|x\|_1 \text{ s.t. } \|\widetilde{\Phi}x - \tilde{y}\|_2 < \epsilon. \tag{8}$$

Part of the intuition for the approach is provided by the fact that, while $\|w\|_2$ is potentially unbounded, $\|\widetilde{w}\|_2$ can be bounded as is shown in the following lemma.

Lemma 1. *If $\|e\|_2 \leqslant \epsilon'$, then $\|\widetilde{w}\|_2 \leqslant \epsilon = \sqrt{\widetilde{M}}(\Delta/2) + \epsilon'$.*

Proof. First recall that $w = R(\Phi x + e) - \Phi x$. Let Γ denote the set of indices corresponding to the measurements that do not saturate, so that $\widetilde{w} = w_\Gamma = R(\Phi^\Gamma x + e_\Gamma) - \Phi^\Gamma x$. Thus we have that

$$\begin{aligned}
\|\widetilde{w}\|_2 &= \left\| R\left(\Phi^\Gamma x + e_\Gamma\right) - \left(\Phi^\Gamma x + e_\Gamma\right) + e_\Gamma \right\|_2 \\
&\leqslant \left\| R\left(\Phi^\Gamma x + e_\Gamma\right) - \left(\Phi^\Gamma x + e_\Gamma\right) \right\|_2 + \|e_\Gamma\|_2 \\
&\leqslant \sqrt{\widetilde{M}}(\Delta/2) + \|e\|_2 \leqslant \sqrt{\widetilde{M}}(\Delta/2) + \epsilon,
\end{aligned}$$

as desired. □

Saturation rejection is also useful in conjunction with processing and inference techniques that work directly on the compressive measurements. For example, in the *smashed filter* for signal detection and classification the key calculation is the inner product $\langle \Phi x, \Phi \zeta \rangle$ between the compressive measurements of a test signal x and a target template signal ζ [38]. If x and ζ are sparse then, thanks to the RIP, this low-dimensional inner product can be used as a proxy for the inner product between x and ζ; that is $\langle \Phi x, \Phi \zeta \rangle \approx \langle x, \zeta \rangle$. Unfortunately, if any of the elements of Φx or $\Phi \zeta$ are saturated, then the approximation no longer holds and the performance of the smashed filter deteriorates.

Specifically, to illustrate the impact of quantization suppose that $e = 0$. Consider $R(\Phi x)$ and $R(\Phi \zeta)$ and let Γ_x and Γ_ζ be the supports of the measurements that do not saturate on each vector, respectively. Then we have that for $\Gamma = \Gamma_x \cap \Gamma_\zeta$ that $\|R(\Phi^\Gamma x) - \Phi^\Gamma x\|_\infty \leqslant \Delta/2$ and $\|R(\Phi^\Gamma \zeta) - \Phi^\Gamma \zeta\|_\infty \leqslant \Delta/2$. Thus, it is straightforward to show that

$$\left|\left\langle R\left(\Phi^\Gamma x\right), R\left(\Phi^\Gamma \zeta\right)\right\rangle - \left\langle \Phi^\Gamma x, \Phi^\Gamma \zeta \right\rangle\right| \leqslant \frac{\Delta^2}{4} + \frac{\Delta}{2}\left|\sum_n \left(\Phi^\Gamma x\right)_n\right| + \frac{\Delta}{2}\left|\sum_n \left(\Phi^\Gamma \zeta\right)_n\right|. \tag{9}$$

Furthermore, the two sums in (9) are likely to concentrate around zero for the class of matrices studied in this paper. The results of Section 4 will furthermore imply that $\langle \Phi^\Gamma x, \Phi^\Gamma \zeta \rangle \approx \langle x, \zeta \rangle$. Thus, discarding the corresponding entries of Φx and $\Phi \zeta$ when one of them saturates makes considerable practical sense.

3.3. Saturation consistency signal recovery via convex optimization

Clearly saturation rejection discards potentially useful signal information, since we know that saturated measurements are large (we just do not know *how* large). In our second approach, we augment a standard convex optimization-based CS recovery algorithm with a set of inequality constraints that enforce signal *consistency* with the saturated measurements.

Algorithm 1 SC-CoSaMP greedy algorithm

Input: y, Φ, and K
Initialize: $\hat{x}^{[0]} \leftarrow \mathbf{0}$, $n \leftarrow 0$
while not converged **do**
  Compute proxy:
  $p \leftarrow \widetilde{\Phi}^T(\tilde{y} - \widetilde{\Phi}\hat{x}^{[n]}) + \overline{\Phi}^T((G-\Delta)\cdot\mathbf{1} - \overline{\Phi}\hat{x}^{[n]})_+$
  Update coefficient support:
  $\Omega \leftarrow$ union of support of largest $2K$ coefficients from p and support of $\hat{x}^{[n]}$
  Estimate new coefficient values:
  $\hat{x}^{[n+1]} \leftarrow \operatorname{argmin}_x \|\tilde{y} - \widetilde{\Phi}_\Omega x\|_2^2 + \|((G-\Delta)\cdot\mathbf{1} - \overline{\Phi}_\Omega x)_+\|_2^2$
  Prune:
  $\hat{x}^{[n+1]} \leftarrow$ keep largest K coefficients of $\hat{x}^{[n+1]}$
  $n \leftarrow n+1$
end while

That is, we constrain the recovered signal $\hat{x}$ so that the magnitudes of the values of $\Phi\hat{x}$ corresponding to the saturated measurements are larger than $G-\Delta$.

