RICE UNIVERSITY

Regime Change:
Sampling Rate vs. Bit-Depth in Compressive Sensing

by

Jason Noah Laska

A Thesis Submitted
in Partial Fulfillment of the
Requirements for the Degree

Doctor of Philosophy

Approved, Thesis Committee:

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

Dr. Kevin F. Kelly, Associate Professor
Electrical and Computer Engineering

Dr. Wotao Yin, Assistant Professor
Computational and Applied Mathematics

Houston, Texas

December 2011

Abstract

The compressive sensing (CS) framework aims to ease the burden on analog-to-digital converters (ADCs) by exploiting inherent structure in natural and man-made signals. It has been demonstrated that structured signals can be acquired with just a small number of linear measurements, on the order of the signal complexity. In practice, this enables lower sampling rates that can be more easily achieved by current hardware designs. The primary bottleneck that limits ADC sampling rates is quantization, i.e., higher bit-depths impose lower sampling rates. Thus, the decreased sampling rates of CS ADCs accommodate the otherwise limiting quantizer of conventional ADCs.

In this thesis, we consider a different approach to CS ADC by shifting towards lower quantizer bit-depths rather than lower sampling rates. We explore the extreme case where each measurement is quantized to just one bit, representing its sign. We develop a new theoretical framework to analyze this extreme case and develop new algorithms for signal reconstruction from such coarsely quantized measurements. The 1-bit CS framework leads us to scenarios where it may be more appropriate to reduce bit-depth instead of sampling rate. We find that there exist two distinct regimes of operation that correspond to high/low signal-to-noise ratio (SNR). In the *measurement compression* (MC) regime, a high SNR favors acquiring fewer measurements with more bits per measurement (as in conventional CS); in the *quantization compression* (QC) regime, a low SNR favors acquiring more measurements with fewer bits per measurement (as in this thesis). A surprise from our analysis and experiments is that in many practical applications it is better to operate in the QC regime, even acquiring as few as 1 bit per measurement.

The above philosophy extends further to practical CS ADC system designs. We propose two new CS architectures, one of which takes advantage of the fact that the sampling and quantization operations are performed by two different hardware components. The former can be employed at high rates with minimal costs while the latter cannot. Thus, we develop a system that discretizes in time, performs CS preconditioning techniques, and then quantizes at a low rate.

Acknowledgements

Every era is measured by a randomly weighted sum of the ideas, philosophies, and knowledge of those around us. This thesis would not have been possible without the advice, support, and collaboration of many people.

I would like to begin by thanking my thesis committee, Richard Baraniuk, Kevin Kelly, Wotao Yin, all of whom I have collaborated with during the course of my program. They have given me significant knowledge and guidance.

Thanks to my collaborators, Richard Baraniuk, Petros Boufounos, Mark Davenport, Laurent Jacques, J. P. Slavinsky, John Treichler, Zaiwen Wen, and Wotao Yin. These brilliant minds have contributed their expertise on recent projects and papers that ultimately make up part of this work. I have noted their collaboration in each chapter.

I would like to thank my advisor Richard Baraniuk. Rich has been a wise and masterful mentor to me in a ridiculous number of ways. He is a bottomless cup of good ideas – his approach to studying problems is something I will try to emulate. He always asks the right questions and suggests new directions that develop ideas further. Clear, meaningful presentation of your work is essential in any field; Rich has provided an amazing example and instruction of how this should be done. Finally, I am not sure I would have made it this far without an advisor that allowed me the time to develop to where I needed to be.

Both Petros Boufounos and Mark Davenport have also been great mentors (and friends). Through several projects, I have learned numerous skills from these guys that have enabled me to drive future projects myself. Marco Duarte also guided me along in my early days at Rice. I would like to thank important collaborators on past projects, Anna, Martin, Justin, Yehia. Particular thanks to Joel for teaching me "computational science" at an early point in my program.

Thanks to Mike, Mark, Chris, the editors and co-founders of *Rejecta Mathematica*.

Oh man, the RichB group was so good during my time here. Thanks to my postdocs: Dror, Petros, Volkan, Jarvis, Aswin, Arian, Christoff; you have provided the fantastic diverse research culture in our group. Thanks to the graduate students (past and present): Shri, Mike, Ryan, Ray, Marco, Mark, Matthew, Mona, Chin, eEva, Manjari, Stephen, Drew, Sriram, Andrew, Amirali; you have been some of the most brilliant and colorful people that could have been collected in one group. And of course thanks to J. P. for being so wise and so talented. Thanks to Liz for making sure the group ran like a well-oiled machine. Thanks to Lee for switching us over to olive oil and home-made Kombucha. I would also like to thank the wonderful people at Lyric Semiconductor (Ben, Bill, Theo, Jeff, Andy, ...) as well as Dan Dudgeon for amazing summers during this segment.

School was made significantly more enjoyable by days in the office and at Valhalla with Ilan, Mark, Manjari, eEva, Gareth, Plediii, Kia, Scott, Brian.

Thanks to my closest friends during this era, Andrea and Ilan, who listened patiently to all my stories, obsessions with coffee, and dealt with various neuroses. Thanksgiving to all my UIUC friends.

Finally, I would like to thank my family who also support me in any endeavour that I choose.

Contents

1 Introduction **1**

1.1 Roadmap and Main Contributions . . . 5

1.2 Compressive Sensing (CS) Toolkit . . . 7

1.2.1 Signal and sensing models . . . 7

1.2.2 The restricted isometry property (RIP) . . . 7

1.2.3 Signal reconstruction via convex optimization . . . 8

1.2.4 Signal reconstruction via greedy and first order algorithms . . . 9

1.2.5 Oracle-assisted signal reconstruction . . . 11

1.2.6 Noise folding . . . 11

2 Quantization and Dynamic Range in Compressive Sensing **13**

2.1 Finite-Range Scalar Quantization . . . 13

2.2 Dynamic Range of CS-Based Acquisition Systems . . . 15

2.2.1 A deterministic approach to dynamic range . . . 15

2.2.2 Dynamic range of a conventional ADC . . . 17

2.2.3 Dynamic range of a CS-based acquisition system . . . 17

2.2.4 Impact of CS on the PAR . . . 19

2.3 Liberating Dynamic Range via Democracy . . . 21

2.3.1 The democratic caucus of random matrices . . . 22

2.3.2 Saturation rejection signal recovery . . . 23

2.3.3 Saturation rejection signal processing . . . 24

2.3.4 Saturation consistency signal recovery via convex optimization . . . 24

2.4 Experimental Performance of CS Dynamic Range 25

3 Single Bit Compressive Sensing 28

3.1 Supersaturated Sensing 28

3.2 The 1-bit CS framework 30

3.3 Immediate Benefits of 1-bit CS 31

3.4 Noiseless Reconstruction Performance 32

3.4.1 Reconstruction performance lower bounds 32

3.4.2 Achievable performance via random projections 34

3.5 Robust 1-bit CS via Binary Stable Embeddings 36

3.5.1 Binary ϵ-stable embeddings 36

3.5.2 Binary ϵ-stable embeddings via random projections 38

3.5.3 Noisy measurements and compressible signals 40

3.6 Saturation-Agnostic Sensing 42

3.7 1-bit CS Reconstruction Algorithms 43

3.7.1 Trust, but Verify: Restricted-step shrinkage (RSS) 43

3.7.2 Binary iterative hard thresholding (BIHT) 49

3.7.3 Convex 1-bit reconstruction formulations 52

3.8 Empirical Verification 55

3.9 A Note on Alternative 1-Bit Frameworks 65

4 Regime Change 66

4.1 Analysis of Quantized CS Systems with Signal Noise 67

4.2 Experiments 69

4.2.1 Setup 70

4.2.2 Oracle-assisted reconstruction 70

4.2.3 Reconstruction performance as a function of $\mathfrak{B}$ 70

4.2.4 Reconstruction performance as a function of ISNR 73

4.3 Relationship to Oversampled ADCs 73

5 Compressive Sensing Architectures 75

5.1 The Random Demodulator (RD) 77

5.1.1 Drawbacks of the RD 78

5.2 The Compressive Multiplexer (CMUX) 79

5.2.1 Custom CMUX reconstruction algorithms 80

5.2.2 The passive averager: A CMUX hardware concept 82

5.2.3 Comparison with random demodulator 83

5.3 The Polyphase Random Demodulator (PRD) 83

5.4 Simulations 85

5.4.1 Exactly sparse recovery (CMUX) 85

5.4.2 Practical RF example (CMUX) 85

5.4.3 Effects of a non-ideal modulator waveform (PRD) 86

6 Discussion 88

6.1 Summary 88

6.2 Open Questions and Future Directions 89

6.2.1 Does the RIP of a matrix imply the BϵSE for that matrix? 89

6.2.2 Does the Fourier basis provide a BϵSE? 90

A Dynamic Range 92

A.1 Lemma 1 92

A.2 Theorem 2 92

B Binary Stable Embeddings 95

B.1 Lemma 5: Intersections of Orthants by Subspaces 95

B.2 Theorem 4: Distributing Signals to Quantization Points 97

B.3 Theorem 5: Optimal Performance via Gaussian Projections 98

B.4 Lemma 8: Concentration of Measure for δ-Balls 99

B.5 Theorem 6: Gaussian Matrices Provide BϵSEs 103

B.6 Lemma 9: Stability with Measurement Noise 104

B.7 Corollary 3: Stability with Compressible Signals 105

C The RSS Algorithm 106

C.1 Quadratic Penalty Framework 106

C.2 Convergence Proof of Algorithm 4 106

C.2.1 Proof of Theorem 7 108

D Regime Change: Proof of Theorem 8 110

List of Figures

2.1 Midrise scalar quantization. 14
2.2 Dynamic range of CS systems. 26

3.1 The geometry of the 1-bit CS framework. 31
3.2 Intersection of orthants by a subspace. 33
3.3 Example trust-region iterations. 44
3.4 The geometry of convex hyperplane-based 1-bit CS. 53
3.5 The geometry of convex hinge-loss-based 1-bit CS. 54
3.6 Average 1-bit CS reconstruction angular error. 56
3.7 1-bit CS reconstruction misses and false alarms. 57
3.8 1-bit CS reconstruction angular error vs. Hamming error. 58
3.9 One-sided ℓ_1 vs. one-sided ℓ_2 BIHT consistency performance. 59
3.10 One-sided ℓ_1 vs. one-sided ℓ_2 BIHT noise performance. 60
3.11 1-bit CS vs. conventional CS. 61
3.12 1-bit CS vs. Nyquist samples. 62
3.13 Noiseless comparison of 1-bit CS algorithms (noiseless). 63
3.14 Comparison of 1-bit CS algorithms in noise. 64
3.15 Saturation-agnostic CS reconstruction. 65

4.1 Oracle-assisted reconstruction bound for noisy quantized measurements. 69
4.2 Oracle-assisted reconstruction simulation for noise quantized measurements. 71
4.3 Reconstruction performance as a function of total bits, for different ISNRs. 72
4.4 Maximum RSNR given a fixed bit-budget $\mathfrak{B}$. 73

5.1 The random demodulator (RD). 76
5.2 Waveforms at preconditioning stages in the RD. 76
5.3 The compressive multiplexer (CMUX). 79
5.4 The passive averager CMUX (PA-CMUX). 82
5.5 The polyphase random demodulator (PRD). 84
5.6 Ideal CMUX performance (number of channels). 85
5.7 Practical CMUX reconstruction performance. 85
5.8 Simulated non-ideal square waveforms. 86
5.9 Reconstruction performance of the RD vs. the PRD 86

6.1 1-bit quantized Nyquist samples (with sparse reconstruction). 90

B.1 The geometry of orthants and spherical caps. 96

List of Algorithms

1 Prototype CS greedy algorithm . . . 10
2 Iterative Hard Thresholding (IHT) . . . 11
3 RSS-outer . . . 46
4 RSS-inner (subroutine) . . . 48
5 Binary Iterative Hard Thresholding (BIHT) . . . 50
6 Trivial Reconstruction . . . 80
7 Block Coordinate Relaxation (BCR) . . . 81

CHAPTER 1

Introduction

The great shift to digital processing over the last few decades has created an insatiable demand for the digitization of ever wider bandwidth signals [1]. In turn, this has led to an increased burden on signal acquisition devices that rely on the Shannon sampling theorem which requires that such devices must sample at least at the Nyquist-rate, twice the bandwidth of the signal, for any bandlimited signal [2, 3].[1] This requirement forces analog-to-digital converters (ADCs) to sample faster to capture wideband signals for later digital processing. It is no longer feasible to build devices that meet our demands for size, weight, power, and bandwidth while still adhering to classical notions of signal acquisition [6, 7].

Thankfully, we have come a long way in our understanding of signals since Shannon's original theory. The class of bandlimited signals is extremely broad, consisting of all signals with some maximum frequency. For example, the sampling theorem enables us to accurately capture an instance of bandlimited noise, even though there may be little utility for such a signal. In fact, most natural and man-made signals have some inherent additional structure beyond bandlimitedness. In particular, in this thesis we are interested in signals that when transformed into some domain (via a linear transformation), have energy that is primarily concentrated among just a few large coefficients, and all other coefficients can be approximated as zero. This particular description of signal structure is extremely practical since transforms exist that firmly put many natural signals such as images [8, 9] and man-made signals [10, 11] in this class. Indeed, the exploitation of this form of structure is the basis for transform coding and compression, e.g., JPEG image compression [12, 13].

The confluence of rigorously defined structured signal models and the desire to circumvent the Shannon-Nyquist limitation has prompted a new signal acquisition framework, *compressive sensing*

[1] While Shannon's theory of communication (ca. 1949) is perhaps the most popularly cited origin of this idea, similar sampling theorems were proven by Whittaker [4] in 1915 and Kotelnikov [5] in 1933.

(CS) [14, 15]. A key insight is rather than attempting to acquire *all* bandlimited signals, CS assumes that we are only interested in signals with the structure described above. By reducing the size of class of signals of interest, we should be able to drive down the number of samples required to ultimately distinguish between the signals. The CS framework harnesses this insight via three fundamental components:

1. *underdetermined linear measurement systems*, i.e. we obtain the measurements
$$y = \Phi x + e, \tag{1.1}$$
of the signal $x \in \mathbb{R}^N$ where Φ is an $M \times N$ matrix with $M \ll N$ that models the linear physical sampling system, and with measurement error $e \in \mathbb{R}^M$;

2. *signal models*, the most simple model being that comprising of all K-sparse signals, i.e., signals for which only K elements are non-zero; and

3. *algorithmic reconstruction*, such as convex optimization or greedy algorithms. Briefly, to reconstruct a signal estimate $\widehat{x}$ from y we generally ask for the sparsest solution such that its measurements, $\Phi\widehat{x}$, are the same as, or within some close distance of the observed measurements y. Such algorithms are non-linear and iterative.

A large body of work has been devoted to the study of each of these components, e.g., by *i*) characterizing conditions on Φ that provide robust mappings of sparse signals to lower dimensions and designing physical sampling systems that satisfy such conditions [16–21]; *ii*) proposing more refined classes of highly structured signals [22–24]; and *iii*) providing reconstruction guarantees and fast solvers for convex programs [25–28] as well as greedy and first order algorithms [29–31]. We will review some of these topics in more detail in Section 1.2.

CS promises to lessen our sampling burden by decreasing sampling rates. The simple consequence of (1.1) is that when the acquisition of each measurement is "expensive," then we benefit by only sensing M values rather than N. For instance, it is possible to design a physical sampling system $\bar{\Phi}$ such that $y = \Phi x = \bar{\Phi}(x(t))$ where x is a vector of Nyquist-rate samples of a bandlimited signal $x(t)$, $t \in \mathbb{R}$. In this case, (1.1) translates to low, sub-Nyquist sampling rates. This is a potential boon for wideband acquisition as mentioned earlier, since it enables current ADC technology to acquire larger bandwidths than was possible before, or alternatively enables a higher precision ADC to be used in current wideband systems.

The significant attention given to reducing the number of acquired measurements only explicitly acknowledges half of the acquisition process. In practice analog-to-digital conversion really comprises of two steps: *i*) discretization in time (sampling), and *ii*) discretization in amplitude (quantization). Thus, just as in any conventional sampling system, CS measurements are quantized, i.e., each measurement is mapped from a real value (over a potentially infinite range) to a discrete value over some finite range. For example, in scalar quantization, a measurement is mapped to one of 2^B distinct values, where B denotes the number of bits per measurement, i.e., the *bit-depth*. The finite range of the quantizer results from a finite number of bits as well as physical limitations of hardware components.

There are several interesting problems and attributes associated with quantization (and physical quantizers) that are not considered by the prototype CS framework described above:

- *The quantizer begets the dynamic range of the system*: Quantization introduces two kinds of error: quantization error and saturation error. The former is the result of measurements that are within the range of the quantizer; this error is bounded. The latter is the result of measurements with amplitudes beyond the range of the quantizer, i.e., *saturated* measurements; this error is unbounded and in many case more detrimental to performance than quantization error.

  The dynamic range of a system is typically defined as the ratio maximum amplitude tone to the minimum amplitude tone that can sampled with some given accuracy. Thus, the finite range of the quantizer places strict limits on the dynamic range of the ADC system.

  Dynamic range tells us how much quieter the "quietest" signal can be than the "loudest." This is a fundamental metric of system performance for many practical applications [32].

- *The quantizer is the ADC bottleneck*: The ADC is beholden to the quantizer [6, 7]. Quantization significantly limits the maximum speed of the analog-to-digital converter (ADC), forcing an exponential decrease in sampling rate as the number of bits is increased linearly [7]. Furthermore, the quantizer is the primary power consumer in an ADC. Thus, more bits per measurement directly translates to slower sampling rates and increased ADC costs.

- *The quantizer is sensitive to noise*: High bit-depth quantization is more susceptible to nonlinear distortion in the ADC electronics [33].

By reducing the sampling rate, the CS framework implicitly assumes we can relieve some of the burdens associated with the quantizer.

In this thesis, we take a unified approach to CS ADCs, considering the sampling rate, finite range quantization, and signal noise when studying CS systems. A driving theme behind this work is that the tradeoff between sampling and quantization can be manipulated in *both* directions; simply put, reducing the bit-depth of the quantizer also reduces our sampling burden and furthers the goals of CS acquisition devices. We develop the following main ideas (each roughly corresponding to a chapter), that ultimately lead to this insight.

CS enables higher dynamic range systems. By reducing the sampling rate, CS enables the use of a higher bit-depth or higher dynamic-range quantizer [7]. If we are to claim any benefit by this fact, then it is of fundamental importance that the CS measurement system Φ can take advantage of any additional dynamic range granted by a better quantizer. We rigorously study the dynamic range of CS systems and determine that indeed it is on the same order as conventional systems for a uniform quantizer at a given bit-depth. This then verifies that by reducing the sampling rate, we may indeed obtain an improvement in the dynamic range of the system.

It is possible to extend the dynamic range of CS systems beyond the claims above. In CS systems, saturated measurements can either be rejected before reconstruction or included in a reconstruction algorithm. Robust signal reconstruction is possible in both cases [34, 35]. Intuitively then, if we increase the input signal gain (equivalent to increasing the scale of the measurements) such that the quantizer saturates significantly, yet we still achieve similar reconstruction performance, then the dynamic range of the system has effectively been increased. The question arises: how many measurements are allowed to saturate?

Saturate all measurements ↔ Quantize measurements to just 1 bit. Although unintuitive, it is possible to saturate all of the CS measurements. In this case, we effectively retain just 1 bit of information per measurement, representing its sign. Thus, an alternative interpretation is that we can drive the depth of quantizer down such that it is a simple comparator, testing for measurements above or below zero. Previous notions of dynamic range no longer apply since any positive scaling of the signal will result in the same set of measurement signs, i.e., the scale of the signal will be obliterated. To reconstruct we search for the sparsest signal that yields the same measurement signs when projected through the measurement system. We call this *consistent* reconstruction. Since the scale of the signal is unknown and arbitrary, we only search for signals with unit energy. We demonstrate that there are is a large class of 1-bit CS mappings that enable stable reconstruction in this way and we further demonstrate that practical algorithms can be designed to solve this reconstruction problem. Finally, we extend the methods for 1-bit CS to measurements that have been quantized at arbitrary bit-depths, or with arbitrary numbers of saturations, unknown a priori. We dub this reconstruction technique saturation-agnostic CS.

Reduce the bit-depth, increase the sampling rate. 1-bit CS provides a fresh perspective on CS ADCs. Driving down the bit-depth to the extreme case of a single bit per measurement enables extremely fast hardware quantizers; now the quantizer is a simple comparator. Thus, in stark contrast to the typical CS assertion that we should reduce the sampling rate and increase the bit-depth of the quantizer we have demonstrated that indeed the reverse possible. We take this yet a step further to answer the question: *when* should we do this?

Signal noise subject to *noise folding* in CS, i.e., it is amplified by underdetermined linear systems [36–38]. This means that as the sampling rate decreases, we incur an increasing penalty due to input signal noise. Employing more bits at the quantizer when there is more noise means the extra precision is not being used efficiently. Sampling at a higher rate with an extremely low bit-depth addresses this problem. Noise folding either becomes less prevalent, or in the oversampled case, is not present at all. Meanwhile, the burdens of higher-rate sampling are still relieved by the low bit-depth of the quantizer.

CS ADCs: Disconnect the sampler from the quantizer. Sampling and quantization are carried out by two distinct hardware components in physical ADCs. Specifically, the *sample and hold* (S/H) component discretizes in time while the hardware quantizer discretizes in amplitude. As previously noted, the quantizer is the main ADC bottleneck. Indeed, S/H components can operate accurately at extremely high speeds and low power, as opposed to the quantizer. We propose two new CS ADC architectures that take advantage of this insight. While previous CS ADC designs use an off-the-shelf ADC at a low rate, we separate the two ADC discretization steps. The sampling components operate at a high rate while the quantizer operates at a low rate. We demonstrate that this yields new CS ADCs that avoid many problems associated with earlier designs.

In this thesis we carefully explore the ideas described above. Along the way we develop new theoretical frameworks for analyzing dynamic range and 1-bit CS, we develop new algorithms for sparse signal reconstruction, and we perform extensive simulations to demonstrate the validity of our claims. We now briefly describe in ever so slightly more detail what topics and results can be found in each chapter.

1.1 Roadmap and Main Contributions

For the remainder of this chapter in Section 1.2, we review and define the key components and results of the CS framework that will be made use of later in this thesis. We cover sparse signal models and undetermined linear sensing models. We then move on to introduce a useful property of the matrix Φ, the *restricted isometry property* (RIP), for which robust signal reconstruction from many algorithms is guaranteed. We review the convex optimization formulations that can be used to reconstruct sparse signals as well as greedy and first order algorithms that are often used in practice and will be adapted for our purposes. Finally we discuss the noise folding effect in these systems.

In **Chapter 2** we analyze the dynamic range of CS systems with finite-range uniform scalar quantizers. Our new contributions are as follows. We begin by defining a rigorous and deterministic notion of dynamic range. This enables us to avoid more heuristic dynamic range analyses that make assumptions about the signal (or measurement) distribution. We then go on to derive the dynamic range of a conventional system and demonstrate that it reasonably similar to the more conventional dynamic range definitions. The dynamic range of conventional systems provides a basis for comparison against CS systems. Thus, given our definition, we next derive the dynamic range for a large class of CS systems (those depending on the RIP of Φ discussed in Section 1.2). Combining these results we can then claim that the dynamic range of CS systems is on the same order as that of conventional sampling systems. We follow up with a short discussion on the peak-to-average ratio (PAR) and how for some CS systems this is improved on average.

We conclude the chapter with a review of the democratic property of random matrices and explain how this can be exploited to further increase the dynamic range of CS systems [34]. We derive an analytical expression for the improvement in terms of the previous analysis. We experimentally verify the claims in the chapter and demonstrate the improvement gained by both increasing the quantizer bit-depth as a function of decreasing measurement rate as well as employing the democratic saturation-robustness techniques. A surprising empirical result leads us towards the philosophy espoused in the next two chapters: we should consider decreasing bit-depth and increasing measurement-rate.

In **Chapter 3** we study 1-bit quantization for CS measurements. We begin by explaining how saturating all measurements corresponds precisely to 1-bit quantization. We then formally define the 1-bit framework as in [39] and explain several benefits of this framework. Our new contributions are as follows. First we provide optimal reconstruction bounds from 1-bit measurements from any mapping. We then demonstrate that Gaussian Φ (among a few others) will enable us to satisfy the previous bounds in the noiseless case. We next introduce a new property for 1-bit CS systems that we dub the *Binary ϵ-Stable Embedding* (BϵSE) and demonstrate how if a system satisfies this property, then robust reconstruction is guaranteed. We further demonstrate that again Gaussian sensing systems satisfy this property with high probability and we derive the number of measurements required for this to hold. We then apply our results to formulate guarantees from noisy measurements and signals that are not strictly sparse. We next derive a new reconstruction formulation that extends the framework to be used not only in the fully saturated case, but can be applied to problems with arbitrary saturations, i.e., saturation-agnostic sensing.

We continue on to more practical aspects of the 1-bit framework by introducing two new algorithms for signal reconstruction: *Restricted-Step Shrinkage* (RSS) and *Binary Iterative Hard Thresholding* (BIHT). For the former algorithm we give convergence guarantees and for the latter algorithm we discuss what problem it is attempting to solve and why the reconstruction error performance may differ between the two. We additionally motivate the formulation of several convex reconstruction algorithms. We conclude the chapter and our contributions with an extensive suite of simulations, comparing the 1-bit algorithms against previously proposed algorithms and studying the performance in comparison with higher bit-depth uniformly quantized CS systems. We also verify the validity of the saturation-agnostic approach.

In **Chapter 4** we explore the tradeoff between bit-depth and measurement rate. We find that by considering signal noise, we expose a regime where 1-bit CS outperforms more conventional CS systems. Our new contributions are as follows. We study the scenario where there is a fixed bit-budget, a scalar quantizer, and input signal noise. We begin by developing a theoretical bound on the reconstruction error from quantized CS measurements. We then show numerically that the minimum of of this bound is attained for lower bit-depths as the input measurement noise is increased. In fact, these simulations demonstrate that 1-bit CS outperforms conventional CS when the input SNR is low enough. Thus, we can categorize CS into two compression regimes, corresponding to the input SNR: *measurement compression* (MC) when input SNR is high, and *quantization compression* (QC) when input SNR is low. The former finds application when measurements are expensive to sense and high bit-depths are inexpensive, while the latter finds application when measurements are inexpensive to sense and high bit-depths are expensive.

In **Chapter 5** we introduce two new CS architectures. First, we introduce the *Compressive Multiplexer* (CMUX) that can be used to acquire signals from a multi-channel signal model. We discuss the benefits of this design over previous designs. We also discuss several algorithms that can be used with this system due to its unique structure. Second, we introduce the *Polyphase Random Demodulator* (PRD), a new take on a more "classical" system the *Random Demodulator* (RD) [18]. The key insight driving this design is that the S/H hardware can be separated from the quantizer, providing significant gains over the RD in calibration and computer modeling for reconstruction. We also discuss the relationship between the CMUX and the PRD. We conclude this chapter with simulations demonstrating the validity of the CMUX and PRD designs.

Without further delay, we now throw down a plain introduction to vanilla CS.

1.2 Compressive Sensing (CS) Toolkit

1.2.1 Signal and sensing models

In the CS framework [14, 15], we acquire a signal $x \in \mathbb{R}^N$ via the linear measurements

$$y = \Phi x + e, \tag{1.2}$$

where the underdetermined matrix $\Phi \in \mathbb{R}^{M \times N}$ models the physical sampling system, $y \in \mathbb{R}^M$ is the vector of measurements acquired, and $e \in \mathbb{R}^M$ is a measurement noise vector.

In the most basic CS setup, we are interested in K-sparse signals, i.e., $x \in \Sigma_K$ where $\Sigma_K := \{x \in \mathbb{R}^N : \|x\|_0 := |\text{supp}(x)| \leq K\}$.[2] However in practice, signals may not be strictly sparse but rather may contain many small coefficients that do not contribute considerable energy to the signal, or when sorted by magnitude, the signal coefficients decay with some power law, i.e., have elements such as $x_n \propto |n|^{-1/p}$ for $p > 1$. Such signals that can be well-approximated by just K largest in magnitude coefficients are called *compressible* signals. We will denote the best K-term approximation of x as x_K. Finally, in many cases x will not be canonically sparse; instead it can be sparse is some orthonormal transform basis Ψ. In this case we write $x = \Psi\alpha$ where $\alpha \in \Sigma_K$. Since we still sense x, the measurements can be written as $y = \Phi\Psi\alpha + e$, and thus the matrix $\Phi\Psi$ is used in reconstruction when solving for a sparse estimate $\widehat{\alpha}$. Unless otherwise noted, for the remainder of this thesis without the loss of generality, we fix $\Psi = I$, the identity matrix, implying that $x = \alpha$.

Recently there has been significant interest in exploiting stronger signal models. In some cases, there is additional structure known *a priori* about the non-zero coefficients. For instance, model-based signal reconstruction algorithms have been proposed for the case when some explicit relationship between the support of the non-zero coefficients is known [24]. This has been used for recovery of spectrally sparse signals [22] and neural spike trains [23]. Another popular signal model has been that of *group-sparsity* where the non-zero coefficients are clustered together [40, 41]. These stronger models further empower the CS framework to produce more accurate estimates with the same set of measurements but are primarily beyond the scope of this thesis. We only mention these models since it is possible that many of the methods described in this thesis can be extended to make use of these models and could be considered in practical instantiations.

1.2.2 The restricted isometry property (RIP)

Not all underdetermined sensing systems Φ are admissible. For instance, it is clear that if any signal $x \in \Sigma_K$ lies in the nullspace of Φ, then it can never be recovered with bounded error. While Φ can sometimes be analyzed in conjunction with a reconstruction algorithm to provide theoretical guarantees [16, 18, 42, 43], we can study a more generic property of Φ, the so-called *restricted isometry property* (RIP); the sufficient condition that the norm of the measurements is close to the norm of the signal for all sparse x, i.e.,

$$(1 - \delta)\|x\|_2^2 \leq \|\Phi x\|_2^2 \leq (1 + \delta)\|x\|_2^2, \tag{1.3}$$

[2] $\|\cdot\|_0$ denotes the ℓ_0 quasi-norm, which simply counts the number of nonzero entries of a vector.

for all $x \in \Sigma_K$ [44]. As a minimal sanity check, notice that under this definition no sparse signal will be in the nullspace of Φ. In words, the RIP requires Φ to act as an approximate isometry on the set of K-sparse vectors. Remarkably, it has been shown that if we set $M \geq C_\delta K \log(N/K)$ (where C_δ is some constant) and draw the elements of Φ from a sub-Gaussian distribution, then these matrices will indeed satisfy the RIP with high probability [45, 46]. Indeed, practical measurement systems with significantly more structure may also be admissible. For instance, hardware inspired designs have also been shown to hold this property [18, 47–50]. We will discuss some of these systems as well as some new architectures in Chapter 5.

The RIP can be expressed in general terms, as a *δ-stable embedding*. Let $\delta \in (0,1)$ and $X, S \subset \mathbb{R}^N$. We say the mapping Φ is a *δ-stable embedding* of X, S if

$$(1-\delta)\|x-s\|_2^2 \leq \|\Phi x - \Phi s\|_2^2 \leq (1+\delta)\|x-s\|_2^2, \tag{1.4}$$

for all $x \in X$ and $s \in S$. The RIP requires that (1.4) holds for $x \pm s \in \Sigma_K$; it is a stable embedding of sparse vectors. Expressing the RIP in this way enables further interpretation of this property. Specifically, it is clear from (1.4) that the RIP ensures that the distance between any two length-N K-sparse vectors is preserved when they are mapped down to the lower dimensional space. This interpretation will be important when we study 1-bit quantized CS measurements in Chapter 3.

1.2.3 Signal reconstruction via convex optimization

To reconstruct an estimate $\widehat{x}$ from y when there is no noise, i.e., $\|e\|_2 = 0$, we could naïvely solve for the sparsest signal that satisfies (1.2),

$$\widehat{x} \leftarrow \underset{x \in \mathbb{R}^N}{\operatorname{argmin}} \|x\|_0 \quad \text{s.t.} \quad y = \Phi x; \tag{ℓ_0-min}$$

however, this non-convex program exhibits combinatorial complexity in the size of the problem [51]. Instead, we solve *Basis Pursuit* (BP) by relaxing the objective in (ℓ_0-min) to the ℓ_1-norm

$$\widehat{x} \leftarrow \underset{x \in \mathbb{R}^N}{\operatorname{argmin}} \|x\|_1 \quad \text{s.t.} \quad y = \Phi x; \tag{BP}$$

the result is a convex, polynomial-time algorithm [52]. A key realization is that, under certain conditions on Φ (e.g., the RIP), the BP solution will be equivalent to that of (ℓ_0-min) [14]. This is remarkable since we seemingly have solved a combinatorial problem in polynomial time. Indeed, this is a key result that generated significant interest during the nascent years of CS.

The RIP suffices to ensure that a variety of other convex optimization algorithms can successfully recover any sparse or compressible signal from noisy measurements. In particular, for bounded errors of the form $\|e\|_2 \leq \epsilon$, the convex program *Basis Pursuit Denoising* (BPDN)

$$\widehat{x} \leftarrow \underset{x \in \mathbb{R}^N}{\operatorname{argmin}} \|x\|_1 \ \text{ s.t. } \|\Phi x - y\|_2 \leq \epsilon \tag{BPDN}$$

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

Theorem 1 (Theorem 1.2 of [44])**.** *Suppose that Φ satisfies the RIP of order $2K$ with $\delta < \sqrt{2} - 1$. Given measurements of the form $y = \Phi x + e$, where $\|e\|_2 \leq \epsilon$, the solution to (BPDN) obeys*

$$\|\widehat{x} - x\|_2 \leq C_0 \epsilon + C_1 \frac{\|x - x_K\|_1}{\sqrt{K}}, \tag{1.5}$$

where

$$C_0 = \frac{4\sqrt{1+\delta}}{1 - \left(\sqrt{2}+1\right)\delta}, \quad C_1 = 2\frac{1 + \left(\sqrt{2}-1\right)\delta}{1 - \left(\sqrt{2}+1\right)\delta}. \tag{1.6}$$

Many other convex formulations for reconstruction from noisy measurements have been proposed with different robustness guarantees depending on the noise model of e. [53, 54]. Furthermore, many fast algorithms have been developed to solve these problems [26–28, 55].

We conclude this subsection on convex reconstruction algorithms by mentioning another extremely popular reconstruction formulation, known as the LASSO [56],

$$\widehat{x} \leftarrow \operatorname*{argmin}_{x \in \mathbb{R}^N} \frac{1}{2}\|y - \Phi x\|_2^2 + \lambda \|x\|_1. \tag{LASSO}$$

For any ϵ in (BPDN), there is an appropriate choice of λ such that the solutions to (BPDN) and (LASSO) are equivalent [57, 58]. A wide range of algorithms have been design to solve this problem rather than (BPDN) [25, 59].

