

RICE UNIVERSITY

**Democracy in Action: Quantization, Saturation,
and Compressive Sensing**

by

Jason N. Laska

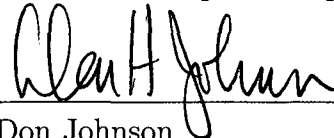
A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Master of Science

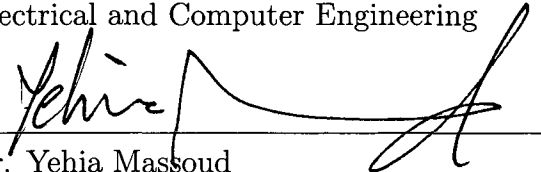
APPROVED, THESIS COMMITTEE:



Dr. Richard G. Baraniuk, Chair
Victor E. Cameron Professor
Electrical and Computer Engineering



Dr. Don Johnson
J. S. Abercrombie Professor
Electrical and Computer Engineering



Dr. Yehia Massoud
Associate Professor
Electrical and Computer Engineering

Houston, Texas

August, 2009

UMI Number: 1486023

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



UMI 1486023

Copyright 2010 by ProQuest LLC.

All rights reserved. This edition of the work is protected against unauthorized copying under Title 17, United States Code.



ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Abstract

Democracy in Action: Quantization, Saturation,
and Compressive Sensing

by

Jason N. Laska

We explore and exploit a heretofore relatively unexplored hallmark of compressive sensing (CS), the fact that certain CS measurement systems are *democratic*, which means that each measurement carries roughly the same amount of information about the signal being acquired. Using this property, we re-think how to quantize the compressive measurements. In Shannon-Nyquist sampling, we scale down the analog signal amplitude (and therefore increase the quantization error) to avoid the gross saturation errors. In stark contrast, we demonstrate a CS system achieves the best performance when we operate 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.

Acknowledgements

The work presented in this thesis would not have been possible if it were not for my collaborators Petros Boufounos and Mark Davenport, and my advisor Richard Baraniuk. In particular, I thank Petros and Rich for always pushing me to achieve more, and Mark for the enlightening conversations and endless patience. I also thank my committee, Don Johnson and Yehia Massoud, for their input on this work.

In addition to those who directly helped with this project, I would like to thank Marco Duarte, who has influenced my research habits in a number of ways and who quickly became a role model in my first years at Rice. I thank my friends in Houston and elsewhere for making life enjoyable. Finally, I thank my family for their unbounded support.

Contents

List of Illustrations	vi
List of Tables	viii
1 Introduction	1
2 Background	6
2.1 Analog-to-digital conversion	6
2.2 Scalar quantization	6
2.3 Compressive sensing (CS)	8
2.4 CS in practice	11
3 Signal recovery from saturated measurements	13
3.1 Unbounded saturation error	13
3.2 Recovery via saturation rejection	14
3.3 Recovery via convex optimization with consistency constraints	15
3.4 Recovery via greedy algorithms with consistency constraints	16
4 Random measurements and democracy	21
4.1 Democracy and recovery	21
4.2 Random measurements are democratic	22
5 Experimental validation	26
5.1 Experimental setup	27
5.2 Reconstruction SNR: K -sparse signals	28

5.3	Reconstruction SNR: Compressible signals	31
5.4	Robustness to saturation	32
6	Extensions	35
6.1	Automatic gain control (AGC) for CS	35
7	Discussion	40
A	The expected error of quantized and saturated measure- ments	42
B	Preservation of inner products	46
C	Consistent recovery via fixed point continuation	49
D	Proof of the democracy of Gaussian matrices	53
D.1	Concentration of measure	53
D.2	Democracy	58
	Bibliography	61

Illustrations

2.1	(a) Midrise scalar quantizer. (b) Finite-range midrise scalar quantizer with saturation level G	7
2.2	Random demodulator compressive ADC.	11
5.1	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 $M/N = 15/16$	29
5.2	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 $p = 1$	30

5.3	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.	33
6.1	Automatic gain control (AGC) for tuning to nonzero saturation rates in CS systems.	36
6.2	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.	39
A.1	Measurement error. Dash-dotted line: Expected measurement error due to quantization and saturation (A.4). Dashed line: Expected measurement error due to quantization only (A.2).	44

Tables

2.1	Quantization parameters.	7
-----	----------------------------------	---

Chapter 1

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 non-zero coefficients in a known transform basis over a fixed time interval [5]. Signals that have few non-zero 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 simultane-

ously acquiring and compressing signals. This leads to lower sampling rates and thus simplified 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, whereas the removal of others will have negligible effect.

Several CS-inspired hardware architectures for acquiring signals, images, and videos have been proposed, analyzed, and in some cases implemented [6–12]. 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 work, we focus on this quantization step.

While the effect of quantization on the CS framework has been previously explored [13, 14], 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 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. Current CS recovery algorithms only provide guarantees for noise that is either bounded or bounded with high probability (for example, Gaussian noise) [15].

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 work, we present two new approaches for mitigating of 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. Be-

cause 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 and show that necessary recovery properties of the measurements hold for any fixed rejection rate as the initial number of measurements becomes large.

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 non-zero 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 non-zero 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 non-zero 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. This AGC uses only the saturation rate to determine the gain, unlike conventional AGCs, since such systems require the saturation rate to be very close to zero.

The organization of this thesis is as follows. In Section 2, we review quantization with saturation and the key concepts of the CS framework. In Section 3, we discuss the problem of unbounded saturation error in CS and define our proposed solutions. In Section 4 we provide theoretical analysis to show that CS measurements are demo-

cratic and that our solutions solve the stated problem. In Section 5, we validate our claims experimentally and show that in many scenarios, we achieve improved performance. In Section 6 we derive a simple AGC for CS systems and in Section 7 we discuss how the democracy property can be useful in other applications. Appendix D contains the proof of democracy for Gaussian matrices. For completeness, we provide additional analysis on the mean squared error of quantized and saturated measurements and the preservation of inner products between two measurement vectors in Appendix A and B, respectively. The main text provides a greedy algorithm for our approach and in Appendix C, we supplement this with an optimization algorithm.

Chapter 2

Background

2.1 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. Although both steps are necessary to represent a signal in the discrete digital world, 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. On the other hand, quantization results in an irreversible loss of information unless the signal amplitudes belong in the discrete set defined by the quantizer. A central ADC system design goal is to minimize the distortion due to quantization.

2.2 Scalar quantization

Scalar quantization is the process of converting the continuous value of individual measurements 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.

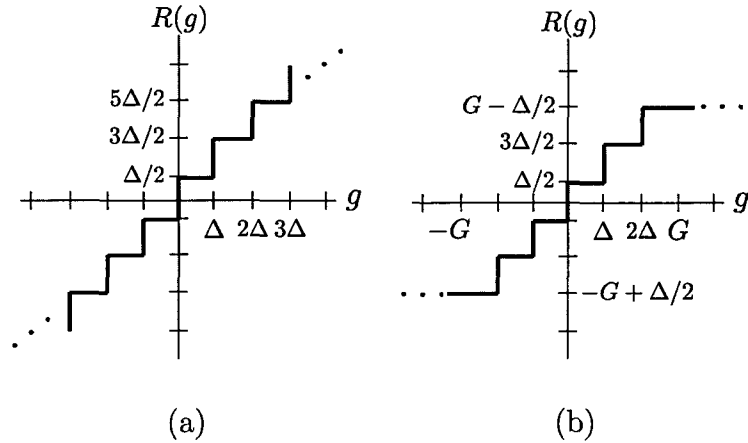


Figure 2.1 : (a) Midrise scalar quantizer. (b) Finite-range midrise scalar quantizer with saturation level G .

Table 2.1 : Quantization parameters.

G	saturation level
B	number of bits
Δ	bin width
$\Delta/2$	maximum error per measurement

In this work, 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 point $R(g) = \operatorname{argmin}_{q_k} |g - q_k|$. This implies that the quantization error per measurement, $|g - R(g)|$, is bounded by $\Delta/2$. Figure 2.1(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 repre-

sents a symmetric range of values $|g| < G$, where $G > 0$ is known as the saturation level [16]. 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 points are $q_k = \Delta/2 + k\Delta$, where $k = -2^{B-1}, \dots, 2^{B-1} - 1$. Any measurement with magnitude greater than G saturates the quantizer, i.e., it quantizes to the quantization point $G - \Delta/2$, implying an unbounded error. Figure 2.1(b) depicts the mapping performed by a finite range midrise quantizer with saturation level G and Table 2.1 summarizes the parameters defined with respect to quantization. An analysis of the average error due to quantization and saturation for Gaussian signals can be found in Appendix A. This analysis demonstrates that on average, saturation error dominates the total error in finite-range quantization.

2.3 Compressive sensing (CS)

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

$$\mathbf{y} = \Phi\mathbf{x} + \mathbf{e}, \quad (2.1)$$

where Φ 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* Ψ , i.e., $\mathbf{x} = \Psi\mathbf{x}$ with $\|\mathbf{x}\|_0 \leq K$,* then one can acquire only $M = O(K \log(N/K))$ measurements and still

* $\|\cdot\|_0$ denotes the ℓ_0 quasi-norm, which simply counts the number of non-zero entries of a vector.

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 [17], Candès and Tao introduced the *restricted isometry property* (RIP) of a matrix Φ and established its important role in CS. Slightly adapted from [17], we say that a matrix Φ satisfies the RIP of order K if there exist constants, $0 < a \leq b < \infty$, such that

$$a\|\mathbf{x}\|_2^2 \leq \|\Phi\mathbf{x}\|_2^2 \leq b\|\mathbf{x}\|_2^2, \quad (2.2)$$

holds for all \mathbf{x} with $\mathbf{x} = \Psi\boldsymbol{\theta}$ and $\|\boldsymbol{\theta}\|_0 \leq K$. 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 given 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 $M = O(K \log(N/K))$ [18]. In this paper, without the loss of generality, we fix $\Psi = \mathbf{I}$, the identity matrix, implying that $\mathbf{x} = \boldsymbol{\theta}$.

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 Φ must satisfy the RIP of order $2K$ with $a > 0$ 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

$$\hat{\mathbf{x}} = \underset{\mathbf{x}}{\operatorname{argmin}} \|\mathbf{x}\|_1 \quad \text{s.t.} \quad \|\Phi\mathbf{x} - \mathbf{y}\|_2 \leq \epsilon \quad (2.3)$$

can recover a sparse or compressible signal \mathbf{x} . The following theorem, a slight modification of Theorem 1.2 from [19], makes this precise by bounding the recovery error of \mathbf{x} with respect to the measurement noise norm, denoted by ϵ , and with respect to the best approximation of \mathbf{x} by its largest K terms, denoted using \mathbf{x}_K .

Theorem 1. *Suppose that $\Phi\Psi$ satisfies the RIP of order $2K$ with $b/a < 1 + \sqrt{2}$. Given measurements of the form $\mathbf{y} = \Phi\Psi\mathbf{x} + \mathbf{e}$, where $\|\mathbf{e}\|_2 \leq \epsilon$, then the solution to (2.3) obeys*

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

where

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

While convex optimization techniques like (2.3) 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 [20, 21]. Furthermore, alternative recovery strategies based on (2.3) have been analyzed in [15, 22]. These methods replace the constraint in (2.3) with an alternative constraint that is motivated by the assumption that the measurement noise is Gaussian in the case of [15] and that is agnostic to the value of ϵ in [22].