More specifically, let S^+ and S^- correspond be the index sets of the positive saturated measurements and negative saturated measurements, respectively. Define the matrix $\overline{\Phi}$ as

$$\overline{\Phi} := \begin{bmatrix} \Phi^{S^+} \\ -\Phi^{S^-} \end{bmatrix}. \tag{10}$$

We estimate $\hat{x}$ via the program

$$\hat{x} = \underset{x}{\operatorname{argmin}} \|x\|_1 \quad \text{s.t.} \quad \|\widetilde{\Phi}x - \tilde{y}\|_2 < \epsilon \tag{11a}$$

$$\text{and} \quad \overline{\Phi}x \geqslant (G-\Delta)\cdot\mathbf{1}, \tag{11b}$$

where $\mathbf{1}$ denotes an $(M-\widetilde{M})\times 1$ vector of ones. In words, we seek the 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 $\tilde{y}$, such as those proposed in [14,15], or alternative techniques for the unsaturated quantized measurements can be used such as those proposed in [16].

In some ADC hardware implementations, the measurements directly following a saturation event can have higher distortion than the other unsaturated measurements. In this case, an additional ℓ_2 constraint can be applied, such as $\|\widetilde{\Phi}^\Gamma x - \tilde{y}_\Gamma\|_2 < \epsilon_1$, where Γ here denotes the indices of the measurements immediately following a saturation event and $\epsilon_1 > \epsilon$. The measurements $\tilde{y}_\Gamma$ can be determined via measured properties of the physical system.

3.4. Saturation consistency signal recovery via greedy algorithms

Greedy algorithms can also be customized to incorporate a saturation consistency constraint. In this subsection, we demonstrate how one popular greedy algorithm, CoSaMP [32], can be modified into *Saturation Consistent CoSaMP* (SC-CoSaMP).

We first briefly summarize CoSaMP [32]. The algorithm recovers a signal estimate $\hat{x}$ by iteratively finding a coefficient support set Ω and then estimating the signal coefficients over that support. To estimate the support Ω at a given iteration, we merge the support corresponding to the largest K coefficients of the signal estimate $\hat{x}$ from the previous iteration with the support corresponding to the largest $2K$ coefficients of the *proxy* vector $p = \Phi^T(y - \Phi\hat{x})$, where the superscript T denotes matrix transpose. We then estimate the signal coefficients by solving the least squares problem $\min_x \|\Phi_\Omega x - y\|_2^2$ over the column support Ω of Φ. These steps are performed successively until the algorithm converges.

We can incorporate saturation consistency into CoSaMP by modifying both the proxy step and the coefficient estimation step. A complete summary is described in Algorithm 1. Our specific modifications are as follows. We form the proxy vector p by computing the sum of two proxy vectors; a proxy from $\tilde{y}$ and a proxy that uses the supports of the saturated measurements. To compute the proxy from $\tilde{y}$, we repeat the same computation as in conventional CoSaMP. To compute the proxy from the support of the saturated measurements, we introduce the *saturation residual* $(G-\Delta)\cdot\mathbf{1} - \overline{\Phi}\hat{x}^{[n]}$. This vector measures how close the elements of $\overline{\Phi}\hat{x}$ are to $G-\Delta$. In consistent reconstruction, the magnitudes of the elements of $\overline{\Phi}\hat{x}$ should be greater than or equal to $G-\Delta$; thus, once these magnitudes are greater than $G-\Delta$, the saturation residual will be zero, contributing no bias to the coefficient location estimate. Thus, consistency is achieved by applying the function $(\cdot)_+$ that sets the negative elements of the saturation residual to zero and retains the positive elements. By combining the two proxies, the new proxy vector becomes

$$p = \widetilde{\Phi}^T\left(\tilde{y} - \widetilde{\Phi}\hat{x}^{[n]}\right) + \overline{\Phi}^T\left((G-\Delta)\cdot\mathbf{1} - \overline{\Phi}\hat{x}^{[n]}\right)_+. \tag{12}$$

In this arrangement, elements of $\overline{\Phi}\hat{x}$ having magnitude smaller than $G-\Delta$ will contribute new information to p; however, elements having magnitude larger than $G-\Delta$ will be set to zero and therefore will not contribute additional information to p. A similar modification can be made to the iterated hard thresholding CS recovery algorithm [33,34].

To reformulate the coefficient estimate step to include the saturation constraint, SC-CoSaMP finds the solution to

$$\hat{x}^{[n+1]} \leftarrow \underset{x}{\operatorname{argmin}} \|\tilde{y}-\widetilde{\Phi}_{\Omega}x\|_2^2 + \left\|\left((G-\Delta)\cdot\mathbf{1}-\overline{\Phi}_{\Omega}x\right)_+\right\|_2^2. \tag{13}$$

This can be achieved via gradient descent or other optimization techniques. By employing the one-sided quadratic in the objective of (13), we ensure a soft application of the constraint (11b) and ensure that the program is feasible even in the presence of noise (for more details, see [39]).

In practice, we have found that the proxy step modification provides a significant increase in performance over the equivalent step in CoSaMP, while the coefficient estimation modification provides only a marginal performance increase.

3.5. Democracy and signal recovery with saturation

To demonstrate that our proposed algorithms will recover signals from saturated CS measurements, we introduce a strong notion of democracy that quantifies the intuition that each measurement contributes a similar amount of information about the signal x to the compressed representation y [24,25,27]. We then demonstrate that if Φ is democratic, then the algorithms developed above can be used for signal recovery. In the next section we will further prove that the random measurement schemes typically advocated for usage in CS are democratic. Our definition and analysis significantly strengthen the informal (and weak) notion of democracy in the existing literature.[1]

The fact that random measurements are democratic seems intuitively obvious: when the entries of Φ are chosen at random, each measurement is a randomly weighted sum of a large fraction (or all) of the entries of x, and since the weights are typically chosen independently, no preference is given to any particular set of entries. 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 Φ 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 y). Then we are left with a length-$\widetilde{M}$ measurement vector $\tilde{y}$ such that each $\tilde{y}_i \sim f_Y$. 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 then deleting D of them, provided that this deletion is done *independently* of the actual values of 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 not independently but by an adversary who can see the measurements. By adaptively deleting the entries of y one can change the distribution of $\tilde{y}$. For example, the adversary can delete the D largest elements of y, thereby skewing the distribution of $\tilde{y}$. This means that $\|\tilde{y}\|_2$ could potentially significantly deviate from $\|x\|_2$, or in other words, could potentially delete the pertinent information about the signal. We can alternatively view the adversary as retaining only measurements that are close to zero. Then x could potentially lie close to the nullspace of $\widetilde{\Phi}$. In many cases, especially if the same matrix Φ is 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.