Although (LASSO) can be thought of as a relaxation of (BPDN) where the constraints have been moved into the objective function, the LASSO actually has its roots in statistical regression and is often interpreted as solving the least squares problem with a sparse penalty, or ℓ_1-regularizer. It has long been known in the statistics community that the ℓ_1 penalty biases in favor of sparse solutions, but a complete analytical framework for signal reconstruction with deterministic guarantees such as those given in Theorem 1.5 are a new result of the CS framework.

1.2.4 Signal reconstruction via greedy and first order algorithms

While convex optimization programs such as (BPDN) are powerful methods for CS signal recovery, there also exist a variety of alternative algorithms that are commonly used in practice and for which performance guarantees comparable to that of Theorem 1 can be established. In particular, greedy algorithms such as CoSaMP [29] and first order optimization algorithms such as iterative hard thresholding (IHT) [30, 60] are known to satisfy similar guarantees under slightly stronger assumptions on the RIP constants. We briefly describe a prototypical greedy CS algorithm and IHT since algorithms in later chapters will inspired by these methods.

Greedy Algorithms. We call an algorithm *greedy* if it makes decisions that are locally optimal in each iteration. While greedy algorithms are popular because they are fast and effective in practice, under certain circumstances these algorithms can produce solutions that are also globally optimal [61]. Algorithm 1 summarizes the general steps employed by a CS greedy algorithm, such as CoSaMP. The basic steps are as follows. Form a signal proxy, usually by computing $h^{s+1} = \Phi^T y^s$ at

Algorithm 1: Prototype CS greedy algorithm

```
s0 Initialize
     Set initial solution $x^0 := 0$
     Set iteration $s := 0$
   while not converged do
s1     Form signal proxy
         $h^{s+1} := \Phi^T y^s$
s2     Update coefficient support set $\Omega$
         e.g., add location of largest element in $h^{s+1}$ to $\Omega$ (in OMP [16, 61])
s3     Update coefficient estimate
         e.g., via pseudo-inverse $\widehat{x}^{s+1}|_\Omega := \Phi_\Omega^\dagger y^s$
s3     Subtract current estimate from measurements
         $y^{s+1} := y^s - \Phi x^{s+1}$
s4     Update iteration count
         Set $s := s + 1$
```

iteration s. This vector looks like a noisy version of the signal and enables fairly accurate detection of signal support, of course depending on $\|e\|_2$ and K. We next refine our support set estimate Ω. In orthogonal matching pursuit (OMP) [16, 61] this is done by simply selecting the support of the single largest in magnitude element of h^{s+1} and adding it to the support set, while in CoSaMP, $2K$ elements are selected simultaneously and added to support set. In the case of CoSaMP this set is later pruned. After updating the support estimate, we refine the coefficient amplitude estimates. Typically to do this, the optimal linear estimator, least squares is performed

$$\widehat{x}|_\Omega \leftarrow \min_{x|_\Omega} \frac{1}{2}\|y - \Phi_\Omega x|_\Omega\|_2^2, \tag{1.7}$$

where Φ_Ω denotes the submatrix of Φ formed by selecting the columns of Φ according to the index set Ω and similarly $x|_\Omega$ represents the corresponding subvector of x. Thus, the estimator is only applied to the non-zero coefficients and the resulting linear system is overdetermined. This can be computed via

$$\widehat{x}^s|_\Omega \leftarrow \Phi_\Omega^\dagger y^s, \quad \widehat{x}^s|_{\Omega^C} = 0, \tag{1.8}$$

where Φ_Ω denotes the submatrix of Φ formed by selecting the columns of Φ according to the index set Ω, $\widehat{x}^s|_\Omega$ is the corresponding subvector of $\widehat{x}^s$, Ω^C is the complement set to Ω, and $\dagger$ denotes the Moore-Penrose pseudo-inverse. This can also be computed using an algorithmic technique such as the conjugate-gradient method [62]. Finally, once the coefficients supports and values are estimated, we subtract their contribution from the measurements, $y^{s+1} = y^s - \Phi_\Omega \widehat{x}^s|_\Omega$.

Iterative Hard Thresholding (IHT). Algorithm 2 summarizes the IHT algorithm. In the first step we add the proxy h (from the greedy algorithm) to the current signal estimate resulting in $a^{s+1} = x^s + \Phi^T(y - \Phi x^s)$, at iteration s. We then simply threshold this estimate by setting all elements of a^{s+1} to zero except for the largest K elements via the function $\eta_K(\cdot)$. The first step is effectively a gradient descent for the function $\frac{1}{2}\|y - \Phi x\|_2^2$. Thus, IHT for CS can be thought of as

Algorithm 2: Iterative Hard Thresholding (IHT) [30, 60]

```
S0 Initialize
       Set initial solution $x^0 := 0$
       Set iteration $s := 0$
   while not converged do
S1     Update estimate
           $a^{s+1} := x^s + \Phi^T\big(y - \Phi x^s\big)$
S2     Hard threshold − select largest $K$
           $x^{s+1} := \eta_K(a^{s+1})$
S3     Update iteration count
           Set $s := s + 1$
```

trying to solve the problem

$$\widehat{x} \leftarrow \underset{x \in \mathbb{R}^N}{\operatorname{argmin}} \ \tfrac{1}{2}\|y - \Phi x\|_2^2 \quad \text{s.t.} \quad \|x\|_0 = K. \tag{1.9}$$

Other first order algorithms such as *approximate message passing* (AMP) proceed in a similar manner, sometimes adding additional terms to the first step and adapting how many coefficients are selected in each iteration [31].

1.2.5 Oracle-assisted signal reconstruction

As we saw in the greedy algorithm, CS reconstruction can be thought of as consisting of two steps: first finding the non-zero coefficient locations (the support) and then estimating the coefficient values. If we can correctly identify the true signal support, then the optimal linear estimate for coefficient values can be computed via least squares. Indeed, if an oracle were to provide the true support Ω, then no linear CS reconstruction algorithm can perform better than (1.8). Thus, reconstruction with known signal support is sometimes called *oracle-assisted* reconstruction [38, 53]. Some of our analysis will be primarily in terms of the performance of this best-case reconstruction algorithm.

1.2.6 Noise folding

Signal noise is amplified by underdetermined linear measurement systems [36]. Specifically, it has been shown that if n is white Gaussian noise with variance σ_n^2, then for CS measurement systems, Φn is also white, Gaussian, and each noise measurement has increased variance $\sigma_{\Phi n}^2 \approx \frac{N}{M}\sigma_n^2$ [37, 38]; this increase in noise power is often called *noise folding*. It can be shown that for Gaussian noise, the oracle reconstruction error is proportional to $\sigma_{\Phi n}^2$, thus the reconstruction incurs a penalty due to the noise folding. Roughly speaking, this result implies that during reconstruction we lose about 3dB of signal-to-noise ratio (SNR) as the number of measurements decreases by half. The key series of results that make this so are as follows.

Suppose that $z = \Phi x - y$, where z is a zero-mean random vector with uncorrelated (white) entries, each having variance σ_z^2. Furthermore suppose that Φ has the RIP of order K, and that x is K-sparse. Then Theorem 4.1 of [38] demonstrates that oracle-assisted reconstruction will have expected error

$$\frac{K\sigma_z^2}{1+\delta} \leq \mathbb{E}(\|x - \widehat{x}\|_2^2) \leq \frac{K\sigma_z^2}{1-\delta}. \tag{1.10}$$

A key component of our analysis in Chapter 4 will be understanding the variance of the noise term z that arises from quantized noisy measurements. The expression (1.10) then gives the intuition that the expected reconstruction error behaves on the order of the variance of the error per measurement σ_z^2.

The variance σ_n^2 of the signal noise can be easily related to the variance of the measured noise $\sigma_{\Phi n}^2$. If n is white with mean zero and variance σ_n^2, and Φ has orthonormal rows, i.e., $\Phi\Phi^T = \frac{N}{M} I_M$,[3] then it is straightforward to show that the measured noise is also white and zero mean and has variance

$$\sigma_{\Phi n}^2 = \frac{N}{M}\sigma_n^2. \tag{1.11}$$

Note that the measured noise is only uncorrelated (i.e., white) when $M \leq N$; indeed, the condition $\Phi\Phi^T = \frac{N}{M} I_M$ can only hold when $M \leq N$.

In [38], the authors combine the results of (1.10) and (1.11) to obtain a bound on the oracle-assisted reconstruction error due to noise folding. In Chapter 4 we will take a similar approach, however we will additionally include the effects of quantization. Furthermore, because our quantization error is not necessarily uncorrelated, we also generalize (1.10) to obtain an upper bound on the oracle reconstruction error with correlated measurement noise.

[3] The so-called *tight frame* condition $\Phi\Phi^T = \frac{N}{M} I_M$ is not overly restrictive, since for any RIP matrix U, a matrix that has both the same row-space as U and the tight frame condition can be derived from U [38].

CHAPTER 2

Quantization and Dynamic Range in Compressive Sensing

Practical, finite range quantization imposes a finite dynamic range on a system, i.e., there is an intimate relationship between the scale and the precision of the signal that can be represented. A fundamental advantage of CS is that it enables a significantly lower sampling rate for sparse signals, which in turn enables the use of higher-resolution ADCs [7]. By exploiting this fact, a CS acquisition system should be able to provide a significantly larger dynamic range than a conventional system. In this chapter,[1] we justify this claim in two ways. First, we define and review finite range scalar quantization. Second, we provide a theoretical justification that the dynamic range of a conventional CS systems is on the same order as for a conventional ADC. We can then conclude that using a lower rate ADC enables higher bit-depth quantizers and thus the dynamic range is increased. Third, we demonstrate that because a large class of randomized CS systems are *democratic*, we can in fact increase the dynamic range of some CS systems in unconventional ways.

2.1 Finite-Range Scalar Quantization

In practice CS measurements are mapped to bits via a physical quantizer. A more precise model of the CS acquisition step (1.2) might be written as

$$y_Q = \mathcal{Q}_B(\Phi(x + n) + e), \tag{2.1}$$

where $\mathcal{Q}_B : \mathbb{R} \rightarrow \mathfrak{A}$ is a B-bit scalar quantization function (applied element-wise in (2.1)) that maps real-valued CS measurements to the discrete alphabet $\mathfrak{A}$ with $|\mathfrak{A}| = 2^B$. We have additionally

[1] This chapter includes work done in collaboration with Mark Davenport, John Treichler, Petros Boufounos, and Richard Baraniuk [34, 38].

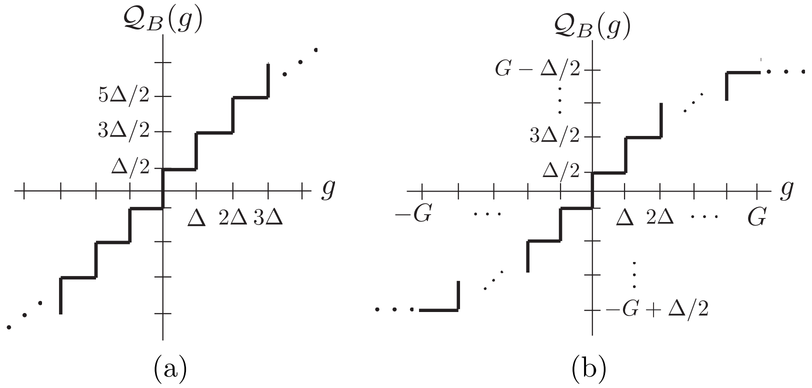


Figure 2.1: (a) Midrise scalar quantizer. (b) Finite-range midrise scalar quantization function $\mathcal{Q}_B$ with saturation level G and quantization interval $\Delta = 2^{-B+1}G$.

included *signal noise* $n \in \mathbb{R}^N$ that we will discuss in more detail in Chapter 4 and will be assumed to be zero unless otherwise noted. Since in a well-designed hardware system the primary source of measurement noise derives from quantization and for clarity of exposition, we will also assume $\|e\|_2 = 0^2$ for the remaining chapters.

In practice, quantizers have a finite dynamic range, dictated by hardware constraints such as the voltage limits of the devices and the finite number of bits per measurement of the quantized representation. Thus, a *finite-range* quantizer represents a symmetric range of values $|g| < G$, where $G > 0$ is known as the saturation level [63]. Values of g between $-G$ and G will not saturate, thus, the quantization interval is defined by these parameters as $\Delta = 2^{-B+1}G$. In this chapter, without loss of generality we assume a midrise B-bit uniform quantizer, i.e., the quantization levels are $q_k = \Delta/2 + k\Delta$, where $k = -2^{B-1}, \ldots, 2^{B-1}-1$. Note that if $|g| \leq G$, then we have that $|g - \mathcal{Q}_B(g)| \leq \Delta/2$, but if $|g| > G$ then $|g - \mathcal{Q}_B(g)| = |g| - (G - \Delta/2)$. Figure 2.1(a) depicts the mapping performed by a midrise quantizer with interval Δ and Figure 2.1(b) depicts a finite range variant with saturation level G.

The quantizer induces two forms of error on the measurement: quantization and saturation (or clipping) error. The former is due to the finite precision of the quantizer and the latter is due to the finite range of the quantizer. One way to account for quantization error is to treat it as bounded noise and employ robust reconstruction algorithms. Alternatively, we might try to reduce the error by choosing the most efficient quantizer for the distribution of the measurements. Several reconstruction techniques that specifically address CS quantization have also been proposed [34, 64–69]. Saturation error is usually avoided by scaling the measurements such that few or no saturations occur. We will see shortly that in CS there are alternative techniques for dealing with saturations.

[2] The general trends presented in this thesis remain unchanged when $\|e\|_2 > 0$, unless otherwise noted.

2.2 Dynamic Range of CS-Based Acquisition Systems

We begin our analysis by first providing a rigorous and general definition of dynamic range. Roughly, we define the dynamic range as the ratio of the maximum to the minimum signal power levels that can be handled with "full fidelity".[3] In order to make this notion precise, as previously stated, we will ignore the effects of any noise or nonlinearities from the other ADC components and examine only the impact of quantization. This is a fair assumption, since a key goal in the design of an ADC is that the quantizer be the only component that limits the device's dynamic range.

Our definition of dynamic range has two properties that aid us in the analysis of CS systems: (*i*) the dynamic range does not depend on a stochastic quantization error model, and (*ii*) any reduction of quantization error yields a corresponding improvement in dynamic range, i.e., the dynamic range of the quantizer effectively determines the dynamic range of the system. With this definition in hand, we examine quantization in both conventional and CS systems and provide lower bounds on the dynamic range of each. Our key finding in this section will be that, all things being equal, the dynamic range of a CS acquisition system is generally no worse than that of a conventional system. Thus, since CS enables lower sampling rates for sparse signals, we can employ a higher-resolution ADC and attain a larger dynamic range.

2.2.1 A deterministic approach to dynamic range

To formulate our definition of dynamic range, we first analyze the error induced by the quantization of x. For a given x, we define the *reconstruction SNR* (RSNR) as

$$\mathrm{RSNR} = \frac{\|x\|_2^2}{\|\widehat{x} - x\|_2^2}, \tag{2.2}$$

where $\widehat{x}$ is the output of our CS reconstruction algorithm and the *signal-to-quantization noise ratio* (SQNR) of the quantizer as

$$\mathrm{SQNR}(x) = \frac{\|x\|_2^2}{\|x - \mathcal{Q}_B(x)\|_2^2}. \tag{2.3}$$

We make the dependence of the SQNR on x explicit, since our definition of dynamic range will be based on how the scaling of x affects the SQNR. First, however, we establish a practical bound on the best SQNR attainable for a given G, Δ, and x.

Lemma 1. *Let $x \in \mathbb{R}^N$ be arbitrary. There always exists a $\beta > 0$ such that*

$$\mathrm{SQNR}(\beta x) \geq \frac{1}{\gamma(x)^2}\left(\frac{2G}{\Delta}\right)^2, \tag{2.4}$$

where

$$\gamma(x) = \frac{\|x\|_\infty}{\|x\|_2/\sqrt{N}}. \tag{2.5}$$

[3] In this section we are analyzing the CS-based receiver's dynamic range *as a system.* This should not be confused with the dynamic range *of a signal*, which in our framework could be quantified as the ratio of the largest to smallest entry in x.

The proof of this lemma can be found in Appendix A.1.

The quantity γ in (2.5) is known as the *peak-to-average ratio* (PAR) of x. Also known as the *crest factor* or *loading factor* [70], it is a measure of the ratio between a signal's "average energy" to its peak.

While the expression in (2.4) may look foreign to some, this bound is similar to standard results for peak SQNR. Recall that $2G/\Delta = 2^B$. Thus, if we express (2.4) in dB, then we observe that by setting β appropriately we can obtain

$$\mathrm{SQNR}(\beta x) \geq 20B\log_{10}(2) - 20\log_{10}(\gamma(x)) \gtrsim 6.02B - 20\log_{10}(\gamma(x)). \tag{2.6}$$

This corresponds to the well-known result that the peak SQNR grows by approximately 6dB per quantizer bit [70]. Furthermore, although the SQNR bound in (2.6) provides only a *lower bound* on the SQNR, it generally agrees with the results in the literature that assume probabilistic models on the signal x and/or the quantization noise. For example, a more conventional probabilistic analysis would assume that the quantization noise has a uniform distribution. In this case, one can derive the expression

$$\mathrm{SQNR}(\beta x) \approx 6.02B - 20\log_{10}(\gamma(x)) + 4.77,$$

where the additive constant 4.77 reflects the improvement made possible over our worst-case bound by placing a uniform distribution on the quantization noise [70]. For our purposes below, a lower bound on the SQNR is sufficient. We view the deterministic nature of our bound as a strength allowing us to avoid any questionable assumptions concerning the quantization noise distribution. It is important to note that by considering only the deterministic, worst-case error as in (2.6), the resulting expressions will generally differ from more standard results based on uniformly distributed quantization noise by 4.77dB.

We now show how we can use the SQNR to offer a concrete definition for dynamic range. Specifically, suppose that we are given a target SQNR C to achieve.[4] We aim to identify the range of scalings β of a given signal x for which $\mathrm{SQNR}(\beta x) \geq C$. More formally, we can always ensure that $\mathrm{SQNR}(\beta x) \geq C$ for all $\beta \in \left[\beta_C^{\min}(x), \beta_C^{\max}(x)\right]$, where $\beta_C^{\min}(x)$ and $\beta_C^{\max}(x)$ are scalars satisfying

$$\beta_C^{\min}(x) \leq G/\left\|x\right\|_\infty \leq \beta_C^{\max}(x). \tag{2.7}$$

In words, $\beta_C^{\max}(x)$ and $\beta_C^{\min}(x)$ define a range of scalings over which we achieve the desired SQNR C.

We define the dynamic range of a conventional acquisition system as

$$\mathrm{DR}_C(x) := \left(\frac{\beta_C^{\max}(x)}{\beta_C^{\min}(x)}\right)^2. \tag{2.8}$$

[4] In our analysis we consider $C \in \left(1, (2G/\Delta)^2/\gamma(x)^2\right]$ to ensure that our definition leads to a meaningful notion of dynamic range. Specifically, once we fix Δ and G, there is an upper limit on the SQNR we can hope to achieve, and for C beyond that limit the dynamic range will be ill-defined. Similarly, if we set $C = 1$ then one can easily achieve infinite dynamic range by quantizing all signals to zero. However, for the range of C considered we can always set $\beta = G/\left\|x\right\|_\infty$ will satisfy $\mathrm{SQNR}(\beta x) \geq C$ (see the proof of Lemma 1).

Hence, the dynamic range of a conventional ADC is the ratio of the maximum input scaling to the minimum input scaling of x such that for both scalings the SQNR is at least C.

At first sight, (2.8) may appear to be a rather complicated way of describing what is at heart an elementary concept — dynamic range is often simply quantified as the ratio of the largest to smallest quantization levels. However, the strength of this definition is that it can easily be extended to quantify the dynamic range of a CS-based ADC in which the measurement and reconstruction processes obscure the impact of finite-range quantization on the final RSNR as given by (2.2). Specifically, given an input signal x we apply a reconstruction algorithm to the quantized CS measurements $\mathcal{Q}_B\left(y\right) = \mathcal{Q}_B\left(\Phi x\right)$ to obtain a reconstruction $\widehat{x}$. We wish to understand the impact of this quantization on the resulting RSNR of $\widehat{x}$. While it might not otherwise be immediately apparent, (2.8) suggests a natural way to extend the definition of dynamic range to the CS setting by simply replacing RSNR with SQNR, i.e., defining $\beta_C^{\min}(x)$ and $\beta_C^{\max}(x)$ by considering the range of scalars β such that $\mathrm{RSNR}(\beta x) \geq C$. Note that for a conventional ADC, since RSNR = SQNR, the definition remains unchanged from (2.8). We now analyze the dynamic range of a conventional acquisition system in Section 2.2.2 and then extend this analysis to the CS setting in Section 2.2.3.

2.2.2 Dynamic range of a conventional ADC

We now provide a simple bound on the dynamic range $\mathrm{DR}_C(x)$ for a conventional ADC.

Theorem 2. *The dynamic range of a quantizer as defined by (2.8) is bounded by*

$$\mathrm{DR}_C(x) \geq \frac{1}{C\gamma(x)^2 - 1}\left(\left(\frac{2G}{\Delta}\right)^2 - 1\right), \tag{2.9}$$

where $\gamma(x)$ is defined as in (2.5).

The proof of this Theorem can be found in Appendix A.2.

For large B, the "-1" term in (2.9) will be negligible, and so by expressing (2.9) in dB we obtain

$$\mathrm{DR}_C(x) \gtrsim 6.02B - 10\log_{10}\left(C\gamma(x)^2 - 1\right). \tag{2.10}$$

This coincides with the familiar rule of thumb that just like the SQNR in (2.6), ADC dynamic range increases by 6dB per quantizer bit [70]. Note, however, that we again have an additive constant that here depends both on the targeted SQNR C as well as the PAR $\gamma(x)$. This is again expected, since a more ambitious required SQNR is more difficult to achieve and since a signal with higher PAR is harder to quantize, which both lead to a more limited dynamic range. We revisit the issue of PAR below in Section 2.2.4.

In summary, our definition of dynamic range (2.8) yields a reasonable expression (2.9) for a conventional ADC that coincides with the traditional "folk wisdom" on dynamic range.

2.2.3 Dynamic range of a CS-based acquisition system

Thus far we have proposed a rigorous and general definition of dynamic range and analyzed a conventional ADC in this context. We now extend the dynamic range analysis to the CS case. Our argument proceeds by first showing that we can always relate RSNR(βx) to SQNR(βy) and then relate SQNR(βy) to SQNR(βx). This allows us to argue that whenever SQNR(βx) $> C$, we have that RSNR(βx) $> C'$ for some C'. In other words, whenever we can achieve a certain SQNR C by directly quantizing x, a CS-based system can also achieve the RSNR C' (where C' is typically comparable to C). Thus, the dynamic range of these systems will be essentially the same. We begin by relating RSNR(βx) to SQNR(βy).

Lemma 2. *Suppose that $y = \Phi x$, where x is K-sparse and Φ satisfies the RIP of order K with constant δ. Let $\widehat{x}$ denote the output of applying a reconstruction algorithm to the quantized measurements $\mathcal{Q}_B(y)$ which satisfies a reconstruction guarantee like that given in Theorem 1, i.e.,*

$$\|\widehat{x} - x\|_2^2 \le \kappa_1^2 \|\mathcal{Q}_B(y) - y\|_2^2 . \tag{2.11}$$

Then

$$\mathrm{RSNR}(\beta x) \ge \frac{\mathrm{SQNR}(\beta y)}{(1+\delta)\kappa_1^2}. \tag{2.12}$$

Proof. Without loss of generality, suppose that $\beta = 1$. From the RIP we have that

$$\|x\|_2^2 \ge \frac{\|\Phi x\|_2^2}{1+\delta}.$$

Combining this with (2.11), we obtain the bound

$$\mathrm{RSNR}(x) = \frac{\|x\|_2^2}{\|\widehat{x} - x\|_2^2} \ge \frac{\|y\|_2^2}{(1+\delta)\kappa_1^2 \|\mathcal{Q}_B(y) - y\|_2^2} = \frac{\mathrm{SQNR}(y)}{(1+\delta)\kappa_1^2},$$

which completes the proof. □

In words, the RSNR(βx) is lower bounded by a constant multiple of the SQNR(βy). This means that we can expect the RSNR to follow the same trend as the SQNR of the measurements. Thus, we can restrict our analysis and comparisons to the measurement SQNR.

We next aim to compare SQNR(βy) here to SQNR(βx) from Section 2.2.1. The following lemma shows that we can bound SQNR(βy) in a manner similar to how Lemma 1 bounds SQNR(βx).

Lemma 3. *Suppose that $y = \Phi x$, where x is K-sparse and Φ satisfies the RIP of order K with constant δ. Then there exists a β such that*

$$\mathrm{SQNR}(\beta y) \ge (1-\delta)\frac{N}{M}\frac{\|x\|_\infty^2}{\|y\|_\infty^2}\frac{1}{\gamma(x)^2}\left(\frac{2G}{\Delta}\right)^2.$$

Proof. We begin by noting that from Lemma 1 we have that for $\beta = G/\|y\|_\infty$ we have that

$$\mathrm{SQNR}(\beta y) \ge \left(\frac{\|y\|_2^2/M}{\|y\|_\infty^2}\right)\left(\frac{2G}{\Delta}\right)^2.$$

Since Φ satisfies the RIP we have that

$$\|y\|_2^2 \geq (1-\delta)\,\|x\|_2^2 .$$

Thus we have that

$$\frac{\|y\|_2^2\,/M}{\|y\|_\infty^2} \geq \frac{(1-\delta)\,\|x\|_2^2}{\|y\|_\infty^2} = (1-\delta)\frac{N}{M}\frac{\|x\|_\infty^2}{\|y\|_\infty^2}\left(\frac{\|x\|_2^2\,/N}{\|x\|_\infty^2}\right) = (1-\delta)\frac{N}{M}\frac{\|x\|_\infty^2}{\|y\|_\infty^2}\frac{1}{\gamma(x)^2},$$

which establishes the lemma. □

Thus, CS has the same 6dB per quantier bit behavior as in (2.6) with

$$\mathrm{SQNR}(\beta y) \gtrsim 6.02B - 20\log_{10}(\gamma(x)) + 20\log_{10}\left(\frac{\sqrt{(1-\delta)\frac{N}{M}}\,\|x\|_\infty}{\|y\|_\infty}\right), \tag{2.13}$$

the only difference being an additional additive constant that we will analyze in more detail in Section 2.2.4.

We are now ready to compute the dynamic range of the CS acquisition system. We retain the same definition of dynamic range as in (2.8), but with $\beta_C^{\max}(x)$ and $\beta_C^{\min}(x)$ defined by substituting the SQNR constraint with the requirement that $\mathrm{RSNR}(\beta x) \geq C$. In this setting, we can repeat the same analysis as in Theorem 2 to obtain

$$\mathrm{DR}_C(x) \geq \frac{1}{C'\gamma(x)^2 - 1}\left(\left(\frac{2G}{\Delta}\right)^2 - 1\right),$$

where

$$C' = \frac{1-\delta}{(1+\delta)\kappa_1^2}\frac{N}{M}\frac{\|x\|_\infty^2}{\|y\|_\infty^2}.$$

Thus, when measured in dB the dynamic range is affected by CS only through an additive constant.

In practice, we can take significant advantage of the fact that, all things being equal, a CS system has the same dynamic range as a conventional Nyquist ADC. Moreover, because the ADC employed in a CS-based system operates at a significantly lower rate than would be required in a conventional system, a slower quantizer with higher bit-depth can be employed [7]. If the gain in effective bits is large, then the 6dB per bit improvement in dynamic range will dominate the additive constant and result in a substantial *increase* in the CS system's dynamic range as compared to a conventional ADC. We explore this idea empirically in Section 2.4.

2.2.4 Impact of CS on the PAR

We conclude this section on dynamic range with one last note regarding the mollifying effect of a CS acquistion system on the PAR. All of our expressions for the SQNR or RSNR as well as the dynamic range of a system depend in some way on the PAR of the signal x or the measurements y, depending on the context. In practice, the PAR has a significant impact on the resulting expressions. However, the PAR of a signal x can vary widely in the range

$$1 \leq \gamma(x) \leq \sqrt{N}, \tag{2.14}$$

which follows from standard norm inequalities. As an example, combining (2.14) with the lower bound on the SQNR of a conventional ADC in (2.6) means that in the best case (which corresponds to an all-constant vector x) the bound in (2.6) reduces to 6dB per bit growth in SQNR with no offset, whereas in the worst case (which corresponds to a $K = 1$ sparse x) we incur an additive penalty of $-10\log_{10}(N)$ dB. As the dimension N grows this penalty can become large, reflecting the fact that as the number of samples grows it becomes possible to construct a signal that has ever larger PAR. This translates to a similarly wide range of possible values for the additive penalty in the bound on dynamic range in (2.10).

Our aim here is to understand how CS impacts PAR. Clearly, we expect the PAR of the CS measurements y to differ from that of the signal x since each measurement typically consists of a weighted sum of the entries of x. Intuitively, such measurements have the potential to average out some of the "spikes" in x resulting in a potentially improved PAR. This appears in the analysis in the expression for SQNR(βy) in (2.13), which shows that SQNR(βy) can be improved over SQNR(βx) in (2.6) if $\frac{N}{M}\left\|x\right\|_\infty^2 / \left\|y\right\|_\infty^2$ is somewhat larger than 1.

In the worst-case, the quantity $\frac{N}{M}\left\|x\right\|_\infty^2 / \left\|y\right\|_\infty^2$ can be a great deal smaller than 1; however, on average we are likely to do significantly better. As an illustration, we describe what can be said when Φ is a matrix with i.i.d. $\pm 1/\sqrt{M}$ (Rademacher) entries.

We begin with the worst-case. By combining the the Cauchy-Schwartz inequality with standard ℓ_p-norm inequalities, we have that for all j, $|y_j| \leq \frac{N}{M}\left\|x\right\|_\infty$. Thus we obtain

$$\frac{N}{M}\frac{\left\|x\right\|_\infty^2}{\left\|y\right\|_\infty^2} \geq \frac{M}{N}.$$

Hence, in the worst-case

$$20\log_{10}\left(\frac{\sqrt{(1-\delta)\frac{N}{M}}\left\|x\right\|_\infty}{\left\|y\right\|_\infty}\right) \approx -10\log_{10}\left(\frac{N}{M}\right),$$

which corresponds to an SQNR loss of 3dB per octave increase in the subsampling factor. However, this bound will be achieved only when x is both constant magnitude and has elements with signs exactly matching one of the (randomly chosen) rows of Φ — a highly unlikely scenario. Furthermore, this bound makes no use of the "dithering" effect promoted by the randomized measurements; a grave omission indeed. Towards this end, we next consider a probabilistic bound to see that we can typically obtain better performance.

Lemma 4. *Suppose that Φ is chosen with i.i.d. entries with variance $1/M$ drawn according to any strictly sub-Gaussian distribution. Then*

$$\frac{N}{M}\frac{\left\|x\right\|_\infty^2}{\left\|y\right\|_\infty^2} \geq \frac{\gamma(x)^2}{4\log(M)} \tag{2.15}$$

with probability at least $1 - 2/M$.

Proof. By combining the union bound (over M measurements) with standard tail bounds on a strictly sub-Gaussian distribution, we obtain

$$\mathbb{P}\left(\|y\|_\infty > t\right) \le 2M \exp\left(-\frac{Mt^2}{2\|x\|_2^2}\right).$$

Thus, the probability that (2.15) does not hold is bounded by

$$\begin{aligned} 2M \exp\left(-\frac{4M\frac{N}{M}\log(M)\|x\|_\infty^2}{2\gamma(x)^2\|x\|_2^2}\right) &= 2M\exp\left(-\frac{2\log(M)\|x\|_\infty^2}{\gamma(x)^2\|x\|_2^2/N}\right) \\ &= 2\exp\left(\log(M) - 2\log(M)\right) = \frac{2}{M}, \end{aligned}$$

which establishes the lemma. □

Thus, in practice we expect our bound for SQNR(y) in (2.13) to differ from our bound for SQNR(x) in (2.6) only by a factor of $\gamma(x)^2/4\log(M)$. Recalling our bound on $\gamma(x)$ we have that

$$\frac{1}{4\log(M)} \le \frac{\gamma(x)^2}{4\log(M)} \le \frac{N}{4\log(M)}.$$

Hence, for x with small PAR, we can expect a potential loss in SQNR when compared to direct quantization of x, while for x with moderate or large PAR we can actually expect a significant improvement.

Finally, we can use Lemma 4 to approximate (2.13) with high probability as

$$\mathrm{SQNR}(\beta y) \gtrsim 6.02B - 20\log_{10}\left(4\log(M)/\sqrt{1-\delta}\right),$$

which implies that CS allows us to essentially eliminate the negative impact of high PAR signals. This is because the randomized measurement procedure of CS will, with high probability, produce measurements having a PAR that is completely independent of the input signal's PAR. For high PAR signals, this results in a substantial improvement.

2.3 Liberating Dynamic Range via Democracy

As previously explained, the limited dynamic range of the system is induced by both the precision and the finite range of the quantizer. An example of how limited dynamic range presents a design challenge in practice is as follows. Error due to saturation is typically considered more detrimental than the error due to quantization. Thus the naïve approach to dealing with saturation is to scale down the amplitude of the signal or its measurements so that saturation never or very rarely occurs. This is the approach pursued in many conventional sensor systems; a typical rule of thumb used with communication system ADCs suggests that one reduce the signal amplitude until only 63 in one million samples saturates [32]. Unfortunately, scaling down the signal amplitude proportionately scales up the amount of quantization noise.

Fortunately, we can exploit the so-called *democracy* property exhibited by many CS systems. Roughly, this property explains that each measurement contains about the same amount of information as any other measurement. Said another way, it is possible to reconstruct sparse signals from any subset of measurements, subject to only a small penalty in reconstruction error. This means that if a measurement saturates with significant error, we may incur less reconstruction error by simply discarding it.

In [34, 71], the authors demonstrated that indeed rejecting saturated measurements can lead to improved performance. Interestingly, these results concluded that the best performance in these systems is achieved when the quantizer range is set low enough to induce a significantly non-zero saturation rate. This is due to the fact that as the quantizer range G decreases (and thus saturation rate increases), the error due to quantization on the remaining measurements decreases since the quantization interval decreases, as expressed by $\Delta = 2^{-B+1}G$. Furthermore, the authors found that the amount of saturation allowed is determined by the sparsity of signal. The implication is clear: the dynamic range of these democratic systems is limited by the complexity of the signal, not the range of the quantizer.

In this section we review the democracy property and some of its implications. We review two reconstruction approaches for dealing with saturation. We then discuss how these approaches lead to increased dynamic range. The approaches detailed in this section will provide significant motivation the ideas and methods found in the next chapter.