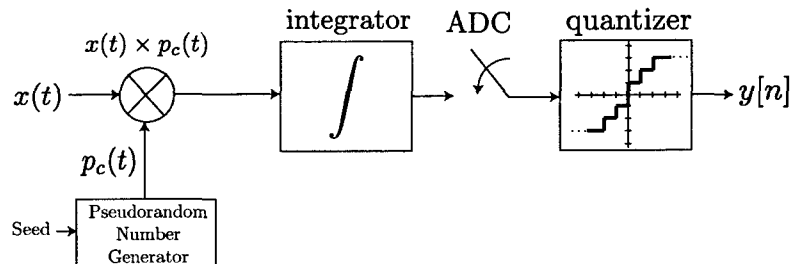


Figure 2.2 : Random demodulator compressive ADC.

2.4 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], and several compressive imaging architectures [10–12].

We briefly describe the random demodulator as an example of such a system [7]. Figure 2.2 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.

Before quantization, systems such as these represent a linear operator mapping the analog input signal to a discrete output vector. It is possible to relate this operator to

a discrete measurement matrix Φ which maps, for example, the Nyquist-rate samples of the input signal to the discrete output vector.

Chapter 3

Signal recovery from saturated measurements

3.1 Unbounded saturation error

A standard CS recovery approach like the program (2.3) 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 sinc-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 is due to both the *democracy* of CS measurements and the *non-linear* nature of CS reconstruction.

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. This is because in each approach, recovery from the measurements that did not saturate must be possible. This implies that the signal-dependent submatrix of Φ , made up of rows corresponding to the measurements that did not saturate, must satisfy RIP. A main result of this work, that we prove below, is that there exists a class of matrices Φ such that an arbitrary subset of their rows will indeed satisfy the RIP.

3.2 Recovery via saturation rejection

An intuitive way to handle saturated measurements is to simply discard them [23].

Denote the vector of the measurements that did not saturate as $\tilde{\mathbf{y}}$ with length \tilde{M} . The matrix $\tilde{\Phi}$ is created by selecting the rows of Φ that correspond to the elements of $\tilde{\mathbf{y}}$. Then, as an example, using (2.3) for reconstruction yields the program:

$$\hat{\mathbf{x}} = \underset{\mathbf{x}}{\operatorname{argmin}} \|\mathbf{x}\|_1 \quad \text{s.t.} \quad \|\tilde{\Phi}\mathbf{x} - \tilde{\mathbf{y}}\|_2 < \epsilon. \quad (3.1)$$

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* [24] for detection, which utilizes the inner products $\langle \Phi \mathbf{u}, \Phi \mathbf{v} \rangle$ between the measurement of vectors \mathbf{u}, \mathbf{v} . Such techniques depend on $\langle \Phi \mathbf{u}, \Phi \mathbf{v} \rangle$ being close to $\langle \mathbf{u}, \mathbf{v} \rangle$. Saturation can induce unbounded errors in $\langle \Phi \mathbf{u}, \Phi \mathbf{v} \rangle$, 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. A specific bound on these inner products is derived in Appendix B.

3.3 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 $\hat{\mathbf{x}}$ so that the magnitudes of the values of $\Phi \hat{\mathbf{x}}$ 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. Let Φ^{S^+} and Φ^{S^-} denote the submatrices of Φ obtained by keeping only the rows of Φ indexed

by S^+ and S^- . Form a new matrix $\mathring{\Phi}$ as

$$\mathring{\Phi} \triangleq \begin{bmatrix} \Phi^{S^+} \\ -\Phi^{S^-} \end{bmatrix}. \quad (3.2)$$

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

$$\hat{\mathbf{x}} = \underset{\mathbf{x}}{\operatorname{argmin}} \|\mathbf{x}\|_1 \quad \text{s.t.} \quad \|\tilde{\Phi}\mathbf{x} - \tilde{\mathbf{y}}\|_2 < \epsilon \quad (3.3)$$

$$\text{and} \quad \mathring{\Phi}\mathbf{x} \geq G \cdot \mathbf{1}, \quad (3.4)$$

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. An algorithm that solves this formulation is presented in Appendix C. Alternative regularization terms that reduce the space of solutions for quantized measurements can be used on $\tilde{\mathbf{y}}$, such as those proposed in [13, 14]. In some hardware systems, the measurements that immediately follow a saturation event can have higher distortion than the other unsaturated measurements. In this case, an additional ℓ_2 constraint, $\|\tilde{\Phi}^*\mathbf{x} - \tilde{\mathbf{y}}^*\|_2 < \epsilon_1$, can be applied where \star denotes the indices of the measurements immediately following a saturation event and where $\epsilon_1 > \epsilon$.

3.4 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 [20].

In this subsection, we introduce *Saturation Consistent CoSaMP* (SC-CoSaMP), a modified version of CoSaMP that performs consistent reconstruction with saturated measurements.

CoSaMP estimates the signal $\hat{\mathbf{x}}$ 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 $\mathbf{p} = \Phi^T(\Phi\hat{\mathbf{x}} - \mathbf{y})$, that allows us to infer large signal coefficients, and hence is called the proxy vector [20], 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 $\hat{\mathbf{x}}$ to produce Ω . Given Ω , CoSaMP estimates the signal coefficients by solving the least squares problem:

$$\hat{\mathbf{x}} = \min_{\mathbf{x}} \|\Phi_{\Omega}\mathbf{x} - \mathbf{y}\|_2^2. \quad (3.5)$$

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 (3.5).

The steps of SC-CoSaMP are displayed in Algorithm 1. In steps 1 and 2, the algorithm initializes by choosing an estimate $\hat{\mathbf{x}}^{[0]} = \mathbf{0}$, an N -dimensional vector of zeros, and 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 $\tilde{\mathbf{y}}$ and a proxy that uses the supports of the saturated measurements. To compute the proxy from $\tilde{\mathbf{y}}$, we repeat the same computation as in CoSaMP, $\tilde{\Phi}^T(\tilde{\mathbf{y}} - \tilde{\Phi}\hat{\mathbf{x}}^{[n]})$, 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 $G \cdot \mathbf{1} - \mathring{\Phi}\hat{\mathbf{x}}^{[n]}$. This vector measures how close the elements of $\mathring{\Phi}\hat{\mathbf{x}}$ are to G . In consistent reconstruction, the magnitude of the elements of $\mathring{\Phi}\hat{\mathbf{x}}$ 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,

$$h(y_i) = \begin{cases} 0, & y_i < 0 \\ y_i, & y_i \geq 0. \end{cases} \quad (3.6)$$

This function is applied element-wise to a vector as $\bar{h}(\mathbf{y}) = \sum_i h(y_i)\mathbf{e}_i$ where \mathbf{e}_i is the i^{th} canonical vector.

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

$$\mathbf{p} = \tilde{\Phi}^T(\tilde{\mathbf{y}} - \tilde{\Phi}\hat{\mathbf{x}}^{[n]}) + \mathring{\Phi}^T\bar{h}(G \cdot \mathbf{1} - \mathring{\Phi}\hat{\mathbf{x}}^{[n]}). \quad (3.7)$$

In this arrangement, the elements of $\mathring{\Phi}\hat{\mathbf{x}}$ 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 [21].

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 $\hat{\mathbf{x}}^{[n]}$. 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 $\|\Phi_{\Omega}\mathbf{x} - \mathbf{y}\|_2^2$ where Φ_{Ω} denotes a submatrix of Φ restricted to columns indexed by Ω . Thus in CoSaMP, new coefficient values are estimated via $\Phi_{\Omega}^{\dagger}\mathbf{y}$, where \dagger denotes the Moore-Penrose pseudo-inverse. However, this can be reformulated to include the saturation constraint. Specifically, step 6 of SC-CoSaMP finds the solution to

$$\hat{\mathbf{x}}^{[n+1]} = \underset{\mathbf{x}}{\operatorname{argmin}} \|\tilde{\Phi}_{\Omega}\mathbf{x} - \tilde{\mathbf{y}}\|_2^2 \quad \text{s.t.} \quad \mathring{\Phi}_{\Omega}\mathbf{x} \geq G \cdot \mathbf{1}. \quad (3.8)$$

This can be achieved via gradient descent or other optimization techniques by employing a one-sided quadratic to the constraint [25].

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.

Algorithm 1 SC-CoSaMP greedy algorithm

1: **Input:** \mathbf{y} , Φ , and K

2: **Initialize:** $\hat{\mathbf{x}}^{[0]} \leftarrow \mathbf{0}$, $n \leftarrow 0$

3: **while** not converged **do**

4: **Compute proxy:**

$$\mathbf{p} \leftarrow \tilde{\Phi}^T (\tilde{\mathbf{y}} - \tilde{\Phi} \hat{\mathbf{x}}^{[n]}) + \mathring{\Phi}^T \bar{h} (G \cdot \mathbf{1} - \mathring{\Phi} \hat{\mathbf{x}}^{[n]})$$

5: **Update coefficient support:**

$\Omega \leftarrow$ union of

- support of largest $2K$ coefficients from \mathbf{p}
- support of $\hat{\mathbf{x}}^{[n]}$

6: **Estimate new coefficient values:**

$$\hat{\mathbf{x}}^{[n+1]} \leftarrow \operatorname{argmin}_{\mathbf{x}} \|\tilde{\Phi}_{\Omega} \mathbf{x} - \tilde{\mathbf{y}}\|_2^2 \quad \text{s.t.} \quad \mathring{\Phi}_{\Omega} \mathbf{x} \geq G \cdot \mathbf{1}$$

7: **Prune:**

$$\hat{\mathbf{x}}^{[n+1]} \leftarrow \text{keep largest } K \text{ coefficients of } \hat{\mathbf{x}}^{[n+1]}$$

8: $n \leftarrow n + 1$

9: **end while**

Chapter 4

Random measurements and democracy

4.1 Democracy and recovery

In this section, we demonstrate that CS measurements have the democracy property, i.e., each measurement contributes a similar amount of information about the signal \mathbf{x} to the compressed representation \mathbf{y} [26–28].*

In this work, we say that Φ is \widetilde{M} -democratic if all $\widetilde{M} \times N$ submatrices Φ^Γ of Φ have the RIP. Thus, the matrix $\widetilde{\Phi}$ as defined in the Section 3 is a specific example of Φ^Γ . We note that this condition on Φ is significantly stronger than drawing a new $\widetilde{M} \times N$ RIP matrix: the democracy property implies that once Φ is drawn, any \widetilde{M} rows of Φ will have RIP.

If Φ is \widetilde{M} -democratic, then both approaches described in Section 3 will recover sparse and compressible signals. It directly follows from [19] and the fact that the democracy property implies that any $\widetilde{M} \times N$ submatrix of Φ has RIP, that the rejection approach (3.1) recovers sparse and compressible signals. Note that the two approaches will not necessarily produce the same solution. This is because the

*The original introduction of this term was with respect to quantization [26, 27], 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 [28].

solution from the rejection approach may not lie in the feasible set of solutions of the consistent approach (3.3). However, the reverse is true. The solution to the consistent approach does lie in the feasible set of solutions to the rejection approach. Because of this, the consistent approach recovers sparse and compressible signals as well.

In general, since the consistent approach is merely incorporating additional knowledge about the signal, we expect that it will perform no worse than the rejection approach.