Since our proposed algorithms do, in fact, remove the largest magnitude measurements, we need a significantly strong definition of democracy that allows us to determine how many measurements can be discarded without losing valuable information about the signal.

Definition 2. Let Φ be an $M \times N$ matrix, and let $\widetilde{M} \leqslant M$ be given. The matrix Φ is $(\widetilde{M}, K, \delta)$-democratic if, for all Γ such that $|\Gamma| \geqslant \widetilde{M}$, the matrix Φ^{Γ} satisfies the RIP of order K with constant δ.

If Φ is $(\widetilde{M}, 2K, \delta)$-democratic, then we can show that a *saturation rejection* algorithm will stably recover sparse and compressible signals in the face of saturated measurements. For example, for the convex optimization example given by (8), we have the following lemma.

Lemma 2. *Let Φ be $(\widetilde{M}, 2K, \delta)$-democratic with $\delta < \sqrt{2}-1$, let $\widetilde{\Phi}$ be any $\widetilde{M} \times N$ submatrix of Φ, and let $\widetilde{w}$ be the corresponding subvector of w. If $\|\widetilde{w}\|_2 < \epsilon$, then the solution to* (8) *obeys* (5)*.*

Proof. Since the democracy property implies that any $\widetilde{M} \times N$ submatrix of Φ has RIP, it immediately follows from Theorem 1 that the saturation rejection program (8) yields a signal estimate with the stability guarantee (5). □

By the same argument, it is straightforward to demonstrate that CoSaMP applied to $\widetilde{\Phi}$ and $\tilde{y}$ will achieve performance given by Theorem A in [32], since CoSaMP relies on the RIP for performance guarantees.

[1] The original introduction of the term democracy was in the context of quantization [24,25] in the sense that a democratic quantizer would ensure that each bit is given "equal weight." This notion was further developed in [26] wherein the authors proposed a simple vector quantization scheme via democratic "Kashin's representations." As the CS theory developed, it became empirically clear that CS systems exhibited this property with respect to compression [27].

Democracy can also be used to demonstrate that the saturation consistency algorithm (11) obeys the same reconstruction error bounds as (8).

Lemma 3. *Let Φ be $(\widetilde{M}, 2K, \delta)$-democratic with $\delta < \sqrt{2}-1$, let $\widetilde{\Phi}$ be any $\widetilde{M} \times N$ submatrix of Φ, and let $\widetilde{w}$ be the corresponding subvector of w. If $\|\widetilde{w}\|_2 < \epsilon$ and $\|e\|_2 \leqslant \epsilon'$, then the solution to* (11) *with $G' = G - \epsilon'$ substituted into* (11b) *obeys* (5)*.*

Proof. From Lemma 2, we have that (11) yields a signal estimate with the stability guarantee (5). This can be seen by observing that the proof of Theorem 1 in [31] essentially depends on only three facts: (i) that the original signal x is in the feasible set, so that we can conclude (ii) that $\|\hat{x}\|_1 \leqslant \|x\|_1$, and finally (iii) that $\|\Phi\hat{x} - \Phi x\|_2 \leqslant \epsilon$, where Φ can be any matrix that satisfies the RIP of order $2K$ with constant $\delta < \sqrt{2}-1$. Since Φ is democratic we have that (iii) holds for $\widetilde{\Phi}$ regardless of whether we incorporate the additional constraints. Since the original signal x will remain feasible in (11), (i) and (ii) will also hold. □

At this point it is useful to note that a saturation rejection algorithm and a saturation consistency algorithm will not necessarily yield the same signal estimate. This is because the solution from the rejection approach may not lie in the feasible set of solutions of the consistency approach (11). However, the reverse is true. The solution to the consistent approach does lie in the feasible set of solutions of the rejection approach. While we do not provide a detailed analysis that compares the performance of these two algorithm classes, one should expect that the consistency approach will outperform the rejection approach in general, since it incorporates additional information about the signal. We provide experimental confirmation of this in Section 5.

4. Random measurements and democracy

4.1. Random matrices are democratic

We now demonstrate that certain randomly generated matrices are democratic. While the theorem below can be extended (with different constants) to the more general class of *sub-Gaussian* matrices (see the methods in [40]), for simplicity we restrict our attention to Gaussian matrices.

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

$$M = C_1(K+D)\log\left(\frac{N+M}{K+D}\right), \tag{14}$$

then with probability exceeding $1 - 3e^{-C_2 M}$ we have that Φ 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 $A = [I\ \Phi]$ formed by appending Φ to the $M \times M$ identity matrix. Theorem 1 from [20] demonstrates that under the assumptions in the theorem statement, with probability exceeding $1 - 3e^{-C_2 M}$ we have that 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 Φ satisfy the RIP of order K with constant $\delta/(1-\delta)$.