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

2.3.1 The democratic caucus of random matrices

We begin by establishing a strong notion of the democratic property of a matrix Φ as in [34, 35, 46].

Definition 1. *Let Φ be an $M \times N$ matrix, and let $\mathfrak{M} \leq M$ be given. The matrix Φ is $(\mathfrak{M}, K, \delta)$-democratic if, for all Γ such that $|\Gamma| \geq \mathfrak{M}$, the matrix Φ^{Γ} satisfies the RIP of order K with constant δ.*

In words, this definition explains that for any RIP matrix Φ, any subset of rows of Φ will satisfy the RIP, with perhaps a different constant δ.

It is possible to show that certain randomly generated matrices will be $(\mathfrak{M}, K, \delta)$-democratic. The following theorem restates a result of [34, 35, 46] for democratic Gaussian matrices, but the analysis can be extended (with different constants) to the more general class of *sub-Gaussian* matrices (see methods in [46]).

Theorem 3. *Let Φ be an $M \times N$ matrix with elements ϕ_{ij} drawn according to $\mathcal{N}(0, 1/M)$ and let $\mathfrak{M} \leq M$, $K < \mathfrak{M}$, and $\delta \in (0,1)$ be given. Define $D = M - \mathfrak{M}$. If*

$$M = C_1(K+D)\log\left(\frac{N+M}{K+D}\right), \tag{2.16}$$

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

Observe that we require roughly $O(D\log(N))$ additional measurements to ensure that Φ is $(\mathfrak{M}, K, \delta)$-democratic compared to the number of measurements required to simply ensure that Φ satisfies the RIP of order K (recall that $D = M - \mathfrak{M}$). This seems intuitive; if we wish to be robust to the loss of any D measurements while retaining the RIP of order K, then we should expect to take *at least* D additional measurements.

Theorem 3 further guarantees the graceful degradation of CS recovery due to loss of measurements. Specifically, the theorem implies that recovery from any subset of CS measurements is stable to the loss of a potentially larger number of measurements than anticipated. To see this, suppose that an $M \times N$ matrix Φ is $(M - D, K, \delta)$-democratic, but consider the situation where $D + \mathfrak{D}$ measurements are dropped. It is clear from the proof of Theorem 3 that if $\mathfrak{D} < K$, then the resulting matrix Φ^{Γ} will satisfy the RIP of order $K - \mathfrak{D}$ with constant δ. Thus, from [72], if we define $\mathfrak{K} = (K - \mathfrak{D})/2$, then the signal recovery error is bounded by

$$\|x - \widehat{x}\|_2 \leq C_3 \frac{\|x - x_{\mathfrak{K}}\|_1}{\sqrt{\mathfrak{K}}}, \tag{2.17}$$

where $x_{\mathfrak{K}}$ denotes the best $\mathfrak{K}$-term approximation of x and C_3 is an absolute constant depending on Φ that can be bounded using the constants derived in Theorem 3. Thus, if $\mathfrak{D}$ is small enough, then the additional error incurred by dropping too many measurements will also be relatively small.

This property and its implications are key to enabling the rejection of saturated measurements.

2.3.2 Saturation rejection signal recovery

A simple and intuitive way to handle saturated measurements is to simply discard them and then run a standard CS recovery algorithm [71]. Using, for instance, (BPDN) for reconstruction yields the program:

$$\widehat{x} \leftarrow \operatorname*{argmin}_{x \in \mathbb{R}^N} \|x\|_1 \quad \text{s.t.} \quad \|\Phi^U x - y^U\|_2 < \epsilon. \tag{2.18}$$

Since the democracy property implies that any $\mathfrak{M} \times N$ submatrix of Φ has RIP, it immediately follows from Theorem 1 that the saturation rejection program (2.18) yields a signal estimate with the stability guarantee (1.5). By the same argument, it is straightforward to demonstrate that other algorithms such as CoSaMP applied to Φ^U and y^U will achieve performance given by Theorem A in [29], as long as they rely on the RIP for performance guarantees.

2.3.3 Saturation rejection signal processing

Saturation rejection is also useful in conjunction with processing and inference techniques that work directly on the compressive measurements. For example, in the *smashed filter* for signal detection and classification the key calculation is the inner product $\langle \Phi x, \Phi v \rangle$ between the compressive measurements of a test signal x and a target template signal v [73]. If x and v are sparse then, thanks to the RIP, this low-dimensional inner product can be used as a proxy for the inner product between x, and v; that is $\langle \Phi x, \Phi v \rangle \approx \langle x, v \rangle$. Unfortunately, if any of the elements of Φx or Φv are saturated, then the approximation no longer holds and the performance of the smashed filter deteriorates.

Consider $\mathcal{Q}_B(\Phi x)$ and $\mathcal{Q}_B(\Phi v)$ and let Γ_x and Γ_v be the supports of the measurements that do not saturate on each vector, respectively. Then we have that for $\Gamma = \Gamma_x \cap \Gamma_v$ that $\|\mathcal{Q}_B(\Phi^\Gamma x) - \Phi^\Gamma x\|_\infty \leq \Delta/2$ and $\|\mathcal{Q}_B(\Phi^\Gamma v) - \Phi^\Gamma v\|_\infty \leq \Delta/2$. Thus, it is straightforward to show that

$$\left|\langle \mathcal{Q}_B(\Phi^\Gamma x), \mathcal{Q}_B(\Phi^\Gamma v)\rangle - \langle \Phi^\Gamma x, \Phi^\Gamma v\rangle\right| \leq \frac{\Delta^2}{4} + \frac{\Delta}{2}\left|\sum_n (\Phi^\Gamma x)_n\right| + \frac{\Delta}{2}\left|\sum_n (\Phi^\Gamma v)_n\right|. \tag{2.19}$$

Furthermore, the two sums in (2.19) are likely to concentrate around zero. The democracy of Φ furthermore implies that $\langle \Phi^\Gamma x, \Phi^\Gamma v\rangle \approx \langle x, v\rangle$. Thus, discarding the corresponding entries of Φx and Φv when one of them saturates makes considerable practical sense.

2.3.4 Saturation consistency signal recovery via convex optimization

Clearly saturation rejection discards potentially useful signal information, since we know that saturated measurements are large (we just do not know *how* large). It is possible to augment a standard convex optimization-based CS recovery algorithm with a set of inequality constraints that enforce signal *consistency* with the saturated measurements. By consistency we mean that the magnitudes of the values of $\Phi\widehat{x}$ corresponding to the saturated measurements are larger than $G - \Delta$, i.e., they are consistent with what we observed.

More specifically, let S^+ and S^- correspond be the index sets of the positive saturated measurements and negative saturated measurements, respectively. Define the matrix Φ^S as

$$\Phi^S := \begin{bmatrix} \Phi^{S^+} \\ -\Phi^{S^-} \end{bmatrix}. \tag{2.20}$$

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

$$\hat{x} \leftarrow \operatorname*{argmin}_{x \in \mathbb{R}^N} \|x\|_1 \qquad \text{s.t.} \qquad \|\Phi^U x - y^U\|_2 < \epsilon \tag{2.21a}$$

$$\text{and} \qquad \Phi^S x \geq (G - \Delta)\cdot \mathbf{1}, \tag{2.21b}$$

where $\mathbf{1}$ denotes an $(M - \mathfrak{M}) \times 1$ vector of ones. In words, we seek the x with the minimum ℓ_1 norm such that the measurements that do not saturate have bounded ℓ_2 error and the measurements that do saturate are consistent with the saturation constraint. The program (2.21) obeys the same reconstruction error bounds as (2.18) [34]. Alternative regularization terms that impose the consistency requirement on the unsaturated quantized measurements can be used on y^U, such as those proposed in [64, 65], or alternative techniques for the unsaturated quantized measurements can be used such as those proposed in [66].

In addition to the convex optimization program (2.21), the authors in [34] proposed a greedy algorithm, *saturation consistent CoSaMP* (SC-CoSaMP) to impose saturation consistency during reconstruction. Some of our simulations will make use of this algorithm since it is fast and has been shown empirically to improve performance from finite range quantized measurements.

We note that a saturation rejection algorithm and a saturation consistency algorithm will not necessarily yield the same signal estimate. This is because the solution from the rejection approach may not lie in the feasible set of solutions of the consistency approach (2.21). However, the reverse is true. The solution to the consistent approach does lie in the feasible set of solutions of the rejection approach. While we do not provide a detailed analysis that compares the performance of these two algorithm classes, one should expect that the consistency approach will outperform the rejection approach in general, since it incorporates additional information about the signal.

2.4 Experimental Performance of CS Dynamic Range

In this section, we conduct an experiment that demonstrates how the dynamic range of a CS system can be increased. We are interested in demonstrating two main points: *i)* as we decrease the sample rate of a system, we can apply a higher bit-depth quantizer. This should improve performance even though the number of measurements is fewer; and *ii)* the saturation consistent approach that utilizes the democracy of CS systems can be used to increase the performance further, by extending the dynamic range of the system.

Any improvement in the SQNR of the CS measurements will translate to an improved dynamic range. Thus, in our experiments, we compute the average RSNR obtained after recovery from quantized CS measurements as a proxy for the dynamic range. Furthermore, we make use of the trends outlined in [7] that show that the number of bits per measurement grows according to $B = \lambda - 10 \log 10(M)/2.3$ where λ is a constant that determines the bit-depth of a Nyquist-rate sampler. The number of bits per measurements then grows linearly with the octaves of subsampling, with slope of about 1.3. This relationship between sample rate and bit-depth is fundamental to understanding the dynamic range benefits of CS systems.

Our experiment proceeds as follows and is depicted in Figure 2.2. The signal to be acquired consists of a single 3.1 kHz-wide unmodulated voice signal single-side-band-upconverted to a frequency within the 1 MHz input bandwidth of the receiver. The signals are noise-free so that we can isolate the impact of quantization noise. Additionally, we employ an ideal *random demodulator* [18] (discussed in Chapter 5) to measure the signals. Performance is measured as a function of the subsampling factor N/M. In each trial we generate a single voice-like signal and compute measurements with the CS receiver. The measurements are further quantized utilizing the full scale of the quantizer in the oracle and conventional CoSaMP cases. In the saturation consistent case, the scale of the signal (and thus measurements) is tuned to maximize the RSNR performance. This optimal performance occurs when a significant number of measurements have saturated. The

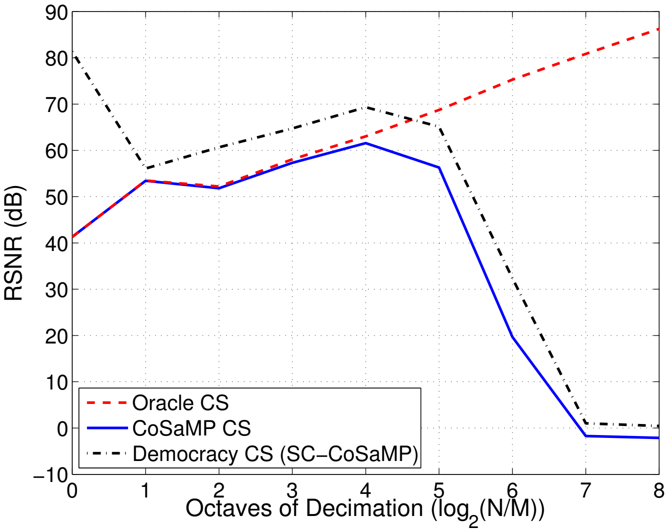


Figure 2.2: RSNR for an environment consisting of a noise-free single unmodulated voice channel and quantized measurements starting at a bit-depth of 4 bits per measurement when $\log_2(N/M) = 0$. We increased the bit-depth as a function of the sample rate according to the trends outlined in [7]. We see a marked improvement in RSNR as a direct result of the sampling rate being decreased. We see further improvement when the gain is tuned to maximize the performance of a saturation consistent algorithm, SC-CoSaMP. Interestingly, the best saturation consistent CS performance occurs when no subsampling has been performed, but when significantly many measurements saturate (even though the quantizer precision is at its lowest). This suggests that it may be beneficial to sample at a high rate and increase the dynamic range by exploiting the democratic nature of CS systems, rather than applying a higher bit-depth quantizer at a lower rate.

measurements were quantized to 4 bits each, and then recovered using CoSaMP (solid line), the oracle recovery algorithm (dashed line), and SC-CoSaMP (dash-dotted). We report the average RSNR for each subsampling factor.

From this experiment we see that in both the oracle and conventional CS cases, the RSNR grows significantly, achieving a 20dB gain at 4 octaves of subsampling over Nyquist sampling. Conventional CS performance then decreases as we move to an undersampled regime where CS recovery is no longer sustainable (too few measurements for the given sparsity). The oracle performance continues to improve as subsampling is further increased. This experiment highlights the very real benefit of reduced sampling rates; easing the sampling rate requirements can allow us to use higher fidelity hardware components, such as high bit-depth quantizers.

The saturation consistent case provides further insight. When the number of measurements is decimated, even by half ($\log_2(N/M) = 1$), the saturation consistent approach achieves about a 5dB to 10dB gain over the conventional CS approach, but follows the same performance trend. However, when there is no decimation, the saturation consistent algorithm exhibits a 40dB gain over conventional CS and the oracle. Indeed the SC-CoSaMP performance at the Nyquist rate is as good as the oracle performance at more than 6 octaves of decimation (i.e., better than using an ADC that is 2^6 times as slow). The implication of this result is that it might be better to take many measurements and drive up the gain such that most of them saturate, rather than attempting to reduce the sampling rate and applying a higher bit-depth quantizer. Abusing the quantizer by saturating most of the measurements leads to the theme of the following chapters– can we saturate *all* of the measurements? Can we expand CS methods to include scenarios where we drive up the sampling rate and drive down the bit-depth of the quantizer?

CHAPTER 3

Single Bit Compressive Sensing

3.1 Supersaturated Sensing

One question that arises from the previous chapter is how many measurements can saturate in practice? The saturation rejection reconstruction approach of Section 2.3.2 will fail when the number of non-saturated measurements is too few; unfortunately the constants for the democratic property of random matrices are not tight enough to predict the precise number of measurements at which this transition occurs. It has been shown that the saturation consistent reconstruction approach of Section 2.3.4 can achieve reasonable performance in the face of significantly more saturation than in the rejection approach; however, even this technique appears to fail when too many measurements have saturated [34].

In this chapter,[1] we consider the most extreme case when all measurements have saturated, i.e., the measurements are *supersaturated*. We ask the question: is signal reconstruction possible in this regime? In supersaturated sensing, the measurements take the value G or $-G$ and information about the true scale of the signal is lost. Indeed, we are only able maintain a single bit of information about each measurement and a lower bound on the signal energy. In fact, if we intend to operate only in the supersaturated regime, the quantizer can be reduced to a simple comparator that tests if values are above or below zero, enabling extremely simple, efficient, and fast quantization.

It is not obvious that the signs of the CS measurements retain enough information for signal reconstruction. For instance, as just explained, we have lost information about the scaling of the signal. Nonetheless, there has been recent empirical evidence that signal reconstruction is possible from just the signs of the measurements, via the 1-bit compressive sensing framework established

[1] This chapter includes work done in collaboration with Laurent Jacques, Petros Boufounos, Zaiwen Wen, Wotao Yin, and Richard Baraniuk [74, 75].

in [33, 39, 76]. This framework suggests that signals can be reconstructed, up to a scale factor, from only the signs of their CS measurements.

The primary contribution of this chapter is a rigorous analysis of the 1-bit CS framework. We provide two flavors of results. First, we determine the best achievable performance of this 1-bit CS framework. We further demonstrate that if the elements of measurement system Φ are drawn randomly from Gaussian distribution or its rows are drawn uniformly from the unit sphere, then it is possible to pose a reconstruction formulation that will have bounded error on the order of the optimal lower bound. Second, we provide conditions on the measurement system that enable us to characterize the reconstruction performance even when some of the measurement signs have changed (e.g., due to noise in the measurements). In other words, we derive the conditions under which robust reconstruction from 1-bit measurements can be achieved. We do so by demonstrating that 1-bit CS systems can be stable embeddings of sparse signals, in similar fashion to the RIP systems of conventional CS. We apply these stable embedding results to the cases where we have noisy measurements and signals that are not strictly sparse. Our guarantees demonstrate that the 1-bit CS framework is on sound footing.

To develop robust reconstruction guarantees, we propose a new tool, the *binary ϵ-stable embedding* (BϵSE), to characterize 1-bit CS systems. The BϵSE implies that the normalized angle between any sparse vectors on the unit sphere is close to the normalized Hamming distance between their 1-bit measurements. We demonstrate that again the quantized measurements from Gaussian measurement matrices exhibit this property when $M \geq C_\epsilon K \log N$ (where C_ϵ is some constant). Thus remarkably, there exist systems such that the BϵSE holds when both the number of measurements M is smaller than the dimension of the signal N and the measurement bit-depth is at minimum.

As a complement to our theoretical analysis, we introduce two algorithms to solve the non-convex reconstruction problem originally posed in this context, as well as several new convex formulations of the reconstruction problem. We present extensive numerical simulations to prove the validity of this framework and these algorithms. Finally, we demonstrate that the 1-bit reconstruction algorithms can be extended to perform consistent reconstruction of multibit quantized measurements with arbitrary numbers of saturations. We thusly provide a complete solution for handling finite range quantized measurements.

This chapter is organized as follows. In Section 3.2 we formally summarize the 1-bit CS framework. In Section 3.3 we describe some additional benefits of 1-bit quantized measurements beyond those described above. In Section 3.4 we provide reconstruction bounds on the performance from noiseless measurements and demonstrate that a large class of measurements matrices will yield such performance. In Section 3.5, we introduce the BϵSE property that ensures robust recovery guarantees. We then prove that such mappings exist and give an example of a class of these matrices. This section also provides reconstruction bounds for noisy measurements and compressible signals. In Section 3.6 we demonstrate how the 1-bit framework can be extended to handle multibit quantized measurements as well as an arbitrarily large or small number of saturations. In Section 3.7 we introduce two new algorithms for solving the reconstruction problem and also pose some convex formulations. In Section 4.2 we perform numerical simulations to validate and characterize the ideas presented in this chapter. We conclude by reviewing some alternative 1-bit frameworks in

Section 3.9.

3.2 The 1-bit CS framework

We briefly describe the 1-bit CS framework proposed in [39]. Measurements of a signal $x \in \mathbb{R}^N$ are computed via

$$y_s = A(x) := \operatorname{sign}(\Phi x). \tag{3.1}$$

Thus, the measurement operator $A(\cdot)$ is a mapping from $\mathbb{R}^N$ to the Boolean cube[2] $\mathcal{B}^M := \{-1, 1\}^M$. At best, we hope to recover signals $x \in \Sigma_K^* := \{x \in S^{N-1} : \|x\|_0 \leq K\}$ where $S^{N-1} := \{x \in \mathbb{R}^N : \|x\|_2 = 1\}$ is the unit hyper-sphere of dimension N. We restrict our attention to sparse signals on the unit sphere since, as previously mentioned, the scale of the signal has been lost during the quantization process. To reconstruct, we enforce *consistency* on the signs of the estimate's measurements, i.e., that $A(\widehat{x}) = A(x)$. Specifically, we define a general non-linear reconstruction algorithm $\Delta^{\text{1bit}}(y_s, \Phi, K)$ such that, for $\widehat{x} = \Delta^{\text{1bit}}(y_s, \Phi, K)$, the solution $\widehat{x}$ is

(i) sparse, i.e., satisfies $\|\widehat{x}\|_0 \leq K = \|x\|_0$; and

(ii) consistent, i.e., satisfies $A(\widehat{x}) = y_s = A(x)$.

With (ℓ_0-min) from CS as a guide, one candidate program for reconstruction is of course

$$\widehat{x} \leftarrow \operatorname*{argmin}_{x \in S^{N-1}} \|x\|_0 \quad \text{s.t.} \quad y_s = \operatorname{sign}(\Phi x). \tag{ℓ_0-min$_{\text{1B}}$}$$

Although the parameter K is not explicit in (ℓ_0-min$_{\text{1B}}$), the property *(i)* above holds because x is a feasible point of the constraint.

Since (ℓ_0-min$_{\text{1B}}$) is computationally intractable, [39] proposes a relaxation that replaces the objective with the ℓ_1-norm and enforces consistency via a linear convex constraint. Specifically, let the matrix Y have the elements of y_s along the diagonal and zero elsewhere. Then we can try to solve

$$\widehat{x} \leftarrow \min_{x \in S^{N-1}} \|x\|_1 \;\text{ s.t. }\; Y\Phi x \geq 0 \;\text{ and }\; \|x\|_2 = 1, \tag{ℓ_1-min$_{\text{1B}}$}$$

rather than (ℓ_0-min$_{\text{1B}}$). The ℓ_1 objective favors sparse solutions while the first constraint enforces consistency between the 1-bit quantized measurements and the solution. However, (ℓ_1-min$_{\text{1B}}$) remains non-convex due to the the unit-sphere requirement. Be that as it may, an algorithm has been developed for the relaxation, as well as a greedy algorithm inspired by the same ideas [39, 76]. The program (ℓ_1-min$_{\text{1B}}$) can also be posed in a convex way as will be discussed in Section 3.7.3, but for the sake of theory we proceed with the unit energy constraint.

Figure 3.1 depicts the geometry of the components of (ℓ_1-min$_{\text{1B}}$) in two dimensions. The hyperplanes (lines) φ_1 and φ_2 correspond to the first and second rows of Φ, respectively. In this figure they are drawn to be perfectly orthogonal but in general this may not be the case.

[2] Generally, the M-dimensional Boolean cube is defined as $\{0, 1\}^M$. Without loss of generality, we use $\{-1, 1\}^M$ instead.

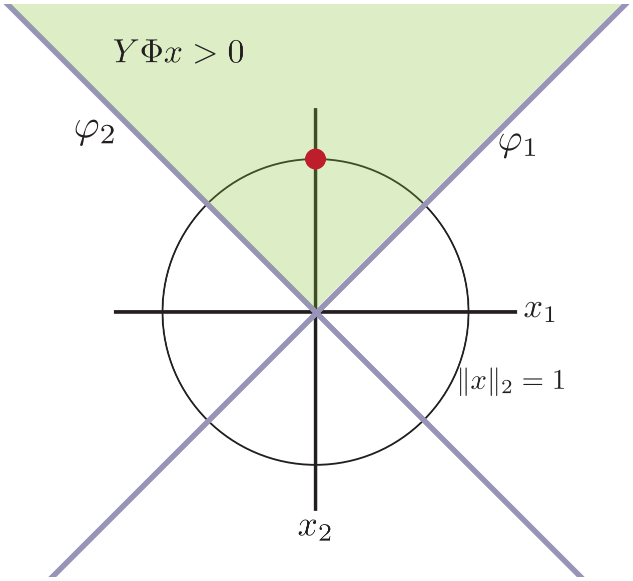


Figure 3.1: The geometry of the components of the 1-bit CS reconstruction formualtion in two dimensions. The hyperplanes (lines) φ_1 and φ_2 correspond to the first and second rows of Φ, respectively. The green shaded region between the planes denotes the feasible region. The circle denotes the unit sphere. The red dot denotes the sparsest feasible solution on the unit sphere.

Indeed, if these rows were drawn randomly from a Gaussian distribution, they will be approximately orthogonal. Furthermore, in this example we choose Φ to be square to clearly depict the relevant concepts in two dimensions. The green shaded region depicts the feasible region, i.e., the set where all x satisfy $Y\Phi x \geq 0$, and thus have measurement signs that are consistent with the diagonal of Y. The unit sphere is represented by the circle labelled $\|x\|_2 = 1$ and thus the only unit norm sparse solution in the feasible region lies at $[0, 1]$, denoted by the red dot. The key feature of this picture is that each row of Φ defines some hyperplane and each measurement sign determines on which side of the hyperplane the solution lies. The feasible region can be though of as a "cone[3]" and our goal during reconstruction is to find the sparsest solution within that region. A major goal of this chapter is to show that all sparse unit norm solutions in this cone are within some small error tolerance of each other.

3.3 Immediate Benefits of 1-bit CS

There are several benefits to obtaining 1-bit quantized measurements. First, efficient hardware quantizers can be built to operate at high speeds, since the quantizer can be a simple comparator that merely tests if a measurement is above or below zero. Indeed, as previously discussed there is an inverse relationship between sample rate and quantization bit-depth, such that the sample rate increases exponentially as the bit-depth is decreased linearly. Second, it has been shown that

[3] This is not a true cone in the geometrical sense.

the program (ℓ_1-min$_{1\mathrm{B}}$) can be used to recover signals with gross non-linearities applied to the measurements [33]. In particular, suppose a non-linearity $f(\cdot)$ is applied to the measurements. If the $f(\cdot)$ preserves the sign of the measurements, then clearly (ℓ_1-min$_{1\mathrm{B}}$) can be still be used to recover x with the same performance as using the non-linearity-free measurements. Additionally, if we assume that the non-linearity preserves the relationship

$$f(x_i) < f(x_{i+1}) \text{ if } x_i < x_{i+1},$$

then the program

$$\widehat{x} \leftarrow \Delta^{1\mathrm{bit}}(\mathrm{sign}(\mathrm{diff}(f(\Phi x))), D\Phi, K), \tag{3.2}$$

can be used to recover x with similar guarantees as (ℓ_1-min$_{1\mathrm{B}}$), where D is a difference matrix with 1's along the diagonal and -1's along the first sub-diagonal, with $\mathrm{diff}(x) = x_{i+1} - x_i,$ for $i = 1, \ldots, N-1$ [33].

3.4 Noiseless Reconstruction Performance

3.4.1 Reconstruction performance lower bounds

In this section, we seek to provide guarantees on the reconstruction error from 1-bit CS measurements. Before analyzing this performance from a specific mapping A with the consistent sparse reconstruction algorithm $\Delta^{1\mathrm{bit}}(y_s, \Phi, K)$, it is instructive to determine the best achievable performance from measurements acquired using any mapping. Thus, in this section we seek a lower bound on the reconstruction error.

We develop the lower bound on the reconstruction error based on how well the quantizer exploits the available measurement bits. A distinction we make in this section is that of measurement bits, which is the number of bits acquired by the measurement system, versus information bits, which represent the true amount of information carried in the measurement bits. Our analysis follows similar ideas to that in [77, 78], adapted to sign measurements.

We first examine how 1-bit quantization operates on the measurements. Specifically, we consider the orthants of the measurement space. An orthant in $\mathbb{R}^M$ is the set of vectors such that all the vector's coefficients have the same sign pattern

$$\mathcal{O}_s = \{x \mid \mathrm{sign}\, x = s\},$$

where s is a vector of ± 1. Any M-dimensional space is partitioned to 2^M orthants. Figure 3.2(a) shows the 8 orthants of $\mathbb{R}^3$ as an example. Since 1-bit quantization only preserves the signs of the measurements, it encodes in which measurement space orthant the measurements lie. Thus, each available quantization point corresponds to an orthant in the measurement space. Any unquantized measurement vector Φx that lies in an orthant of the measurement space will quantize to the corresponding *quantization point* of that orthant and cannot be distinguished from any other measurement vector in the same orthant. To obtain a lower bound on the reconstruction error, we begin by bounding the number of quantization points (or equivalently the number of orthants) that are used to encode the signal.

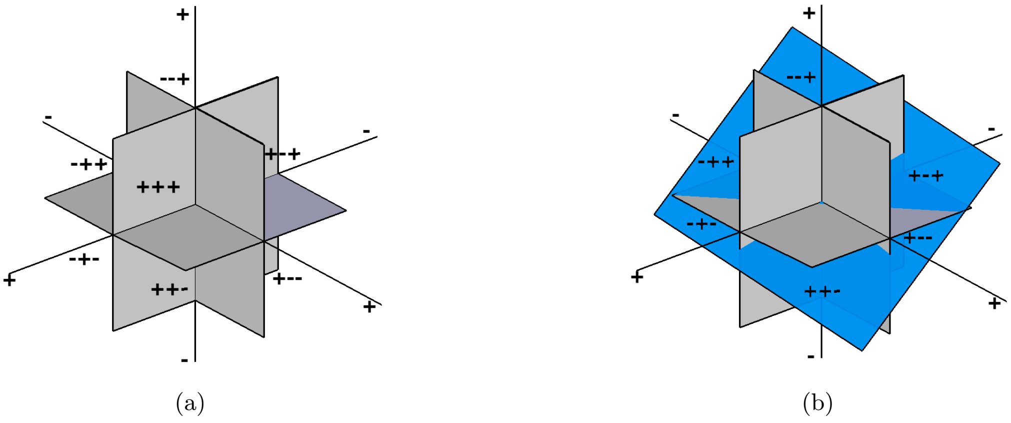


Figure 3.2: (a) The 8 orthants in $\mathbb{R}^3$. (b) Intersection of orthants by a 2-dimensional subspace. At most 6 of the 8 available orthants are intersected.

While there are generally 2^M orthants in the measurement space, the space formed by measuring all sparse signals occupies a small subset of the available orthants. We determine the number of available orthants that can be intersected by the measurements in the following lemma:

Lemma 5. *Let $x \in \mathcal{S} := \bigcup_{i=1}^{L} \mathcal{S}_i$ belong to a union of L subspaces $\mathcal{S}_i \subset \mathbb{R}^N$ of dimension K, and let M 1-bit measurements y_s be acquired via the mapping $A : \mathbb{R}^N \to \mathcal{B}^M$ as defined in (3.1). Then the measurements y_s can effectively use at most $L\binom{M}{K}2^K$ quantization points, i.e., carry at most $K \log_2(eLM/K)$ information bits.*

Proof. A K-dimensional subspace in an M dimensional space cannot lie in all the 2^M available octants. For example, as shown in Fig. 3.2(b), a 2-dimensional subspace of a 3-dimensional space can intersect at most 6 of the available octants. In Appendix B.1, we demonstrate that one arbitrary K-dimensional subspace in an M-dimensional space intersects at most $\binom{M}{K}2^K$ orthants of the 2^M available. Since Φ is a linear operator, any K-dimensional subspace $\mathcal{S}_i$ in the signal space $\mathbb{R}^N$ is mapped through Φ to a subspace $\mathcal{S}'_i = \Phi\mathcal{S}_i \subset \mathbb{R}^M$ that is also at most K-dimensional and therefore follows the same bound. Thus, if the signal of interest belongs in a union $\mathcal{S} := \bigcup_{i=1}^{L} \mathcal{S}_i$ of L such K-dimensional subspaces, then $\Phi x \in \mathcal{S}' := \bigcup_{i=1}^{L} \mathcal{S}'_i$, and it follows that at most $L\binom{M}{K}2^K$ orthants are intersected. This means that at most $L\binom{M}{K}2^K$ effective quantization points can be used, i.e., at most $K \log_2(eLM/K)$ information bits can be obtained. □

Since K-sparse signals in any basis $\Psi \in \mathbb{R}^{N \times N}$ belong to a union of at most $\binom{N}{K}$ subspaces in $\mathbb{R}^N$, using Lemma 5 we can obtain the following corollary.

Corollary 1. *Let $x = \Psi\alpha \in \mathbb{R}^N$ be K-sparse in a certain basis $\Psi \in \mathbb{R}^{N \times N}$, i.e., $\alpha \in \Sigma_K$. Then the measurements $y_s = A(x)$ can effectively use at most $\binom{N}{K}\binom{M}{K}2^K$ 1-bit quantization points, i.e, carry at most $2K \log_2(e\sqrt{NM}/K)$ information bits.*

The set of signals of interest to be encoded is the set of unit-norm K-sparse signals Σ_K^*. Since unit-norm signals of a K-dimensional subspace form a K-dimensional unit sphere in that subspace,

Σ_K^* is a union of $\binom{N}{K}$ such unit spheres. The $Q = \binom{N}{K}\binom{M}{K}2^K$ available quantization points partition Σ_K^* into Q smaller sets, each of which contains all the signals that quantize the same point.

To develop the lower bound on the reconstruction error we examine the optimal such partition, with respect to the worst-case error, given the number of quantization points used. The measurement and reconstruction process maps each signal in Σ_K^* to a finite set of quantized signals $\mathcal{Q} \subset \Sigma_K^*, |\mathcal{Q}| = Q$. At best this map ensures that the worst case reconstruction error is minimized, i.e.,

$$\epsilon_{\text{opt}} \;=\; \max_{x\in\Sigma_K^*}\min_{q\in\mathcal{Q}}\|x-q\|_2, \tag{3.3}$$

where ϵ_{opt} denotes the worst-case quantization error and q each of the available quantization points. The optimal lower bound is achieved by designing $\mathcal{Q}$ to minimize (3.3) without considering whether the measurement and reconstruction process actually achieve this design. Thus, designing the set $\mathcal{Q}$ becomes a set covering problem.

Using this intuition and Lemma 5, Appendix B.2 proves the following statement about a set of unit-norm signals in a union of L, K-dimensional subspaces, specifically $x \in \Sigma_K^*$.

Theorem 4. *Let the mapping $A : \mathbb{R}^N \to \mathcal{B}^M$ and measurements y_s be defined as in (3.1) and let $x \in \Sigma_K^*$. Then the estimate from the reconstruction algorithm $\Delta^{1\text{bit}}(y_s, \Phi, K)$ has error defined by (3.3) of at least*

$$\epsilon_{\text{opt}} \geq \frac{K}{2eM} = \Omega\left(\frac{K}{M}\right).$$

Thus, the worst-case error cannot decay at a rate faster than $\Omega(1/M)$ as a function of the number measurements, no matter what reconstruction algorithm is used. The bound in the theorem is independent of L, but similarly to the relation between Lemma 5 and Corollary 1, K-sparse signals are a special case with $L = \binom{N}{K}$.

This result assumes noiseless acquisition and provides no guarantees of robustness and noise resiliency. This is in line with existing results on scalar quantization in oversampled representations and CS that state that the distortion due to scalar quantization of noiseless measurements cannot decrease faster than the inverse of the measurement rate [77–81]. To improve the rate vs. distortion trade-off, alternative quantization methods must be used, such as Sigma-Delta quantization [82–88] or non-monotonic scalar quantization [89].

Theorem 4 bounds the best possible performance of a consistent reconstruction over all possible mappings A. However, it is straightforward to construct mappings A that do not behave as the lower bound suggests. In the next section we identify one class of matrices such that the mapping A admits an almost optimal upper bound on the reconstruction error from a general algorithm $\Delta^{1\text{bit}}$.

3.4.2 Achievable performance via random projections

In this section we describe a class of matrices Φ such that the consistent sparse reconstruction algorithm $\Delta^{1\text{bit}}(y_s, \Phi, K)$ can indeed achieve error decay rates of optimal order, described by Theorem 4, with the number of measurements growing linear in the sparsity K and logarithmically in the dimension N, as is required in conventional CS. We first focus our analysis on Gaussian matrices, i.e., Φ such that each element $\phi_{i,j}$ is randomly drawn i.i.d. from the standard Gaussian distribution, $\mathcal{N}(0,1)$. We use the short notation $\Phi \sim \mathcal{N}^{M\times N}(0,1)$ for characterizing such matrices, and we write $\varphi \sim \mathcal{N}^{N\times 1}(0,1)$ for describing equivalent random vectors in $\mathbb{R}^N$ (e.g., the rows of Φ). For these matrices Φ, we prove the following in Appendix B.3.