4.2 Random measurements are democratic

We now demonstrate that if Φ is generated according to a Gaussian distribution, then the measurements are \widetilde{M} -democratic. We begin by analyzing the concentration properties of $\Phi^\Gamma \mathbf{x}$.

Lemma 1. *Suppose that Φ is an $M \times N$ matrix whose entries $\phi_{ij} \sim \mathcal{N}(0, 1)$. Let $\alpha \in (0, 1)$, $\beta \in (1, \infty)$, and $0 < \widetilde{M} \leq M$ be given. Then for any $\mathbf{x} \in \mathbb{R}^N$, we have that*

$$\alpha \widetilde{M} \|\mathbf{x}\|_2^2 \leq \|\Phi^\Gamma \mathbf{x}\|_2^2 \leq \beta \widetilde{M} \|\mathbf{x}\|_2^2 \quad (4.1)$$

holds for all sets Γ with $|\Gamma| = \widetilde{M}$ that index the rows of Φ with probability exceeding $1 - P_\alpha - P_\beta$, where

$$P_\alpha \leq \widetilde{M} \sqrt{\frac{2}{\pi}} \left(\frac{e^{\lambda\alpha}}{\sqrt{1+2\lambda}} \right)^{\widetilde{M}} \quad (4.2)$$

$$\times \int_0^\infty \frac{(1 - 2Q(u\sqrt{1+2\lambda}))^{\widetilde{M}}}{(1 - 2Q(u))^{\widetilde{M}+1}} e^{-u^2/2} \quad (4.3)$$

$$\times \mathcal{B}_{\widetilde{M}}(M, 1 - 2Q(u)) du,$$

for all $\lambda > 0$ and

$$P_\beta \leq (M - \widetilde{M}) \sqrt{\frac{2}{\pi}} \left(\frac{e^{-\lambda\beta}}{\sqrt{1-2\lambda}} \right)^{\widetilde{M}} \quad (4.4)$$

$$\times \int_0^\infty \frac{(Q(t\sqrt{1-2\lambda}))^{\widetilde{M}}}{Q(t)^{\widetilde{M}}(1-2Q(t))} e^{-t^2/2} \quad (4.5)$$

$$\times \mathcal{B}_{\widetilde{M}}(M, 2Q(t)) dt,$$

for all $\lambda \in (0, \frac{1}{2})$. In each bound

$$\mathcal{B}_d(n, p) = \binom{n}{d} p^d (1-p)^{n-d} \quad (4.6)$$

is the Binomial distribution function and $Q(z) = \frac{1}{\sqrt{2\pi}} \int_z^{+\infty} e^{-t^2/2} dt$ is the tail integral of the standard Gaussian distribution.

The proof of this lemma can be found in Appendix D.1. Our approach determines the upper and lower bounds by analyzing the case where the smallest measurements are selected and the case where the largest measurements are selected respectively. For the lower bound, we condition the probability that the norm of the \widetilde{M} smallest measurements is less than $\alpha\widetilde{M}$ on the value of the \widetilde{M} -th largest element u , i.e., $\mathbb{P}(\|U(\mathbf{y})\|_2^2 \leq \alpha\widetilde{M} | u)$ where $U(\mathbf{y})$ denotes the \widetilde{M} smallest entries of \mathbf{y} . This probability is then bounded using Markov's inequality and other standard concentration of measure techniques. A similar proof is conducted for the upper bound.

For these results to be useful, we require that the bounds approach zero as M grows. The integrals (D.3) and (D.5) cannot be solved in closed form; thus, we demonstrate that the bounds on P_α and P_β are meaningful by analyzing the behavior of each bound for a fixed ratio $\frac{\widetilde{M}}{M}$ as $M \rightarrow \infty$.

We begin by examining the part of the bound on P_α given by (D.2). For a fixed ratio $\frac{\widetilde{M}}{M}$ and for any $\alpha < 1$, the parameter $\lambda > 0$ can be chosen such that

$$\lim_{M \rightarrow \infty} \widetilde{M} \sqrt{\frac{2}{\pi}} \left(\frac{e^{\lambda\alpha}}{\sqrt{1+2\lambda}} \right)^{\widetilde{M}} = 0. \quad (4.7)$$

This can be achieved by choosing λ such that

$$e^{2\lambda\alpha} < 1 + 2\lambda, \quad (4.8)$$

thus, if $W(\lambda) = \{w : \lambda = we^w\}$, the Lambert W -function, then the valid range of λ is

$$0 < \lambda < -\frac{1}{2} - \frac{W(-\alpha e^{-\alpha})}{2\alpha}. \quad (4.9)$$

For a fixed ratio $\frac{\widetilde{M}}{M}$, λ can be chosen so that as $M \rightarrow \infty$, the integral (D.3) of the bound on P_α evaluates to a finite number. The integral will be finite if λ is chosen to be a function of \widetilde{M} that decays fast, such as $\lambda = (1/\widetilde{M})^{\widetilde{M}}$. Thus, the bound on P_α converges to 0 as $M \rightarrow \infty$.

For P_β , we first note that when $\widetilde{M} = M$, the bound is trivially 0. For $\widetilde{M} < M$, we have similar results as for P_α . For instance, for a fixed ratio $\frac{\widetilde{M}}{M}$ and for any $\beta > 1$, the parameter $\lambda \in (0, 1/2)$ can be chosen such that

$$\lim_{M \rightarrow \infty} (M - \widetilde{M}) \sqrt{\frac{2}{\pi}} \left(\frac{e^{-\lambda\beta}}{\sqrt{1-2\lambda}} \right)^{\widetilde{M}} = 0. \quad (4.10)$$

The valid λ lie in the range

$$0 < \lambda < \frac{1}{2} + \frac{W(-\beta e^{-\beta})}{2\beta}. \quad (4.11)$$

Additionally, if λ decays quickly as \widetilde{M} grows, the integral (D.5) is finite; thus the bound converges to 0 for large M and some λ .

We have shown that the concentration property holds for all Φ^Γ with high probability for any $\alpha < 1$, $\beta > 1$, as $M \rightarrow \infty$. Using this result, we now show that all submatrices Φ^Γ will satisfy the RIP with high probability.

Theorem 2. [Democracy] *Suppose that Φ is an $M \times N$ matrix with entries $\phi_{ij} \sim \mathcal{N}(0, 1/\widetilde{M})$, where $0 < \widetilde{M} \leq M$. Let $a > 0$ and $b > a$ be given. Then with probability at least $1 - P_F$, we have that all $\widetilde{M} \times N$ submatrices Φ^Γ of Φ satisfy*

$$a\|\mathbf{x}\|_2^2 \leq \|\widetilde{\Phi}\mathbf{x}\|_2^2 \leq b\|\mathbf{x}\|_2^2 \quad (4.12)$$

for all $\mathbf{x} \in \Sigma_K$, where

$$P_F \leq \left(\frac{3eN}{\epsilon K}\right)^K (P_\alpha + P_\beta), \quad (4.13)$$

for $0 < \epsilon < (\sqrt{b} - \sqrt{a})/2\sqrt{b}$.

The proof of this Theorem can be found in Appendix D.2. It uses the results of Lemma 1 with the procedure found in [18] to obtain the result.

This bound implies that for large enough N , $CK \log(N/K) < M < N$ can be chosen such that for a fixed ratio \widetilde{M}/M , the RIP will hold for any of the \widetilde{M} measurements. This is because P_α and P_β are not dependent on N and go to zero for large \widetilde{M} .

Chapter 5

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 (2.3);
2. the *rejection approach*, discarding saturated measurements before reconstruction with (3.1); and
3. the *consistent approach*, incorporating saturated measurements as a constraint in the program (3.3), (3.4).

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 non-zero.
- 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.

5.1 Experimental setup

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

- K -sparse: in each trial, K non-zero 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

$$x_n = v_n n^{-1/p}, \quad (5.1)$$

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, shows that our results are robust to large

variety of measurement matrix classes such as 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):

$$\text{SNR} \triangleq 10 \log_{10} \left(\frac{\|\mathbf{x}\|_2^2}{\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2} \right), \quad (5.2)$$

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

5.2 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 Φ according to the details in Section 5.1 and compute $\mathbf{y} = \Phi\mathbf{x}$. 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 [29, 30], a general purpose

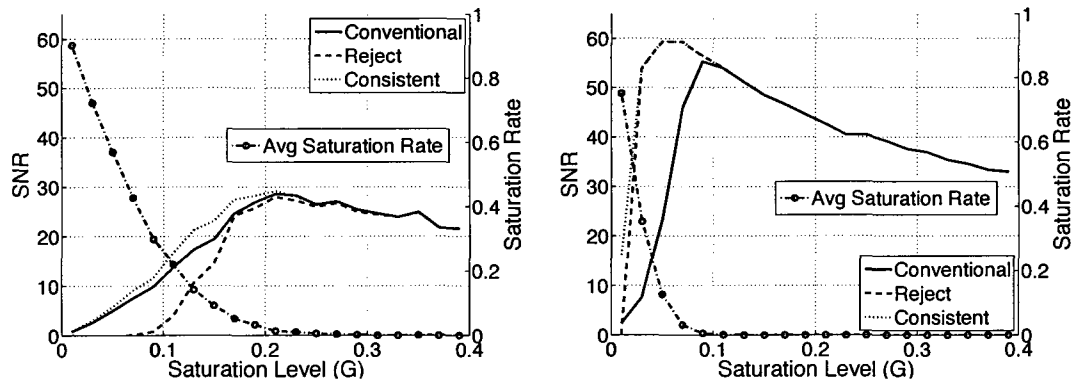
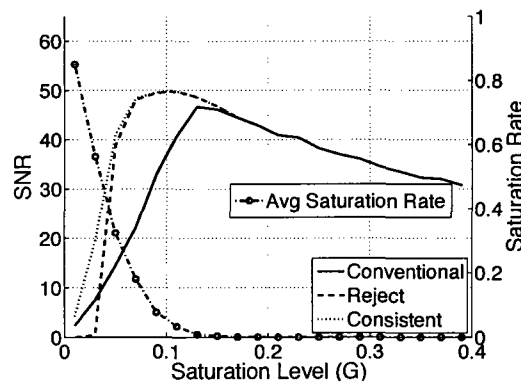
(a) $M/N = 2/16$ (c) $M/N = 15/16$ (b) $M/N = 6/16$

Figure 5.1 : 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 $M/N = 15/16$.

optimization package.