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

$$P_\Lambda \triangleq A_\Lambda A_\Lambda^\dagger, \tag{15}$$

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

$$P_\Lambda^\perp \triangleq I - P_\Lambda, \tag{16}$$

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

$$P_\Lambda = I_\Lambda I_\Lambda^\dagger = I_\Lambda \left(I_\Lambda^T I_\Lambda\right)^{-1} I_\Lambda^T = I_\Lambda I_\Lambda^T = I(\Lambda),$$

where we use $I(\Lambda)$ 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 $I(\Lambda)$ from the $M \times D$ matrix I_Λ — 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

$$P_\Lambda^\perp = I - P_\Lambda = I(\Gamma).$$

Thus, we observe that the matrix $P_\Lambda^\perp A = I(\Gamma)A$ is simply the matrix A with zeros replacing all entries on any row i such that $i \notin \Gamma$, i.e., $(P_\Lambda^\perp A)^\Gamma = A^\Gamma$ and $(P_\Lambda^\perp A)^\Lambda = 0$. Furthermore, Theorem 5 from [41] states that for A satisfying the RIP of order $K + D$ with constant δ, we have that

$$\left(1 - \frac{\delta}{1-\delta}\right)\|u\|_2^2 \leqslant \left\|P_\Lambda^\perp A u\right\|_2^2 \leqslant (1+\delta)\|u\|_2^2 \tag{17}$$

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

4.2. Robustness and stability

Observe that we require roughly $O(D\log(N))$ additional measurements to ensure that Φ is $(\widetilde{M}, K, \delta)$-democratic compared to the number of measurements required to simply ensure that Φ satisfies the RIP of order K (recall that $D = M - \widetilde{M}$). 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 behavior is not unique to the CS framework. For instance, by *oversampling*, that is, sampling faster than the minimum required Nyquist rate, uniform sampling systems can also improve their robustness with respect to the loss of measurements. Recovery can be performed in principle on the remaining non-uniform grid of unsaturated samples, as long as the remaining samples satisfy the Nyquist rate on average [42]. However, signal recovery from a non-uniform grid is known to be unstable, and the required reconstruction kernels are difficult to compute. Even under conditions where stability is achievable, the associated computation can become expensive, and the set of samples that can be discarded is restricted [43]. In CS, on the other hand, democracy allows us to discard an arbitrary subset D of the set of CS measurements without compromising the stability of the recovery, independently of which measurements are discarded (even if they are discarded adversarially).

Theorem 2 further guarantees the graceful degradation of CS recovery due to loss of measurements. Specifically, the theorem implies that recovery 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 an $M \times N$ matrix Φ 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 Φ^Γ will satisfy the RIP of order $K - \widetilde{D}$ with constant δ. Thus, from [44], if we define $\widetilde{K} = (K - \widetilde{D})/2$, then the signal recovery error is bounded by

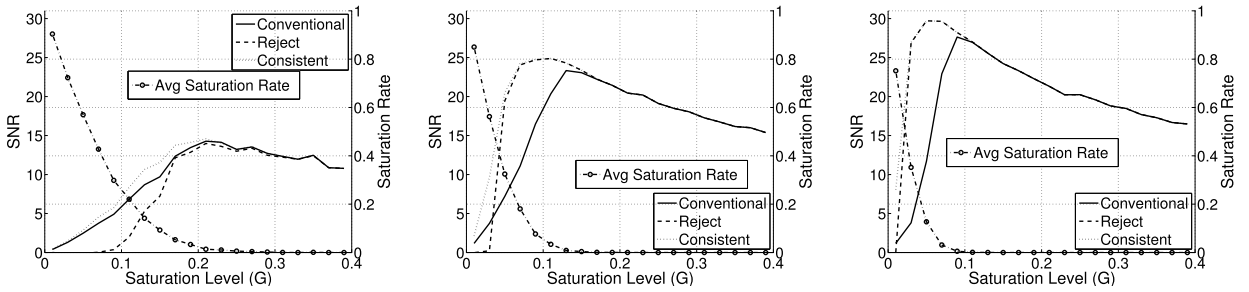
$$\|x - \hat{x}\|_2 \leqslant C_3 \frac{\|x - x_{\widetilde{K}}\|_1}{\sqrt{\widetilde{K}}}, \tag{18}$$

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

5. Simulations

We now conduct a series of computational experiments to validate the performance of the saturation rejection approach and the saturation consistency approach versus what we define as the *conventional approach*, which scales each signal so that its measurement saturation rate is zero and then reconstructs using either the program (4) or a greedy algorithm such as CoSaMP. Overall, we find that, on average, the consistency approach outperforms the other approaches for a given saturation level G. Note that for a scalar quantizer with a fixed number of bits per measurement, varying the quantizer saturation level G is exactly equivalent to varying the signal gain and keeping the saturation level G constant. Our particular findings include:

- In many cases the optimal performance for the rejection and consistency approaches is significantly superior to the optimal performance for the conventional approach, and this occurs when the saturation rate is manifestly nonzero.
- The difference in optimal performance between the rejection and consistency approaches is small for a given ratio of M/N.
- The consistency approach can achieve high recovery accuracy at higher saturation rates than the rejection approach.



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

Fig. 2. Comparison of recovery approaches in the face of measurement saturation using CVX for K-sparse signals with $N = 1024$, $K = 20$, and $B = 4$. Solid line depicts recovery SNR for the conventional approach via the program (4). Dotted line depicts recovery SNR for the consistency approach via the program (11). Dashed line depicts recovery SNR for the rejection approach via (8). SNR curves are measured on the left vertical axis. The dashed-circled line, measured on the right vertical axis, corresponds to the average saturation rate when averaged over 100 trials. Each plot represents a different measurement regime: (a) low $M/N = 2/16$, (b) medium $M/N = 6/16$, and (c) high $M/N = 15/16$.

5.1. Experimental setup

Signal model: Recall that, without loss of generality, we assume that the signal x is sparse in the canonical basis. We use two signal classes in the simulations:

- K-sparse: in each trial, K nonzero elements $x(n)$ at random locations are drawn from an i.i.d. Gaussian distribution;
- weak ℓ_p-compressible: in each trial, the elements $x(n)$ are generated according to the power law

$$x(n) = \theta_n n^{-1/p}, \tag{19}$$

where $p \leqslant 1$ and θ_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 add no additional noise.

Measurement matrix: For each trial we generate a measurement matrix from an i.i.d. Gaussian distribution with mean zero and variance $1/M$. Extensive additional experiments not presented here due to space considerations produce similar findings across a wide variety of measurement matrices, including i.i.d. ± 1 Rademacher matrices and other sub-Gaussian matrices, as well as the random demodulator [7] and random time-sampling [45].