Theorem 5. *Let Φ be matrix generated as $\Phi \sim \mathcal{N}^{M\times N}(0,1)$, and let the mapping $A : \mathbb{R}^N \to \mathcal{B}^M$ be defined as in (3.1). Fix $0 \le \eta \le 1$ and $\epsilon_o > 0$. If the number of measurements is*

$$M \ge \tfrac{1}{\epsilon_o}\left(2K\,\log(N) + 4K\log(\tfrac{16}{\epsilon_o}) + \log\tfrac{1}{\eta}\right), \tag{3.4}$$

then for all $x, s \in \Sigma^*_K$ we have that

$$\|x-s\|_2 > \epsilon_o \;\Rightarrow\; A(x) \ne A(s), \tag{3.5}$$

or equivalently

$$A(x) = A(s) \;\Rightarrow\; \|x-s\|_2 \le \epsilon_o,$$

with probability higher than $1-\eta$.

The Theorem demonstrates that if we use Gaussian matrices in the mapping A, then, given a fixed probability level η, the reconstruction algorithm $\Delta^{1\mathrm{bit}}(y_s, \Phi, K)$ will recover signals with optimal error order

$$\epsilon_o \;=\; O\Big(\big(\tfrac{K}{M}\big)^{1-\alpha}\log N\Big),$$

for arbitrarily small $\alpha > 0$; the presence of the $\log(1/\epsilon_0)$ term in (3.4) prevents us from setting $\alpha = 0$.

A similar result has been very recently shown for sign measurements of non-sparse signals in the context of quantization using frame permutations [90]. Specifically, it has been shown that reconstruction from sign measurements of signals can be achieved (almost surely) with a $O((1/M)^{1-\alpha})$ error rate decay for arbitrarily small $\alpha > 0$. Our main contribution here is extending this result to K-sparse vectors in $\mathbb{R}^N$. Our results, in addition to introducing the almost linear dependence on K, also show that if the signal is sparse then we pay a logarithmic penalty in N. This is consistent with results in CS, but seems not to be necessary from the lower bound in the previous section. We will see in Section 4.2 that for Gaussian matrices, the optimal error behavior is empirically exhibited on average. Finally, we note that for a constant ϵ_0, the number of measurements required to guarantee (3.5) is $M = O(K\log N/K)$, nearly the same as order in conventional CS.

We note a few minor extensions of the Theorem. We can multiply the rows of Φ with a positive scalar without changing the signs of the measurements. By normalizing the rows of the Gaussian matrix, we obtain a matrix with rows drawn uniformly from the unit ℓ_2 sphere in $\mathbb{R}^N$. It is thus straightforward to extend the Theorem to such matrices with such rows as well. Furthermore, note that these projections are “universal,” meaning that the theorem remains valid for sparse signals in Ψ, i.e., for x, s belonging to $\Sigma^*_{\Psi,K} := \{u = \Psi\alpha \in \mathbb{R}^N : \alpha \in \Sigma^*_K\}$. This is true since for any orthonormal basis $\Psi \in \mathbb{R}^{N\times N}$, $\Phi' = \Phi\Psi \sim \mathcal{N}^{M\times N}(0,1)$ when $\Phi \sim \mathcal{N}^{M\times N}(0,1)$.

We can also view the binary measurements as a *hash* or a *sketch* of the signal. With this interpretation of the result we guarantee with high probability that no sparse vectors with Euclidean distance greater than ϵ_o will "hash" to the same binary measurements. In fact, similar results play a key role in *locality sensitive hashing* (LSH), a technique that aims to efficiently perform approximate nearest neighbors searches from quantized projections [91–94]. Most LSH results examine the performance on point-clouds of a discrete number of signals instead of the infinite subspaces that we explore in this chapter. Furthermore, the primary goal of the LSH is to preserve the structure of the nearest neighbors with high probability. Instead, in this chapter we are concerned with the ability to reconstruct the signal from the hash, as well as the robustness of this reconstruction to measurement noise and signal model mismatch. To enable these properties, we require a property of the mapping A that preserves the structure (geometry) of the entire signal set. Thus, in the next section we seek an embedding property of A that preserves geometry for the set of sparse signals and thus ensures robust reconstruction.

3.5 Robust 1-bit CS via Binary Stable Embeddings

3.5.1 Binary ϵ-stable embeddings

In this section we establish an embedding property for the 1-bit CS mapping A that ensures that the sparse signal geometry is preserved in the measurements, analogous to the RIP for real-valued measurements. This robustness property enables us to upper bound the reconstruction performance even when some measurement signs have been changed due to noise. Conventional CS achieves robustness via the δ-stable embeddings of sparse vectors (1.4) discussed in Section 1.2. This embedding is a restricted *quasi-isometry* between the metric spaces $(\mathbb{R}^N, d_X)$ and $(\mathbb{R}^M, d_Y)$, where the distance metrics d_X and d_Y are the ℓ_2-norm in dimensions N and M, respectively, and the domain is restricted to sparse signals.[4] We seek a similar definition for our embedding; however, now the signals and measurements lie in the different spaces S^{N-1} and $\mathcal{B}^M$, respectively. Thus, we first consider appropriate distance metrics in these spaces.

The Hamming distance is the natural distance for counting the number of unequal bits between two measurement vectors. Specifically, for $y, v \in \mathcal{B}^M$ we define the *normalized Hamming distance* as

$$d_H(y,v) = \frac{1}{M}\sum_{i=1}^{M} y_i \oplus v_i,$$

where $\oplus$ is the XOR operation such that $a \oplus b$ equals 0 if $a = b$ and 1 otherwise. The distance is normalized such that $d_H \in [0,1]$. In the signal space we only consider unit-norm vectors, thus, a natural distance is the angle formed by any two of these vectors. Specifically, for $x, s \in S^{N-1}$, we consider

$$d_S(x,s) := \frac{1}{\pi}\arccos\langle x, s\rangle.$$

[4] A function $A : X \to Y$ is called a *quasi-isometry* between metric spaces (X, d_X) and (Y, d_Y) if there exists $C > 0$ and $D \geq 0$ such that $\frac{1}{C} d_X(x,s) - D \leq d_Y(A(x), A(s)) \leq C d_X(x,s) + D$ for $x, s \in X$, and $E > 0$ such that $d_Y(y, A(x)) < E$ for all $y \in Y$ [95]. Since $D = 0$ for δ-stable embeddings, they are also called bi-Lipschitz mappings.

As with the Hamming distance, we normalize the true angle $\arccos\langle x, y\rangle$ such that $d_S \in [0, 1]$. Note that since both vectors have the same norm, the inner product $\langle x, s\rangle$ can easily be mapped to the ℓ_2-distance using the polarization identity.

Using these distance metrics we define the binary stable embedding.

Definition 2 (Binary ϵ-Stable Embedding)**.** *Let $\epsilon \in (0, 1)$. A mapping $A : \mathbb{R}^N \to \mathcal{B}^M$ is a **binary ϵ-stable embedding** (BϵSE) of order K for sparse vectors if*

$$d_S(x, s) - \epsilon \leq d_H(A(x), A(s)) \leq d_S(x, s) + \epsilon$$

for all $x, s \in S^{N-1}$ with $x \pm s \in \Sigma_K$.

Our definition describes a specific quasi-isometry between the two metric spaces (S^{N-1}, d_S) and $(\mathcal{B}^M, d_H)$, restricted to sparse vectors. While this mirrors the form of the δ-stable embedding for sparse vectors, one important difference is that the sensitivity term ϵ is additive, rather than multiplicative, and thus the BϵSE is not bi-Lipschitz. This is a necessary side-effect of the loss of information due to quantization.

A stated in the next Lemma, the BϵSE enables robustness guarantees on any reconstruction algorithm extracting a sparse signal $\widehat{x}$ from the mapping $A(x)$.

Lemma 6. *Let $A : \mathbb{R}^N \to \mathcal{B}^M$ be a BϵSE of order $2K$ for sparse vectors and let $x \in \Sigma_K^*$. A sparse, unit norm estimate $\widehat{x}$ of x with Hamming error $d_H\big(A(x), A(\widehat{x})\big)$ from any reconstruction algorithm has angular error bounded by*

$$d_S(x, \widehat{x}) \ \leq d_H\big(A(x), A(\widehat{x})\big) + \epsilon.$$

Proof. If $\widehat{x}$ is K-sparse ($\|\widehat{x}\|_0 \leq K$) and unit norm, then the result follows from the lower bound in Definition 2. □

In other words, the reconstruction error is bounded by a small quantity more than the Hamming error. Thus, if an algorithm returns a unit norm sparse solution with measurements that are not consistent (i.e., $d_H(A(x), A(\widehat{x})) > 0$), as is the case with several algorithms [39, 75, 76], then the the worst-case angular reconstruction error is close to Hamming distance between the estimate's measurements' signs and the original measurements' signs. Section 4.2 verifies this behavior with simulation results. Furthermore, in Section 3.5.3 we use the BϵSE property to guarantee that if measurements are corrupted by noise or if signals are not exactly sparse, then the reconstruction error is bounded.

Note that if A is a BϵSE, then the angular error of any $\Delta^{\text{1bit}}(y_s, \Phi, K)$ algorithm is bounded by ϵ since in that case $d_H\big(A(x), A(\widehat{x})\big) = 0$. As we have seen earlier this is to be expected because, unlike conventional noiseless CS, quantization fundamentally introduces uncertainty and exact recovery cannot be guaranteed. This is an obvious consequence of the mapping of the infinite set Σ_K^* to a discrete set of quantized values.

We next identify a class of matrices Φ for which A is a BϵSE.

3.5.2 Binary ϵ-stable embeddings via random projections

As is the case for conventional CS systems with RIP, designing a Φ for 1-bit CS such that A has has the BϵSE property is a computationally intractable task. Fortunately, an overwhleming number of "good" matrices do exist. Specifically we again focus our analysis on Gaussian matrices, i.e., $\Phi \sim \mathcal{N}^{M\times N}(0,1)$ such that each element $\phi_{i,j}$ is randomly drawn i.i.d. from $\mathcal{N}(0,1)$, as in as in Section 3.4.2. As motivation that this choice of Φ will indeed enable robustness, we begin with a classical concentration of measure result for binary measurements from a Gaussian matrix.

Lemma 7. *Let Φ be a matrix generated as $\Phi \sim \mathcal{N}^{M\times N}(0,1)$, and let the mapping $A : \mathbb{R}^N \to \mathcal{B}^M$ be defined as in (3.1). Fix $\epsilon > 0$. For any $x, s \in S^{N-1}$, we have*

$$\mathbb{P}\Big(\Big| d_H\big(A(x), A(s)\big) \;-\; d_S(x,s) \Big| \;\le\; \epsilon \Big) \;\ge\; 1 - 2\,e^{-2\epsilon^2 M}, \tag{3.6}$$

where the probability is with respect to the generation of Φ.

Proof. This lemma is a simple consequence of Lemma 3.2 in [96] which shows that, for one measurement, $\mathbb{P}[A_j(x) \neq A_j(s)] = d_S(x,s)$. The result then follows by applying Hoeffding's inequality to the binomial random variable $M d_H\big(A(x), A(s)\big)$ with M trials. □

In words, Lemma 7 implies that the Hamming distance between two binary measurement vectors $A(x), A(s)$ tends to the angle between the signals x and s as the number of measurements M increases. In [96] this fact is used in the context of randomized rounding for max-cut problems; however, this property has also been used in similar contexts as ours with regards to preservation of inner products from binary measurements [97, 98].

The expression (3.6) indeed looks similar to the definition of the BϵSE, however, it only holds for a fixed pair of arbitrary (not necessarily sparse) signals, chosen prior to drawing Φ. Our goal is to extend (3.6) to cover the entire set of sparse signals. Indeed, concentration results similar to Lemma 7, although expressed in terms of norms, have been used to demonstrate the RIP [45]. These techniques usually demonstrate that the cardinality of the space of all sparse signals is sufficiently small, such that the concentration result can be applied to demonstrate that distances are preserved with relatively few measurements.

Unfortunately, due to the non-linearity of A we cannot immediately apply Lemma 7 using the same procedure as in [45]. To briefly summarize, [45] proceeds by covering the set of all K-sparse signals Σ_K with a finite set of points (with covering radius $\delta > 0$). A concentration inequality is then applied to this set of points. Since any sparse signal lies in a δ-neighborhood of at least one such point, the concentration property can be extended from the finite set to Σ_K by bounding the distance between the measurements of the points within the δ-neighborhood. Such an approach cannot be used to extend (3.6) to Σ_K, because the severe discontinuity of our mapping does not permit us to characterize the measurements $A(x+s)$ using $A(x)$ and $A(s)$ and obtain a bound on the distance between measurements of signals in a δ-neighborhood.

To resolve this issue, we extend Lemma 7 to include all points within Euclidean balls around the vectors x and s inside the (sub) sphere $\Sigma^*(T) = \{u \in S^{N-1} : \operatorname{supp} u \subset T\}$ for some fixed support set $T \subset \{1, \cdots, N\}$ of size $|T| = D$. Define the δ-ball $B_\delta(x) := \{a \in S^{N-1} : \|x - a\|_2 < \delta\}$ to be the ball of Euclidean distance δ around x, and let $B^*_\delta(x) = B_\delta(x) \cap \Sigma^*(T)$.

Lemma 8. *Given $T \subset \{1, \cdots, N\}$ of size $|T| = D$, let Φ be a matrix generated as $\Phi \sim \mathcal{N}^{M\times N}(0,1)$, and let the mapping $A : \mathbb{R}^N \to \mathcal{B}^M$ be defined as in (3.1). Fix $\epsilon > 0$ and $0 \leq \delta \leq 1$. For any $x, s \in \Sigma^*(T)$, we have*

$$\mathbb{P}\Big(\Big| d_H\big(A(u), A(v)\big) \; - \; d_S(x,s) \Big| \; \leq \; \epsilon + \sqrt{\tfrac{\pi}{2} D}\,\delta \Big) \; \geq \; 1 - 2\, e^{-2\epsilon^2 M}$$

for all $u \in B^*_\delta(x)$ and $v \in B^*_\delta(s)$.

The proof of this result is given in Appendix B.4.

In words, if the width δ is sufficiently small, then the Hamming distance between the 1-bit measurements $A(u)$, $A(v)$ of any points u, v within the balls $B^*_\delta(x)$, $B^*_\delta(s)$, respectively, will be close to the angle between the centers of the balls.

Lemma 8 is key for providing a similar argument to that in [45]. We now simply need to count the number of pairs of K-sparse signals that are euclidean distance δ apart. The Lemma can then be invoked to demonstrate that the angles between all of these pairs will be approximately preserved by our mapping.[5] Thus, with Lemma 8 under our belt, we demonstrate in Appendix B.5 the following result.

Theorem 6. *Let Φ be a matrix generated as $\Phi \sim \mathcal{N}^{M\times N}(0,1)$ and let the mapping $A : \mathbb{R}^N \to \mathcal{B}^M$ be defined as in (3.1). Fix $0 \leq \eta \leq 1$ and $\epsilon > 0$. If the number of measurements is*

$$M \; \geq \; \tfrac{4}{\epsilon^2}\Big(K \, \log(N) + 2K \, \log(\tfrac{50}{\epsilon}) + \log(\tfrac{2}{\eta})\Big), \tag{3.7}$$

then with probability exceeding $1 - \eta$, the mapping A is a BϵSE of order K for sparse vectors.

By choosing $\Phi \sim \mathcal{N}^{M\times N}(0,1)$ with $M = O(K \log N)$, with high probability we ensure that the mapping A is a BϵSE. Additionally, from (3.7) we find that the error decreases as

$$\epsilon \; = \; O\Big((K/M)^{(1-\alpha)/2}\,(\log N)^{1/2}\Big), \tag{3.8}$$

for arbitrarily small $\alpha > 0$. Unfortunately, this decay is at a slower rate (roughly by a factor of $\sqrt{K/M}$) than the lower bound on the error given in Section 3.4.1. This error rate results from an application of the Chernoff-Hoeffding inequality in the proof of Theorem 6. An open question is whether it is possible to obtain a tighter bound (with optimal error rate) for this robustness property.

As mentioned in Section 3.3, it may be advantageous to reconstruct a signal from the signs of the differences of the measurements. As suggested by (3.2), in this case we interested in applying the sparse consistent reconstruction algorithm to the measurement matrix $D\Phi$, where D is a difference matrix and Φ is the original measurement matrix. When Φ is a Gaussian matrix, this is indeed possible with the number of measurements on the same order as before, as explained by the following Corollary.

[5] We note that the covering argument in the proof of Theorem 5 also employs δ-balls in similar fashion but only considers the probability that $d_H = 0$, rather than the concentration inequality.

Corollary 2. *Let Φ be a matrix generated as $\Phi \sim \mathcal{N}^{M \times N}(0,1)$, let D be an $M-1 \times M$ difference matrix, and let the mapping $A : \mathbb{R}^N \to \mathcal{B}^M$ be defined as in (3.1) with the matrix $D\Phi$ instead of Φ. Fix $0 \leq \eta \leq 1$ and $\epsilon > 0$. If the number of measurements is*

$$M \ \geq \ \tfrac{8}{\epsilon^2}\Big(K \ \log(N) + 2K \ \log(\tfrac{50}{\epsilon}) + \log(\tfrac{2}{\eta})\Big), \tag{3.9}$$

then with probability exceeding $1-\eta$, the mapping A is a BϵSE of order K for sparse vectors.

Proof. Let the $(M-1)/2 \times M$ matrix E be obtained by selecting every other row of the matrix D. Then the matrix $E\Phi$ has i.i.d. Gaussian entries, since it is obtained by summing disjoint sets of independent Gaussian entries in Φ. Note that the entries of $E\Phi$ will no longer have unit variance but are still zero mean, i.e., they are just scaled Gaussians. As previously discussed, scaling the entries of Φ has no effect on BϵSE property (or its probability of occurance). Thus, if given the signs of the measurements from $D\Phi$, we can perform reconstruction with a subset of measurements and $E\Phi$. To obtain the final result, we note that we have half as many valid measurements as in Theorem 6. □

Note that the only difference between (3.9) and (3.7) is that the minimum number of required measurements is now double of what was required in Theorem 6, and thus is on the same order as in (3.8). This is because there are half as many independent measurements in this case.

Besides robustness to non-linearities as discussed in Section 3.3, this technique can also be used for 1-bit quantization of measurements that are all positive, such as those acquired by the single-pixel-camera [19].

As with Theorem 5, Gaussian matrices provide a universal mapping, i.e., the result remains valid for sparse signals in a basis $\Psi \in \mathbb{R}^{N \times N}$. Moreover, Theorem 6 can also be extended to rows of Φ that are drawn uniformly on the sphere, since the rows of Φ in Theorem 6 can be normalized without affecting the outcome of the proof. Note that by normalizing the Gaussian rows of Φ, is is as if they had been drawn from a uniform distribution of unit-norm signals.

We have now established a large class of robust BϵSEs: 1-bit quantized Gaussian projections. We now make use of this robustness by considering an example where the measurements are corrupted by Gaussian noise.

3.5.3 Noisy measurements and compressible signals

In practice, hardware systems may be inaccurate when taking measurements; this is often modeled by additive noise. The mapping A is robust to noise in an unusual way. After quantization, the measurements can only take the values -1 or 1. Thus, we can analyze the reconstruction performance from corrupted measurements by considering how many measurements flip their signs. For example, we analyze the specific case of Gaussian noise on the measurements prior to quantization, i.e.,

$$A_n(x) := \text{sign}\,(\Phi x + n), \tag{3.10}$$

where $n \in \mathbb{R}^M$ has i.i.d. elements $n_i \sim \mathcal{N}(0, \sigma^2)$. In this case, we demonstrate, via the following lemma, a bound on the Hamming distance between the corrupted and ideal measurements with the BϵSE from Theorem 6 (see Appendix B.6).

Lemma 9. *Let Φ be a matrix generated as $\Phi \sim \mathcal{N}^{M\times N}(0,1)$, let the mapping $A : \mathbb{R}^N \to \mathcal{B}^M$ be defined as in (3.1), and let $A_n : \mathbb{R}^N \to \mathcal{B}^M$ be defined as in (3.10). Let $n \in \mathbb{R}^M$ be a Gaussian random vector with* i.i.d. *components $n_i \sim \mathcal{N}(0, \sigma^2)$. Fix $\gamma > 0$. Then for any $x \in \mathbb{R}^N$, we have*

$$\begin{aligned} \mathbb{E}\left(d_H\big(A_n(x), A(x)\big) \right) \ &\leq \ e(\sigma, \|x\|_2), \\ \mathbb{P}\left(d_H\big(A_n(x), A(x)\big) > \ e(\sigma, \|x\|_2) + \gamma \right) &\leq \ e^{-2M\gamma^2}, \end{aligned}$$

where $e(\sigma, \|x\|_2) = \frac{1}{2}\frac{\sigma}{\sqrt{\|x\|_2^2+\sigma^2}} \leq \frac{1}{2}\frac{\sigma}{\|x\|_2}$.

If $\widehat{x_n}$ is the estimate from a sparse consistent reconstruction algorithm $\Delta^{1\text{bit}}(A_n(x), \Phi, K)$ from the measurements $A_n(x)$, then it immediately follows from Lemma 9 and Theorem 6 that

$$d_S(\widehat{x_n}, x) \ \leq \ d_H\big(A_n(x), A(x)\big) + \epsilon \ \leq \ \tfrac{1}{2}\tfrac{\sigma}{\|x\|_2} + \gamma + \epsilon, \tag{3.11}$$

with high probability (depending on M and γ). Given alternative noise distributions, e.g., Poisson noise, a similar analysis can be carried out to determine the likely number of sign flips and thus provide a bound on the error due to noise.

Another practical consideration is that real signals are not always strictly K-sparse. Indeed, it may be the case that signals are *compressible*; i.e., they can be closely approximated by a K-sparse signal. Lemma 9 can be extended to compressible signals. To do this, we consider the small coefficients, i.e., the "tail" of the residual of a best K-term approximation of x, to be a source of Gaussian noise on the measurements and then apply Lemma 9. This is possible due to our particular Gaussian choice of Φ and the fact that for binary measurements, we are only concerned with the number of measurements that change sign.

Corollary 3. *Let Φ be a matrix generated as $\Phi \sim \mathcal{N}^{M\times N}(0,1)$ and let the mapping $A : \mathbb{R}^N \to \mathcal{B}^M$ be defined as in (3.1). Furthermore, let Φ have RIP constant δ_K. Let $\gamma > 0$. Then for any $x \in \mathbb{S}^{N-1}$ we have*

$$\begin{aligned} \mathbb{E}\left(d_H\big(A(x), A(x_K)\big) \right) \ &\leq \ \tfrac{1}{2}\tfrac{\|x - x_K\|_2}{\|x\|_2}, \\ \mathbb{P}\left(d_H\big(A(x), A(x_K)\big) > \ \tfrac{1}{2}\tfrac{\|x - x_K\|_2}{\|x\|_2} + \gamma \right) &\leq \ e^{-2M\gamma^2}, \end{aligned}$$

where x_K is the best K-term approximation of x.

The proof is given in Appendix B.7. In similar fashion to (3.11), this result implies that with high probability (depending on M and γ), the angular reconstruction error of $\widehat{x} = \Delta^{1\text{bit}}(A(x), \Phi, K)$ for any signal x (sparse or compressible) is bounded as

$$d_S(\widehat{x}, x) \leq \frac{1}{2}\frac{\|x - x_K\|_2}{\|x\|_2} + \gamma + \epsilon,$$

Much like conventional CS results, the reconstruction error on the order of the best K-term approximation error of the signal.

Thus far we have demonstrated a lower bound on the reconstruction error from 1-bit measurements (Theorem 5) and introduced a condition on the mapping A that enables stable reconstruction in noiseless, noisy, and compressible settings (Definition 2). We have furthermore demonstrated that a large class of random matrices—specifically matrices with coefficients draw from a Gaussian distribution and matrices with rows drawn uniformly from the unit sphere—provide good mappings (Theorem 6). We now demonstrate how the above ideas can be extended to perform saturation-agnostic (and multi-bit) reconstruction.

3.6 Saturation-Agnostic Sensing

It is possible to use the sparse consistent reconstruction algorithm $\Delta^{\text{1bit}}(y_s, \Phi, K)$ to recover measurements that have been quantized at bit depths higher than one and with arbitrary numbers of saturation events. To do this, we extend the idea that signals can be recovered from the signs of the pair-wise differences of the measurements as in Corollary 2. However, instead of considering only the relationship between any two consecutive pairs, we consider the unique relationships between *all* pairs of measurements. We can represent this by an overdetermined difference matrix $D_M \in \{-1, 0, 1\}^{\binom{M}{2} \times M}$. For example, for $M = 4$ we would have the 6×4 matrix

$$D_M = \begin{bmatrix} 1 & \text{-1} & 0 & 0 \\ 1 & 0 & \text{-1} & 0 \\ 1 & 0 & 0 & \text{-1} \\ 0 & 1 & \text{-1} & 0 \\ 0 & 1 & 0 & \text{-1} \\ 0 & 0 & 1 & \text{-1} \end{bmatrix}, \tag{3.12}$$

and the measurements $y = [1, -4, 3, 6]^T$ would quantize to $[1, -1, -1, -1, -1, -1]^T = \text{sign}(D_M y)$. Thus, we can perform the following procedure:

1. Acquire real-valued measurements;
2. Quantize the measurements (this may induce an unknown amount of saturation);
3. Apply D_M to the quantized, saturated measurements and compute the resulting signs.

We can then perform reconstruction via $\Delta^{\text{1bit}}(\text{sign}(D_M y), D_M \Phi, K)$. A similar idea has been proposed for quantization of frame coefficients in a non-CS context [99]. This can also be thought of as a specific application of some of the ideas presented in [33].

The key benefits of this technique are that it *i*) provides a simple way to perform consistent[6] reconstruction from multi-bit quantized measurements; and *ii*) is agnostic to the number of saturations. Indeed, when all measurements saturate, this technique reduces to the signed differences

[6] In this case consistency is defined in terms of the signs of the differences of the measurements, not the absolute quantization intervals in which the measurements lie.

reconstruction problem with guarantees given in Corollary 2 and problem formulation given by (3.2), i.e., for all practical purposes it is equivalent to the 1-bit CS case. It is thus expected that we can maintain robust reconstruction performance regardless of how many measurements saturate. We will see empirical validation of this idea in Section 4.2. This may be useful in situations where the saturation rate may be hard to control or the gains of the input signals are unpredictable.

3.7 1-bit CS Reconstruction Algorithms

3.7.1 Trust, but Verify: Restricted-step shrinkage (RSS)

Background on trust-region algorithms

One approach to solving optimization problems like (ℓ_1-min$_{1\mathrm{B}}$) and (3.2) is to adapt standard CS optimization algorithms to seek a solution on the sphere. However, since these algorithms are intended to solve convex problems and the sphere constraint is non-convex, computational performance may suffer. In particular, the choice of an appropriate step-size is elusive. Common methods for choosing adaptive step-sizes, such as Barzilai-Borwein (BB) steps, do not necessarily perform well with a unit sphere constraint, since they were designed for unconstrained convex optimization [100]. In addition, to enforce the sphere constraint, many approaches must introduce an additional step that renormalizes intermediate solutions. It is not obvious that such approaches will converge.

The methods used in this section are inspired by a particular class of restricted step-size algorithms called *trust-region* methods [101]. Given the unconstrained nonlinear programming problem

$$\min_{x \in \mathbb{R}^N} f(x), \tag{3.13}$$

trust-region methods compute the next trial point iteratively by finding the minimizer of the approximation $m_s(x)$ of $f(x)$ within a trust-region defined by a ball centered at the current point x^s with radius Δ^s;[7] that is,

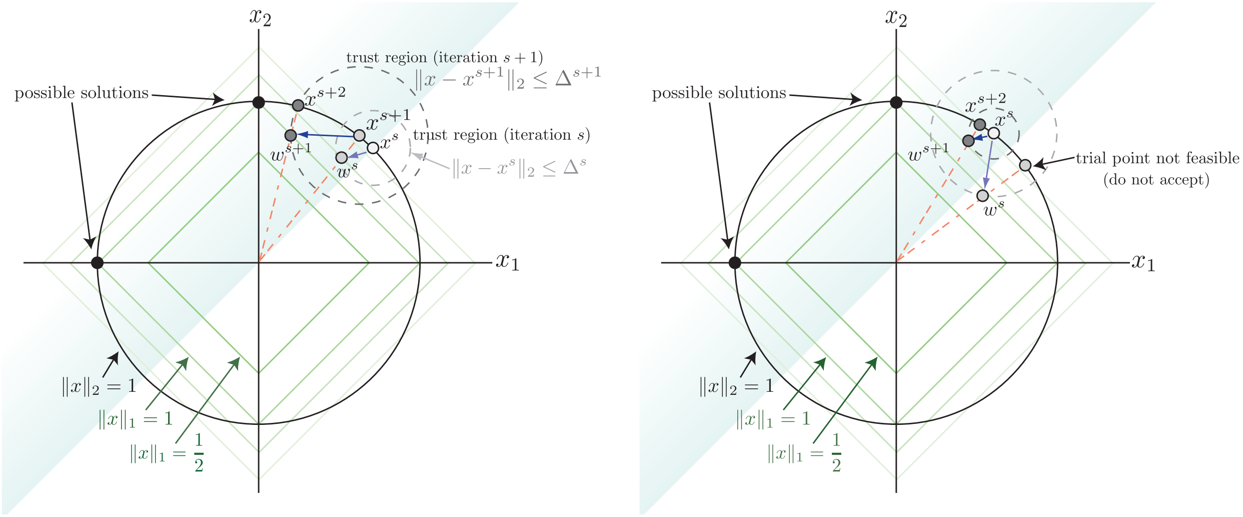
$$\min_{x \in \mathbb{R}^N} m_s(x) \quad \text{s.t.} \quad \|x - x^s\|_2 \leq \Delta^s. \tag{3.14}$$

The size of the trust-region Δ^s is increased or decreased automatically according to the performance of the model (3.14) during previous iterations. These methods choose step directions and lengths simultaneously, and they have been proven to be reliable for solving difficult non-convex problems [101]. Additionally, these algorithms often have provable convergence guarantees. These algorithms can also be used for constrained optimization, for example, by linearizing the constraints and applying a conventional constrained optimization technique. For more details on trust region methods and their adaptation for constrained optimization, we refer the reader to [101, 102].

To motivate the use of trust-region methods in 1-bit CS, consider the following simple example program:

$$\min_{x \in S^1} \|x\|_1 \quad \text{s.t.} \quad x_1 \leq x_2 \quad \text{and} \quad \|x\|_2 = 1. \tag{3.15}$$

[7] In this section Δ^s refers to the trust region radius, not the quantization width or 1-bit reconstruction algorithm.



(a) successful iteration example

(b) failure iteration example

Figure 3.3: Example scenarios of trust region algorithm iterations to solve (3.15). The goal is find an x with the minimum ℓ_1-norm such that x has unit ℓ_2-norm and $x_1 \leq x_2$. The shaded region denotes the feasible constraint region and the iteration number is denoted by s, with initial point x^s. The light dashed circle denotes trust region at iteration s and the dark dashed circle denotes trust region at iteration $s+1$. w^s and w^{s+1} denote steps taken before projecting onto the unit circle. (a) During a successful iteration, the trial point falls within the feasible region and thus is accepted, denoted by x^{s+1}, and the radius of the trust region is enlarged. (b) During a failure iteration, the trust region radius is too large and the trial point falls outside the feasible region. In this case, the trust region radius is reduced and a new trial step is taken from the initial point x^s.

The behavior of the method can be best explained by examining both a successful iteration and a failure iteration of the algorithm applied to (3.15). Examples of these cases are depicted in Figure 3.3. The first constraint is depicted by the shaded area. The initial point is denoted by x^s, where s is the iteration number. The algorithm will take a step in a direction specified by an approximation $m_s(x)$ (not depicted) to point w and then project the result onto the unit sphere. The light dashed sphere depicts the trust region at iteration s while the dark dashed sphere depicts the trust region at iteration $s+1$. Depending on the success of the trial point, the trust region will expand or contract.

During a successful iteration, as depicted in Figure 3.3(a), the algorithm takes a step to point w^s and projects the point onto the sphere. This is depicted by the red dashed line. Since the result is within the feasible region, the point is accepted and denoted by x^{s+1}. The trust region radius is expanded and the procedure repeats. During a failure iteration, as depicted in Figure 3.3(b), the trust region radius is too large and the trial point on the circle is not within the feasible region. Thus, we do not accept this trial point and take a new step from the initial point x^s, this time with a smaller trust region radius. In this example, the new step results in a feasible point.

The program (3.14) is generally not solvable in closed form. This includes the case studied in this section where $f(x)$ is the ℓ_1-norm. However, by relaxing the problem, a closed form optimal solution can often be obtained, resulting in lower cost computation at each iteration. In this section, rather than solving (3.14), we iteratively solve a sequence of problems of the form

$$\min_{x\in S^{N-1}} m_s(x) + \frac{\lambda^s}{2}\|x - x^s\|_2^2, \tag{3.16}$$

where the parameter λ^s essentially plays a role like the trust-region radius Δ^s in model (3.14). In fact, the solutions of (3.14) and (3.16) are the same under some properly chosen λ^s and Δ^s. We will show that our adaptation of this algorithm indeed also has guaranteed convergence, as with conventional trust region algorithms.

The restricted step shrinkage algorithm for 1-bit CS

In this section, we derive an algorithm for the generalized formulation of (ℓ_1-$\min_{1\mathrm{B}}$) and (3.2)

$$\min_{x\in S^{N-1}} \|x\|_1 \text{ s.t. } Ax \geq b \text{ and } \|x\|_2 = 1. \tag{3.17}$$

Our strategy is as follows. First, using the augmented Lagrangian framework, we formulate an algorithm that solves (3.17) and denote it as RSS-outer. We choose the augmented Lagrangian framework since many state-of-the-art CS reconstruction algorithms are formulated this way [26, 103, 104]. Second, a step within RSS-outer requires that we solve a non-convex subproblem of the form

$$\min_{x\in S^{N-1}} \zeta_\mu(x) = \|x\|_1 + \mu f(x) \text{ s.t. } \|x\|_2 = 1, \tag{3.18}$$

where $f(x) : \mathbb{R}^N \to \mathbb{R}$ is differentiable and $\mu > 0$. We solve (3.18) with a trust-region-like subroutine, denoted as RSS-inner. The total procedure obtained by combining RSS-outer and RSS-inner is called the RSS algorithm.