Figures 5.1(a), 5.1(b), and 5.1(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

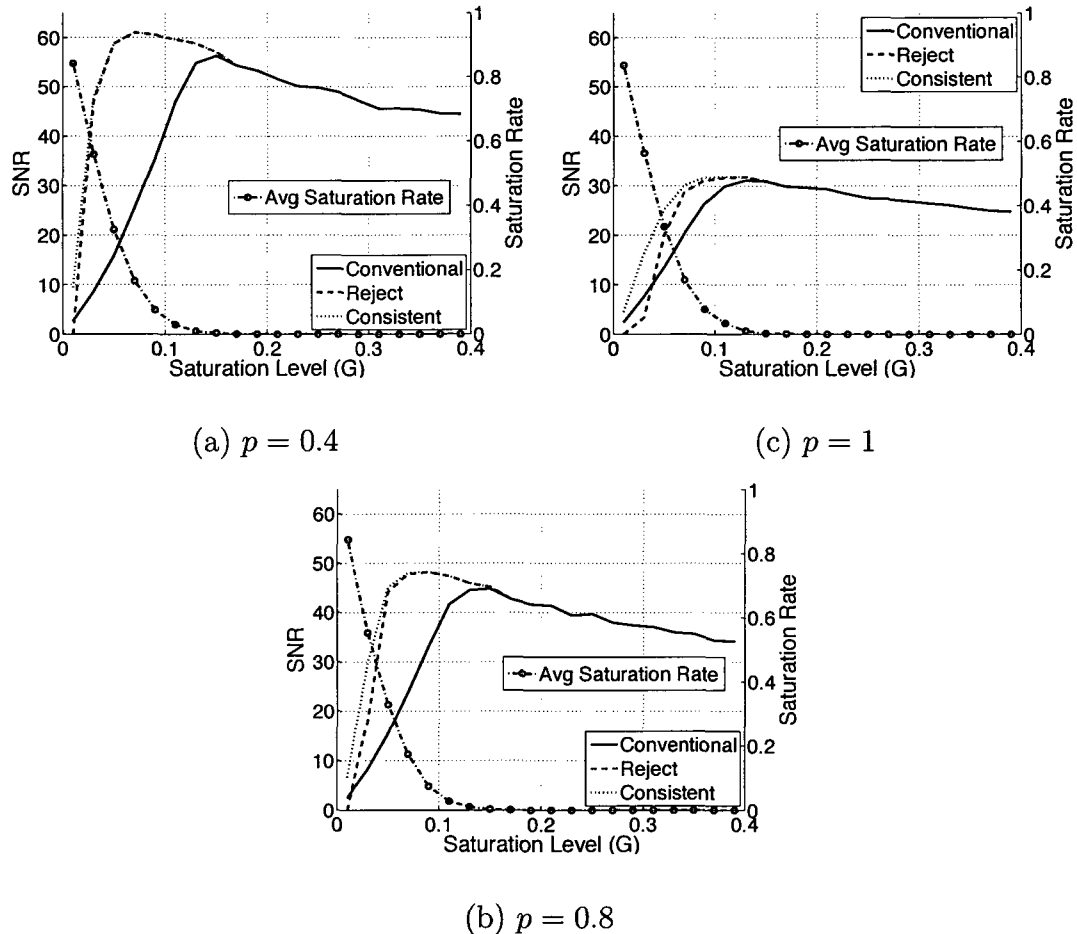


Figure 5.2 : Comparison of reconstruction approaches using CVX for weak ℓ_p compressible signals with $N = 1024$, $M/N = 6/16$, and $B = 4$. Solid line depicts reconstruction for the conventional approach. Dotted line depicts reconstruction for the consistent approach. Dashed line depicts reconstruction 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 $p = 1$.

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 5.1(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 5.1(b), the three lines meet at $G = 0.15$. Both the consistent and the rejection approaches achieve their optimal performance at around $G = 0.09$, where the saturation rate is 0.2. In Figure 5.1(c), the three lines meet at $G = 0.1$ and both the consistent and rejection approaches achieve their optimal performance at $G = 0.06$.

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.

5.3 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 exhibit p in (5.1) in this range, for instance, it has been shown that the wavelet coefficients of natural images have decay rates between $p = 0.3$ and $p = 0.7$ [31].

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 5.2 reflect this intuition. Figures 5.2(a), 5.2(b), and 5.2(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 25%, while for $p = 0.8$ the saturation rate can only be 22%, and for $p = 1$ the highest SNR occurs at a saturation rate of 10%. This means that the smaller the p , the more the measurements should be allowed to saturate.

5.4 Robustness to saturation

We also compare the optimal performance between 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 3.4 Algorithm 1.

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

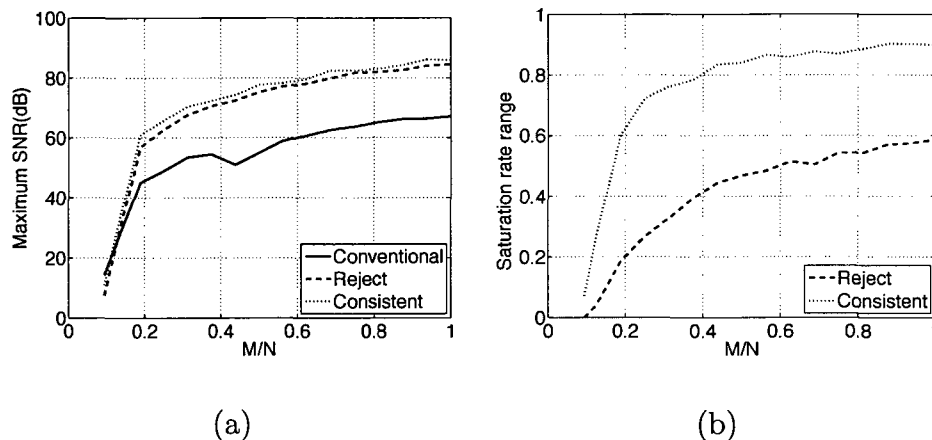


Figure 5.3 : SNR performance using SC-CoSaMP for $N = 1024$, $K = 20$, and $B = 4$. (a) Best-achieved average SNR vs. M/N . (b) Maximum saturation rate such that average SNR performance is as good or better than the best average performance of the conventional approach. For best-case saturation-level 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.

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 5.1 but for a larger range of M .

Figure 5.3(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 num-

ber 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.

Usually, in practice the saturation level that achieves the maximum SNR cannot be efficiently determined or maintained. 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 5.3(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.

The results of this experiment are depicted in Figure 5.3(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 $[0, 0.55]$ while the consistent approach achieves a range of $[0, 0.8]$. Thus, these experiments show that the consistent approach is more robust to saturation rate.

Chapter 6

Extensions

6.1 Automatic gain control (AGC) for CS

Most CS reconstruction approaches (with the exception of [32]) consider finite-length signals \mathbf{x} . However, in many applications of CS the measured signal is a 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

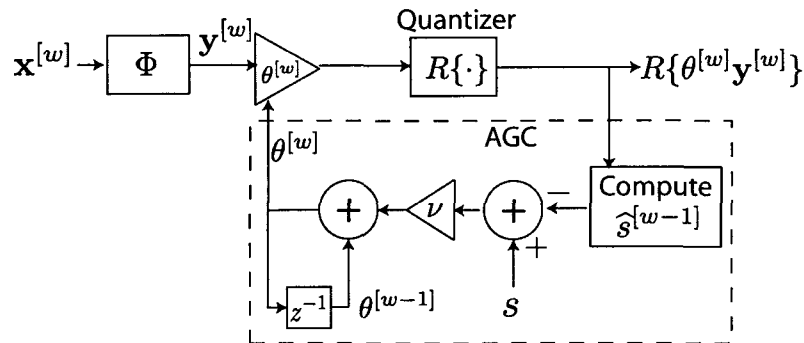


Figure 6.1 : Automatic gain control (AGC) for tuning to nonzero saturation rates in CS systems.

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 easier and in earlier stages of the signal acquisition systems, compared to other measures 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 Φ 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 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 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 6.1 and operates 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.,

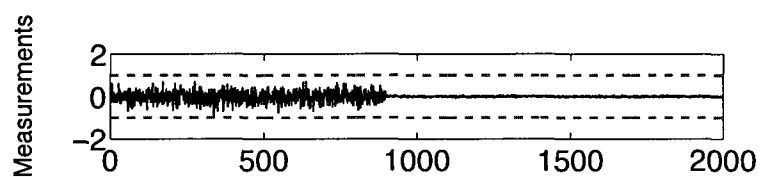
$$\theta^{[w]} = \theta^{[w-1]} + \nu(s - \hat{s}^{[w-1]}), \quad (6.1)$$

where $\nu > 0$ is constant. This negative feedback system is BIBO stable under fairly general conditions on ν [33].

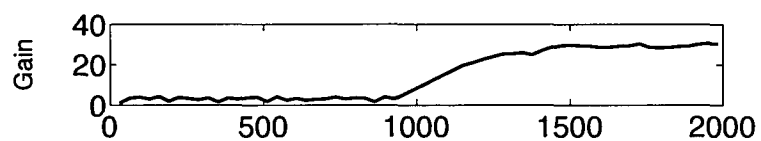
To demonstrate that this AGC is sensitive to both increases in signal strength as well as decreases, we perform an experiment where the signal strength drops suddenly and significantly. The experiment is depicted in Figure 6.2 and was performed as follows. We generated a signal such that the parameters 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 6.2(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 6.2(b) shows the gain that the AGC applies as it receives each measurement. Figure 6.2(c) shows the resulting output signal with quantizer range, and Figure 6.2(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.

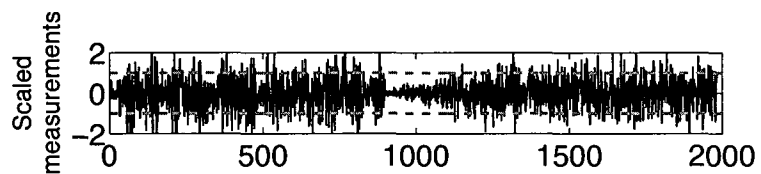
Of course more elaborate gain update loops can be considered to provide better adaptability and more rapid updates to the gain from block to block. Such methods are beyond the scope of this work.



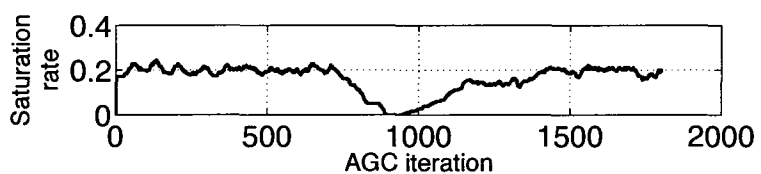
(a)



(b)



(c)



(d)

Figure 6.2 : 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.

Chapter 7

Discussion

In this work, 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.

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 allow some saturation.

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 CS 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.

The democracy property can be used in additional 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 [34] or when applying CS as an multiple description coding (MDC) [35] code.

Appendix A

The expected error of quantized and saturated measurements

In this Appendix, we analyze the mean squared error of Gaussian measurements due to quantization and saturation. Although the CS measurements described in this work are technically not Gaussian, since both the signal \mathbf{x} and the matrix Φ are deterministic, if we were to suppose that the coefficients of the input signal \mathbf{x} are drawn from a random distribution, then by the Lapuyanov variant of the central limit theorem [36], for large enough K , the measurements will be Gaussian. Thus, to motivate why saturation is undesirable, we demonstrate that the expected error on the saturated measurements quickly becomes large for decreasing saturation levels.

The setup is as follows. Let each measurement y be drawn according to $\mathcal{N}(0, \sigma)$ and set the saturation level to be G . We denote the error on the measurements that are below the saturation level as ϵ_Q , the error above the saturation level as ϵ_S , and the total measurement error as $\epsilon = \epsilon_Q + \epsilon_S$. Without loss of generality, we choose $\sigma = 1$.