Recovery algorithms: To ensure a fair comparison between algorithms and because of the theoretical recovery guarantees of (8) and (11), the simulations in Sections 5.2 and 5.3 were conducted using the general-purpose convex optimization package CVX [46,47] to implement (8) and (11). However, the simulations in Section 5.4 were conducted using SC-CoSaMP, since its faster running time was more conducive to the large number of trials required. In experiments not presented in this paper, we found that SC-CoSaMP performs with similar trends to (11) with CVX, but at significantly faster speeds.

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

$$\mathrm{SNR} \triangleq 10 \log_{10}\left(\frac{\|x\|_2^2}{\|x - \hat{x}\|_2^2}\right), \tag{20}$$

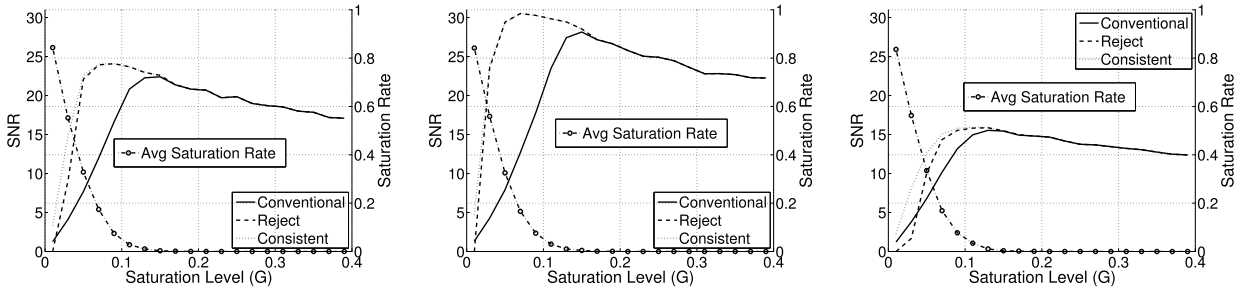
where $\hat{x}$ denotes the recovered signal.

5.2. Recovery SNR: K-sparse signals

We compare the recovery performance of the three approaches by applying each to the same set of measurements. We fixed the parameters $N = 1024$, $K = 20$, and $B = 4$ and varied the saturation 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, hence those experiments are not displayed here.

The experiments were performed as follows. For each trial we drew a new sparse signal x and a new matrix Φ according to Section 5.1 and computed $y = \Phi x$. We quantized the measurements using saturation level G and then used them to reconstruct the signal. For each approach, we set $\epsilon = \|y - R(y)\|_2$ or $\epsilon = \|\tilde{y} - R(\tilde{y})\|_2$, where applicable. This choice of ϵ can be thought of as an "oracle" value since it would be impossible to estimate exactly in practice; however, since it obtains the best possible results, it is therefore an upper bound on practical performance with these settings.

Figs. 2(a), 2(b), and 2(c) display the recovery SNR performance of the three approaches in dB for $M/N = 2/16$, $M/N = 6/16$, $M/N = 15/16$. The solid line depicts the conventional approach (4), the dashed line depicts the rejection approach (8), and the dotted line depicts the consistency approach (11). Each of these lines follows the scale on the left vertical axis. The



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

Fig. 3. Comparison of recovery approaches in the face of measurement saturation using CVX for weak ℓ_p-compressible signals with $N = 1024$, $M/N = 6/16$, and $B = 4$. Solid line depicts recovery SNR for the conventional approach via the program (4). Dotted line depicts recovery SNR for the consistency approach via the program (11). Dashed line depicts recovery SNR for the rejection approach via (8). SNR curves are measured on the left vertical axis. The dashed-circled line, measured on the right vertical axis, represents the average saturation rate when averaged over 100 trials. 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 $p = 1$.

dashed-circled line denotes the average saturation rate, $(M - \widetilde{M})/M$, and corresponds to the right vertical axis. In Fig. 2(a), the three lines meet at $G = 0.25$, as expected, because there the saturation rate is effectively zero. This is the operating point for the conventional approach and is the largest SNR value for the solid line. In this case, only the consistency approach obtains SNRs greater than the conventional approach. In Fig. 2(b), the three lines meet at $G = 0.15$. Both the consistency and the rejection approaches achieve their optimal performance at around $G = 0.1$, where the saturation rate is 0.09 (that is, 9%). In Fig. 2(c), the three lines meet at $G = 0.1$, and both the consistency and rejection approaches achieve their optimal performance at $G = 0.06$.

The implications of this experiment are threefold: First, saturation consistency offers the best recovery performance. 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 at a manifestly positive saturation rate.

5.3. Recovery SNR: Compressible signals

In addition to sparse signals, we also compare the recovery performance of the three approaches with compressible signals. As in the strictly sparse experiments, we used CVX for recovery and we chose the parameters, $N = 1024$, $M/N = 6/16$, and $B = 4$ and varied the saturation level parameter G over the range $[0, 0.4]$. The decay parameter p was varied in the range $[0.4, 1]$, but we will discuss only three decays $p = 0.4$, 0.8, and 1. Many signals are known to exhibit p in (19) in this range; for instance, it has been shown that the wavelet coefficients of natural images have decay rates between $p = 0.3$ and $p = 0.7$ [48]. For each parameter combination, we performed 100 trials, and computed the average performance. The experiments were 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 Fig. 3 reflect this intuition. Figs. 3(a), 3(b), and 3(c) depict the results for $p = 0.4$, $p = 0.8$, and $p = 1$, respectively. The highest SNR for $p = 0.4/p = 0.8/p = 1$ is achieved at a saturation rate of 17%/13%/5%, respectively. Hence, the more compressible the signal (smaller p), the more the measurements should be allowed to saturate in order to maximize recovery performance.

5.4. Robustness to saturation

We also compare the optimal performance of the rejection and consistency recovery 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.