The RSS-inner subroutine is the main contribution of this section. Thus, we choose to describe RSS-inner in terms of the general program (3.18). Algorithm frameworks other than the augmented Lagrangian can be used to formulate an algorithm for (3.17), and in some cases may employ the RSS-inner subroutine. As an example, the quadratic penalty formulation to this problem is given in Appendix C.1. This formulation is simpler to implement, but does not perform as fast in practice.

For the remainder of this section, we will use the following terms. A *stationary point* of an optimization problem is a point that satisfies the Karush-Kuhn-Tucker (KKT) first-order optimality conditions [102]. By *convergence* we mean that an algorithm converges to a stationary point of the objective from any starting point, but not necessarily to a global minimizer of the objective. We say a point x is a *cluster point* of sequence $\{x_s\}_{s\in\mathbb{N}}$ if for any $\epsilon > 0$ there exist an infinite number of points of $\{x_s\}$ lying in the ϵ-ball of x. Note that the sequence $\{x_s\}$ may not converge. A *feasible solution* is a solution such that all constraints are satisfied. The *subgradient* ∂f of function $f(x)$ at point x_0 is defined as any vector z such that

$$f(x) - f(x_0) \geq z(x - x_0). \tag{3.19}$$

Algorithm 3: RSS-outer

```
S0 Initialize
     Given initial solution $x^0$
     Choose initial step-size $\mu^0$ and $\kappa > 0$
     Set iteration $s := 0$, Lagrangian multiplier $\lambda^0 = 0$
   while not converged do
S1     Compute next estimate (via RSS-inner)
         Set $x^{s+1} = \min \mathcal{L}(x, \lambda^s, \mu^s)$ s.t. $\|x\|_2 = 1$,
         where the objective is given by (3.21).
S2     Update multiplier and $\mu$
         Set $\lambda^{s+1} := \max\{\lambda^s - \mu^s(Ax^{s+1} - b), 0\}$
         Set $\mu^{s+1} := \kappa\mu^s$
S3     Update iteration count
         Set $s := s + 1$
```

Augmented Lagrangian formulation of (3.17) **(RSS-outer)** We first formulate an algorithm to solve (3.17) using augmented Lagrangian framework. Starting from $\lambda^0 = 0$, at each iteration s we solve the Lagrangian function

$$\min_{x \in S^{N-1}} \mathcal{L}(x, \lambda^s, \mu^s) \text{ s.t. } \|x\|_2 = 1, \tag{3.20}$$

for x^{s+1}, where $\lambda \in \mathbb{R}^M$ and $\mu > 0$. We then set $\mu^{s+1} := \kappa\mu^s$, with $\kappa > 0$, and updates the Lagrangian multipliers λ^{s+1} according to

$$\lambda^{s+1} := \max\left\{\lambda^s - \mu^s(Ax^{s+1} - b), 0\right\}.$$

The augmented Lagrangian function for (3.17) is

$$\mathcal{L}(x, \lambda, \mu) := \|x\|_1 + \sum_{i=1}^{m} \rho((Ax - b)_i, \lambda_i, \mu), \tag{3.21}$$

where

$$\rho(t, \sigma, \mu) := \begin{cases} -\sigma t + \frac{1}{2}\mu t^2, & \text{if } t - \frac{\sigma}{\mu} \leq 0, \\ -\frac{1}{2\mu}\sigma^2, & \text{otherwise.} \end{cases} \tag{3.22}$$

Thus, the intermediate problem (3.20) is of the form of (3.18) and will be solved with RSS-inner. The complete augmented Lagrangian procedure, and how it relies on the RSS-inner subroutine is summarized in Algorithim 3.

Restricted-step subroutine to solve (3.18) **(RSS-inner)** The RSS-inner subroutine finds the solution to the subproblem (3.18) and proceeds as follows. We begin with an initial signal estimate x^0 and an initial step-size τ^0. At iteration s, from the point x^s, we compute a smooth approximation $m_s(x)$ to the original objective function $\zeta_\mu(x)$ in (3.18). The approximation is formed by adding the first-order Taylor expansion of $\mu f(x)$ and a proximal term with respect to x^s to the ℓ_1-norm of x

$$m_s(x) = \|x\|_1 + \mu f(x^s) + \mu \left(g^s\right)^\top (x - x^s) + \frac{\tau^s}{2}\|x - x^s\|_2^2,$$

where the step size $\tau^s > 0$ and g^s is the gradient of $f(x)$. Next, we find the optimal solution to the smoothed approximation

$$z^s := \arg\min_{x\in\mathbb{R}^n} m_s(x) \text{ s.t. } \|x\|_2 = 1. \tag{3.23}$$

The relationship between the optimal solution z^s of the subproblem (3.23) and its subgradient $\partial\|z^s\|_1$, together with the norm constraint, implies that z^s can be expressed explicitly. In fact, z^s can be expressed in terms of the shrinkage ("soft threshold") operator, defined for any $\alpha \in \mathbb{R}^N$, as

$$\mathcal{S}(\alpha, T) := \mathrm{sgn}(\alpha) \odot \max\left\{|\alpha| - T, 0\right\}, \tag{3.24}$$

where $\odot$ denotes the element-wise product between two vectors and $|\cdot|$ denotes the magnitude of each element in the vector. This is demonstrated in the following Lemma.

Lemma 10. *Suppose that x^s is not a stationary point of ζ_μ.*

1. *If $\mathcal{S}^s := \mathcal{S}\left(\tau^s x^s - \mu g^s, 1\right) \neq 0$, then the closed-form solution of the subproblem* (3.23) *is*

$$z^s = \frac{\mathcal{S}^s}{\|\mathcal{S}^s\|_2}. \tag{3.25}$$

2. *If $|\tau^s x_i^s - \mu g_i^s| < 1$ for $i = 1, \dots, n$, then $z_i^s = 0$ for all i except that $z_i^s = \mathrm{sgn}(\tau_i^s x^s - \mu g_i^s)$, where $i = \arg\max_{k=1,\dots,n} |\tau^s x_k^s - \mu g_k^s|$ (select only one i if there are multiple solutions).*

3. *Otherwise, the optimal Lagrangian multiplier λ with respect to $\|x\|_2 = 1$ satisfies $\tau^s - \lambda = 0$, $|\tau^s x^s - \mu g^s| \leq 1$, and the set $\{i \mid |\tau^s x_i^s - \mu g_i^s| = 1\}$ is not empty and the closed-form solutions of the subproblem* (3.23) *satisfy $\|z^s\|_2 = 1$ and*

$$\begin{cases} z_i^s \in (0, +\infty), & \textit{if } \tau^s x_i^s - \mu g_i^s = 1, \\ z_i^s \in (-\infty, 0), & \textit{if } \tau^s x_i^s - \mu g_i^s = -1, \\ z_i^s = 0, & \text{otherwise.} \end{cases} \tag{3.26}$$

The proof of this Lemma can be found in Appendix C.2. The Lemma implies that the next trial point z^s can be computed in closed form via the ratio (3.25).

We now present our strategy for choosing the step-size τ^s and updating the new iterate x^{s+1} from z^s. We first calculate the difference between the actual reduction of the objective function $\zeta_\mu(x)$ and predicted reduction

$$\delta(x^s, z^s) = \|x^s\|_1 - \|z^s\|_1 - \mu(g^s)^\top(z^s - x^s)$$

```
Algorithm 4: RSS-inner (subroutine)
S0 Initialize
     Given initial solution $x^0$ and initial step-size $\tau^0$
     Choose $0 < \eta_1 \leq \eta_2 < 1$ and $0 < \gamma_1 \leq \gamma_2 < 1 < \gamma_3$
     Set iteration $s := 0$
   while not converged do
S1     Compute step
         Compute a new trial point $z^s$ via (3.25)
         Compute the ratio $r_s$ via (3.27)
S2     Accept or reject the trial point
         If $r_s \geq \eta_1$, then set $x^{s+1} = z^s$
         otherwise, set $x^{s+1} = x^s$
S3     Adapt step-size
         Update $\tau^s$ according to (3.28)
S4     Update iteration count
         Set $s := s + 1$
```

and then compute the ratio

$$r_s = \frac{\zeta_\mu(x^s) - \zeta_\mu(z^s)}{\delta(x^s, z^s)} \tag{3.27}$$

to decide whether to accept the trial point z^s as well as if the step-size should be updated. Specifically, if $r_s \geq \eta_1 > 0$, then the iteration was successful and we set $x^{s+1} = z^s$; otherwise, the iteration was not successful and we set $x^{s+1} = x^s$. Finally, the step-size τ^s is updated as

$$\tau^{s+1} \in \begin{cases} [\gamma_1 \tau^s, \gamma_2 \tau^s], & \text{if } r_s \geq \eta_2, \\ [\gamma_2 \tau^s, \tau^s], & \text{if } r_s \in [\eta_1, \eta_2), \\ [\gamma_3 \tau^s, \tau_{\max}], & \text{if } r_s \leq \eta_1, \end{cases} \tag{3.28}$$

where $0 < \eta_1 \leq \eta_2 < 1$ and $0 < \gamma_1 \leq \gamma_2 < 1 < \gamma_3$. The parameters $\eta_1, \eta_2, \gamma_1, \gamma_2, \gamma_3$ determine how aggressively the step-size is increased when an iteration is successful and how aggressively it is decreased when an iteration was unsuccessful. In practice, the performance of RSS-inner is not sensitive to the actual values of the parameters.

The complete RSS-inner procedure to solve subproblem (3.18) is summarized in Algorithm 4.

Convergence We next demonstrate that the RSS algorithm converges. Recall that by convergence we mean that the algorithm will converge to a stationary point of (3.18). Before proceeding, we first note that there exists $\lambda \in \mathbb{R}$ such that the first-order optimality conditions of (3.18) hold; that is,

$$p + \mu g(x) - \lambda x = 0, \quad \|x\|_2 = 1, \quad p \in \partial \|x\|_1, \tag{3.29}$$

where $g(x) = \nabla f(x)$. In addition, we make the following assumption on $g(x)$,

Assumption 1. *The gradient $g(x)$ of $f(x)$ is Lipschitz continuous with constant L:*

$$\|g(x) - g(y)\|_2 \leq L\|x - y\|_2.$$

Note that this assumption is valid for the objective function in (3.18).

We are now ready to establish convergence of the RSS-inner algorithm.

Theorem 7. *Suppose that for (3.18) $\mathcal{S}\left(\tau^s x^s - \mu g^s\right) \neq 0$ for every iteration. If the RSS-inner algorithm has finitely many successful iterations, then it converges to a stationary point. If the RSS-inner algorithm has infinitely many successful iterations, then there exists at least one cluster point of the sequence $\{x^s\}$ and every cluster point is a stationary point.*

To prove this, we first demonstrate that an iteration is successful if the step size at that iteration is sufficiently large. The remainder of the proof is by contradiction, by checking how much the objective function value of (3.18) decreases, for the successful iterations. The detailed proof can be found in Appendix C.2. The convergence of RSS-outer follows from the standard theory for non-smooth optimization [102].

In summary, in this section our goal was to solve (3.17). To do this, we formulated an algorithm for (3.17) using the augmented Lagrangian framework as given by Algorithm 3. Within the algorithm, we must solve the subproblem (3.20). Because (3.20) is of the form (3.18), it can be efficiently solved by the RSS-inner subroutine, as given by Algorithm 4, with provable convergence.

3.7.2 Binary iterative hard thresholding (BIHT)

Problem formulation and algorithm definition

We now introduce a simple first-order algorithm for the reconstruction of sparse signals from 1-bit compressive measurements. Our algorithm, *Binary Iterative Hard Thresholding* (BIHT), is a simple modification of IHT, the real-valued algorithm from which is takes its name [30]. The IHT algorithm has recently been extended to handle measurement non-linearities [105]; however, these results do not apply to quantized measurements since quantization does not satisfy the requirements in [105].

We briefly recall that the IHT algorithm consists of two steps that can interpreted as follows. The first step can be thought of as a gradient descent to reduce the least squares objective $\|y - \Phi x\|_2^2/2$. Thus, at iteration s, IHT proceeds by setting $a^{s+1} = x^s + \Phi^T(y - \Phi x^s)$. The second step imposes a sparse signal model by projecting a^{s+1} onto the "ℓ_0 ball", i.e., selecting the K largest in magnitude elements. Thus, IHT for CS can be thought of as trying to solve the problem

$$\widehat{x} \leftarrow \underset{x \in \mathbb{R}^N}{\operatorname{argmin}} \ \tfrac{1}{2}\|y - \Phi x\|_2^2 \quad \text{s.t.} \quad \|x\|_0 = K. \tag{3.30}$$

The BIHT algorithm simply modifies the first step of IHT to instead minimize a consistency-enforcing objective. Specifically, given an initial estimate $x^0 = 0$ and 1-bit measurements y_s, at iteration s BIHT computes

$$a^{s+1} = x^s + \frac{\tau}{2}\Phi^T\big(y_s - A(x^s)\big), \tag{3.31}$$

Algorithm 5: Binary Iterative Hard Thresholding (BIHT)

```
S0 Initialize
     Set initial solution $x^0 := 0$
     Set iteration $s := 0$
   while not converged do
S1     Update estimate (note that this is quite different from IHT proper)
         $a^{s+1} := x^s + \frac{1}{M}\,\Phi^T\big(y_s - A(x^s)\big)$
S2     Hard threshold and project onto unit sphere
         $x^{s+1} := U\big(\eta_K(a^{s+1})\big)$
S3     Update iteration count
         Set $s := s + 1$
```

$$x^{s+1} = \eta_K(a^{s+1}), \tag{3.32}$$

where A is defined as in (3.1), τ is a scalar that controls gradient descent step-size, and the function $\eta_K(v)$ computes the best K-term approximation of v by thresholding. Once the algorithm has terminated (either consistency is achieved or a maximum number of iterations have been reached), we then normalize the final estimate to project it onto the unit sphere. Section 3.7.2 discusses several variations of this algorithm, each with different properties. A quick summary is given in Algorithm 5.

The key to understanding BIHT lies in the formulation of the objective. The following Lemma shows that the term $\Phi^T\big(y_s - A(x^s)\big)$ in (3.31) is in fact the negative subgradient of a convex objective $\mathcal{J}$. Let $[\cdot]_-$ denote the negative function, i.e., $([u]_-)_i = [u_i]_-$ with $[u_i]_- = u_i$ if $u_i < 0$ and 0 else, and $u \odot v$ denote the Hadamard product, i.e., $(u \odot v)_i = u_i v_i$ for two vectors u and v.

Lemma 11. *The quantity $\frac{1}{2}\,\Phi^T\big(A(x) - y_s\big)$ in (3.31) is a subgradient of the convex one-sided ℓ_1-norm*

$$\mathcal{J}(x) = \|[y_s \odot (\Phi x)]_-\|_1,$$

Thus, BIHT aims to decrease $\mathcal{J}$ at each step (3.31).

Proof. We first note that $\mathcal{J}$ is convex. We can write $\mathcal{J}(x) = \sum_i \mathcal{J}_i(x)$ with each convex function $\mathcal{J}_i$ given by

$$\mathcal{J}_i(x; y_s, \Phi) = \begin{cases} |\langle \varphi_i, x\rangle|, & \text{if } A_i(x)\,(y_s)_i < 0, \\ 0, & \text{else,} \end{cases}$$

where φ_i denotes a row of Φ and $A_i(x) = \operatorname{sign}\langle \varphi_i, x\rangle$. Moreover, if $\langle \varphi_i, x\rangle \neq 0$, then the gradient of $\mathcal{J}_i$ is

$$\nabla \mathcal{J}_i(x; y_s, \Phi) = \tfrac{1}{2}(A_i(x) - (y_s)_i)\,\varphi_i = \begin{cases} A_i(x)\,\varphi_i & \text{if } (y_s)_i\,A_i(x) < 0, \\ 0, & \text{else} \end{cases}$$

while if $\langle \varphi_i, x\rangle = 0$, then the gradient is replaced by the subdifferential set

$$\nabla \mathcal{J}_i(x; y_s, \Phi) = \Big\{\tfrac{\xi}{2}(A_i(x) - (y_s)_i)\,\varphi_i : \xi \in [0,1]\Big\} \ni \tfrac{1}{2}(A_i(x) - (y_s)_i)\,\varphi_i.$$

Thus, by summing over i we conclude that $\frac{1}{2}\,\Phi^T\big(A(x) - y_s\big) \in \nabla J(x; y_s, \Phi)$. □

Consequently, the BIHT algorithm can be thought of as trying to solve the problem:

$$x^* \ = \ \underset{x}{\operatorname{argmin}} \ \tau \, \|[y_s \odot (\Phi x)]_-\|_1 \quad \text{s.t.} \quad \|x\|_0 = K, \, \|x\|_2 = 1.$$

Observe that since $y_s \odot (\Phi x)$ simply scales the elements of Φx by the signs y_s, minimizing the one-sided ℓ_1 objective enforces a positivity requirement,

$$y_s \odot (\Phi x) \geq 0, \tag{3.33}$$

that, when satisfied, implies consistency.

Previous 1-bit CS algorithms (such as the RSS algorithm of the previous section) have used a one-sided ℓ_2-norm to impose consistency [33, 39, 75, 76]. Specifically, they have applyied a constraint or objective that takes the form $\|[y_s \odot (\Phi x)]_-\|_2^2/2$. Both the one-sided ℓ_1 and ℓ_2 functions imply a consistent solution when they evaluate to zero, and thus, both approaches are capable of enforcing consistency. However, the choice of the ℓ_1 vs. ℓ_2 penalty term makes a significant difference in performance depending on the noise conditions. We explore this difference in the experiments in Section 4.2.

BIHT shifts

Several modifications can be made to the BIHT algorithm that may improve certain performance aspects, such as consistency, reconstruction error, or convergence speed. We believe that such variations exhibit interesting and useful properties that should be mentioned.

Projection onto sphere at each iteration. We can enforce that every intermediate solution have unit ℓ_2 norm. To do this, we modify the “impose signal model” step (3.32) by normalizing after choosing the best K-term approximation, i.e., we compute

$$x^{s+1} = U\big(\eta_K(a^{s+1})\big), \tag{3.34}$$

where $U(v) = v/\|v\|_2$. While this step is necessary for previous algorithms such as [39, 75, 76], it is in general not necessary in the BIHT case.

If we choose to impose the projection, Φ must be appropriately normalized or, equivalently, the step size of the gradient descent must be carefully chosen. Otherwise, the algorithm will not converge. Empirically, we have found that for a Gaussian matrix, an appropriate scaling is $1/(\sqrt{M}\|\Phi\|_2)$, where the $1/\|\Phi\|_2$ controls the amplification of the estimate from Φ^T in the gradient descent step (3.31) and the $1/\sqrt{M}$ ensures that $\|y_s - A(x^s)\|_2 \leq 2$. Similar gradient step scaling requirements have been imposed in the conventional IHT algorithm and other sparse recovery algorithms as well (e.g., [25]).

Minimizing hinge loss. The one-sided ℓ_1-norm is related to the *hinge-loss* function in the machine learning literature, which is known for its robustness to outliers [106]. Binary classification algorithms seek to enforce the same consistency function as in (3.33) by minimizing a function

$\sum[\kappa - y_s \odot (\Phi x)]_+$, where $[\cdot]_+$ sets negative elements to zero. When $\kappa > 0$, the objective is both convex and has a non-trivial solution. Further connections and interpretations are discussed in Section 4.2. Thus, rather than minimizing the one-sided ℓ_1 norm, we can instead minimize the hinge-loss. The hinge-loss can be interpreted as ensuring that the minimum value that an unquantized measurement $(\Phi x)_i$ can take is bounded away form zero, i.e., $|(\Phi x)_i| \geq \kappa$. This requirement is similar to the sphere constraint in that it avoids a trivial solution; however, will perform differently than the sphere constraint. In this case, in the gradient descent step (3.31), we instead compute

$$a^{s+1} = x^s - \tau\Psi^T(\mathrm{sign}(\Psi x^s - \kappa) - 1)/2$$

where $\Psi = (y_s \odot \Phi)$ scales the rows of Φ by the signs of y_s. Again, the step size must be chosen appropriately, this time as $C_\kappa/\|\Phi\|_2$, where C_κ is a parameter that depends on κ.

Minimizing other one-sided objectives. In general, any function $\mathcal{R}(x) = \sum \mathcal{R}_i(x_i)$, where $\mathcal{R}_i$ is continuous and has a negative gradient for $x_i \leq 0$ and is 0 for $x_i > 0$, can be used to enforce consistency. To employ such functions, we simply compute the gradient of $\mathcal{R}$ and apply it in (3.31).

As an example, the previously mentioned one-sided ℓ_2-norm has been used to enforce consistency in several algorithms. We can use it in BIHT by computing

$$a^s = x^s + \tau\Phi^T[y_s \odot \Phi x^s]_+$$

in (3.31). We compare and contrast the behavior of the one-sided ℓ_1 and ℓ_2 norms in Section 4.2.

As another example, in similar fashion to the Huber norm [107], we can combine the ℓ_1 and ℓ_2 functions in a piecewise fashion. One potentially useful objective is $\sum \mathcal{R}_i(x)$, where $\mathcal{R}_i$ is defined as follows:

$$\mathcal{R}_i(x) = \begin{cases} 0, & x_i \geq 0, \\ |x_i|, & -\frac{1}{2} \leq x_i < 0, \\ x_i^2 + \frac{1}{4}, & x_i < -\frac{1}{2}. \end{cases} \tag{3.35}$$

While similar, this is not exactly a one-sided Huber norm. In a one-sided Huber-norm, the square (ℓ_2) term would be applied to values near zero and the magnitude (ℓ_1) term would be applied to values significantly less than zero, the reverse of what we propose here.

This objective can provide different robustness properties or convergence rates than the previously mentioned objectives. Specifically, during each iteration it may allow us to take advantage of the shallow gradient of the one-sided ℓ_2 cost for large numbers of measurement sign discrepancies and the steeper gradient of the one-sided ℓ_1 cost when most measurements have the correct sign. This objective can be applied in BIHT as with the other objectives, by computing its gradient and plugging it into (3.31).

3.7.3 Convex 1-bit reconstruction formulations

The world is flat: From hyperspheres to hyperplanes

The RSS and BIHT algorithms adhere closely to the theoretical 1-bit framework (Sections 3.2–3.5) in that they attempt to find a solution that lies on the unit sphere. As previously described, this

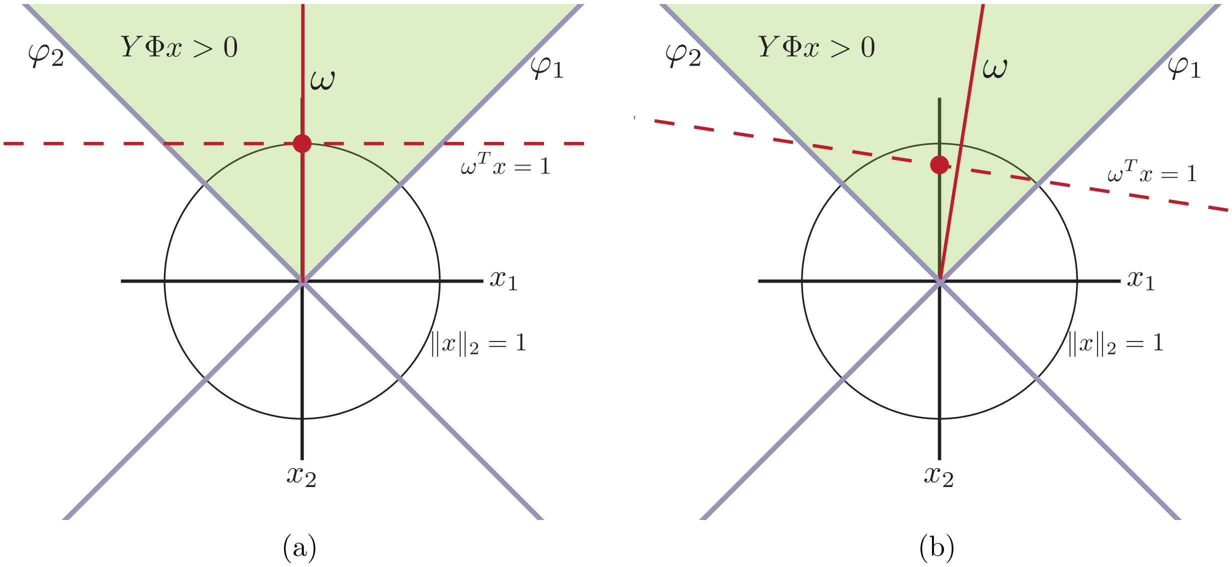


Figure 3.4: The geometry of the convex reconstruction formulation in two dimensions. The hyperplanes (lines) φ_1 and φ_2 correspond to the first and second rows of Φ, respectively. The green shaded region between the planes denotes the feasible region. The circle denotes the unit sphere. The red solid line denotes the centroid vector ω of the feasible region and the dashed red line denotes the constraint $\omega^T x = 1$. In this case the plane lies tangent to the sphere since the solution and the centroid are the same. The red dot denotes the optimal solution before normalization. (a) ω is the centroid, and (b) ω is an approximation to the centroid.

problem is not convex and therefore at best we can hope to prove that the algorithm converges to some local minimum as in the RSS algorithm, but not guarantee that we have found a feasible (i.e., consistent) solution. In these algorithms, we can only check if the solution is feasible.

However, it is possible to formulate the 1-bit reconstruction problem as a convex program. The key insight is that we want to find any non-trivial sparse solution inside the feasible region and the solution does not necessarily have to live on the unit sphere, since any non-trivial solution can be normalized. One convex approach would then choose a hyperplane that cuts through the feasible region. For example, we can solve

$$\widehat{x} \leftarrow \min_{x \in \mathbb{R}^N} \|x\|_1 \;\; \text{s.t.} \;\; Y\Phi x \geq 0 \;\; \text{and} \;\; \omega^T x = 1, \qquad (\ell_1\text{-min}_{1\text{B,LP}})$$

where ω is the centroid of the hyperplanes defined by the rows of Φ. We may also choose ω to be an approximation to the vector, obtained by summing the rows of Φ, i.e., $\omega = \sum \varphi_i$. This is a linear program (LP).[8]

The presence of the linear constraint $\omega^T x = 1$ in (ℓ_1-min$_{1\text{B,LP}}$) ensures that we avoid a trivial solution. The constraint furthermore defines a plane of possible solutions. Figure 3.4 depicts the geometry of the components of (ℓ_1-min$_{1\text{B,LP}}$) formulation in two dimensions. Specifically, the elements of the diagram are the same as in Figure 3.1 with the addition of the vector ω (which is

[8] During the final preparation of this manuscript a similar program was proposed and analyzed [108].

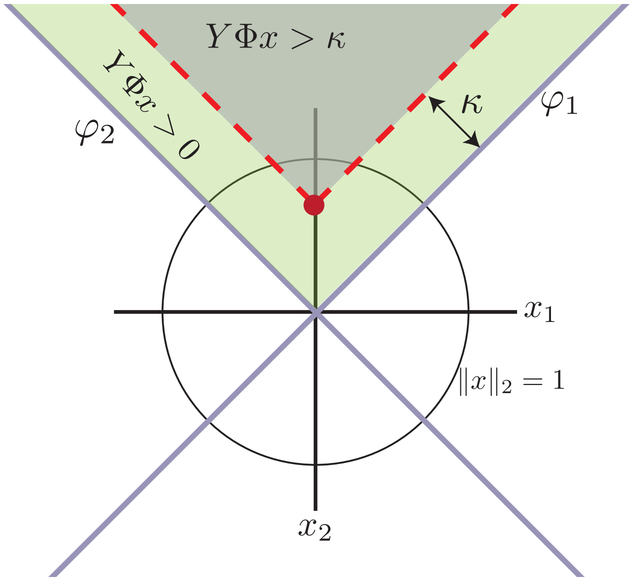


Figure 3.5: The geometry of the hinge-loss inspired reconstruction formulation in two dimensions. The hyperplanes (lines) φ_1 and φ_2 correspond to the first and second rows of Φ, respectively. The green shaded region between the planes denotes the region of consistent solutions. The circle denotes the unit sphere. The dark gray shaded region denotes the feasible region. Note that the feasible region does not include the trivial solution. The red dot denotes the optimal solution before normalization.

exact in (a) and approximated in (b)), denoted by the solid red line and the hyperplane $\omega^T x = 1$ denoted by the dashed red line. In this example, due to the orientation of the vectors φ_1 and φ_2, the true centroid ω aligns with the x_2-axis exactly. Since the centroid is on the same axis as the sparsest solution, the plane lies tangent to the unit sphere, i.e., $\omega^T x = x^T x = \|x\|_2^2 = 1$. In general this need not be the case. To see this, suppose we could only approximate ω, as in Figure 3.4(b). Because it is convex, this program is guaranteed to return a feasible, and thus consistent solution.

If the LP returns a strictly sparse solution, then it can be normalized (projected onto the unit sphere) and thus if Φ is a BϵSE, then we can guarantee stable recovery by Lemma 6. However, this program solves for the minimum ℓ_1-norm and thus will not necessarily return a strictly K-sparse solution.

Doors opened by the hinge-loss

As we saw in the introduction of the BIHT algorithm, we can use the hinge-loss or square-loss to enforce consistency. This can be applied as a constraint in a convex optimization problem. Specifically, the hinge loss reconstruction would be formulated as

$$\widehat{x} \leftarrow \min_{x \in \mathbb{R}^N} \|x\|_1 \;\; \text{s.t.} \;\; Y\Phi x - \kappa \geq 0 \tag{3.36}$$

for $\kappa > 0$, or the relaxation

$$\widehat{x} \leftarrow \min_{x \in \mathbb{R}^N} \|x\|_1 + \lambda \sum [\kappa - y_s \cdot (\Phi x)_+]. \qquad (\ell_1\text{-min}_{\text{1B,hinge}})$$

with $\lambda > 0$. Similarly the square-loss would be formulated as

$$\widehat{x} \leftarrow \min_{x \in \mathbb{R}^N} \|x\|_1 + \frac{\lambda}{2} \sum (\kappa - y_s \cdot (\Phi x)_+)^2. \qquad (\ell_1\text{-min}_{\text{1B,square}})$$

These programs are convex and return non-trivial solutions.

We can interpret these programs as making the assumption that the minimum amplitude of the true measurements Φx was greater than κ. To illustrate this, Figure 3.5 depicts geometry of the hinge-loss inspired formulation in two dimensions. We see from the figure that the addition of the term κ is akin to "lifting" the feasible region away from the intersection with the trivial solution (the origin). In practice the minimum amplitude assumption is not precisely true, however, if there is noise present on the measurement before quantization, it may be a reasonable assumption that measurements below the noise floor had quantized to the wrong values anyway. Thus, by tweaking κ we can adjust the tolerance to noise before quantization.

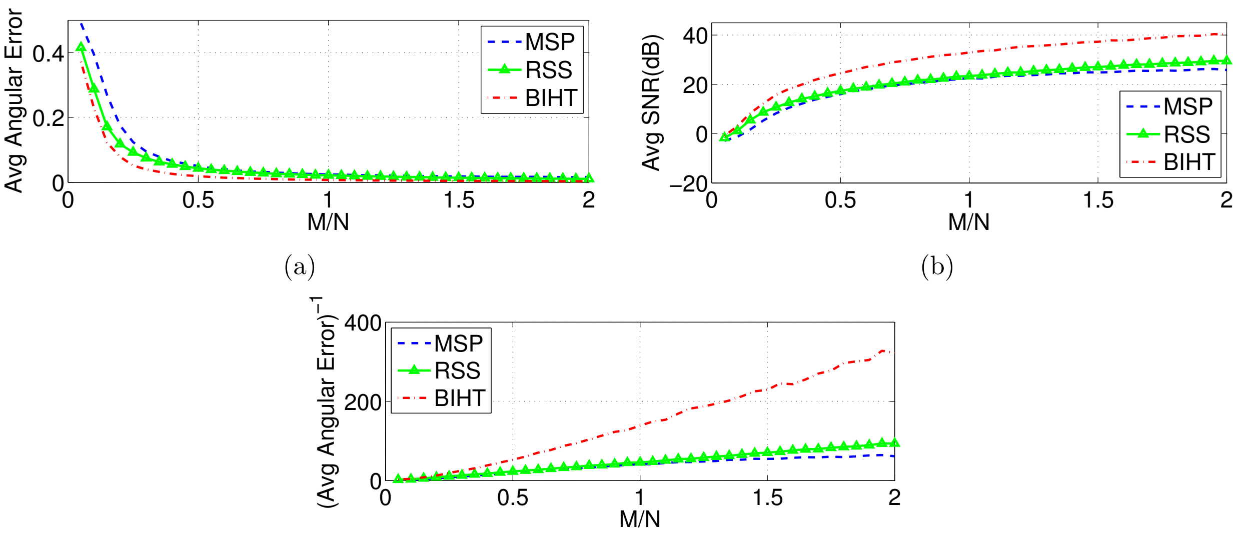
3.8 Empirical Verification

In this section we explore the performance of the RSS and BIHT algorithms and compare them to the performance of previous algorithms for 1-bit CS. We also explore the performance of the convex formulations as well as the multi-bit formulations decribed earlier.

The experimental setup is as follows. For each data point, we draw a length-N, K-sparse signal with the non-zero entries drawn uniformly at random on the unit sphere, and we draw a new $M \times N$ matrix Φ with each entry $\phi_{ij} \sim \mathcal{N}(0, 1)$. We then compute the binary measurements y_s according to (3.1). Reconstruction of $\widehat{x}$ is performed from y_s with three algorithms: *matching sign pursuit* (MSP) [76], *restricted-step shrinkage* (RSS), and BIHT; the algorithms will be depicted by dashed, dotted, and triangle lines, respectively. Each reconstruction in this setup is repeated for 1000 trials and with a fixed $N = 1000$ and $K = 10$ unless otherwise noted. Furthermore, we perform the trials for M/N within the range $[0, 2]$. Note that when $M/N > 1$, we are acquiring more measurements than the ambient dimension of the signal. While the $M/N > 1$ regime is not interesting in conventional CS, it may be very practical in 1-bit systems that can acquire sign measurements at extremely high, super-Nyquist rates.

Average error. We begin by measuring the average reconstruction angular error ϵ_{sim} over the 1000 trials. The results of this are depicted in Figure 3.6. We display the results of this experiment three different ways: *i*) the true angular error in Figure 3.6(a), which we denote as ϵ_{sim}, to demonstrate typical values achieved, *ii*) the signal-to-noise ratio (SNR)[9] in Figure 3.6(b), to demonstrate that the performance of these techniques is practical (since the angular error is unintuitive to most

[9] We define the reconstruction SNR in decibels as $\text{SNR}(x) := 10 \log_{10}(\|x\|_2^2 / \|x - \widehat{x}\|_2^2)$. Note that this metric uses the standard euclidean error and not angular error.