We now compute the error on quantized measurements below and above the saturation level. Using the quantization interval $\Delta = 2^{-B+1}G$, the distribution of the

error of the quantized measurements can be written as:

$$f_{\tilde{\epsilon}_Q} = \frac{1}{\sqrt{2\pi}} \sum_{n=-(2^{B-1})}^{2^{B-1}} \exp \left\{ -\frac{\left(\frac{nG}{2^{B-1}} + \tilde{\epsilon}_Q\right)^2}{2} \right\}, \quad (\text{A.1})$$

a wrapped truncated Gaussian random variable, and bounded by $\pm\Delta/2$. For small quantization intervals the distribution can be well approximated by a uniform distribution in the same interval, with variance $\Delta^2/12$ [37]. Applying this assumption, and the fact that the expected saturation rate is $2MQ(G)$, the expected squared norm of the quantization error on the measurements that do not saturate is

$$E\|\tilde{\epsilon}_Q\|_2^2 = M(1 - 2Q(G))\Delta^2/12 \quad (\text{A.2})$$

$$= 2^{-2B} M\sigma^2(1 - 2Q(G))G^2/3. \quad (\text{A.3})$$

If we keep the saturated measurements, the expected measurement error is equal to:

$$E\|\epsilon\|_2^2 = M \left((1 - 2Q(G))G^2\frac{\Delta^2}{12} + 2Q(G)\sigma_t^2 \right), \quad (\text{A.4})$$

$$= M \left((1 - 2Q(G))G^2\frac{2^{-2B}}{3} + 2Q(G)\sigma_t^2 \right) \quad (\text{A.5})$$

where

$$\sigma_t^2 = \frac{-G}{\sqrt{2\pi}} \exp \left\{ \frac{-G^2}{2} \right\} + (1 + G^2) Q(G), \quad (\text{A.6})$$

the variance of the tail distribution for a standard Normal random variable, as truncated by the saturation. This result can be found in [16] and the explicit derivation of this is given by the Proposition at the end of this appendix.

In Figure A.1, the dash-dotted line depicts the expected error per measurement due to both saturation and quantization and given by (A.4). The dashed line rep-

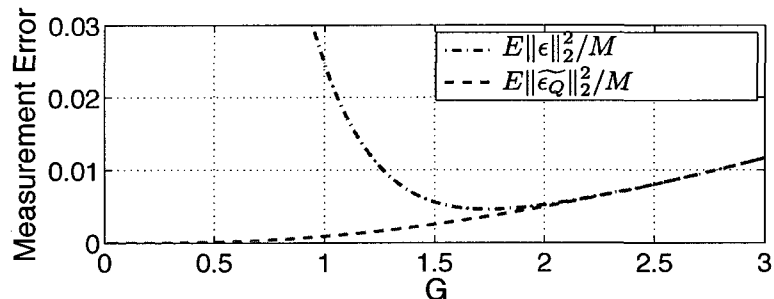


Figure A.1 : Measurement error. Dash-dotted line: Expected measurement error due to quantization and saturation (A.4). Dashed line: Expected measurement error due to quantization only (A.2).

resents the expected error per measurement for the unsaturated quantized measurements only, given by (A.2). Both quantities are depicted as a function of G . We see that as the saturation rate increases, the expected measurement error increases significantly, however, the expected error for just the measurements that do not saturate decreases steadily.

Proposition 1. *Let each element y_i of \mathbf{y} be drawn from a Gaussian distribution with mean zero and variance σ^2 . Let S^+ and S^- correspond to the indices of the positive and negative saturated measurements, respectively, with magnitude above the saturation level G and let \mathbf{y}^{S^+} and \mathbf{y}^{S^-} denote those measurements. We define the vector*

$$\mathring{\mathbf{y}} \triangleq \begin{bmatrix} \mathbf{y}^{S^+} \\ -\mathbf{y}^{S^-} \end{bmatrix}. \quad (\text{A.7})$$

and denote the cardinality of $\mathring{\mathbf{y}}$ as ζ . Then the expected squared error due to saturation

is

$$E \{ \|\dot{\mathbf{y}} - G\|_2^2 \} = \zeta \sigma^2 \left[\frac{-G}{\sqrt{2\pi}} \exp \left\{ \frac{-T^2}{2\sigma^2} \right\} + \left(1 + \frac{G^2}{\sigma^2} \right) Q \left(\frac{G}{\sigma} \right) \right]. \quad (\text{A.8})$$

Proof.

$$E \{ \|\dot{\mathbf{y}} - G\|_2^2 \} = \zeta E \{ (y - G)^2 \}, \quad y > G \quad (\text{A.9})$$

$$= \zeta \int_G^\infty (y - T)^2 p(y) dy \quad (\text{A.10})$$

$$= \zeta \int_G^\infty \frac{y^2 - 2Gy + G^2}{\sigma\sqrt{2\pi}} e^{\frac{-y^2}{2\sigma^2}} dy \quad (\text{A.11})$$

$$= \zeta [(A.15) + (A.18) + (A.19)] \quad (\text{A.12})$$

$$= \zeta \sigma^2 \left[\frac{-G}{\sqrt{2\pi}} \exp \left\{ \frac{-T^2}{2\sigma^2} \right\} + \left(1 + \frac{G^2}{\sigma^2} \right) Q \left(\frac{G}{\sigma} \right) \right]. \quad (\text{A.13})$$

Where the first term from (A.11) is,

$$\frac{1}{\sigma\sqrt{2\pi}} \int_G^\infty y^2 \exp \left\{ \frac{-y^2}{2\sigma^2} \right\} dy = \frac{1}{\sigma\sqrt{2\pi}} \left[\left(-\sigma^2 y \exp \left\{ \frac{-y^2}{2\sigma^2} \right\} \right) \Big|_G^\infty + \int_G^\infty -\sigma^2 \exp \left\{ \frac{-y^2}{2\sigma^2} \right\} dy \right] \quad (\text{A.14})$$

$$= 0 + \frac{\sigma^2 G}{\sqrt{2\pi}} \exp \left\{ \frac{-G^2}{2\sigma^2} \right\} + \sigma^2 Q \left(\frac{G}{\sigma} \right), \quad (\text{A.15})$$

the second term from (A.11) is,

$$\frac{-2G}{\sigma\sqrt{2\pi}} \int_G^\infty y \exp \left\{ \frac{-y^2}{2\sigma^2} \right\} dx = \frac{-2G}{\sigma\sqrt{2\pi}} \left(-\sigma^2 \exp \left\{ \frac{-y^2}{2\sigma^2} \right\} \right) \Big|_G^\infty \quad (\text{A.16})$$

$$= \frac{2G\sigma^2}{\sqrt{2\pi}} \left(0 - \exp \left\{ \frac{-G^2}{2\sigma^2} \right\} \right) \quad (\text{A.17})$$

$$= \frac{-2G\sigma^2}{\sqrt{2\pi}} \exp \left\{ \frac{-G^2}{2\sigma^2} \right\}, \quad (\text{A.18})$$

and the third term from (A.11) is,

$$\frac{G^2}{\sigma\sqrt{2\pi}} \int_G^\infty \exp \left\{ \frac{-y^2}{2\sigma^2} \right\} dy = G^2 Q \left(\frac{G}{\sigma} \right). \quad (\text{A.19})$$

□

Appendix B

Preservation of inner products

Theorem 3. *Given an $M \times N$ matrix Φ with K -RIP and vectors \mathbf{u} and \mathbf{v} of length N such that $\mathbf{u} - \mathbf{v}$ is K -sparse, then the difference of the inner products is bounded as*

$$\left| \frac{2}{b+a} \langle \Phi \mathbf{u}, \Phi \mathbf{v} \rangle - \langle \mathbf{u}, \mathbf{v} \rangle \right| \leq \frac{b-a}{b+a} \|\mathbf{u}\|_2 \|\mathbf{v}\|_2. \quad (\text{B.1})$$

Proof. The derivation of (B.1) is as follows. Consider the following property of Φ applied to vectors \mathbf{u}, \mathbf{v} :

$$a \|\mathbf{u} - \mathbf{v}\|_2^2 \leq \|\Phi \mathbf{u} - \Phi \mathbf{v}\|_2^2 \leq b \|\mathbf{u} - \mathbf{v}\|_2^2. \quad (\text{B.2})$$

By the RIP of Φ , this is true for \mathbf{z} that are K -sparse, where $\mathbf{z} = \mathbf{u} - \mathbf{v}$. We assume that $\|\mathbf{u}\|_2 = \|\mathbf{v}\|_2 = 1$. Additionally, we have the property

$$\|\mathbf{u} \pm \mathbf{v}\|_2^2 = \|\mathbf{u}\|_2^2 + \|\mathbf{v}\|_2^2 \pm 2\langle \mathbf{u}, \mathbf{v} \rangle, \quad (\text{B.3})$$

and thus from (B.2) we have

$$a \leq \frac{\|\Phi \mathbf{u} \pm \Phi \mathbf{v}\|_2^2}{2 \pm 2\langle \mathbf{u}, \mathbf{v} \rangle} \leq b. \quad (\text{B.4})$$

The parallelogram identity states that,

$$|\langle \Phi \mathbf{u}, \Phi \mathbf{v} \rangle| \leq \frac{1}{4} \left| \|\Phi \mathbf{u} + \Phi \mathbf{v}\|_2^2 - \|\Phi \mathbf{u} - \Phi \mathbf{v}\|_2^2 \right| \quad (\text{B.5})$$

$$\leq \frac{1}{4} |2b + 2b\langle \mathbf{u}, \mathbf{v} \rangle - 2a + 2a\langle \mathbf{u}, \mathbf{v} \rangle| \quad (\text{B.6})$$

$$\leq \frac{b-a}{2} + \frac{b+a}{2} \langle \mathbf{u}, \mathbf{v} \rangle. \quad (\text{B.7})$$

By replacing $\|\Phi\mathbf{u} + \Phi\mathbf{v}\|_2^2$ with $\|\Phi\mathbf{u} - \Phi\mathbf{v}\|_2^2$ and vice versa in (B.5), we can also find that

$$|\langle \Phi\mathbf{u}, \Phi\mathbf{v} \rangle| \leq \frac{b-a}{2} - \frac{b+a}{2} \langle \mathbf{u}, \mathbf{v} \rangle, \quad (\text{B.8})$$

and thus, achieve the bound:

$$\left| \frac{2}{b+a} \langle \Phi\mathbf{u}, \Phi\mathbf{v} \rangle - \langle \mathbf{u}, \mathbf{v} \rangle \right| \leq \frac{b-a}{b+a}. \quad (\text{B.9})$$

Since \mathbf{u} and \mathbf{v} are unit norm, we can write this expression as

$$\left| \frac{2}{b+a} \left\langle \Phi \frac{\mathbf{u}}{\|\mathbf{u}\|_2}, \Phi \frac{\mathbf{v}}{\|\mathbf{v}\|_2} \right\rangle - \left\langle \frac{\mathbf{u}}{\|\mathbf{u}\|_2}, \frac{\mathbf{v}}{\|\mathbf{v}\|_2} \right\rangle \right|, \quad (\text{B.10})$$

and thus, from the bilinearity of the inner product we obtain the result:

$$\left| \frac{2}{b+a} \langle \Phi\mathbf{u}, \Phi\mathbf{v} \rangle - \langle \mathbf{u}, \mathbf{v} \rangle \right| \leq \frac{b-a}{b+a} \|\mathbf{u}\|_2 \|\mathbf{v}\|_2. \quad (\text{B.11})$$

□

Corollary 1. *If the elements of the $\Phi\mathbf{u}$ and $\Phi\mathbf{v}$ are uniformly quantized with quantization width Δ , denoted by $R(\Phi\mathbf{u})$ and $R(\Phi\mathbf{v})$, then the difference of the inner products is bounded by*

$$\left| \frac{2}{b+a} \langle R(\Phi\mathbf{u}), R(\Phi\mathbf{v}) \rangle - \langle \mathbf{u}, \mathbf{v} \rangle \right| \leq \frac{b-a}{b+a} \|\mathbf{u}\|_2 \|\mathbf{v}\|_2 + \left(b\Delta\sqrt{M} + \frac{M\Delta^2}{4} \right) \|\mathbf{u}\|_2 \|\mathbf{v}\|_2. \quad (\text{B.12})$$

Proof. The derivation of (B.12) is as follows. We will show that

$$|\langle R(\Phi\mathbf{u}), R(\Phi, \mathbf{v}) \rangle| \leq \langle \Phi\mathbf{u}, \Phi\mathbf{v} \rangle + b\Delta\sqrt{M} + M\frac{\Delta^2}{4}, \quad (\text{B.13})$$

and thus,

$$\left| \frac{2}{b+a} \langle R(\Phi \mathbf{u}), R(\Phi \mathbf{v}) \rangle - \langle \mathbf{u}, \mathbf{v} \rangle \right| \leq \quad (\text{B.14})$$

$$\left| \frac{2}{b+a} \langle \Phi \mathbf{u}, \Phi \mathbf{v} \rangle - \langle \mathbf{u}, \mathbf{v} \rangle \right| + b\Delta\sqrt{M} + M\frac{\Delta^2}{4}. \quad (\text{B.15})$$

We then obtain the stated result by applying (B.1) to (B.14).