We experimentally measured, by tuning G, the best SNR achieved on average for the three strategies. The experiment was performed as follows. Using the same parameters as in the K-sparse experiments in Section 5.2, for each value of M and for each approach, we searched for the saturation level G that yielded 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 Fig. 2 but over a larger range of M.

Fig. 4(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 saturation consistency. For these parameters, in the best case, saturation rejection can improve performance over the conventional approach by 10 dB, while saturation consistency can improve performance over the conventional approach by 13 dB.

There are two important implications of this experiment. First, when the number of measurements exceeds some threshold, intentionally saturating measurements can greatly improve performance. Second, in terms of the maximum achievable

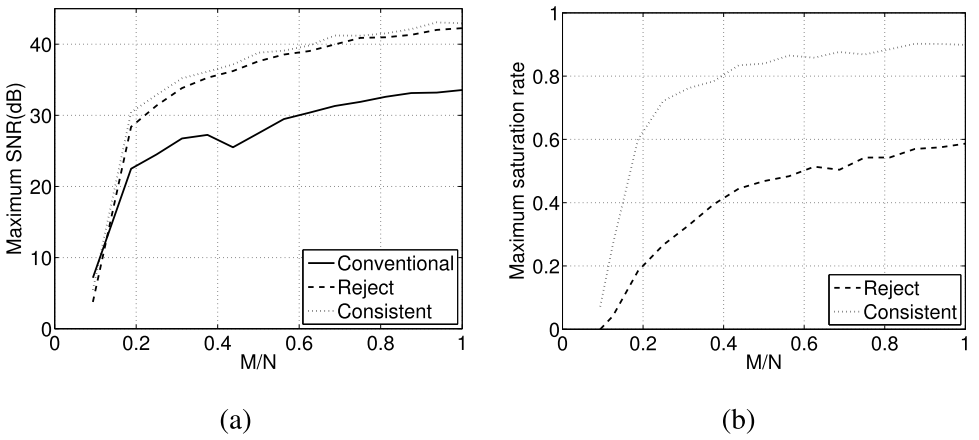


Fig. 4. Comparison of the SNR performance of CoSaMP and SC-CoSaMP in the face of measurement saturation for $N = 1024$, $K = 20$, and $B = 4$. Solid line depicts recovery SNR for the conventional approach via the program (4). Dotted line depicts recovery SNR for the consistency approach via the program (11). Dashed line depicts recovery SNR for the rejection approach via (8). (a) Maximum SNR over all saturation levels 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 consistency approaches can achieve SNRs exceeding the conventional SNR performance by up to 10 dB. The best performance between the rejection and consistency approaches is similar, differing only by 3 dB, but the range of saturation rates for which they achieve high performance is much larger for the consistency approach. Thus, the consistency approach is more robust to saturation.

SNR, the consistency 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. To this end, we now 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 determined the maximum SNR achieved by the conventional approach (i.e., the solid curve in Fig. 4(a)). Then, for the other approaches, we increased the saturation rate by tuning the saturation level (between 0 and 0.4 as before). We continued to increase the saturation rate until the SNR was lower than the best SNR of the conventional approach. The results of this experiment are depicted in Fig. 4(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 consistency approach. At best, the rejection approach achieves a range of $[0, 0.6]$ while the consistency approach achieves a range of $[0, 0.9]$. Thus, these experiments show that the consistency approach is more robust to the saturation rate.

5.5. Automatic gain control (AGC) for CS

In many ADC applications, the signal power does not remain constant but changes over time, which greatly complicates setting the signal scaling to minimize quantization and saturation errors. An automatic gain control (AGC) is a device that monitors signal properties such as the number of saturation events and uses these statistics to tune the signal gain. In a conventional uniform sampling system, the AGC typically targets an extremely low saturation rate (for example, 63 saturation events per 1 million samples [23]), which is quite difficult to robustly estimate and track over time.

The experimental results above suggest that the situation for CS measurements is quite different, since, per Figs. 2, 3, and 4, optimal CS recovery performance is attained at saturation rates well above 1 per 100, which is much easier to robustly estimate and track.

To illustrate the immediate practical benefits of these results, we performed a simulation of the CS AGC in Fig. 5 that automatically tunes the signal scaling based on the local measurement saturation rate to maximize the CS signal recovery performance. In our setup, the signal x is split into consecutive blocks of length N, and Φ is applied to each block separately. For each block, a gain $\theta^{[w]}$ is applied to the measurements and then they are quantized. In different hardware implementations, the gain could be applied before, after, or within the measurement matrix Φ; 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 assume that the signal energy does not deviate significantly between consecutive blocks.

The AGC works as follows. We compute the saturation rate of the previous block of measurements, $\hat{s}^{[w-1]}$, after quantization. The new gain is then computed by adding the error between s and $\hat{s}^{[w-1]}$ to the previous gain, i.e.,

$$\theta^{[w]} = \theta^{[w-1]} + \nu\big(s - \hat{s}^{[w-1]}\big), \tag{21}$$

where $\nu > 0$ is constant. The term ν will determine the sensitivity of the system (this is sometimes called the *proportional* term in control systems theory). This negative feedback system is bounded-input, bounded-output (BIBO) stable for any finite positive ν with $0 < s < 1$ [49].

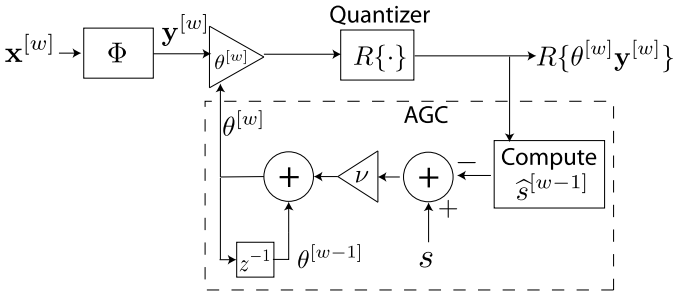


Fig. 5. Automatic gain control (AGC) for tuning to nonzero saturation rates in CS systems.