(a)

(b)

(c)

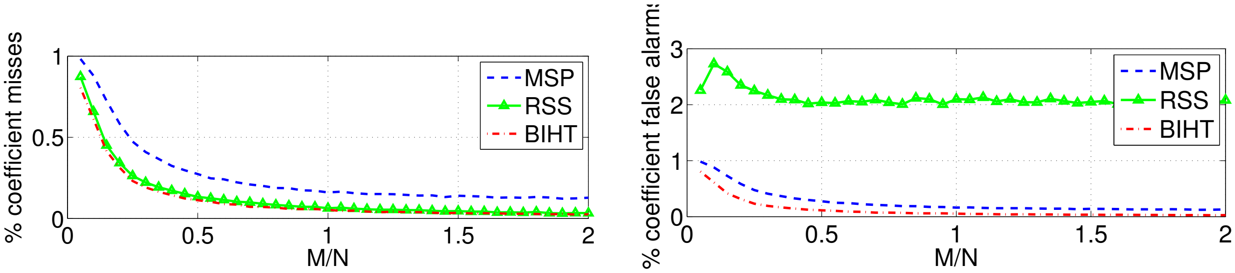
Figure 3.6: Average reconstruction angular error ϵ_{sim} vs. M/N, plotted three ways. (a) Angular error ϵ_{sim}, (b) SNR in decibels, and (c) Inverse angular error $\epsilon_{\text{sim}}^{-1}$. The plot demonstrates that BIHT yields a considerable improvement in reconstruction error, achieving an SNR as high as 40dB when $M/N = 2$. Furthermore, we see that the error behaves according $\epsilon_{\text{sim}} = O\left(1/M\right)$, implying that on average we achieve the optimal performance rate given in Theorem 4.

observers), and *iii*) the inverse of the angular error squared, i.e., $\epsilon_{\text{sim}}^{-1}$ in Figure 3.6(c), to compare with the performance predicted by Theorem 5.

We begin by comparing the performance of the algorithms. While the angular error of each algorithm appears to follow the same trend, BIHT obtains smaller error (or higher SNR) than the others, significantly so when M/N is greater than 0.35. The discrepancy in performance could be due to difference in the algorithms themselves, or perhaps, differences in their formulations for enforcing consistency. This is explored later in this section.

We now consider the actual performance trend. We see from Figure 3.6(c) that, above $M/N = 0.35$ each line appears fairly linear, albeit with a different slope, implying that with all other variables fixed, $\epsilon_{\text{sim}} = O\left(1/M\right)$. This is on the order of the optimal performance as given by the bound given in Theorem 4 and predicted by Theorem 5 for Gaussian matrices.

Misses and false alarms. We dig a little deeper into the source of errors by examining the reconstruction "misses," i.e., those coefficients that were identified as zero that are non-zero in the true signal, as well as the "false-alarms", i.e., those coefficients that were identified as non-zero that are zero in the true signal. The results are depicted in Figure 3.7(a) and (b), respectively. In both cases, BIHT outperforms the other algorithms, although it is very close to the RSS algorithm in the number of misses. While both RSS and MSP have significantly more false-alarms than BIHT, by design, MSP can return at most K non-zero coefficients and thus cannot have more than K false alarms. Meanwhile, the RSS algorithm may have many coefficients that are significantly close to zero but are numerically counted as non-zeros.



(a)

(b)

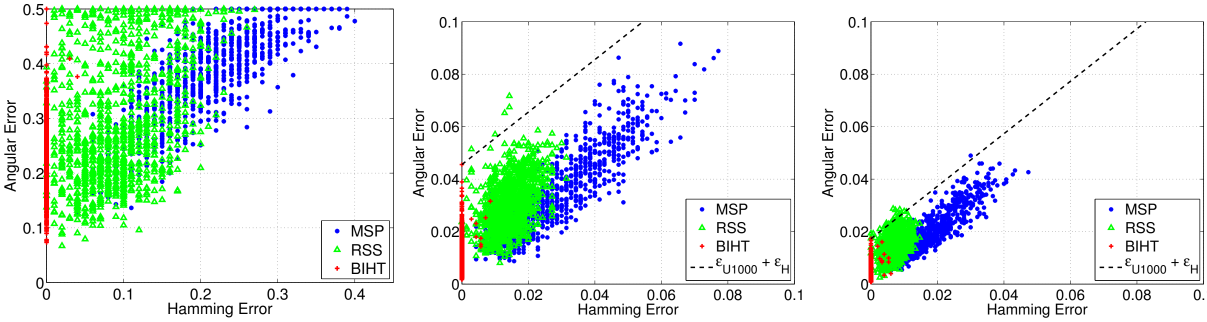
Figure 3.7: Reconstructed signal coefficient (a) misses, and (b) false-alarms. The MSP algorithm is most likely to miss a coefficient, while RSS and BIHT perform comparably. The RSS algorithm returns a large number of coefficients that are close to zero and thus performs poorly in the false-alarms metric. Both BIHT and MSP are restricted to have at most K false alarms by design.

Consistency. We also expose the relationship between the Hamming distance $d_H(A(x), A(\widehat{x}))$ between the measurements of the true and reconstructed signal and the angular error of the true and reconstructed signal. Figure 3.8 depicts the Hamming distance vs. angular error for three different values of M/N. The particularly striking result is that BIHT returns significantly more consistent reconstructions than the two other algorithms. This is clear from the fact that most of the red (plus) points lie on the y-axis while the majority of blue (dot) or green (triangle) points do not. We find that, even in significantly "under-sampled" regimes like $M/N = 0.1$, where the BϵSE is unlikely to hold, BIHT is likely to return a consistent solution (albeit with high variance of angular errors). We also find that in "over-sampled" regimes such as $M/N = 1.7$, the range of angular errors on the y-axis is small.

We can infer an interesting performance trend from Figures 3.8(b) and (c), where the BϵSE property may hold. Since the RSS and MSP algorithms often do not return a consistent solution, we can visualize the relationship between angular error and hamming error. Specifically, on average the angular reconstruction error is a linear function of hamming error, $\epsilon_H = d_H(A(x), A(\widehat{x}))$, as similarly expressed by the reconstruction error bound provided by BϵSE. Furthermore, if we let ϵ_{1000} be the largest angular error (with consistent measurements) over 1000 trials, then we can suggest an empirical upper bound for BIHT of $\epsilon_{1000} + \epsilon_H$. This upper bound is denoted by the dashed line in Figures 3.8(b) and (c).

One-sided ℓ_1 vs. one-sided ℓ_2 objectives. As demonstrated in Figures 3.6 and 3.8, the BIHT algorithm achieves significantly improved performance over MSP and RSS in both angular error and Hamming error (consistency). A significant difference between these algorithms and BIHT is that MSP and RSS seek to impose consistency via a one-sided ℓ_2-norm, as described in Section 3.7.2. Minimizing either the one-sided ℓ_1 or one-sided ℓ_2 objectives will enforce consistency on the measurements of the solution; however, the behavior of these two terms appears to be significantly different, according to the previously discussed experiments.

To test the hypothesis that this term is the key differentiator between the algorithms, we implemented BIHT-ℓ_2, a one-sided ℓ_2 variation of the BIHT algorithm that enabled a fair comparison of the one-sided objectives (see Section 3.7.2 for details). We compared both the angular error and



(a) M/N = 0.1

(b) M/N = 0.7

(c) M/N =1.5

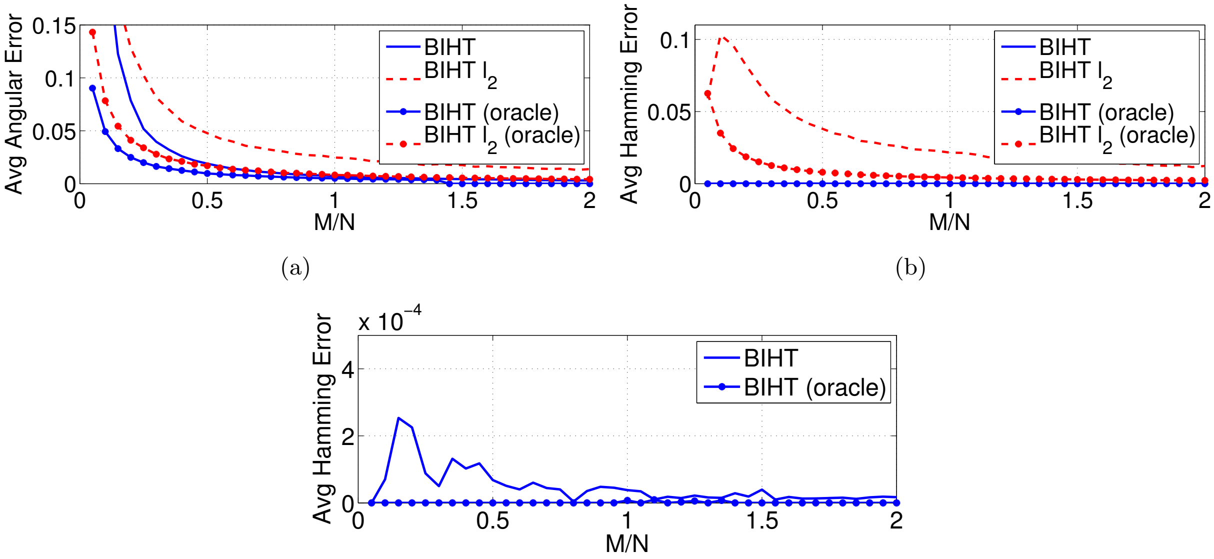
Figure 3.8: Reconstruction angular error ϵ_{sim} vs. measurement Hamming error ϵ_H. BIHT returns a consistent solution in most trials, even when the number of measurements is too low to permit a small angular error (see (a) M/N = 0.1). For larger M/N regimes, we see a linear relationship $\epsilon_{\text{sim}} \approx C + \epsilon_H$ between the average angular error ϵ_{sim} and the hamming error ϵ_H where C is constant (see (b) and (c)). The BϵSE formulation in Definition 2 predicts that the angular error is bounded by the hamming error ϵ_H in addition to an offset ϵ. The dashed line $\epsilon_{U1000} + \epsilon_H$ denotes the empirical upper bound for 1000 trials.

Hamming error performance of BIHT and BIHT-ℓ_2. Furthermore, we implemented *oracle assisted* variations of these algorithms where the true support of the signal is given a priori, i.e., η_K in (3.32) is replaced by an operator that always selects the true support, and thus the algorithm only needs to estimate the correct coefficient values. The oracle assisted case can be thought of as a "best performance" bound for these algorithms. Using these algorithms, we perform the same experiment detailed at the beginning of the section.

The results are depicted in Figure 3.9. The angular error behavior of BIHT-ℓ_2 is very similar to that of MSP and RSS and underperforms when compared to BIHT. We see the same situation with regards to Hamming error: BIHT finds consistent solutions for the majority of trials, but BIHT-ℓ_2 does not. Thus, the results of this simulation suggest that the one-sided term plays a significant role in the quality of the solution obtained.

One way to explain the performance discrepancy between the two objectives comes from observing the deep connection between our reconstruction problem and binary classification. As explained previously, in the classification context, the one-sided ℓ_1 objective is similar to the hinge-loss, and furthermore, the one-sided ℓ_2 objective is similar to the so-called *square-loss*. Previous results in machine learning have shown that for typical convex loss functions, the minimizer of the hinge loss has the tightest bound between expected risk and the Bayes optimal solution [109] and good error rates, especially when considering robustness to outliers [109, 110]. Thus, the hinge loss is often considered superior to the square loss for binary classification.[10] One might suspect that since the one-sided ℓ_1-objective is very similar to the hinge loss, it too should outperform other objectives in our context. Understanding why in our context, the geometry of the ℓ_1 and ℓ_2 objectives results in

[10] Additional "well-behaved" loss functions (e.g., the Huber-ized hinge loss) have been proposed [56] and a host of classification algorithms related to this problem exist [56, 110–113], both of which may prove useful in the 1-bit CS framework in the future.



(a)

(b)

(c)

Figure 3.9: Enforcing consistency: One-sided ℓ_1 vs. one-sided ℓ_2 BIHT. When BIHT attempts to minimize a one-sided ℓ_2 instead of a one-sided ℓ_1 objective, the performance significantly decreases. We find this to be the case even when an oracle provides the true signal support a priori. Note: (c) is simply a zoomed version (b).

different performance is an interesting open problem.

We probed the one-sided ℓ_1/ℓ_2 objectives further by testing the two versions of BIHT on noisy measurements. We flipped a number of measurement signs at random in each trial. For this experiment, $N = M = 1000$ and $K = 10$ are fixed, and we performed 100 trials. We varied the number of sign flips between 0% and 5% of the measurements. The results of the experiment are depicted in Figure 3.10. We see that for both the angular error in Figure 3.10(a) and Hamming error in Figure 3.10(b), that the one-sided ℓ_1 objective performs better when there are only a few errors and the one-sided ℓ_2 objective performs better when there are significantly more errors. This is expected since the ℓ_1 objective promotes sparse errors. This experiment implies that BIHT-ℓ_2 (and the other one-sided ℓ_2-based algorithms) may be more useful when the measurements contain significant noise that might cause a large number of sign flips, such as Gaussian noise.

Performance with a fixed bit-budget. In some applications we are interested in reducing the total number of bits acquired due to storage or communication costs. Thus, given a fixed total number of bits, an interesting question is how well 1-it CS performs in comparison to conventional CS quantization schemes and algorithms. For the sake of brevity, we give a simple comparison here between the 1-bit techniques and uniform quantization with *Basis Pursuit DeNoising* (BPDN) [72] reconstruction. While BPDN is not the optimal reconstruction technique for quantized measurements, it (and its variants such as the LASSO [56]) is considered a benchmark technique for reconstruction from measurements with noise and furthermore, is widely used in practice.

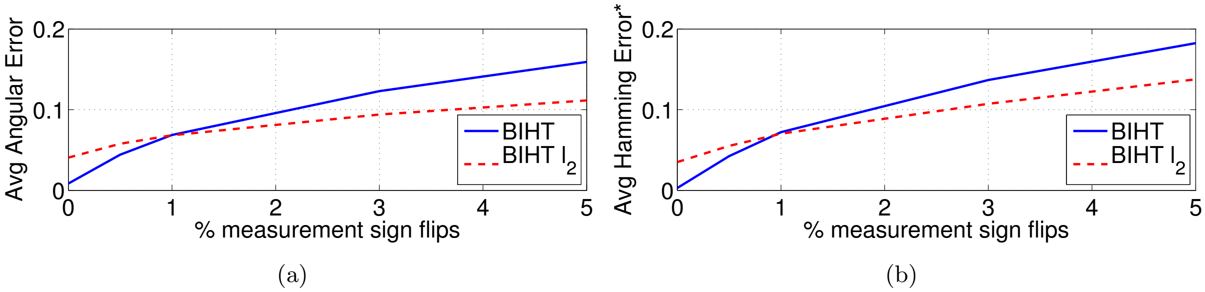


Figure 3.10: Enforcing consistency with noise: One-sided ℓ_1 vs. one-sided ℓ_2 BIHT. When BIHT attempts to minimize a one-sided ℓ_2 instead of the one-sided ℓ_1 objective, the algorithm is more robust to flips of measurement signs. *Note that the Hamming error in (b) is measured with regard to the noisy measurements, e.g., a Hamming error of zero means that we reconstructed the signs of the noisy measurements exactly.

The experiment proceeds as follows. Given a total number of bits and a (uniform) quantization bit-depth B (i.e., number of bits per measurement), we choose the number of measurements as $M =$ total bits$/B$, $N = 2000$, and the sparsity $K = 20$. The remainder of the experiment proceeds as described earlier (in terms of drawing matrices and signals). For bit depth greater than 1, we reconstruct using BPDN with an optimal choice of noise parameter and we scale the quantizer to such that signal can take full advantage of its dynamic range.

The results of this experiment are depicted in Figure 3.11. We see a common trend in each line: lackluster performance until "sufficient" measurements are acquired, then a slow but steady increase in performance as additional measurement are added, until a performance plateau is reached. Thus, since lower bit-depth implies that a larger number of measurements will be used, 1-bit CS reaches the performance plateau earlier than in the multi-bit case (indeed, the transition point is achieved at a higher number of total bits as the bit-depth is increased). This enables significantly improved performance when the rate is severely constrained and higher bit-rates per measurements would significantly reduce the number of available measurements. For higher bit-rates, as expected from the analysis in [78], using fewer measurements with refined quantization achieves better performance.

It is also important to note that, regardless of trend, the BIHT algorithm performs strictly better than BPDN with 4 bits per measurement and uniform quantization for the parameters tested here. This gain is consistent with similar gains observed in [39, 76]. A more thorough comparison of additional CS quantization techniques with 1-bit CS is a subject for future study.

Comparison to quantized Nyquist samples. In our next experiment, we compare the performance of the 1-bit CS technique to the performance of a conventional uniform quantizer applied to uniform Nyquist-rate samples. Specifically, in each trial we draw a new Nyquist-sampled signal in the same way as in our previous experiments and with fixed $N = 2000$ and $K = 20$; however, now the signals are sparse in the discrete cosine transform (DCT) domain. We consider four reconstruction experiments. First, we quantize the Nyquist-rate signal with a bit-depth of β bits per sample (and optimal quantizer scale) and perform linear reconstruction (i.e., we just use the quantized samples as sample values). Second, we apply BPDN to the quantized Nyquist-rate samples with optimal choice of noise parameter, thus denoising the signal using a sparsity model.

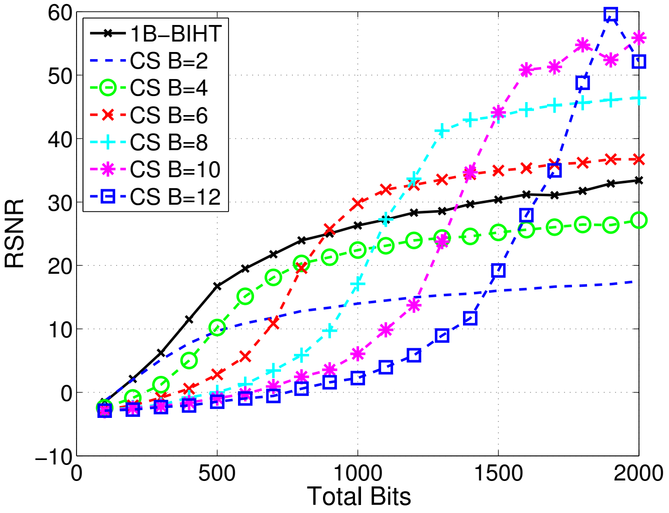


Figure 3.11: Comparison of BIHT to conventional CS multibit uniform scalar quantization (multibit reconstructions performed using BPDN [72]). BIHT is competitive with standard CS working with multibit measurements when the total number of bits is severely constrained. In particular, the BIHT algorithm performs strictly better than CS with 4 bits per measurement.

Third, we draw a new Gaussian matrix with $M = N$, quantize the measurements to β bits, again at optimal quantizer scale, and reconstruct using BPDN. Fourth, we draw a new Gaussian matrix with $M = \beta N$ and compute measurements, quantize to one bit per measurement by maintaining their sign, and perform reconstruction with BIHT. Note that the same total number of bits is used in each experiment.

Figure 3.12 depicts the average SNR obtained by performing 100 of the above trials. The linear, BPDN, Gaussian measurements with BPDN, and BIHT reconstructions are depicted by solid, dashed, dash-circled, and dash-dotted lines, respectively. The linear reconstruction has a slope of 6.02dB/bit-depth, exhibiting a well-known trade-off for conventional uniform quantization. The BPDN reconstruction (without projections) follows the same trend, but obtains an SNR that is at least 10dB higher than the linear reconstruction. This is because BPDN imposes the sparse signal model to denoise the signal. We see about the same performance with the Gaussian projections at $M = N$, although it performs slightly worse than without projections since the Gaussian measurements require a slightly larger quantizer range. Similarly to the results in Fig. 3.11, in low Nyquist bit-depth regimes ($\beta < 6$), 1-bit CS achieves a significantly higher SNR than the other two techniques. When $6 < \beta < 8$, 1-bit CS is competitive with the BPDN scenario. This simulation demonstrates that for a fixed number of bits, 1-bit CS is competitive to conventional sampling with uniform quantization, especially in low bit-depth regimes.

Comparison of BIHT, BIHT-ℓ_2, RSS, and the LP. In our next experiment, we compare the performance of several 1-bit CS reconstruction algorithms for a fixed bit budget. Specifically, we compare the BIHT, BIHT-ℓ_2, and RSS algorithms with the convex (ℓ_1-$\min_{\mathrm{1B,LP}}$) formulation. The LP formulation was implemented using MATLAB's built-in LP solver. Additionally, we include

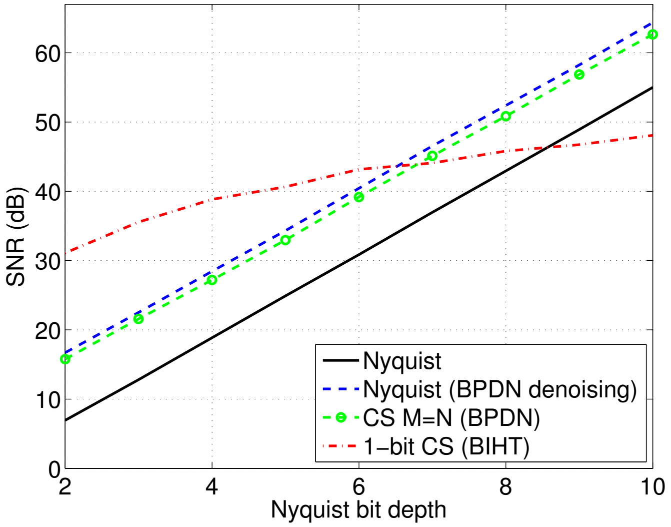


Figure 3.12: Comparison of uniformly quantized Nyquist-rate samples with linear reconstruction (solid) and BPDN denoising (dashed), CS with $M = N$ and BPDN reconstruction (dash-circle), and 1-bit quantized CS measurements with BIHT reconstruction (dash-dotted). Nyquist samples were quantized with bit-depth $\beta \in [2, 10]$ and 1-bit CS used $M = \beta N$ measurements; the same number of bits is used in each reconstruction. The Nyquist-rate lines have the classical 6.02dB/bit-depth slope, as expected. For a fixed number of bits, 1-bit CS does not follow this slope and outperforms conventional quantization when $\beta < 6$.

the performance when the true signal support is known *a priori* for both BIHT and the LP. Our choice of extending these two algorithms will become clear shortly.

We choose the number of measurements as $M =$ total bits, $N = 2000$, and the sparsity $K = 20$. The experiment proceeds in the same fashion as in Figure 3.11, however, now we only consider 1-bit measurements and performance across different algorithms. The BIHT, BIHT-ℓ_2, RSS algorithms, and LP are denoted by solid (black), dotted triangle (black), dash-dotted (blue), and dashed (red) lines, respectively. The known support enhanced variations of the algorithms are marked with circles.

As in the previous noiseless experiments, we find that BIHT significantly outperforms the other non-oracle-assisted algorithms, in this case, when the total number of bits is greater than 400. As a general trend, beyond 500 total bits, we find that in order of decreasing performance we have the LP, RSS, and BIHT-ℓ_2.

Since the LP formulation and BIHT performed the best, we compared them when the support is known *a priori*. Doing so measurements the best performance we can hope to achieve with these algorithms (in similar fashion to the oracle-assisted reconstruction of Section 1.2). These additional experiments are denoted by the lines with hollow circles on them. Although both algorithms perform with the same general trend, we see that even when the support is known, BIHT still outperforms the convex program. Also note that the convex program will always return a consistent solution and that BIHT returns a consistent solution most of the time. Thus, we can draw the conclusion

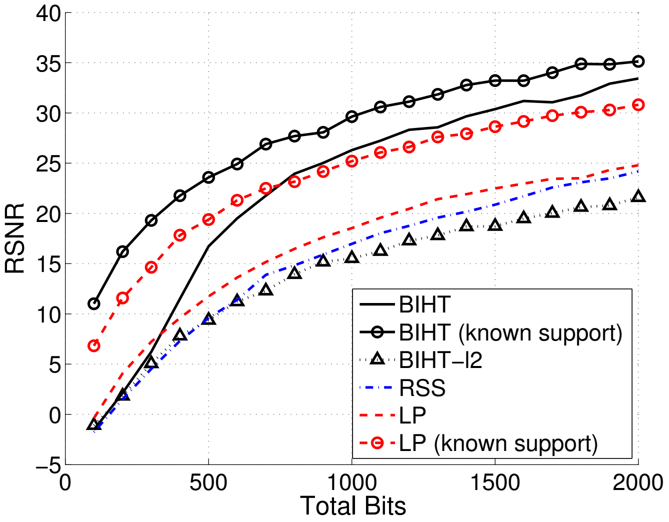


Figure 3.13: Comparison of BIHT, BIHT-ℓ_2, RSS, and the convex LP formulation (ℓ_1-$\min_{1B,LP}$) for a fixed bit budget (and fixed $N = 2000$, $K = 20$). In the noiseless setting even when the supports are known, the BIHT algorithms outperforms all other methods and LP performs second best.

that in the noiseless case, on average, BIHT provides a solution closer to the true solution inside the feasible region than does the LP, and that perhaps consistency isn't everything. It could be possible that the improved performance of BIHT in this case has to do with the distribution of the signals that were drawn in these experiments.

Comparison of BIHT, BIHT-ℓ_2, RSS, and the LP in noise. We also compare the performance with noisy measurements between the different algorithms and formulations proposed earlier. We performed an experiment where in each trial, we add zero-mean Gaussian noise e to the measurements before quantization, i.e.,

$$y_s = \text{sign}(\Phi x + e). \tag{3.37}$$

We use the parameters $N = 1000$, $K = 10$, $M = 2N$ and scale the noise so that the measurement SNR varies between 0 dB and 40 dB. Once the measurements are quantized, we perform reconstruction. The same algorithms are compared as in the previous experiment and again, the LP was implemented using MATLAB's built-in tools.

Figure 3.14 depicts the results of this experiment. As we have seen before, the BIHT algorithm performs best when the noise is very low (and the SNR is very high). However, below 30dB the other algorithms start to outperform BIHT. We also see that BIHT-ℓ_2 generally performs better than RSS but the two algorithms perform about the same when we only keep the K largest returned RSS coefficients (recall that RSS generally does not return a sparse solution). The RSS, LP, and BIHT-ℓ_2 algorithms perform similarly because they employ similar formulations to enforce consistency, similar to that of the square-loss. Intuitively, the one-sided ℓ_1 consistency formulation should bias toward sparse sign error and thus it makes sense that it performs better in lower noise scenarios.

Saturation-agnostic reconstruction. In our final experiment in this section, we explore the

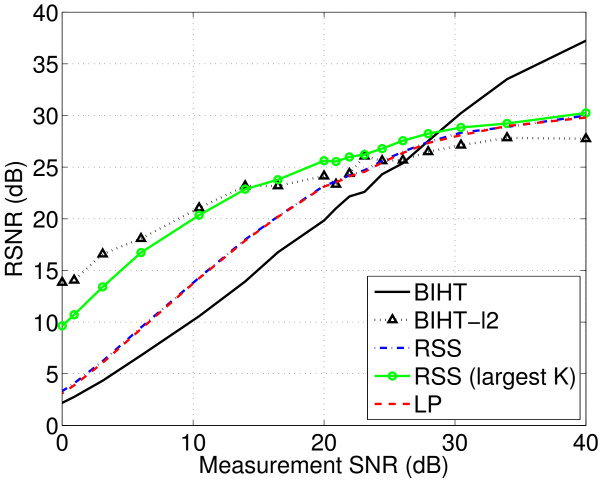


Figure 3.14: Reconstruction SNR as a function of measurement SNR (before quantization). Reconstructions performed with BIHT, BIHT-ℓ_2, RSS, and the LP for fixed $N = 1000$, $K = 10$, $M = 2N$ and measurement SNR between 0 dB and 40 dB. The BIHT algorithm outperforms the others in the high SNR regime (greater than 30dB) but underperforms in lower SNR regimes. The RSS algorithm achieves competitive performance to the BIHT-ℓ_2 algorithm when only the K largest coefficients are saved. The convex algorithm does not appear to outperform the RSS algorithm, despite its potentially nice properties.

benefits of saturation-agnostic reconstruction via the 1-bit algorithms, as explained in Section 3.6. We are interested in testing this technique in the context of measurement saturation as explained in the introduction to this chapter. Specifically, we are interested in demonstrating that signal reconstruction is possible for an arbitrary amount of saturation via a single consistent formulation.

The setup of the experiment is as follows. In each trial we generate an length $N = 1000$, $K = 10$ sparse unit-norm signal. We also draw a new $M \times N$ Gaussian measurement matrix with variance $1/M$, for fixed $M = N/2$. After computing measurements, we apply a 4-bit quantizer with finite range G, as described in Section 2.1. We vary G between 0 and 0.07. We perform reconstruction with two algorithms: *i*) the saturation-agnostic reconstruction formulation of Section 3.6 with the BIHT algorithm, and *ii*) the greedy *saturation consistent* CoSaMP (SC-CoSaMP) used in the democratic method and detailed in Chapter 2. SC-CoSaMP is known to break down when too much saturation is incurred on the measurements. We chose the BIHT algorithm to implement the saturation-agnostic approach since it produced the best noiseless reconstruction performance in the previous experiments. Results are expressed in RSNR and reflect the average over 100 trials. We chose the signal to have unit norm for the sake of comparison but this is not required in general.

Figure 3.15 depicts the results of this experiment. The dashed (blue) line depicts the average reconstruction performance obtained by the multibit technique and the solid (black) line depicts the performance obtained when applying SC-CoSaMP. For large quantizer range G, little saturation occurs and the SC-CoSaMP algorithm slightly outperforms the BIHT algorithm. However, this may be due to the differences in the algorithms and not the formulation itself. When G is decreased significantly, in this case below 0.02, the SC-CoSaMP algorithm drops steeply in performance,

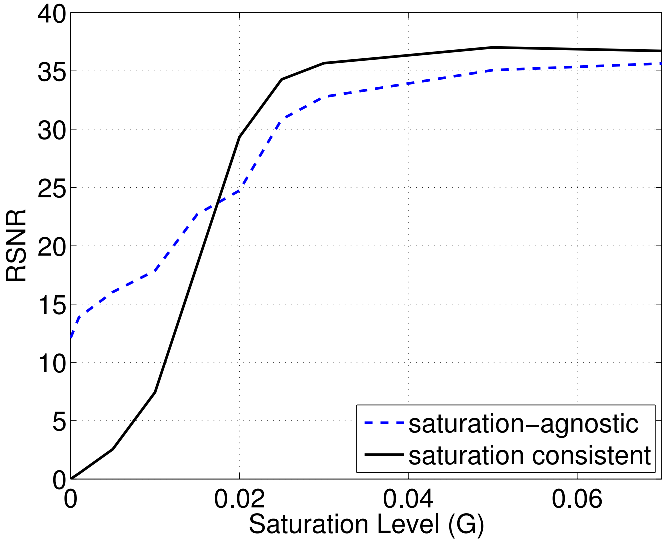


Figure 3.15: Saturation-agnostic sensing in action. A 4-bit quantizer with finite range G was applied to CS measurements as described in Section 2.1. The saturation-agnostic curve corresponds to the technique described in Section 3.6 and the saturation consistent curve was generated with the SC-CoSaMP, noted in Chapter 2. This simulation demonstrates that the saturation-agnostic formulation can achieve a significantly non-zero SNR in regimes where most measurements are saturated, in fact, even when all measurements have saturated. This technique provides robust performance when the saturation level cannot be controlled.

eventually obtaining an SNR of 0 when $G = 0$. However, the multibit BIHT algorithm, although it moderately decreases in performance around the same $G = 0.02$, maintains a significantly non-zero SNR, even at $G = 0$. Indeed, $G = 0$ corresponds to the “supersaturated” case, i.e., 1-bit CS.

The conclusion of this experiment, and indeed this chapter, is that we are in fact able to stably recover signals regardless of saturation level (or even bit-depth). The algorithm is agnostic to the number of saturations and provides reasonable performance where previous algorithms fail.

3.9 A Note on Alternative 1-Bit Frameworks

Two alternative approaches have been introduced to acquire 1-bit measurements and recover sparse signals. In [114], the authors propose a convolution-based imaging system with 1-bit measurements. Reconstruction is performed using total variation (TV) minimization and a gradient descent algorithm. In addition, the authors introduce a convex regularization parameter that simultaneously enforces both sign consistency and non-zero signal norm. In [98], the authors propose both non-adaptive and adaptive 1-bit sensing schemes. The non-adaptive scheme, which most closely relates to the framework presented here, relies on knowledge of the dynamic range of the signal, as well as an assumption about the distribution of the values of the nonzero coefficients.

CHAPTER 4

Regime Change

We now return to the multibit scalar quantizer (2.1) from Chapter 2:

$$y_Q = \mathcal{Q}_B(\Phi(x+n)+e), \tag{4.1}$$

where the *signal noise* is denoted by $n \in \mathbb{R}^N$, and $\mathcal{Q}_B : \mathbb{R} \to \mathfrak{A}$ is a B-bit scalar quantization function (applied element-wise in (2.1)) that maps real-valued CS measurements to the discrete alphabet $\mathfrak{A}$ with $|\mathfrak{A}| = 2^B$.

Since the quantizer is scalar, we can write the bit-budget constraint as

$$\mathfrak{B} = MB. \tag{4.2}$$

This fixed bit-budget $\mathfrak{B} = MB$ and the signal noise n impose a competing performance tradeoff as a function of M. On the one hand, since $B = \mathfrak{B}/M$, we can increase the bit-depth as we decrease the number of measurements, thereby increasing the precision of each measurement. On the other hand, signal noise is amplified due to *noise folding* as we decrease the number of measurements, thereby decreasing the precision of each measurement [36]. Thus, we find ourselves in somewhat of a conundrum: as we take fewer measurements we can allocate more bits per measurement (good), but noise folding increases the risk of wasting these bits on already imprecise measurements (bad).

We can gain more insight into this conundrum through a back-of-the-envelope calculation of the optimal total acquisition error, which comprises the expected mean-squared distortion due to a scalar quantizer for Gaussian measurements $O(\|x\|_2^2 2^{-2B})$ and the expected reconstruction error due to measurement noise $O\left(\frac{N}{M}\sigma_n^2\right)$. Equating these noise levels to minimize the total mean square error (MSE) leads to

$$B \approx \frac{1}{2}\log_2\left(\frac{\|x\|_2^2}{\sigma_n^2}\frac{M}{N}\right).$$

This expression can also be found using classical rate-distortion bounds in terms of the signal-to-noise ratio (SNR) [115, 116]. Imposing the fixed bit-budget $B = \mathfrak{B}/M$ and rearranging terms, we find that the MSE is minimized when

$$\log_2\left(\frac{\|x\|_2^2}{N\sigma_n^2}\right) \approx \frac{2\mathfrak{B}}{M} - \log_2(M). \tag{4.3}$$

The term on the left is the logarithm of the SNR of the input signal. For fixed $\mathfrak{B}$ and N, (4.3) implies that there are two operational regimes that correspond roughly to "high" input SNR and "low" input SNR. At high input SNR, the MSE is minimized by taking a small number of measurements M with large bit-depth; we call this the *measurement compression* (MC) regime. At low input SNR, the MSE is minimized by taking a large number of measurements M with small bit-depth; we call this the *quantization compression* (QC) regime. The exact SNR at which the transition between the two regimes occurs is a function of the total bit-budget.