Without the loss of generality, assume that $\|\mathbf{u}\|_2 = \|\mathbf{v}\|_2 = 1$. The quantized measurements can be written as $R(\Phi \mathbf{u}) = \Phi \mathbf{u} + \mathbf{e}_1$ and $R(\Phi \mathbf{v}) = \Phi \mathbf{v} + \mathbf{e}_2$ where \mathbf{e}_1 and \mathbf{e}_2 are error vectors with each element bounded between $-\Delta/2$ and $\Delta/2$. Thus, we can write

$$\langle R(\Phi \mathbf{u}), R(\Phi \mathbf{v}) \rangle = \langle \Phi \mathbf{u} + \mathbf{e}_1, \Phi \mathbf{v} + \mathbf{e}_2 \rangle \quad (\text{B.16})$$

$$= \langle \Phi \mathbf{u}, \Phi \mathbf{v} \rangle + \langle \Phi \mathbf{u}, \mathbf{e}_2 \rangle \quad (\text{B.17})$$

$$+ \langle \Phi \mathbf{v}, \mathbf{e}_1 \rangle + \langle \mathbf{e}_1, \mathbf{e}_2 \rangle. \quad (\text{B.18})$$

The magnitude of the cross term $\langle \Phi \mathbf{u}, \mathbf{e}_2 \rangle$ can be bounded as

$$|\langle \Phi \mathbf{u}, \mathbf{e}_2 \rangle| \leq b\|\mathbf{u}\|_2 \frac{\Delta}{2} \sqrt{M} = b\frac{\Delta}{2} \sqrt{M}, \quad (\text{B.19})$$

and from $\langle \Phi \mathbf{v}, \mathbf{e}_1 \rangle$ we can obtain the same upper bound. Finally, we can bound the inner product of the two errors as

$$|\langle \mathbf{e}_1, \mathbf{e}_2 \rangle| \leq M\frac{\Delta^2}{4}. \quad (\text{B.20})$$

and thus, our claim in (B.13) holds and we achieve the desired result. \square

Appendix C

Consistent recovery via fixed point continuation

The algorithm *fixed point continuation* (FPC) [38, 39] has been used for recovery of sparse signals within the CS framework. In this appendix, we demonstrate how this algorithm can be modified for consistent recovery with saturated measurements. Our approach closely follows that of [25] by employing a one-sided quadratic function to ensure consistent reconstruction. We define $\mathring{\Phi}$, $\mathbf{1}$, and elements with $\widetilde{(\cdot)}$ as in Section 3.

To solve the program defined by (3.3), saturation consistent FPC (SC-FPC) finds the solution to

$$\hat{\mathbf{x}} = \underset{\mathbf{x}}{\operatorname{argmin}} \|\mathbf{x}\|_1 + \frac{\mu}{2} \|\widetilde{\Phi}\mathbf{x} - \widetilde{\mathbf{y}}\|_2^2 + \mu\bar{h} \left(\mathring{\Phi}\mathbf{x} - G \cdot \mathbf{1} \right) \quad (\text{C.1})$$

where $h(\cdot)$ is a one-sided quadratic penalty defined as

$$h(y_i) = \begin{cases} \frac{y_i^2}{2}, & y_i < 0 \\ 0, & y_i \geq 0. \end{cases} \quad (\text{C.2})$$

This function is applied element-wise to a vector as $\bar{h}(\mathbf{y}) = \sum_i h(y_i)\mathbf{e}_i$ where \mathbf{e}_i is the i^{th} canonical vector. Thus, in words, this program seeks find the \mathbf{x} with minimum ℓ_1 norm, with a quadratic penalty on the measurements that did not saturate and a one-sided element-wise quadratic penalty on those that did.

The steps of the SC-FPC are enumerated in Algorithm 2. The initialization and stopping criteria are taken from [39]. The latter rely on proven FPC convergence

results. The algorithm consists of two loops, the outer loop which serves to update the parameter μ , used in an ℓ_1 gradient descent, and the inner loop which performs a gradient descent on the quadratic penalties. The primary modification to the original FPC is in the calculation of the gradient step, but we provide the details of each step of the algorithm for completeness. A detailed analysis of this algorithm is given in [38]. The inner loop steps are as follows.

Step 5 computes the gradient of the quadratic components of the cost function with respect to \mathbf{x} . Specifically, this is

$$\frac{\partial}{\partial \mathbf{x}} \left(\frac{1}{2} \|\tilde{\Phi} \mathbf{x} - \tilde{\mathbf{y}}\|_2^2 + \bar{h} \left(\dot{\Phi} \mathbf{x} - G \cdot \mathbf{1} \right) \right) = \tilde{\Phi}^T \left(\tilde{\Phi} \hat{\mathbf{x}} - \tilde{\mathbf{y}} \right) + \dot{\Phi}^T \bar{h}' \left(\dot{\Phi} \mathbf{x} - G \cdot \mathbf{1} \right), \quad (\text{C.3})$$

where the each element of the derivative of the one-sided quadratic is

$$\bar{h}'(\mathbf{y})_i = \begin{cases} y_i, & y_i < 0 \\ 0, & y_i \geq 0. \end{cases} \quad (\text{C.4})$$

Step 6 performs gradient descent on the current estimate $\hat{\mathbf{x}}^{[n]}$ with respect to the cost function to produce an intermediate vector $\mathbf{b} = \hat{\mathbf{x}}^{[n]} - \tau \mathbf{g}$. We choose the parameter τ to be the same as the heuristics given for the original FPC.

Step 7 performs the descent on ℓ_1 component of the cost function by finding the solution to

$$\hat{\mathbf{x}} = \min_{\mathbf{x}} \|\mathbf{x}\|_1 + \frac{\mu}{2\tau} \|\mathbf{x} - \mathbf{b}\|_2^2. \quad (\text{C.5})$$

This can be efficiently computed via the shrinkage function,

$$\text{shrink}(\mathbf{b}, \tau/\mu) = \text{sign}((\mathbf{b})_i) \cdot \max \left\{ |(\mathbf{b})_i| - \frac{\tau}{\mu}, 0 \right\}, \quad (\text{C.6})$$

applied to each element $(\mathbf{b})_i$ of the vector \mathbf{b} .

These steps are repeated until the convergence criteria are satisfied.

Algorithm 2 Saturation consistent FPC

- 1: **Input:** \mathbf{y} , Φ , G , $\bar{\mu}$, $xtol$, and $gtol$
 - 2: **Initialize:** $\tau \leftarrow 2 - \epsilon$, $\eta \leftarrow 4$ $\hat{\mathbf{x}}^{[0]} \leftarrow \tau \cdot \Phi^T \mathbf{y}$, $\mu \leftarrow \frac{1}{\|\Phi^T \mathbf{y}\|_\infty}$, $n \leftarrow 0$
 - 3: **while** $\mu < \bar{\mu}$ **do**
 - 4: **while** $\sqrt{\frac{\mu}{\bar{\mu}}} \cdot \frac{\|\mathbf{x}^{[n]} - \mathbf{x}^{[n-1]}\|_2}{\|\mathbf{x}^{[n-1]}\|_2} > xtol$ or $\mu \cdot \|\mathbf{g}\|_\infty > gtol$ **do**
 - 5: **Compute gradient:**

$$\mathbf{g} \leftarrow \tilde{\Phi}^T \left(\tilde{\Phi} \hat{\mathbf{x}}^{[n]} - \tilde{\mathbf{y}} \right) + \dot{\Phi}^T \bar{h}' \left(\dot{\Phi} \hat{\mathbf{x}}^{[n]} - G \cdot \mathbf{1} \right)$$
 - 6: **Gradient descent:**

$$\mathbf{b} \leftarrow \hat{\mathbf{x}}^{[n]} - \tau \cdot \mathbf{g}$$
 - 7: **Shrinkage:**

$$\mathbf{x}^{[n+1]} \leftarrow \text{sign}((\mathbf{b})_i) \cdot \max \left\{ |(\mathbf{b})_i| - \frac{\tau}{\mu}, 0 \right\}$$
 - 8: **Update iteration:**

$$n \leftarrow n + 1$$
 - 9: **end while**
 - 10: **Update μ :**

$$\mu \leftarrow \min\{\eta \cdot \mu, \bar{\mu}\}$$
 - 11: **end while**
-

Appendix D

Proof of the democracy of Gaussian matrices

D.1 Concentration of measure

In the proof of Lemma 1 we make use of *order statistics*. Given a sequence of i.i.d. random variables that have been sorted from smallest to largest, an order statistic is the distribution of the variable at a particular position. To make use of this, we first present the formal definition with notation of an order statistic and then the proof.

Definition 1. Let y_i for $i = 1, \dots, M$ be a sequence of i.i.d. random variables. Denote $y_{\widetilde{M}:M}$ to be the \widetilde{M} -th largest element when the sequence is ordered from smallest to largest. Then $y_{\widetilde{M}:M}$ is called the \widetilde{M} -th order statistic of the M variables.