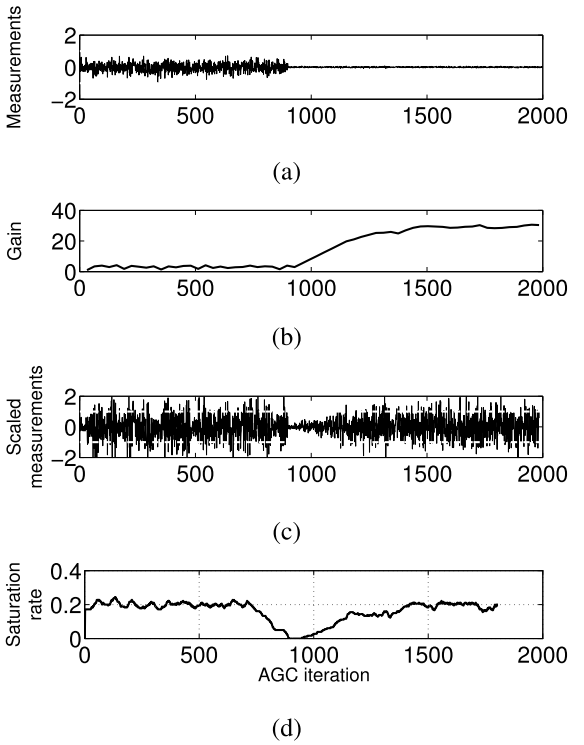


Fig. 6. CS AGC in action. (a) CS measurements with no saturation. The signal strength drops by 90% at measurement 900. (b) Output gain θ from AGC. (c) Measurements scaled by gain from AGC. (d) Saturation rate of scaled measurements.

To demonstrate that this AGC is sensitive to both increases in signal strength as well as decreases, we performed an experiment where the signal strength drops suddenly and significantly. The experiment is depicted in Fig. 6 and was performed as follows. We generated a signal of length 63×512, split it into 63 blocks, and computed 32 measurements per block. The measurements are depicted in Fig. 6(a) where the dashed lines represent the quantizer range $[-1, 1]$. At measurement 900, the signal strength drops by 90%. With $\nu = 12$, we set a desired saturation rate of $s = 0.2$ and ran the AGC. Fig. 6(b) shows the gain that the AGC applied as it received each measurement. Fig. 6(c) shows the resulting output signal with quantizer range, and Fig. 6(d) shows the estimated output saturation rate. We achieve the desired saturation rate within approximately 10 iterations. The system adapted 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.

6. Discussion

Since CS measurements must be quantized in practice, saturation error is inevitable. In this paper, we have developed theory and algorithms for better understanding and mitigating the effect of saturation on CS recovery. Both of our main approaches — saturation rejection and saturation consistency — rely on the democratic nature of CS measurements, a concept that we have made rigorous. In particular, we proved that CS measurements are $\widetilde{M}$-democratic for a large class of random measurement matrices. This means that once an $M \times N$ random matrix is drawn, every $\widetilde{M} \times N$ submatrix has the RIP. The

democracy property is of independent interest and could have a wide range of potential applications. 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 [50] or when applying CS as a multiple description coding (MDC) code [51].

Our extensive experiments have indicated that, given enough measurements, the rejection and consistency approaches outperform conventional CS recovery methods with saturated measurements. Interestingly, we find the best performance of our methods occurs when the saturation rate is manifestly positive, in stark contrast with the conventional uniform sampling theory.

Finally, our recovery methods are not limited to measurements that have been quantized with saturation. Any application where highly corrupted measurements can be easily detected can employ similar techniques to those described here. 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. The algorithms developed in this paper should offer significantly improved performance with such sensors.