In this chapter,[1] we argue for the distinction between the MC and QC regimes in two ways. First, we formalize the back-of-the-envelope calculation in (4.3) by analyzing the reconstruction MSE that results from the combined effects of quantization and signal noise folding. Specifically we provide an upper bound on this MSE for an optimal non-uniform scalar quantizer that roughly predicts the trends of the optimal bit-depth for different signal noise powers and bit-budgets. Second, we provide a suite of simulations for a specific setup frequently encountered in practice: the acquisition of sparse signals from uniformly quantized measurements. Surprisingly, at certain practical SNRs, our simulations suggest that a 1-bit quantizer (using the reconstruction techniques developed in [74]) exhibits better performance than larger bit-depth quantizers.

4.1 Analysis of Quantized CS Systems with Signal Noise

In this section we derive a new upper bound on the oracle-assisted reconstruction error due to both noise and quantization, making the back of the envelope calculation (4.3) more rigorous. This bound enables us to argue that, for a fixed bit-budget $\mathfrak{B} = MB$, it may be better to quantize to fewer bits per measurement B than take fewer measurements M. The following theorem is proved in Appendix D.

Theorem 8. *Suppose that $y_Q = \mathcal{Q}_B(\Phi(x+n))$. Let the signal $x \in \mathbb{R}^N$ be sparse with support $\Omega \in \{1,\ldots,N\}$ and $|\Omega| = K$, where the elements Ω are chosen uniformly at random and the amplitudes of the non-zero coefficients are drawn according to $x_j \in \Omega \sim \mathcal{N}(0,\sigma_x^2)$. Let the signal noise $n \in \mathbb{R}^M$ be a random, white, zero-mean vector with variance σ_n^2. Furthermore, let the $M \times N$ matrix Φ satisfy the RIP of order K with constant δ, $\Phi\Phi^T = \frac{N}{M} I_M$, and $M < N$. Choose $\mathcal{Q}_B$ to be the optimal scalar quantizer with $B > 1$ that minimizes the MSE for the distribution of the measurements $\Phi(x+n)$. Then for a fixed bit-budget of $\mathfrak{B} = MB$, the MSE of the oracle-assisted reconstruction estimate $\widehat{x}$ satisfies*

$$\mathbb{E}\left(\|x-\widehat{x}\|_2^2\right) \leq \frac{2K}{\mathfrak{B}(1-\delta)}\left(K\sigma_x^2 B 2^{-2B} + N\sigma_n^2 B\left(1+2^{-2B}\right)\right) + \frac{K}{(1-\delta)}\left(\frac{\mathfrak{B}}{B}-1\right)\mathfrak{S}, \tag{4.4}$$

[1] This chapter includes work done in collaboration with Richard Baraniuk [117].

where $\mathfrak{S} = \max_{i \neq j} |\mathbb{E}(\mathcal{Q}_B(\Phi x + \Phi n)_i \mathcal{Q}_B(\Phi x + \Phi n)_j)|$ *is the correlation between the quantized measurements.*

Each component of the bound (4.4) is fairly intuitive. The term $K\sigma_x^2 B 2^{-2B}$ reflects the error due to quantizing the measurements. The term $N\sigma_n^2 B\left(2^{-2B} + 1\right)$ reflects both the error due to measured signal noise as well as the quantization of that noise. The reconstruction error is effectively proportional to these two terms. The final term $\left(\frac{\mathfrak{B}}{B} - 1\right)\mathfrak{S}$ reflects an additional error due to the correlation between the quantized measurements. In many CS scenarios we expect this term to be close to zero, and furthermore for large B it has been shown that this term can be accurately approximated as zero [118]. Thus, choosing the optimal B primarily comes down to balancing the terms inside the parentheses.

The bound in (4.4) applies to strictly sparse signals immersed in signal noise. However, it may also be of interest to consider so-called *compressible signals*, i.e., signals that are not strictly sparse but that can be reasonably approximated by retaining their K largest magnitude coefficients. For such signals, the "tail" part of the signal that we do no expect to recover, i.e., the subset of the smallest $N - K$ entries, is also subject to noise folding. Theorem 8 can be extended to handle compressible signals by inflating the second term to account for the additional correlation between the quantized measurements. The general performance trends will be similar to sparse signals in noise; i.e., signals that are "less compressible" will induce the same regime as signals with low input SNR.

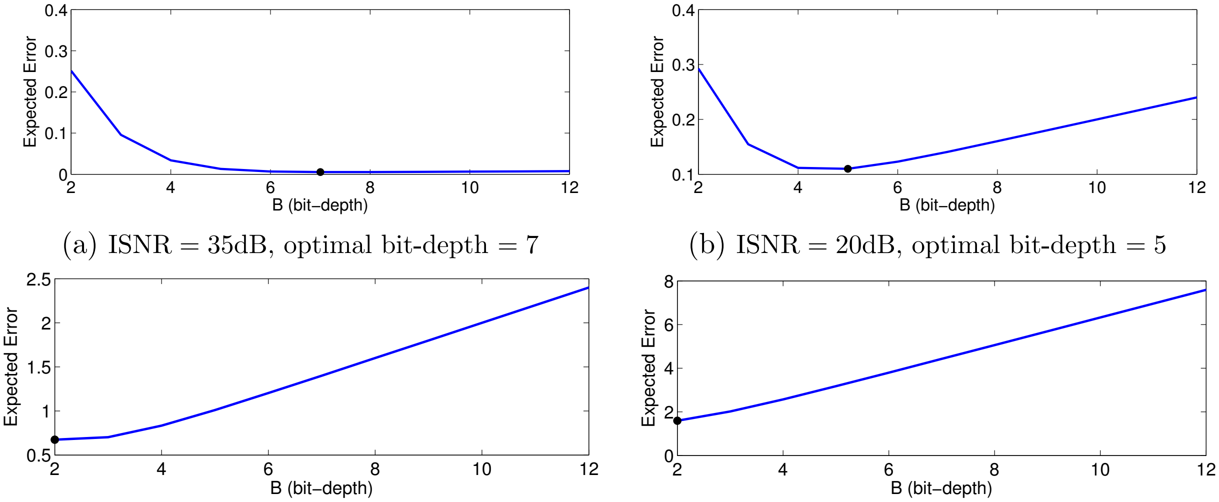
The bound in (4.4) is pessimistic, since we do not take into account the benefits accrued by increasing the number of measurements, for instance by improving the RIP constants of Φ. Furthermore, when the quantization error is large enough to dominate the measurement noise, the measurement noise terms may not play an active role in the true behavior of the system. Again, this is not reflected by the bound. Finally, the bound does not apply to 1-bit quantization or the case where $M > N$.

To use the bound (4.4) to support our argument that there are both MC and QC regimes in CS, we examine the behavior of the oracle-assisted reconstruction error as a function of the bit-depth B (or equivalently the number of measurements M since $\mathfrak{B} = MB$). Since the solution for the optimal B cannot be computed in closed form without resorting to tabulated functions, we evaluate the bound over some interesting parameters. The evaluation of the bound is depicted in Figure 4.1, where plots (a)–(d) correspond to input signal-to-noise ratios (ISNRs) of 35dB, 20dB, 10dB, and 5dB, respectively. We define the *input SNR* (ISNR) in dB as

$$\text{ISNR} := 10 \log_{10}\left(\frac{\mathbb{E}(\|x\|_2^2)}{\mathbb{E}(\|n\|_2^2)}\right). \tag{4.5}$$

where $\mathbb{E}(\|x\|_2^2) = K\sigma_x^2$ and $\mathbb{E}(\|n\|_2^2) = N\sigma_n^2$.

Since we are primarily concerned with the performance trend of (4.4) as a function of B and the ISNR, we make a few simplifications when plotting the bound. First, we only evaluate the term inside the parenthesis; this term is proportional to the error on the measurements and does not depend on the RIP constant, the sparsity K, or the correlation between the quantization errors. Second, by only evaluating the term inside the parenthesis in (4.4), we do not take into account



(a) ISNR = 35dB, optimal bit-depth = 7

(b) ISNR = 20dB, optimal bit-depth = 5

(c) ISNR = 10dB, optimal bit-depth = 2

(d) ISNR = 5dB, optimal bit-depth = 2

Figure 4.1: Upper bound on the oracle-assisted reconstruction error as a function of bit-depth B and ISNR. The term inside the parenthesis in the bound (4.4) was computed. Black dots denote the minimum point on each curve.

the effect of M on the RIP constants (δ decreases as M increases). The minimum error point in each curve is denoted by a solid black dot.

The message from Figure 4.1 is clear. The tradeoff between the number of measurements M and bit-depth B empirically follows a convex curve, i.e., the error not only increases when B is too small, but the error also increases when B is too large. In other words, more bits per measurement is not necessarily optimal. Furthermore, as expected, the minimum reconstruction error occurs for smaller B as the ISNR decreases. For the high ISNR of 35dB, the bound is minimized at a bit-depth of approximately 7 bits per measurement. The is an example of the MC regime, where larger bit-depths and thus lower M yield the best performance. For the low ISNR of 10dB, the bound is minimized at a bit-depth of approximately 2 bits per measurement. This is an example of the QC regime, where larger bit-depths and thus higher M yield the best performance.

4.2 Experiments

In the previous section we have argued that the QC regime exists by deriving an upper bound on the oracle-assisted reconstruction error. In this section we perform a suite of simulations to empirically study for which input noise levels and bit-budgets this regime will occur in practical systems. Specifically our simulations *i*) validate the theoretical result in Theorem 8, *ii*) demonstrate the performance achieved in practice when combining quantization and signal noise, and finally *iii*) prove the existence of the QC regime. A surprising additional result emerges from the simulations: when nontrivial signal noise is present, 1-bit CS systems perform competitively with, if not better than conventional CS with uniform multibit quantization.

4.2.1 Setup

Our simulations were performed using canonically (identity) sparse signals x.[2] The signals were measured with i.i.d. Gaussian matrices, i.e., $y = \Phi(x+n)$ with $\Phi \sim \mathcal{N}^{M \times N}(0, 1/M)$. The measurements were quantized uniformly with quantization interval $\Delta = G2^{-B+1}$, where G is the dynamic range of the quantizer. In all simulations, we chose $G = \|\Phi x\|_\infty$ to maximize the range of the quantizer and ensure that for any noiseless measurement $|(\Phi x)_i - \mathcal{Q}_B((\Phi x)_i)| \leq \Delta/2$.

In each trial we drew a new $M \times N$ sensing matrix Φ and a new signal x. The non-zero coefficients of x were chosen according to a Gaussian distribution, and their positions were chosen at random. We additionally added Gaussian noise to x to obtain the desired ISNR. For $B > 1$, reconstruction of the estimate $\widehat{x}$ was performed using the oracle-assisted reconstruction algorithm (1.10) for Section 4.2.2 and (BPDN) with an oracle value of $\epsilon = \|y - \mathcal{Q}_B(y)\|_2$ for the remaining subsections. For $B = 1$, reconstruction was performed using both the *binary iterative hard thresholding* (BIHT-ℓ_1) and BIHT-ℓ_2 algorithms; the former generally performs better in lower noise scenarios and the latter performs better in higher noise scenarios [74]. We report the *reconstruction SNR* (RSNR)

$$\text{RSNR} := 10 \log_{10} \left(\frac{\|x\|_2^2}{\|x - \widehat{x}\|_2^2} \right) \tag{4.6}$$

in dB unless otherwise noted. Recall that the number of measurements and bit-depth are constrained by $\mathfrak{B} = MB$. We average our results over 100 trials for each parameter tuple $(N, K, \mathfrak{B}, B, \text{ISNR})$.

4.2.2 Oracle-assisted reconstruction

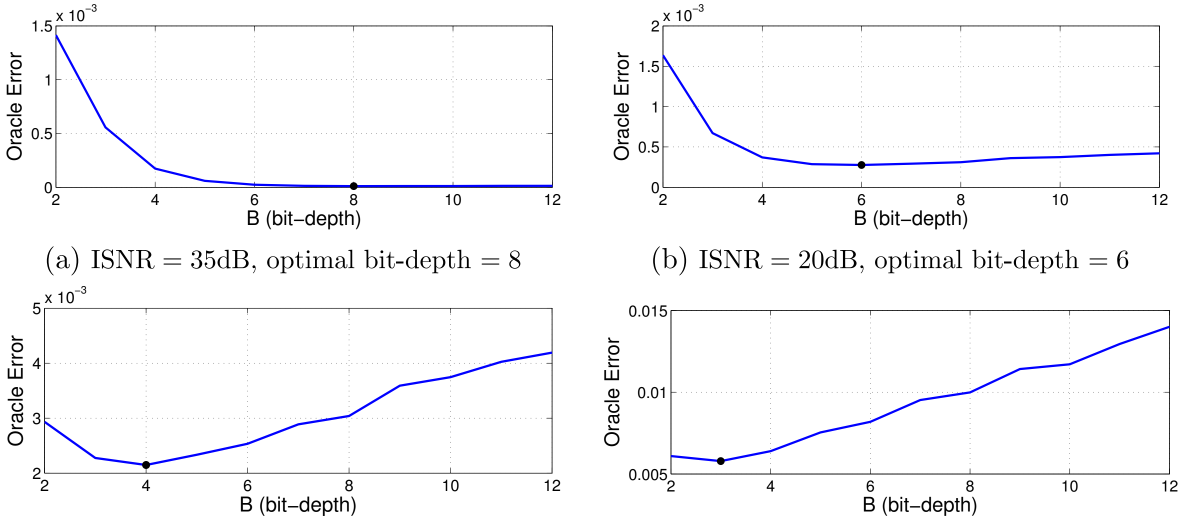
We begin by validating the message from Theorem 8, i.e., we examine the solution to the oracle-assisted reconstruction algorithm to see how the empirical performance relates to the bound (4.4). Our goal is to compare the performance of our simulations to the theory-based plots in Figure 4.1. The experiments were performed as described previously with the oracle-assisted reconstruction algorithm. We plot the reconstruction error $\|x - \widehat{x}\|_2^2$ for bit-depths between 2 and 12 for a fixed bit-budget $\mathfrak{B} = 3N$. We compared bit-depths of 2 and higher, since (4.4) does not hold for lower bit-depths. Furthermore, unlike the statement of Theorem 8, recall that we used a uniform quantizer and not an optimal quantizer for the Gaussian measurements. Figures 4.2(a)–(d) depict the results for ISNR = 35dB, 20dB, 10dB, and, 5dB, respectively.

The plots generally follow the same trends as in Figure 4.1; however the minimum error occurs for a slightly higher bit-depth in each case. The plots demonstrate that, as claimed in Section 4.1, the best performance is obtained for smaller bit-depths as the ISNR decreases.

4.2.3 Reconstruction performance as a function of $\mathfrak{B}$

We next explore the performance achieved using practical algorithms instead of oracle-assisted reconstruction. The experiments were performed as explained previously, for $N = 1000$ and $K =$

[2] The results of simulations did not change when the signals were DCT-sparse.



(a) ISNR = 35dB, optimal bit-depth = 8

(b) ISNR = 20dB, optimal bit-depth = 6

(c) ISNR = 10dB, optimal bit-depth = 4

(d) ISNR = 5dB, optimal bit-depth = 3

Figure 4.2: Oracle-assisted reconstruction error (compare to the analytical upper bound plotted in Figure 4.1) for $N = 1000$, $K = 10$, and $\mathfrak{B} = 3N$. As predicted by (4.4), the minimum reconstruction error (denoted by black dots) is achieved by smaller bit-depths as the ISNR decreases.

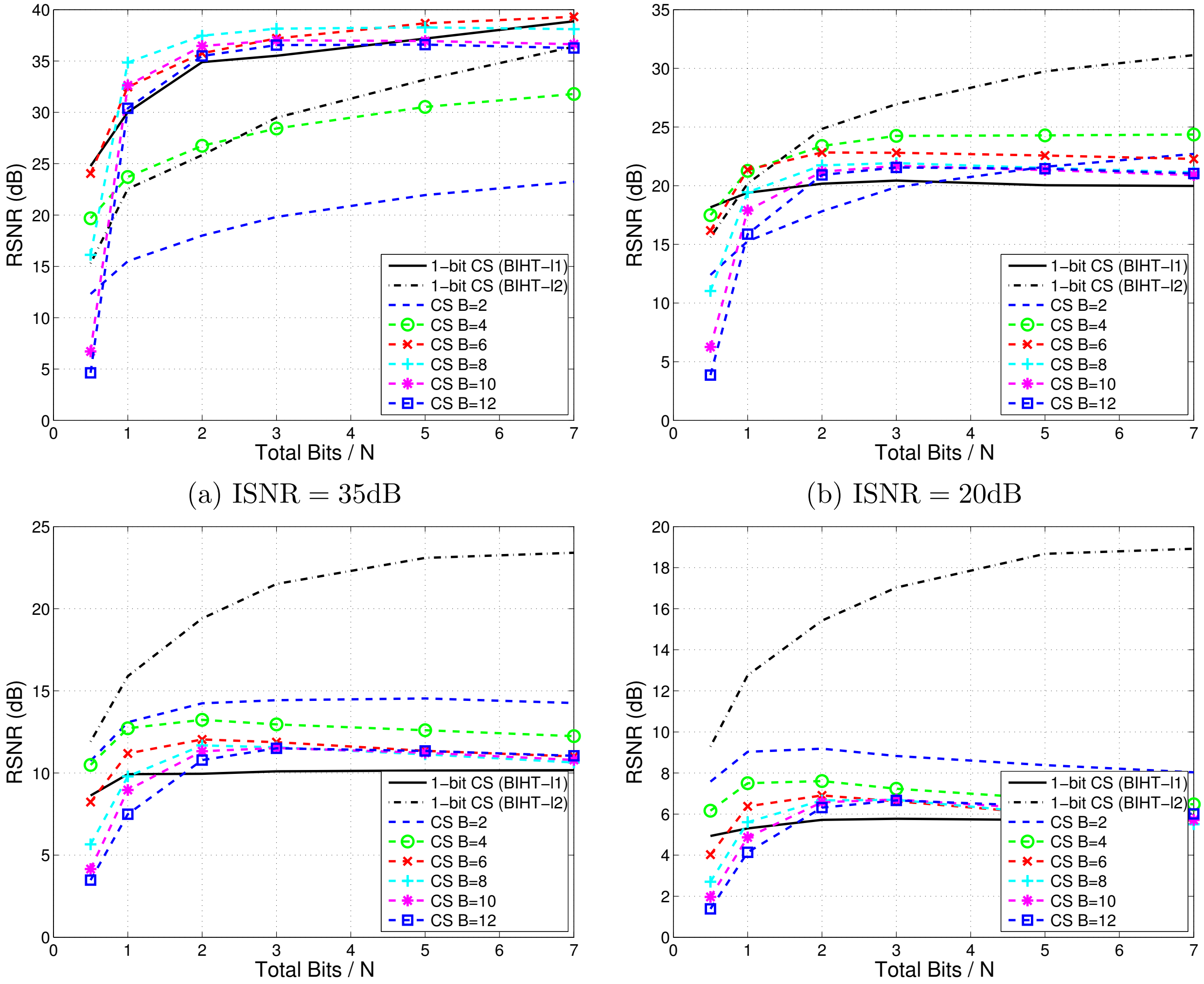
10, bit-depths $B = 1, 2, 4, 6, 8, 10, 12$, and for bit-budgets $\mathfrak{B} \in [N/2, 7N]$, with the BPDN and BIHT algorithms. Figures 4.3(a)–(d) depict the experiment for the input ISNR $= 35, 20, 10, 5$dB, respectively.

In the high ISNR regime of 35dB, bit-depths of $B = 1, 6, 8, 10$, and 12 obtain similar RSNRs of around 35dB, while smaller bit-depths result in poorer performance. This is to be expected; since when the signal noise is fairly small, we will generally do better by using more bits per measurement.

The performance of BIHT in this case is consistent with previous results showing that the 1-bit techniques can outperform even 4-bit uniformly quantized CS measurements with BPDN recovery. This trend starts to reverse for lower signal ISNRs. Indeed for ISNRs of 10dB and 5dB, we see that 2 and 4 bit-depth quantization outperforms larger bit-depths for all budgets. Strikingly, the best performance for input SNRs of 20dB, 10dB, and 5dB is achieved by acquiring just 1 bit per measurement and reconstructing with the BIHT-ℓ_2 algorithm.

In addition to the simulations presented here, we also performed the similar simulations with $N = 1000$ and $K = 60$. We found that all of the curves in Figure 4.3 dropped in SNR by roughly the same constant (that depends on K). The relationship between the 1-bit curves and the others was about the same for $\mathfrak{B} = 2N$ and lower. For $\mathfrak{B} > 2N$, the 1-bit reconstructions still outperformed the others; however the performance disparity was not as great as for $K = 10$.

These simulations demonstrate two points. First, they verify that the intuition provided by the upper bound (4.4) is indeed correct: *for lower ISNRs it is beneficial to choose smaller bit-depths B and more measurements M.* This validates the distinction between the QC and MC regimes. Second, the 1-bit CS setup performs significantly better than the multi-bit setup for low ISNRs



(a) ISNR $= 35$dB

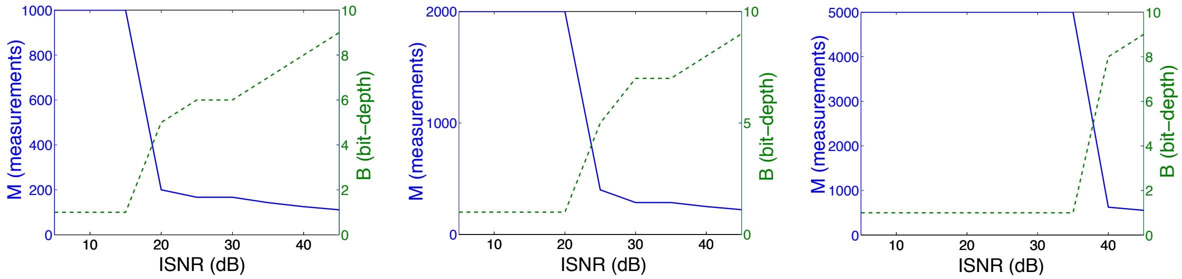
(b) ISNR $= 20$dB

(c) ISNR $= 10$dB

(d) ISNR $= 5$dB

Figure 4.3: Reconstruction performance as a function of total bits, for different ISNRs. Plots depict RSNR for different bit-depths B for different ISNR with parameters $N = 1000$ and $K = 10$, and reconstruction via BPDN. The figure demonstrates that as the ISNR is decreased, smaller bit-depths achieve better performance. Additionally, 1-bit CS techniques perform competitively with or better than BPDN for all ISNRs tested.

and is competitive with the multi-bit setup for moderate ISNRs. There are several reasons for this. When the quantization error dominates the measurement noise, the reconstruction error is primarily due to the quantization error only. This case arises when B is small; i.e., we can likely satisfy $\mathcal{Q}_B(x + n) = \mathcal{Q}_B(x)$ for increasing values of $|n_i|$ as B decreases. Furthermore, in this case consistent reconstruction of the 1-bit algorithms may have an advantage. Consistency could be presumably added to multibit reconstruction to improve performance but this is a topic left for future research.



(a) $\mathfrak{B} = N$

(b) $\mathfrak{B} = 2N$

(c) $\mathfrak{B} = 5N$

Figure 4.4: Maximum RSNR given a fixed bit-budget $\mathfrak{B}$ for parameters $N = 1000$, $K = 10$. The left side of each plot corresponds to the QC regime, while the right side corresponds to the MC regime. The solid line (blue) corresponds to the number of measurements M, while the dashed line (green) corresponds to the bit-depth B.

4.2.4 Reconstruction performance as a function of ISNR

In this set of experiments, we varied the ISNR between 5dB and 45dB and searched for the (M, B) pair that maximized the RSNR, for a fixed bit-budget $\mathfrak{B}$ and parameters $N = 1000$ and $K = 10$. As demonstrated by the previous experiment, the RSNR will not be the same for each bit-budget.

Figures 4.4(a)–(c) depict the results of this experiment for $\mathfrak{B} = N$, $2N$, and $5N$, respectively. The left axis and solid line (blue) corresponds to the number of measurements M, while the right axis and dashed line (green) corresponds to the bit-depth B. As always, we have that $\mathfrak{B} = MB$. The QC regime is represented on the left side of the plots (low ISNR), while the MC regime is represented on the right side of the plots (high ISNR). For example, for a bit-budget of $\mathfrak{B} = 2N$, if the ISNR is 30dB, then we are operating in the MC regime and should set the bit-depth to approximately 7, resulting in the measurement ratio of approximately $M/N = 0.29$. However, for the same bit-budget, if the ISNR is 15dB, then we are operating in the QC regime and should set the bit-depth to 1, resulting in a measurement ratio of $M/N = 2$.

In each plot in Figure 4.4 there is a sharp transition between optimal bit-depth being high ($B \geq 5$) and low ($B \leq 2$). This transition is centered at the ISNRs 19dB, 23dB, and 38dB, for the bit-budgets $\mathfrak{B} = N$, $2N$, and $5N$, respectively. This implies that the transition occurs at higher ISNRs for higher bit-budgets. Thus, we infer that, for higher bit-budgets $\mathfrak{B}$, it is better to choose low B, even when the input ISNR is fairly high. The bottom line then is that, for moderate ISNR, the MC regime can be assumed when the bit-budget $\mathfrak{B}$ is small, while the QC regime can be assumed when the bit-budget is large.

4.3 Relationship to Oversampled ADCs

Choosing a low bit-depth quantizer to reduce hardware complexity while driving up the sampling rate, as is recommended for the QC regime, is not a new idea. Indeed, this same principle is the motivational force behind sigma-delta ADCs [77, 82, 84, 85] and other non-CS oversampled ADC architectures [81, 119, 120]. However, the ideas presented here differ significantly from previous oversampled ADC architectures in the following ways: *i*) CS is compressive: Small bit-depth CS systems are expected to be used in cases where the bit-budget is significantly lower than in a conventional oversampled ADC system. The use of sparse signal models enables *compression*, i.e., a reduction in the total number of acquired bits, as opposed to just efficient sampling. *ii*) CS is non-adaptive: As described earlier, CS measurement systems are non-adaptive, meaning they do not depend on the input signal. This is true even for the 1-bit CS case. Almost all previous oversampled ADCs require some kind of feedback during quantization to produce stable representations. These differences place low bit-depth CS systems in a unique class of their own. In a few words, CS, like physics has "plenty of room at the bottom [121]."

CHAPTER 5

Compressive Sensing Architectures

A central question of the CS framework is how do we design a good sensing system. As previously explained in Section 1.2 from an analysis perspective, a variety of different conditions on Φ can be used to ensure that robust signal recovery is possible [16, 21, 42, 44, 122]. From a practical perspective, we wish to design a physical sampling system for which, when modeled by Φ, the aforementioned conditions are provably satisfied. To this end, several hardware architectures have been proposed that theoretically enable CS to be used in practical settings with analog signals. Examples include the random demodulator, random filtering, and random convolution for time-varying signals [18, 47, 48, 123], and several compressive imaging architectures [19, 124, 125]. Other compressive frameworks exist that have yielded similar types of architectures, for instance for multi-band signal acquisition [49, 126] or within the finite rate of innovation area [127, 128]. In this chapter,[1] we introduce two new practical CS acquisition architectures. We first motivate our new designs.

Beyond theoretical requirements, the aim of most CS-ADCs is to exploit the fact that fewer measurements are required to represent the signal, for instance, by reducing the sampling rate of a conventional ADC. This is often achieved through analog hardware that preconditions the signal before it is sampled at a sub-Nyquist rate by a conventional ADC.

One such architecture that exemplifies this philosophy is the random demodulator (RD) [18, 21, 130, 131]. Figure 5.1 depicts the block diagram of the random demodulator. The four key components are a pseudo-random ±1 "chipping sequence" $p(t)$ 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, i.e., preconditioned, and integrated. The output of the integrator is sampled, and the integrator is reset after each

[1] This chapter includes work done in collaboration with J. P. Slavinsky, Mark Davenport, and Richard Baraniuk [20, 129]

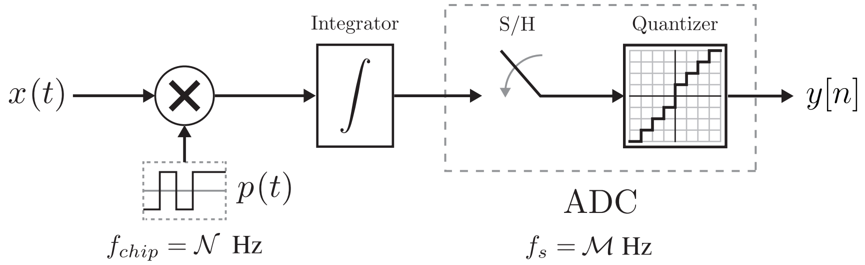


Figure 5.1: The random demodulator (RD) [18]. The analog signal $x(t)$ is "preconditioned" i.e., modulated by a ±1 square-waveform $p(t)$ that is determined by a pseudo random sequence. The result is integrated and sampled at a sub-Nyquist rate by a conventional ADC.

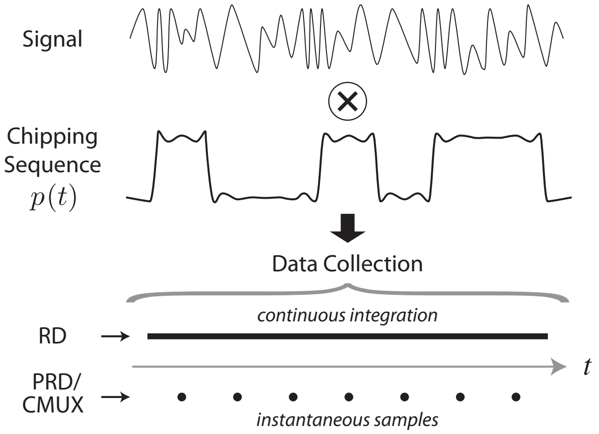


Figure 5.2: Waveforms at preconditioning stages in the RD and polyphase random demodulator (PRD). The analog signal $x(t)$ is modulated by a non-ideal analog square-wave. The RD then integrates the entire result before sampling and quantization. The CMUX and PRD integrate only instantaneous samples of the modulated signal before quantization.

sample. The output measurements from the ADC are then quantized.

There are additional desirable properties that CS-ADCs must have if we wish to execute them in practice:

1. precise calibration with computational models;
2. few additional sources of hardware noise; and
3. efficient computational implementation of recovery.

As we will discuss in this chapter, these requirements preclude the use of certain kinds of analog components, such as analog filters, that are commonly found in current designs for wideband signal acquisition, such as the RD [18, 49, 130, 131].

We briefly describe two instances in which analog filtering hinders the RD. *i*) In practice the integrator is implemented by an analog low pass filter. The impulse response of this filter must be accurately modeled in the discrete implementation of Φ. Precise calibration of this model with physical hardware is time consuming and can vary significantly between devices due to temperature and other operating conditions. *ii*) As shown in Figure 5.2, the rise and fall times of $p(t)$ fluctuate significantly away from the ideal square waveform. These fluctuations are integrated into the RD measurements. Such non-idealities are extremely difficult to calibrate for, and largely behave as a source of noise in the system [132].

In this chapter, we introduce two new CS-ADC architectures, namely the *compressive multiplexer* (CMUX) and the *polyphase random demodulator* (PRD), for wideband signal acquisition that addresses these practical considerations.

5.1 The Random Demodulator (RD)

The random demodulator (RD) was originally proposed in [18] and further studied in [21, 130, 131, 133]. Its primary goal is to acquire a wideband signal $x(t)$ while reducing the hardware costs associated with a high-rate Nyquist ADC. We refer the reader to Figure 5.1 for a diagram and detailed description of the RD. We briefly review how the RD fits into the previously described CS framework, the kinds of signals it aims to acquire, and further improvements to its design.

The RD can be thought of as integrating short time windows of a pseudo-randomly modulated signal. That is, each measurement y_m before quantization can be written as

$$y_m = \int_{m/\mathcal{M}}^{(m+1)/\mathcal{M}} (p(t) + n(t))x(t)dt,\ m = 0, \ldots, M-1, \tag{5.1}$$

where $\mathcal{M}$ is the measurement rate of the RD S/H, $p(t)$ is the ideal square-waveform

$$p(t)|_{t=n/\mathcal{N}}^{(n+1)/\mathcal{N}} = p_n \in \{-1, 1\}, \tag{5.2}$$

according to the "chipping sequence" p, $n(t)$ is the error associated with practical chipping sequences, and $\mathcal{N}$ is the Nyquist rate for $x(t)$. For the purposes of this chapter, we consider a fixed time window of T seconds. We denote $N = \mathcal{N}T$ to be the number of Nyquist samples in the time window and $M = \mathcal{M}T$ to be the number of CS measurements in the time window.

For a fixed time window T, we can rewrite (5.1) as

$$y_m = \sum_{n=mN/M}^{(m+1)N/M} p_n x_n + e_m, \tag{5.3}$$

where

$$e_m = \int_{m/M}^{(m+1)/M} n(t)x(t)dt \tag{5.4}$$

since, for bandlimited $x(t)$, $x_n = \int_{n/N}^{(n+1)/N} x(t)dt$ are simply the Nyquist samples x of the signal $x(t)$. For an example of $N = 6$, $M = 3$, and a sequence $p = [-1, 1, -1, -1, 1, -1]$, if the square wave-form is ideal, i.e., $n(t) = 0$, then we have the discrete measurement matrix

$$\Phi = \begin{bmatrix} -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix} \tag{5.5}$$

such that $y = \Phi x$.

The most immediate way to apply the RD to the CS framework is via the so-called *discrete multi-tone* signal model. In this case, the signal $\mathbf{x}$ as composed of a linear combination of pure tones. Specifically, suppose that F is the $N \times N$ DFT matrix with elements

$$f_{n,\omega} = \frac{1}{\sqrt{N}} \exp\left\{-2\pi i \omega n / N\right\}, \tag{5.6}$$

where $n \in [0, N-1]$ and $\omega = 0, \pm 1, \ldots, \pm N/2 - 1$. We call F the sparsity basis. The observed signal is then represented as $\mathbf{x} = F\mathbf{s}$ and we solve for a sparse $\mathbf{s}$ from the underdetermined matrix ΦF.

In [18] it was shown that the program (BPDN) applied to recover $\mathbf{s}$ with ideal ΦF will observe similar robustness guarantees as described earlier if the number of measurements satisfies

$$M \geq CK \log^6 N, \tag{5.7}$$

with high probability depending on N and statistical properties of the chipping sequence. Further results were later shown for Φ with other sparsity bases, using analysis based on coherence [21].