Lemma 1. Suppose that Φ is an $M \times N$ matrix whose entries $\phi_{ij} \sim \mathcal{N}(0, 1)$. Let $\alpha \in (0, 1)$, $\beta \in (1, \infty)$, and $0 < \widetilde{M} \leq M$ be given. Then for any $\mathbf{x} \in \mathbb{R}^N$, we have that

$$\alpha \widetilde{M} \|\mathbf{x}\|_2^2 \leq \|\Phi^\Gamma \mathbf{x}\|_2^2 \leq \beta \widetilde{M} \|\mathbf{x}\|_2^2 \quad (\text{D.1})$$

holds for all sets Γ with $|\Gamma| = \widetilde{M}$ that index the rows of Φ with probability exceeding

$1 - P_\alpha - P_\beta$, where

$$P_\alpha \leq \widetilde{M} \sqrt{\frac{2}{\pi}} \left(\frac{e^{\lambda\alpha}}{\sqrt{1+2\lambda}} \right)^{\widetilde{M}} \quad (\text{D.2})$$

$$\times \int_0^\infty \frac{(1 - 2Q(u\sqrt{1+2\lambda}))^{\widetilde{M}}}{(1 - 2Q(u))^{\widetilde{M}+1}} e^{-u^2/2} \quad (\text{D.3})$$

$$\times \mathcal{B}_{\widetilde{M}}(M, 1 - 2Q(u)) du,$$

for all $\lambda > 0$ and

$$P_\beta \leq (M - \widetilde{M}) \sqrt{\frac{2}{\pi}} \left(\frac{e^{-\lambda\beta}}{\sqrt{1-2\lambda}} \right)^{\widetilde{M}} \quad (\text{D.4})$$

$$\times \int_0^\infty \frac{(Q(t\sqrt{1-2\lambda}))^{\widetilde{M}}}{Q(t)^{\widetilde{M}}(1 - 2Q(t))} e^{-t^2/2} \quad (\text{D.5})$$

$$\times \mathcal{B}_{\widetilde{M}}(M, 2Q(t)) dt,$$

for all $\lambda \in (0, \frac{1}{2})$. In each bound

$$\mathcal{B}_d(n, p) = \binom{n}{d} p^d (1-p)^{n-d} \quad (\text{D.6})$$

is the Binomial distribution function and $Q(z) = \frac{1}{\sqrt{2\pi}} \int_z^{+\infty} e^{-t^2/2} dt$ is the tail integral of the standard Gaussian distribution.

Proof: First observe that it suffices to prove the lemma for the case where $\|\mathbf{x}\|_2 = 1$ since both the norm and submatrices of Φ are linear. Thus, assume without loss of generality that $\|\mathbf{x}\|_2 = 1$. We now wish to obtain upper and lower bounds on $\|\Phi^\Gamma \mathbf{x}\|_2^2$ where Φ^Γ is an arbitrary $\widetilde{M} \times N$ submatrix of Φ . A key observation is that in order to establish (D.1) we do not need to consider every possible submatrix, since any submatrix can be bounded by two special cases. For the lower bound we need only consider the matrix obtained by selecting the \widetilde{M} rows of Φ corresponding to the

\widetilde{M} entries of $\Phi \mathbf{x}$ with smallest magnitude, since by removing the largest entries we decrease the norm by the maximum amount possible. Similarly, for the upper bound we need only consider the matrix obtained by selecting the \widetilde{M} rows of Φ corresponding to the entries of $\Phi \mathbf{x}$ with largest magnitude. We will let $\mathbf{y} = \Phi \mathbf{x}$ denote the random vector obtained by retaining all rows of Φ .

We begin by deriving the lower bound. Let the function $U(\mathbf{y})$ map \mathbf{y} to the \widetilde{M} smallest magnitude elements of \mathbf{y} and let $f_{\widetilde{M}:M}(u)$ be the PDF of $|y_i|_{\widetilde{M}:M}$, the order statistic of the \widetilde{M} -th largest magnitude element of \mathbf{y} . If for a particular instance of \mathbf{y} , u is the value of \widetilde{M} -th largest magnitude element of \mathbf{y} , then we have that $U(\mathbf{y}) = \{y_i : |y_i| \leq u\}$. We begin by considering

$$P_\alpha = \mathbb{P}(\|\mathbf{y}\|_2^2 \leq \alpha \widetilde{M}) \quad (\text{D.7})$$

$$= \int_0^\infty \mathbb{P}(\|\mathbf{y}\|_2^2 \leq \alpha \widetilde{M} | u) f_{\widetilde{M}:M}(u) du. \quad (\text{D.8})$$

We will estimate $\mathbb{P}(\|\mathbf{y}\|_2^2 \leq \alpha \widetilde{M} | u)$ using Markov's inequality, from which we observe that for any $\lambda > 0$

$$\mathbb{P}(\|\mathbf{y}\|_2^2 \leq \alpha \widetilde{M} | u) \quad (\text{D.9})$$

$$\begin{aligned} &= \mathbb{P}\left(e^{\lambda \|\mathbf{y}\|_2^2} \leq e^{\lambda \alpha \widetilde{M}} | u\right) \\ &\leq \frac{\mathbb{E}\left(e^{-\lambda \|\mathbf{y}\|_2^2} | u\right)}{e^{-\lambda \alpha \widetilde{M}}} \\ &= \frac{\prod_{i=1}^{\widetilde{M}} \mathbb{E}\left(e^{-\lambda y_i^2} | u\right)}{e^{-\lambda \alpha \widetilde{M}}} \\ &= \frac{\mathbb{E}\left(e^{-\lambda y_1^2} | u\right)^{\widetilde{M}}}{e^{-\lambda \alpha \widetilde{M}}}, \end{aligned} \quad (\text{D.10})$$

where the last two steps follow since the $y_i | u$ are independent and identically dis-

tributed. We now wish to compute $\mathbb{E}(e^{-\lambda y_1^2} | u)$. In order to do so, we must determine the distribution of $y_1 | u$. We begin by observing that since $\phi_{ij} \sim \mathcal{N}(0, 1)$ and $\|x\|_2 = 1$, we have that $y_1 \sim \mathcal{N}(0, 1)$. Thus, the PDF of $y_1 | u$ is given by

$$f(y_1 | u) = \tag{D.11}$$

$$= \begin{cases} \frac{1}{\sqrt{2\pi \cdot (1-2Q(u))}} \cdot e^{-y_1^2/2}, & |y_1| \leq u \\ 0, & |y_1| > u, \end{cases} \tag{D.12}$$

where $Q(z) = \frac{1}{\sqrt{2\pi}} \int_z^{+\infty} e^{-t^2/2} dt$ is the tail integral of the standard Gaussian distribution.

Returning to (D.10), we can now write that

$$\begin{aligned} \mathbb{E}\left(e^{-\lambda y_1^2} \mid u\right) &= \\ &= \int_{-\infty}^{\infty} e^{-\lambda y_1^2} f(y_1 | u) dy_1 \\ &= \int_{-u}^u e^{-\lambda y_1^2} \frac{e^{-y_1^2/2}}{\sqrt{2\pi(1-2Q(u))}} dy_1 \\ &= \frac{1-2Q(u\sqrt{1+2\lambda})}{(1-2Q(u))\sqrt{1+2\lambda}} \\ &\quad \times \int_{-u}^u \frac{\sqrt{1+2\lambda} e^{-y_1^2(1+2\lambda)/2}}{\sqrt{2\pi(1-2Q(u\sqrt{1+2\lambda}))}} dy_1 \\ &= \frac{1-2Q(u\sqrt{1+2\lambda})}{(1-2Q(u))\sqrt{1+2\lambda}}, \end{aligned}$$

for any $\lambda \geq 0$, where the last equality follows since the integrand is the PDF of a truncated Gaussian random variable, and hence integrates to 1.

Thus, by substituting the bound (D.9) we obtain

$$P_\alpha \leq \int_0^\infty \left(\frac{1-2Q(u\sqrt{1+2\lambda})}{(1-2Q(u))\sqrt{1+2\lambda}} \right)^{\tilde{M}} f_{\tilde{M}:M}(u) du. \tag{D.13}$$

We complete the proof of the bound on P_α by applying the expression for $f_{\widetilde{M}:M}(u)$,

$$\begin{aligned} & \sqrt{\frac{2}{\pi}} \frac{M!}{(\widetilde{M}-1)!(M-\widetilde{M})!} \\ & \times (1-2Q(u))^{\widetilde{M}-1} (2Q(u))^{M-\widetilde{M}} e^{-u^2/2}. \end{aligned} \quad (\text{D.14})$$

We derive this PDF using the fact that the magnitudes of the elements are distributed as $|y_i| \sim \chi_1$, a standard chi distribution, and the standard formula for the PDF of an order statistic [40].

In order to establish P_β , we define $T(\mathbf{y})$ to be a function that maps \mathbf{y} to the \widetilde{M} greatest magnitude elements of \mathbf{y} and $f_{(M-\widetilde{M}):M}(t)$ to be the PDF of $|y_i|_{(M-\widetilde{M}):M}$, the order statistic of the $(M-\widetilde{M})$ -th largest magnitude element of \mathbf{y} . To find $P_\beta = \mathbb{P}(\|T(\mathbf{y})\|_2^2 \geq \beta\widetilde{M})$, we again apply Markov's inequality to obtain that for any $\lambda > 0$

$$\mathbb{P}(\|\mathbf{y}\|_2^2 \geq \beta\widetilde{M} \mid t) \leq \frac{\mathbb{E}\left(e^{\lambda y_1^2} \mid t\right)^{\widetilde{M}}}{e^{\lambda\beta\widetilde{M}}} \quad (\text{D.15})$$

where t denotes the value of the $M-\widetilde{M}$ -th largest (or \widetilde{M} -th smallest) magnitude element of \mathbf{y} . This bound follows from the same argument used to establish (D.10).

In this case, y_1 has the PDF given by

$$f(y_1|t) = \begin{cases} 0, & |y_1| < t \\ \frac{1}{\sqrt{2\pi(2Q(t))}} e^{-y_1^2/2}, & |y_1| \geq t. \end{cases} \quad (\text{D.16})$$

One can now use the same approach as before to establish that

$$\mathbb{E}\left(e^{\lambda y_1^2} \mid t\right) = \frac{Q(t\sqrt{1-2\lambda})}{Q(t)\sqrt{1-2\lambda}}, \quad (\text{D.17})$$

for all $\lambda \in (0, \frac{1}{2})$. Thus, P_β is bounded as

$$P_\beta \leq \int_0^\infty \left(\frac{Q(t\sqrt{1-2\lambda})}{Q(t)\sqrt{1-2\lambda}} \right)^{\widetilde{M}} \frac{1}{e^{\lambda\beta}} f_{(M-\widetilde{M}):M}(t) dt. \quad (\text{D.18})$$

We substitute the expression for $f_{(M-\widetilde{M}):M}(t)$,

$$\begin{aligned} & \sqrt{\frac{2}{\pi}} \frac{M!}{(M-\widetilde{M}-1)!\widetilde{M}!} \\ & \times (1-2Q(t))^{M-\widetilde{M}-1} (2Q(t))^{\widetilde{M}} e^{-t^2/2}, \end{aligned} \quad (\text{D.19})$$

and use the identity

$$\mathcal{B}_{M-\widetilde{M}}(M, 1-2Q(t)) = \mathcal{B}_{\widetilde{M}}(M, 2Q(t)), \quad (\text{D.20})$$

to complete the proof. \square

D.2 Democracy

Theorem 2. [Democracy] *Suppose that Φ is an $M \times N$ matrix with entries $\phi_{ij} \sim \mathcal{N}(0, 1/\widetilde{M})$, where $0 < \widetilde{M} \leq M$. Let $a > 0$ and $b > a$ be given. Then with probability at least $1 - P_F$, we have that all $\widetilde{M} \times N$ submatrices Φ^Γ of Φ satisfy*

$$a\|\mathbf{x}\|_2^2 \leq \|\widetilde{\Phi}\mathbf{x}\|_2^2 \leq b\|\mathbf{x}\|_2^2 \quad (\text{D.21})$$

for all $\mathbf{x} \in \Sigma_K$, where

$$P_F \leq \left(\frac{3eN}{\epsilon K}\right)^K (P_\alpha + P_\beta), \quad (\text{D.22})$$

for $0 < \epsilon < (\sqrt{b} - \sqrt{a})/2\sqrt{b}$.