References

[1] D. Healy, Analog-to-information, 2005, BAA #05-35.
[2] R. Walden, Analog-to-digital converter survey and analysis, IEEE J. Select. Areas Comm. 17 (1999) 539–550.
[3] D. Donoho, Compressed sensing, IEEE Trans. Inform. Theory 6 (2006) 1289–1306.
[4] E. Candès, Compressive sampling, in: Proc. Int. Congress Math., Madrid, Spain.
[5] M. Vetterli, P. Marziliano, T. Blu, Sampling signals with finite rate of innovation, IEEE Trans. Signal Process. 50 (2002) 1417–1428.
[6] J. Laska, S. Kirolos, M. Duarte, T. Ragheb, R. Baraniuk, 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.
[7] J. Tropp, J. Laska, M. Duarte, J. Romberg, R. Baraniuk, Beyond Nyquist: Efficient sampling of sparse bandlimited signals, IEEE Trans. Inform. Theory 56 (1) (2010) 520–544.
[8] J. Romberg, Compressive sensing by random convolution, SIAM J. Imaging Sciences 2 (2009) 1098–1128.
[9] J. Tropp, M. Wakin, M. Duarte, D. Baron, R. Baraniuk, Random filters for compressive sampling and reconstruction, in: Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP), Toulouse, France.
[10] M. Duarte, M. Davenport, D. Takhar, J. Laska, T. Sun, K. Kelly, R. Baraniuk, Single-pixel imaging via compressive sampling, IEEE Signal Process. Mag. 25 (2008) 83–91.
[11] R. Robucci, L. Chiu, J. Gray, J. Romberg, P. Hasler, 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.
[12] R. Marcia, Z. Harmany, R. Willett, Compressive coded aperture imaging, in: Proc. SPIE Symp. Elec. Imaging: Comput. Imaging, San Jose, CA.
[13] M. Mishali, Y. Eldar, From theory to practice: Sub-Nyquist sampling of sparse wideband analog signals, preprint, 2009.
[14] L. Jacques, D. Hammond, M. Fadili, Dequantizing compressed sensing: When oversampling and non-Gaussian constraints combine, preprint, 2009.
[15] W. Dai, H. Pham, O. Milenkovic, Distortion-rate functions for quantized compressive sensing, preprint, 2009.
[16] A. Zymnis, S. Boyd, E. Candès, Compressed sensing with quantized measurements, preprint, 2009.
[17] J. Sun, V. Goyal, Quantization for compressed sensing reconstruction, in: Proc. Sampling Theory and Applications (SampTA), Marseille, France.
[18] E. Candès, T. Tao, The Dantzig selector: Statistical estimation when p is much larger than n, Ann. Statist. 35 (2007) 2313–2351.
[19] R. Carrillo, K. Barner, T. Aysal, Robust sampling and reconstruction methods for compressed sensing, in: Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP), Taipei, Taiwan.
[20] J. Laska, M. Davenport, R. Baraniuk, Exact signal recovery from corrupted measurements through the pursuit of justice, in: Proc. Asilomar Conf. on Signals Systems and Computers, Asilomar, CA.
[21] Z. Harmany, R. Marcia, R. Willett, Sparse Poisson intensity reconstruction algorithms, in: Proc. IEEE Work. Stat. Signal Processing (SSP), Cardiff, Wales.
[22] I. Rish, G. Grabarnik, Sparse signal recovery with exponential-family noise, in: Proc. Allerton Conf. Comm., Control, and Comput., Monticello, IL.
[23] J. Triechler, Personal communication, 2009.
[24] A. Calderbank, I. Daubechies, The pros and cons of democracy, IEEE Trans. Inform. Theory 48 (2002) 1721–1725.
[25] S. Güntürk, Harmonic analysis of two problems in signal compression, Ph.D. thesis, Program in Applied and Computation Mathematics, Princeton University, Princeton, NJ, 2000.
[26] Y. Lyubarskii, R. Vershynin, Uncertainty principles and vector quantization, IEEE Trans. Inform. Theory 56 (2010) 3491–3501.
[27] E. Candès, Integration of sensing and processing, IMA Annual Program Year Work, 2005.
[28] G. Gray, G. Zeoli, Quantization and saturation noise due to analog-to-digital conversion, IEEE Trans. Aerospace Elec. Systems 7 (1971) 222–223.
[29] E. Candès, T. Tao, Decoding by linear programming, IEEE Trans. Inform. Theory 51 (2005) 4203–4215.
[30] R. Baraniuk, M. Davenport, R. DeVore, M. Wakin, A simple proof of the restricted isometry property for random matrices, Constr. Approx. 28 (2008) 253–263.
[31] E. Candès, The restricted isometry property and its implications for compressed sensing, C. R. Acad. Sci. Paris, Sér. I 346 (2008) 589–592.
[32] D. Needell, J. Tropp, CoSaMP: Iterative signal recovery from incomplete and inaccurate samples, Appl. Comput. Harmon. Anal. 26 (2009) 301–321.
[33] T. Blumensath, M. Davies, Iterative hard thresholding for compressive sensing, Appl. Comput. Harmon. Anal. 27 (2009) 265–274.
[34] J. Haupt, R. Nowak, Signal reconstruction from noisy random projections, IEEE Trans. Inform. Theory 52 (2006) 4036–4048.
[35] P. Wojtaszczyk, Stability and instance optimality for Gaussian measurements in compressed sensing, Found. Comput. Math. 10 (1) (2010), doi:10.1007/s10208-009-9046-4.
[36] J.P. Slavinsky, J. Laska, M. Davenport, R. Baraniuk, The compressive mutliplexer for multi-channel compressive sensing, in: Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP), Prague, Czech Republic.
[37] J. Laska, P. Boufounos, R. Baraniuk, Finite-range scalar quantization for compressive sensing, in: Proc. Sampling Theory and Applications (SampTA), Marseille, France.
[38] M. Davenport, M. Duarte, M. Wakin, J. Laska, D. Takhar, K. Kelly, R. Baraniuk, The smashed filter for compressive classification and target recognition, in: Proc. SPIE Symp. Elec. Imaging: Comput. Imaging, San Jose, CA.
[39] P. Boufounos, R. Baraniuk, 1-bit compressive sensing, in: Proc. Conf. Inform. Science and Systems (CISS), Princeton, NJ.
[40] M. Davenport, Random observations on random observations: Sparse signal acquisition and processing, 2010.
[41] M. Davenport, P. Boufounos, M. Wakin, R. Baraniuk, Signal processing with compressive measurements, J. Select. Topics Signal Proc. 4 (2010) 445–460.
[42] F. Beutler, Error-free recovery of signals from irregularly spaced samples, SIAM Rev. 8 (1966) 328–335.
[43] A. Aldroubi, K. Gröchenig, Nonuniform sampling and reconstruction in shift-invariant spaces, SIAM Rev. 43 (2001) 585–620.

[44] E. Candès, J. Romberg, T. Tao, Stable signal recovery from incomplete and inaccurate measurements, Comm. Pure Appl. Math. 59 (2006) 1207–1223.
[45] E. Candès, J. Romberg, Sparsity and incoherence in compressive sampling, Inverse Problems 23 (2006) 969–985.
[46] M. Grant, S. Boyd, CVX: Matlab software for disciplined convex programming, available online at http://stanford.edu/~boyd/cvx.
[47] M. Grant, S. Boyd, Graph implementations for nonsmooth convex programs, in: V. Blondel, S. Boyd, H. Kimura (Eds.), Recent Advances in Learning and Control, in: Lecture Notes in Control and Inform. Sci., Springer, 2008, pp. 95–110.
[48] R. DeVore, B. Jawerth, B. Lucier, Image compression through wavelet transform coding, IEEE Trans. Inform. Theory 38 (1992) 719–746.
[49] A. Oppenheim, A. Willsky, Signals and Systems, Prentice-Hall, 1996.
[50] D. MacKay, Fountain codes, IEE Proc. Comm. 152 (2005) 1062–1068.
[51] V. Goyal, Multiple description coding: Compression meets the network, IEEE Signal Process. Mag. 18 (2001) 74–93.