Proof: First note that it is enough to prove (D.21) in the case $\|\mathbf{x}\|_2 = 1$, since all submatrices of Φ are linear. Next, fix an index set $I \subset \{1, 2, \dots, N\}$ with $|I| = K$, and let X_I denote the K -dimensional subspace spanned by the columns of Φ indexed by I . We choose a finite set of points S_I such that $S_I \subseteq X_I$, $\|\mathbf{s}\|_2 \leq 1$ for all $\mathbf{s} \in S_I$,

and for all $\mathbf{x} \in X_I$ with $\|\mathbf{x}\|_2 \leq 1$ we have

$$\min_{\mathbf{s} \in S_I} \|\mathbf{x} - \mathbf{s}\|_2 \leq \epsilon. \quad (\text{D.23})$$

One can show (see Ch. 15 of [41]) that such a set S_I exists with $|S_I| \leq (3/\epsilon)^K$.

We then repeat this process for each possible index set I , and collect all the sets S_I together

$$S = \bigcup_{I:|I|=K} S_I. \quad (\text{D.24})$$

There are $\binom{N}{K} \leq (eN/K)^K$ possible index sets I , and hence $|S| \leq (3eN/\epsilon K)^K$. We now use the union bound to apply Lemma 1 to this set of points with $\alpha = (\sqrt{a} + \epsilon\sqrt{b})^2$ and $\beta = b(1 - \epsilon)^2$, with the result that, with probability exceeding (D.22) we have

$$\alpha \|\mathbf{s}\|_2^2 \leq \|\Phi^\Gamma \mathbf{s}\|_2^2 \leq \beta \|\mathbf{s}\|_2^2, \quad \text{for all } \mathbf{s} \in S.$$

One can easily check that provided that $\epsilon < (\sqrt{b} - \sqrt{a})/2\sqrt{b}$, α and β satisfy

$$\alpha < \left(\frac{\sqrt{b} - \sqrt{a}}{2} \right)^2 < \beta.$$

We now define B as the smallest number such that

$$\|\Phi^\Gamma \mathbf{x}\|_2^2 \leq B \|\mathbf{x}\|_2^2, \quad \text{for all } \mathbf{x} \in \Sigma_K, \|\mathbf{x}\|_2 \leq 1. \quad (\text{D.25})$$

Our goal is to show that $B \leq b$. For this, we recall that for any $\mathbf{x} \in \Sigma_K$ with $\|\mathbf{x}\|_2 \leq 1$, we can pick a $\mathbf{s} \in S$ such that $\|\mathbf{x} - \mathbf{s}\|_2 \leq \epsilon$ and such that $\mathbf{x} - \mathbf{s} \in \Sigma_K$ (since if $\mathbf{x} \in X_I$, we can pick $\mathbf{s} \in S_I \subset X_I$ satisfying $\|\mathbf{x} - \mathbf{s}\|_2 \leq \epsilon$). In this case we have

$$\|\Phi^\Gamma \mathbf{x}\|_2 \leq \|\Phi^\Gamma \mathbf{s}\|_2 + \|\Phi^\Gamma (\mathbf{x} - \mathbf{s})\|_2 \leq \sqrt{\beta} + \sqrt{B}\epsilon.$$

Since by definition B is the smallest number for which (D.25) holds, we obtain $\sqrt{B} \leq \sqrt{\beta} + \sqrt{B}\epsilon$, which upon rearranging yields $\sqrt{B} \leq \sqrt{\beta}/(1 - \epsilon) = \sqrt{b}$ as desired. We have thus proven the upper inequality in (D.21). The lower inequality follows from this since

$$\|\Phi^\Gamma \mathbf{x}\|_2 \geq \|\Phi^\Gamma \mathbf{s}\|_2 - \|\Phi^\Gamma(\mathbf{x} - \mathbf{s})\|_2 \geq \sqrt{\alpha} - \sqrt{b}\epsilon = \sqrt{a},$$

which completes the proof. □

Bibliography

- [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. H. Walden, “Analog-to-digital converter survey and analysis,” *IEEE Journal on Selected Areas in Communications*, vol. 17, no. 4, pp. 539–550, April 1999.
- [3] D. Donoho, “Compressed sensing,” in *IEEE Trans. on Information Theory*, vol. 6, no. 4, April 2006, pp. 1289 – 1306.
- [4] E. Candès, “Compressive sampling,” in *Int. Congress of Mathematics*, vol. 3, 2006, pp. 1433–1452.
- [5] M. Vetterli, P. Marziliano, and T. Blu, “Sampling signals with finite rate of innovation,” *IEEE Trans. on Signal Processing*, vol. 50, no. 6, pp. 1417–1428, June 2002.
- [6] J. N. Laska, S. Kirolos, M. F. Duarte, T. Ragheb, R. G. Baraniuk, and Y. Massoud, “Theory and implementation of an analog-to-information converter using random demodulation,” in *IEEE Int. Symp. on Circuits and Systems (ISCAS)*, New Orleans, Louisiana, 2007.
- [7] J. A. Tropp, J. N. Laska, M. F. Duarte, J. K. Romberg, and R. G. Baraniuk,

- “Beyond Nyquist: Efficient sampling of sparse, bandlimited signals,” in *IEEE Trans. on Information Theory*, submitted 2009.
- [8] J. Romberg, “Compressive sensing by random convolution,” Preprint, 2008.
- [9] J. A. Tropp, M. Wakin, M. F. Duarte, D. Baron, and R. G. Baraniuk, “Random filters for compressive sampling and reconstruction,” in *IEEE Trans. on Acoustics, Speech, and Signal Processing*, May 2006.
- [10] M. F. Duarte, M. Davenport, D. Takhar, J. N. Laska, T. Sun, K. Kelly, and R. G. Baraniuk, “Single-pixel imaging via compressive sampling,” in *IEEE Signal Processing Magazine*, vol. 25, no. 2, March 2008, pp. pp. 83–91.
- [11] R. Robucci, L. K. Chiu, J. Gray, J. Romberg, P. Hasler, and D. Anderson, “Compressive sensing on a cmos separable transform image sensor,” in *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 2008.
- [12] R. Marcia, Z. Harmany, and R. Willett, “Compressive coded aperture imaging,” in *SPIE electronic imaging*, 2009.
- [13] L. Jacques, D. K. Hammond, and M. J. Fadili, “Dequantizing compressed sensing: When oversampling and non-gaussian constraints combine,” Preprint, 2009. [Online]. Available: <http://arxiv.org/abs/0902.2367>
- [14] W. Dai, H. V. Pham, and O. Milenkovic, “Distortion-rate functions for quantized compressive sensing,” Preprint, January 2009.

- [15] E. J. Candès and T. Tao, “The dantzig selector: statistical estimation when p is much larger than n ,” in *Annals of Statistics*, vol. 35, no. 6, 2007, pp. 2313–2351.
- [16] G. A. Gray and G. W. Zeoli, “Quantization and saturation noise due to analog-to-digital conversion,” in *IEEE Trans. on Aerospace and Electronic Systems*, January 1971, pp. 222–223.
- [17] E. J. Candès and T. Tao, “Decoding by linear programming,” *IEEE Trans. Inform. Theory*, vol. 51, pp. 4203–4215, Dec. 2005.
- [18] R. G. Baraniuk, M. A. Davenport, R. DeVore, and M. B. Wakin, “A simple proof of the Restricted Isometry Property for random matrices,” in *Constructive Approximation*, vol. 28(3), December 2008, pp. 253–263.
- [19] E. J. Candès, “The restricted isometry property and its implications for compressed sensing,” in *Compte Rendus de l’Academie des Sciences, Paris, Series I*, vol. 346, 2008, pp. 589–592.
- [20] D. Needell and J. A. Tropp, “CoSaMP: Iterative signal recovery from incomplete and inaccurate samples,” *Accepted to Appl. Comp. Harmonic Anal.*, 2008.
- [21] T. Blumensath and M. E. Davies, “Iterative hard thresholding for compressive sensing,” Preprint, July 2008.
- [22] P. Wojtaszczyk, “Stability and instance optimality for Gaussian measurements in compressed sensing,” Preprint, 2008.

- [23] J. N. Laska, P. Boufounos, and R. G. Baraniuk, "Finite-range scalar quantization for compressive sensing," in *Conf. on Sampling Theory and Applications (SampTA)*, 2009.
- [24] M. Davenport, M. F. Duarte, M. Wakin, J. N. Laska, D. Takhar, K. Kelly, and R. G. Baraniuk, "The smashed filter for compressive classification and target recognition," in *Computational Imaging V at SPIE Electronic Imaging*, January 2007.
- [25] P. Boufounos and R. G. Baraniuk, "1-bit compressive sensing," in *Conf. on Info. Science and Systems (CISS)*, Princeton, New Jersey, 2008.
- [26] A. R. Calderbank and I. Daubechies, "The pros and cons of democracy," in *IEEE Trans. on Information Theory*, vol. 48, no. 6, 2002.
- [27] 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.
- [28] E. Candès, "Integration of sensing and processing," *IMA annual program year workshop*, Dec. 2005.
- [29] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming (web page and software)," in *Online.*, February 2009. [Online]. Available: <http://stanford.edu/~boyd/cvx>
- [30] —, *Graph implementations for nonsmooth convex programs*, *Recent*

- advances in learning and control (a tribute to M. Vidyasagar)*, V. Blondel, S. Boyd, and H. Kimura, Eds. Springer, 2008. [Online]. Available: http://stanford.edu/~boyd/graph_dcp.html
- [31] R. DeVore, B. Jawerth, and B. Lucier, “Image compression through wavelet transform coding,” in *IEEE Trans. on Information Theory*, vol. 38, no. 2, March 1992.
- [32] M. Mishali and Y. C. Eldar, “Blind multi-band signal reconstruction: compressed sensing for analog signals,” *IEEE Trans. on Signal Processing*, vol. 57, no. 3, pp. 993–1009, March 2009.
- [33] A. V. Oppenheim and A. S. Willsky, *Signals and systems*. Prentice-Hall, 1996.
- [34] D. J. C. MacKay, “Fountain codes,” in *IEE Proceedings Communications*, vol. 152, no. 6, 2005, pp. pp. 1062–1068.
- [35] V. K. Goyal, “Multiple description coding: compression meets the network,” in *IEEE Signal Processing Magazine*, 2001.
- [36] R. B. Ash and C. A. Doléans-Dade, *Probability and measure theory*. Academic Press, 1999.
- [37] A. B. Sripad and D. L. Snyder, “A necessary and sufficient condition for quantization errors to be uniform and white,” in *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. ASSP-25, 1977, pp. 442 – 448.

- [38] E. T. Hale, W. Yin, and Y. Zhang, "A fixed-point continuation method for ℓ_1 regularized minimization with applications to compressed sensing," Preprint, 2007.
- [39] E. T. Hale, "Fixed-point continuation for compressed sensing signal reconstruction," in *Rice university computational and applied mathematics colloquium*, 2007. [Online]. Available: http://www.caam.rice.edu/~optimization/L1/fpc/fpc_num_p_caam.pdf
- [40] B. C. Arnold, N. Balakrishnan, and H. N. Nagaraja, *A first course in order statistics*. Wiley-Interscience, 1992.
- [41] G. G. Lorentz, M. v. Golitschek, and Y. Makovoz, *Constructive approximation: advanced problems*. Springer-Verlag, 1996. [Online]. Available: <http://www.mi.uni-erlangen.de/at-net/BULL/lorentz.ps>