A oft-noted drawback of the DFT-based model is that if the signal contains a tone that is not perfectly "on-grid," i.e., not one of the columns of the DFT matrix, then the signal will not be truly sparse. To remedy this, several algorithms have been proposed to recover non-integral tones, enhancing the richness of the discrete multi-tone model [22].

Finally, we mention one recent notable modification of the RD. In [133], the authors apply run-length limited (RLL) codes to generate the chipping sequence. These codes enable a potentially slower maximum chipping rate than the conventional RD since the sequences avoid consecutive sign changes.

5.1.1 Drawbacks of the RD

A non-ideal chipping waveform alone presents a serious problem for the RD. The measurement error e_m in (5.4) is dependent on the input signal and thus, if $n(t)$ were modelled as random, then it will still be correlated to the input. However, in practice $n(t)$ is likely to be dependent on the waveform $p(t)$, and thus the measurement error is dependent on both the input signal and the chipping sequence that determines the measurement matrix Φ. Furthermore, we have that

$$\|e\|_2^2 \leq \|x\|_2^2 \int_1^N n(t)dt, \tag{5.8}$$

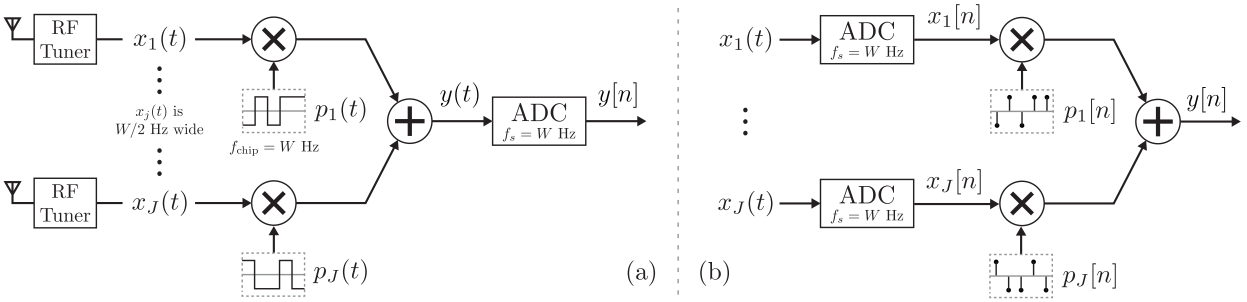


Figure 5.3: (a) CMUX system diagram. Each of the J channels is spread by a different chipping sequence, and then summed and sampled. (b) CMUX equivalent system. The sampling operation is moved to the front of the system for the sake of analysis.

implying that, in the worst case, the measurement error can have more energy than the input signal itself. Luckily this case only occurs when $n(t) = \nu x(t),\ \nu \in \mathbb{R}$, which should be unlikely in an appropriately designed system.

To make matters worse, in practice the integrator will be implemented as a low pass filter, in which case the rows of the matrix (5.5) should be the modulated, sampled impulse response of the filter. Accurate models of the filter impulse response can be difficult to obtain in practice, hard to calibrate for because such devices fluctuate under different temperatures, and computationally inefficient.

5.2 The Compressive Multiplexer (CMUX)

We now introduce the CMUX for acquisition of multi-channel signals [20]. By multi-channel signals we mean that we have multiple disjoint sections of a signal that we wish to measure simultaneously. Each section of the signal alone may not be sparse, our only requirement is that the total signal formed by appending each of the channels to each other is sparse. As an example, consider multiple (potentially discontiguous) channels in the RF spectrum. We may find that some channels contain no energy while others do. Our goal is to measure each of these channels together and recover the signals that are present.

The CMUX acquires J independent signal channels, each of bandwidth $W/2$ Hz, into a single stream of samples running at the Nyquist rate (W Hz) of any one channel. As shown in Figure 5.3(a), each channel is first mixed down to baseband to obtain $x_j(t)$ and then modulated by a pseudo-random ± 1 chipping sequence $p_j(t)$ with chipping frequency W Hz. The spread channels are then summed and sampled once per chip by a single ADC. It is important to note that the summation occurs over the channels and not over time (in contrast to previous systems [18, 47, 49]).

Without loss of generality, the CMUX can be written as a $W \times JW$ matrix Φ, formed by concatenating diagonal $W \times W$ submatrices $\Phi_j,\ j = 1, \cdots, J$. For the sake of analysis, we will consider the elements along the diagonals to be ± 1 Rademacher variables. As an example, let $J = 3$

```
Algorithm 6: Trivial Reconstruction
s1 Demodulate channel $i$
     Set $\widehat{x}_i = \Phi_i y$
```

and $W = 3$. Then the Φ matrix might look like

$$\Phi = \left[\underbrace{\begin{matrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{matrix}}_{\Phi_1} \underbrace{\begin{matrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{matrix}}_{\Phi_2} \underbrace{\begin{matrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{matrix}}_{\Phi_3}\right] \tag{5.9}$$

We consider signals that are *jointly sparse* over the combined bandwidth of the spectrum channels. The sparsity basis Ψ for this model is a $JW \times JW$ block diagonal matrix with $W \times W$ DFT bases along the diagonal. Thus, we aim to recover a K-sparse vector $\alpha \in \mathbb{R}^{JW}$ such that $y = A\alpha$, where A is the union of orthonormal bases

$$A = [\Phi_1\mathcal{F}, \Phi_2\mathcal{F}, \cdots, \Phi_J\mathcal{F}], \tag{5.10}$$

and where $\mathcal{F}$ is the $W \times W$ unitary DFT matrix[2]. For the remainder of this section, the subscript j denotes the submatrix (or subvector) corresponding to channel j and the subscript $\backslash j$ denotes the submatrix (or subvector) corresponding to all channels except for j.

It has recently been demonstrated by Romberg that A of this form satisfy the RIP [134]. We present a modified version of the statement of the theorem (as suggested in [134]) for completeness:

Theorem 9 (Theorem 3.1 in [134])**.** *Let A be defined as in (5.10), and fix $\delta \in (0,1)$. Then there exists C_0 such that when*

$$W \geq C_1 K \log^4(JW) \tag{5.11}$$

A satisfies the RIP of order K as in (1.3) with probability $1 - C_0^2/\delta C_1^2$, where C_0 is constant.

Note that the constant C_0 is the same as that in [134], and improved bounds on the probability may be obtained [135]. It is clear from this statement that for the total bandwidth $N = JW$, the number of possible channels can be upper bounded as $J \leq \frac{N}{K}\frac{1}{C_1 \log^4 N}$.

5.2.1 Custom CMUX reconstruction algorithms

Trivial reconstruction We can trivially produce an approximate recovery of any input channel i by multiplying that channel's chipping sequence against the output samples, i.e., $\widehat{x}_i = \Phi_i y$. It is clear from

$$\widehat{x}_i = \Phi_i y = \Phi_i \left(\sum_j \Phi_j x_j\right) = x_i + \sum_{i \neq j} \Phi_i \Phi_j x_j, \tag{5.12}$$

[2] Note that this is not the same A as in Chapter 3.

Algorithm 7: Block Coordinate Relaxation (BCR) [136]

```
S0 Initialize
     Set $r = y$
     Set initial estimate $\alpha = 0$
   while not converged do
S1     Choose a new block
         Block index $j \in \{1, \cdots, J\}$
S2     Subtract current estimate contribution (except from current block)
         Compute $r = y - A_{\setminus j}\alpha_{\setminus j}$
S3     Update the current block coefficients
         Via soft-thresholding $\alpha_j = \mathcal{S}(A_j^T r)$, where $\mathcal{S}(z) = z(|z| - \lambda)_+/|z|$
```

that this approach yields the original channel x_j, plus a noise term that is the sum of the other channels spread by a new ±1 sequence $\Phi_i\Phi_j$. Note that exact recovery is achieved when all non-zero coefficients are in a single channel and there are no noise sources. Furthermore, the trivial reconstruction can either be used by algorithms resilient to noise (correlation routines, PLLs, etc.). This one step algorithm is summarized in Algorithm 6.

Block coordinate relaxation (BCR) The trivial reconstruction technique can be extended to perform joint reconstruction of all channels. One approach would be to approximate one channel as above, transform and threshold to keep the largest coefficients, subtract that channel's contributions from the measurements, and repeat this process with the other channels. Indeed, this is roughly the procedure of *block coordinate relaxation* (BCR) [136].

BCR provably solves the (LASSO) program when A is a union of orthonormal bases. We initialize by setting $r = y$ and the initial estimate $\alpha = 0$. The remaining steps are as follows: *i*) Choose a new block $j \in \{1, \cdots, J\}$. *ii*) Subtract the contribution of the current estimate (except from current block) from the measurements to update the residual, $r = y - A_{\setminus j}\alpha_{\setminus j}$. *iii*) Update the current block coefficients by soft-thresholding the DFT coefficients of the trivial reconstruction, $\alpha_j = \mathcal{S}(A_j^T r)$, where $\mathcal{S}(z) = z(|z| - \lambda)_+/|z|$, element-wise. The BCR algorithm is summarized in Algorithm 7.

Note that for the CMUX, BCR uses exactly J FFTs and one soft-threshold operation of dimension W per iteration. Most other CS algorithms compute $A^T(y - Ax)$ in each iteration; thus these algorithms will require *at least* twice as many FFTs per iteration. Furthermore, the total number of iterations in BCR can be reduced by adaptively adjusting λ [137].

The soft-thresholding step of BCR projects the current channel estimate onto the ℓ_1-ball, thus "sparsely approximating" or "denoising" the channel estimate. We can extend this algorithm to recover non-integral frequencies, i.e., those oustide of the set of frequencies defined by the length-W DFT, by employing spectral CS [22] or BPDN-analysis [138] in place of soft-thresholding.

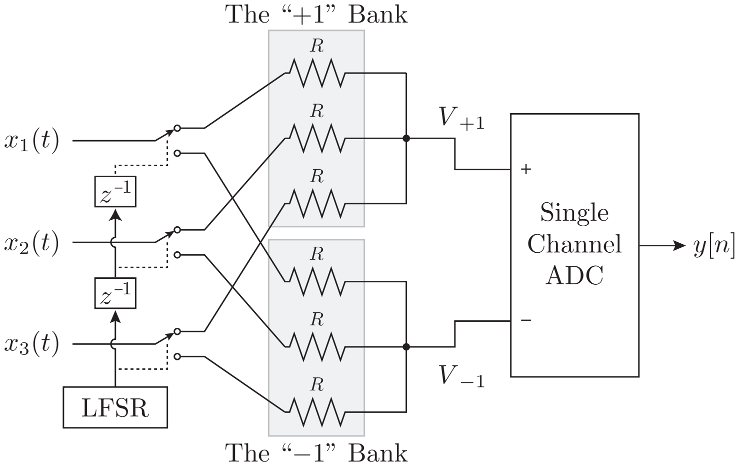


Figure 5.4: Passive Averager CMUX (PA-CMUX) for $J = 3$.

5.2.2 The passive averager: A CMUX hardware concept

A major goal in implementing randomized CS hardware is to reduce the possible sources of hardware noise (i.e., achieve a simple design) so that it does not obscure the benefits achieved through sample rate reduction. To this end, we propose the Passive Averager CMUX (PA-CMUX).

As depicted in Figure 5.4, the PA-CMUX uses a single linear feedback shift register (LFSR) to generate the chipping sequence, J analog switches, two banks of resistors, and a single-channel ADC to achieve the chipping sequence multiply and the instantaneous sum. The J uncorrelated chipping sequences are formed from delays of a single chipping sequence. Depending on the sign of the chipping sequence applied to it, each input signal $x_j(t)$ is routed by an analog switch to either a "+1" or "−1" bank of resistors. Each resistor bank consists of J resistors of nominally the same resistance; in practice, discrete resistors are unnecessary as each switch has a controlled output impedance.

Using Kirchoff's voltage and current laws, the voltage output from each bank, V_{+1}, V_{-1}, equals the average of the voltages that are fed into the bank, hence, inducing passive averaging. These voltages can be written as

$$V_{+1}[n] = \frac{\sum_{j\in P_{n,+1}} x_j[n]}{|P_{n,+1}|}, \quad V_{-1}[n] = \frac{\sum_{j\in P_{n,-1}} x_j[n]}{|P_{n,-1}|},$$

where $P_{n,+1}$ and $P_{n,-1}$ are the sets of channel indices being routed to the $+1$ bank or -1 bank at sample index n, respectively. Thus, we must rescale these voltages to obtain equivalent CMUX samples:

$$y[n] = |P_{n,+1}|\, V_{+1}[n] - |P_{n,-1}|\, V_{-1}[n].$$

While this could be achieved with two ADCs and gain components, we can also invert the averaging scale factors not during measurement, but during reconstruction. We simply compute the difference of the two averages, $y[n] = V_{+1}[n] - V_{-1}[n]$, and apply the averaging weights in Φ during reconstruction. The system designer can also calibrate the system by measuring the actual resistances along each signal pathway; these non-ideal values turn our averages into weighted averages.

The PA-CMUX relies on the fact that a single-channel ADC natively computes the difference between two voltages. In a typical setup, one of these voltages would be ground. However, in the PA-CMUX we can use it to compute the difference between the two averages.

5.2.3 Comparison with random demodulator

The CMUX compares favorably against the RD [18] on a number of fronts. CMUX computational models are more accurate and easier to calibrate due to absence of the analog filter. Additionally, since summation is performed over the channels and does not take place over time, the summation hardware is simpler. We refer the reader to Figure 5.2 for a depiction comparing the integration schemes in the RD and the CMUX. Additionally, this sampling characteristic translates to relaxing the requirements on individual CMUX hardware components between samples (e.g. switching times); the RD's integrator enforces strict time-oriented performance requirements. The ability to bandpass sample significantly reduces the chipping and sampling frequencies for the CMUX as opposed to the RD. Even when ignoring the bandpass sampling issue, the CMUX's chipping sequences operate at a lower rate than the RD for the same total bandwidth. As power requirements for these components typically rise with the square of the frequency, a meaningful savings is achieved.

The multi-channel nature of the CMUX also brings benefits. The CMUX can grow its total bandwidth by adding channels without increasing the chipping and sampling rates. In RF scenarios, splitting the CMUX's target bandwidth across multiple RF tuners matches the fact that commercially-available tuners don't produce arbitrarily large bandwidths. And with access to multiple independent tuners, the CMUX can also allocate its bandwidth capacity where it is needed in the spectrum. The CMUX can also turn off unoccupied channels to improve performance; at an extreme, the CMUX reverts to a Nyquist sampler when all but one input channel is disabled.

There are of course some disadvantages. The CMUX undersampling factor is restricted more than in the RD. This factor is fixed at $J - J_{\text{off}}$, where J_{off} is the number of disabled input channels. Also, non-idealities inherent to the RF tuners (or equivalent) means that signals can fall out of coverage at channel edges.

5.3 The Polyphase Random Demodulator (PRD)

We can add additional components to the CMUX to extend its behavior to be the same as the RD. A key insight to our design is that while ADC comprises both a sample-and-hold (S/H) step (discretization in time) and a quantization step (discretization in amplitude), the former requires significantly simpler hardware and consumes less power than the latter. Additionally, S/H can be performed with high precision at high speeds while quantization cannot [7]. Our PRD design employs a bank of parallel S/H components at an early stage, enabling precise preconditioning of discrete-time analog-valued signals and then employs a single quantizer that operates at a sub-

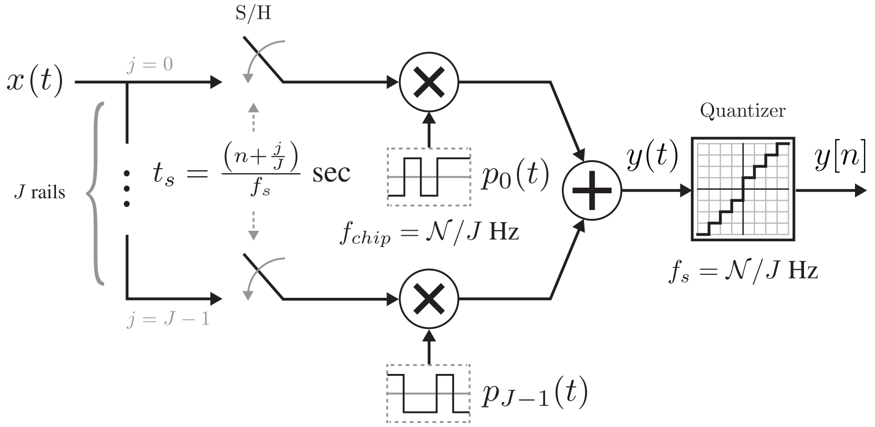


Figure 5.5: The polyphase random demodulator (PRD). J Sample-and-hold (S/H) circuits sample the signal $x(t)$ at time t_s. The outputs of the S/H circuits are each modulated by a different pseudo-random "chipping sequence" (analog waveform) $p_j(t)$, combined, and quantized by a single quantizer. All components operate at the sub-Nyquist rate of $\mathcal{N}/J$ where $\mathcal{N}$ is the Nyquist rate of $x(t)$.

Nyquist rate.

We briefly describe the new architecture, depicted in Figure 5.5. The signal analog $x(t)$ is input into a bank of J S/H "rails" (previously called channels in the CMUX context). Each S/H circuit samples the signal at a rate of $\mathcal{N}/J$ Hz, where $\mathcal{N}$ Hz is the Nyquist rate for $x(t)$. Furthermore, the j-th rail's S/H circuit is delayed by $j/\mathcal{N}$ from the sampling rate clock, where $j = 0, \cdots, J-1$ corresponds to each rail. For a given rail, each held value is scaled by +1 or −1, according to a pseudo-random "chipping" sequence $p_j(t)$. In practice $p_j(t)$ is an analog waveform that approximates a square wave. The rails are then summed together and the result is quantized, again at a rate of $\mathcal{N}/J$ Hz.

As an example, for $N = 6$ Nyquist samples in a time window and $J = 2$ S/H rails, the resulting Φ matrix might look like

$$\Phi = \begin{bmatrix} -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix}. \tag{5.13}$$

Thus, the computational model Φ of the PRD is identical to that of the RD with an ideal square-wave and ideal integrator, and thus satisfies the theoretical guarantees of [18, 21].

Note that starting from the first column, every J-th column corresponds to the first S/H rail (and similarly, starting from the second column, every J-th columns corresponds to the second S/H rail). Since the summation is taken across the rails at a given time instant, the PRD avoids integrating non-ideal fluctuations in the waveforms of the chipping sequences. Because of this, the calibration process only requires that we look at the scalings of the outputs of each rail, a significant improvement over previous designs. Furthermore, since no analog filter is used, we do not need to model an impulse response, leading to both accurate software modeling and significant speedups in recovery computations.

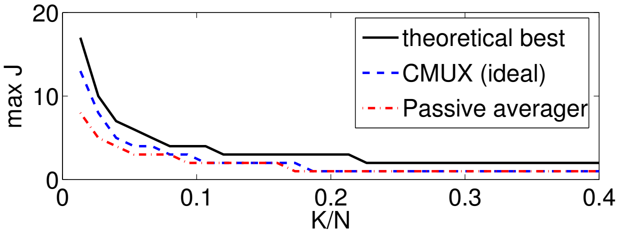


Figure 5.6: Maximum number of channels J for fixed bandwidth $N = JW = 5000$.

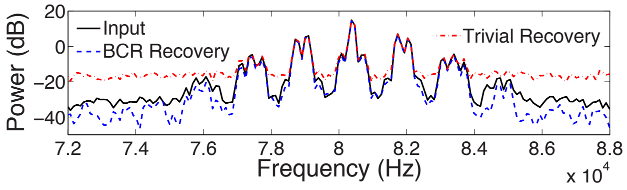


Figure 5.7: Power spectral density of one 12 kHz-wide signal.

5.4 Simulations

5.4.1 Exactly sparse recovery (CMUX)

We wish to characterize the maximum number of channels required for exact recovery of sparse signals in simulation and compare this with the theoretical bound in Section 5.2. We fix $N = JW = 5000$ and vary K/N between 0 and 0.3. We perform 1000 reconstructions using SPGL1 [28] for each choice of J and record the maximum J such that 90% of the reconstructions yielded exact recovery.

Figure 5.6 demonstrates the results of this experiment. The dashed line depicts the experimental performance for an ideal CMUX given by (5.9); the dash-dotted line depicts the performance of the PA-CMUX given in Section 5.2.2 with resistors deviating randomly up to 20% from their intended values. The solid depicts the curve $J = N/(K \log(N/K))$ exactly, the best possible performance for a CS system, (note that this is better than the bound given in Section 5.2). This simulation demonstrates that in both cases, typical CMUX behavior is close to an ideal CS system thus appears to outperform the theoretical guarantees.

5.4.2 Practical RF example (CMUX)

In this example we simulate two FM modulated voice signals, each approximately 12 kHz wide. The voice signals live in two different 400 kHz wide channels. There are 5 total input channels, making the total observed bandwidth 2 MHz. All channels have noise such that the voice signals have an SNR of 30 dB.

PSDs of the signal in one channel are shown in Figure 5.7. The original signal is depicted by solid lines, the trivial recovery is depicted by a dash-dotted line, and the CS recovery is performed with

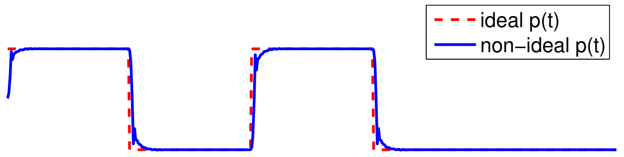


Figure 5.8: Simulated analog $p(t)$ and non-ideal waveform with an SNR of 13 dB.

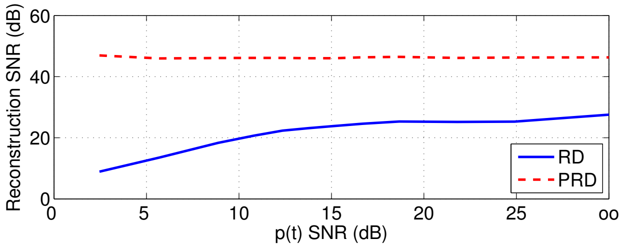


Figure 5.9: Reconstruction performance of RD vs. the PRD. The RD significantly degrades as a function of the chipping waveform $p(t)$ SNR, while the PRD is not affected for the parameters tested.

BCR as described earlier. The plot demonstrates that the largest energy portion of the spectrum can be recovered, even by the trivial method. However, the BCR recovery is significantly more accurate and has a lower noise floor than even the original signal (due to its sparse approximation feature).

5.4.3 Effects of a non-ideal modulator waveform (PRD)

To demonstrate the power of the PRD design over the RD, we conducted a simple experiment highlighting only the effects of a non-ideal chipping waveform. The experiment was performed as follows. We chose a length $N = 5000$ signal x that was $K = 50$ sparse in the discrete cosine transform (DCT) domain. We oversampled this signal by $R = 100$ times to simulate an analog waveform $x(t)$. Similarly, we generated a chipping sequence of length 5000 and oversampled this signal by R as well (obtaining a square wave $p(t)$ at the high rate). We then simulated a non-ideal square wave by applying a causal low pass filter to band-limit the wave. An example non-ideal $p(t)$ is given in Fig. 5.8. Finally, we compute the measurements that both the RD and the PRD would produce given $x(t)$ and $p(t)$ with 10 times undersampling (i.e., $J = 10$ rails). For the PRD, we chose the "instantaneous" sample point to be at the center of the square-wave. Reconstruction of $\widehat{x}$ was performed using SPGL1 [28] and results are reported in terms of $\text{SNR}(z, \widehat{z}) := 20\log_{10}(\|z\|_2/\|z - \widehat{z}\|_2)$. To gauge the fidelity of the analog chipping sequence, we also report $p(t)$ in terms of SNR.

The results of this experiment are depicted in Figure 5.9. There are two clear messages from this example. First, even when the chipping sequence is ideal ($p(t)$ SNR $= \infty$), the PRD outperforms the RD. This is likely an artifact of the simulation since, integrating any R values that simulate $x(t)$ is not equivalent to the corresponding value in x. Second, the reconstruction performance of the RD significantly degrades as SNR of $p(t)$ decreases. In stark contrast, the performance of the

PRD is seemingly not dependent on the the SNR of $p(t)$ as simulated. This is because for the parameters tested, the center of the chip is still ± 1. This highlights the primary advantage of the instantaneous sampling of the PRD.

CHAPTER 6

Discussion

6.1 Summary

Quantization is the primary bottleneck in analog-to-digital conversion. Specifically, as the bit-depth of the quantizer increases, the sampling rate of the ADC must decrease. In this thesis we set out to exploit this relationship between sampling rates and quantizer bit-depth. We demonstrated that while conventional CS dogma espouses that sampling rates be decreased to ease the burden of the ADC, it is also possible to decrease the bit-depth for the same purposes, and in some cases the latter approach will perform better than the former. To review, we have shown the following facts.

We first verified that conventional CS does obtain increased the dynamic ranges versus conventional ADCs. We introduced a deterministic definition of dynamic range and demonstrated that it is meaningful. We further showed that CS systems effectively have the same dynamic range as conventional sampling systems. Thus, since CS enables sampling at lower rates, a high bit-depth quantizer can be employed to increase the overall dynamic range of the system. Our results are general for any CS systems that have the RIP.

We next demonstrated that it is possible to quantize CS measurements to only a single bit, representing their signs. In this case the quantizer is reduced to a simple comparator enabling extremely fast sampling speeds. To develop robust reconstruction guarantees we introduced the binary ϵ-stable embedding property of 1-bit CS systems. This property explains that Hamming distances between the binary measurements are approximately the same as angular distances between the input vectors. We proved that random matrices, specifically those with elements drawn from a Gaussian distribution and those drawn with rows uniformly from the unit sphere, enable this property. We further developed two new algorithms to solve the reconstruction problem and demonstrated their feasibility in simulation. Finally, we connected the 1-bit CS to the context of saturation and dynamic range in CS systems and introduced a method for saturation-agnostic CS.

We next showed that in some situations it may be more beneficial to compress via the quantizer and sample at slightly higher rates. In particular we found two fundamental CS regimes, the measurement compression (MC) regime and the quantization compression (QC) regime . The former is defined by scenarios where measurements are expensive and input signal noise is low. The latter is defined by scenarios where sampling rates are cheap, the quantization process is expensive, and there is high input signal noise. A key realization is that the inverse relationship between sampling rate and quantizer bit-depth enables the development of practical hardware systems that take advantage of the QC regime.

Finally, we developed two new CS architectures for practical signal acquisition. In particular we developed the compressive multiplexer (CMUX) for multichannel signal acquisition and a new random demodulator, the polyphase random demodulator (PRD). The latter architecture explicitly takes advantage of the fact that the quantizer is the main bottleneck in CS systems; the sample-and-hold hardware is separated from the quantizer and thus discrete-in-time measurements (but not in amplitude) are manipulated via analog "pre-processing" before finally being digitized.

6.2 Open Questions and Future Directions

The research contained in this volume open several interesting questions that may be useful to answer in the future.

6.2.1 Does the RIP of a matrix imply the BϵSE for that matrix?

Matrices that have the RIP have been studied extensively since its introduction in CS. Thus, an obvious desire would be for many of these sensing systems to also be used as 1-bit CS systems. While it may be true that many of them can, we can point to at least two counterexamples of matrices that have the RIP but not the BϵSE.

The most trivial example is the identity matrix. To see this consider the length-N vectors $x = [1, 0, \ldots, 0]^T$ and $s = [\sqrt{1-\nu^2}, \nu, 0, \cdots, 0]^T$ and let $\Phi = I_N$ be the $N \times N$ identity matrix. By definition, the identity matrix satisfies the RIP since it is an exact isometry; indeed, it is square and thus not even compressive. Furthermore, it is clear that $\Phi x = x$ and $\Phi s = s$. Without loss of generality, define sign(0) = 1. Then note the following relationship

$$d_H(A(x), A(s)) = \begin{cases} 0, & 0 \leq \nu \leq 1, \\ 1/N, & -1 \leq \nu < 0. \end{cases} \tag{6.1}$$

In words, the Hamming distance between the measurements of the sparse signals only depends on the sign of the second element of s, meaning that a large number of angles between x and s will yield the same Hamming distance. Indeed, it is also possible to construct many orthogonal vectors to x satisfying the same relationship. In fact, even if s were designed to be parallel to x, the Hamming distance of the measurements would only depend on the relative direction.

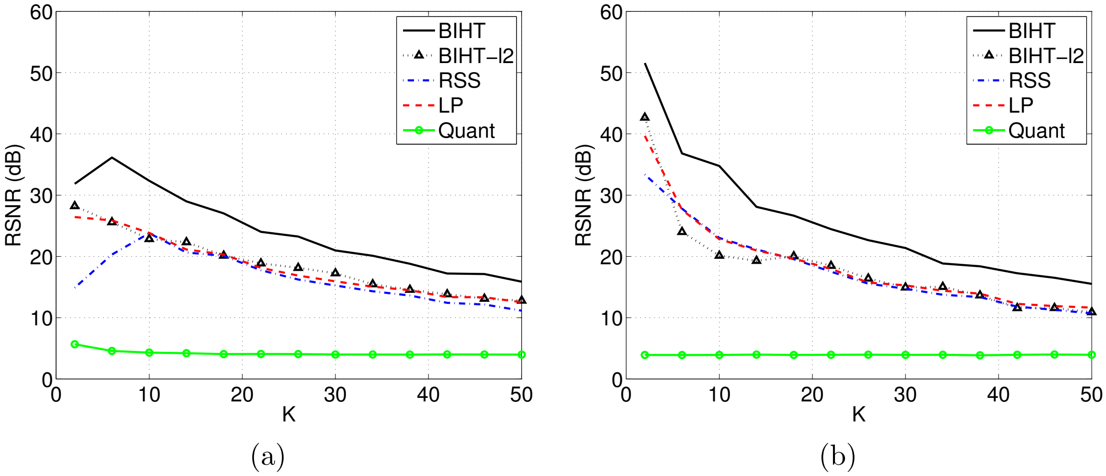


Figure 6.1: (a) Reconstruction SNR as a function of sparsity 1-bit quantized Nyquist samples of a DCT-sparse signal. (b) Reconstruction SNR as a function of sparsity from 1-bit quantized Nyquist samples measured with an $N \times N$ Gaussian matrix. The plots demonstrate that the average reconstruction between the two scenarios is similar, suggesting that the DCT matrix provides something like a weak average-case BϵSE.

A second example is matrices with ± 1 entries, such as those with entries drawn from the Rademacher distribution, i.e., $+1$ or -1 with probability $1/2$. The Rademacher distribution is subGaussian and thus can be shown to have the RIP with high probability [46]. In this case suppose that we choose $x = [\cos(\nu), 0, \ldots, 0]^T$ and $s = [\cos(\nu), \sin(\nu), 0, \cdots, 0]^T$. Then for the hamming distance between $\text{sign}(\Phi x)$ and $\text{sign}(\Phi s)$ to be non-zero, the column $\sin(\nu)\phi_2$ must have elements at least as big as $\cos(\nu)$. Thus, if $\nu < \pi/4$ and thus $\sin(\nu) < \cos(\nu)$, we have

$$d_H(A(x), A(s)) = 0, \tag{6.2}$$

for any ± 1 matrix. This relationship will be true no matter how large we increase M, meaning that the precision of the signals preserved by this mapping cannot be increased. We have found in simulation that indeed Rademacher matrices do not support good performance of reconstruction of canonically sparse signals.[1]

6.2.2 Does the Fourier basis provide a BϵSE?

Although the identity matrix does not provide a BϵSE, it may be possible that other orthonormal bases do. For example, recent simulations suggest that if the $N \times N$ DFT matrix may indeed provide a BϵSE or some weaker notion of this property with a randomized signal model.

To test this hypothesis, we performed an experiment where we generated DCT-sparse signals (i.e., Φ is the DCT basis, the signal we wish to recover x is sparse) and quantized to 1-bit per Nyquist sample. We then performed reconstruction using the 1-bit CS algorithms described in Chapter 3.

[1] A nearly identical example was discussed in [108] which appeared during the final preparation of this manuscript.

The results of this experiment are depicted in Figure 6.1(a). For comparison we provide the same experiment but with an $N \times N$ Gaussian matrix Φ in Figure 6.1(b).

This simulation suggests that we achieve the same performance as if we had used a Gaussian measurement system, however by quantizing the Nyquist samples directly. This implies that for some classes of signals, we maybe be able to perform 1-bit quantization directly. Indeed, approaching this problem from the 1-bit CS perspective may yield theoretical guarantees for previous 1-bit ADC architectures such as some of those in [120].

APPENDIX A

Dynamic Range

A.1 Lemma 1

Proof. Begin by taking $\beta = G/\left\|x\right\|_\infty$. Observe that

$$\left\|\beta x\right\|_\infty = \beta \left\|x\right\|_\infty = G,$$

and thus no entries of βx exceed the saturation level G. Hence, we can bound the quantization error as

$$\left\|\beta x - \mathcal{Q}_B\left(\beta x\right)\right\|_2^2 \leq N\left(\frac{\Delta}{2}\right)^2. \tag{A.1}$$

We also have that

$$\left\|\beta x\right\|_2^2 = \beta^2 \left\|x\right\|_2^2 = \frac{G^2 \left\|x\right\|_2^2}{\left\|x\right\|_\infty^2}. \tag{A.2}$$

Combining (A.1) and (A.2), we obtain that

$$\mathrm{SQNR}(\beta x) = \frac{\left\|\beta x\right\|_2^2}{\left\|\beta x - \mathcal{Q}_B\left(\beta x\right)\right\|_2^2} \geq \frac{G^2 \left\|x\right\|_2^2 / \left\|x\right\|_\infty^2}{N\left(\Delta/2\right)^2},$$

which simplifies to yield the desired result. □

A.2 Theorem 2

Proof. We begin by considering $\beta_C^{\min}(x)$. Recall that for all scalings $\beta < G/\left\|x\right\|_\infty$ we have that $\left\|\beta x\right\|_\infty < G$, so that there are no saturations. Thus we can bound the SQNR as

$$\mathrm{SQNR}(\beta x) \geq \frac{\beta^2 \left\|x\right\|_2^2}{N(\Delta/2)^2}.$$

Thus, if we ensure that

$$\frac{\beta^2 \left\|x\right\|_2^2}{N(\Delta/2)^2} \geq C$$

then we also guarantee that $\mathrm{SQNR}(\beta x) \geq C$. This will occur provided that

$$\beta^2 \geq \frac{CN}{\left\|x\right\|_2^2}(\Delta/2)^2.$$

Thus we can set

$$\beta_C^{\min}(x) = \sqrt{\frac{CN}{\left\|x\right\|_2^2}(\Delta/2)^2}.$$

We now turn to $\beta_C^{\max}(x)$. Since we are now considering $\beta > G/\left\|x\right\|_\infty$, there will be at least one entry of x that takes a value greater than G and thus saturates. Furthermore, the saturated value is guaranteed to have error greater than $\Delta/2$ since our quantizer represents a maximum value of $G - \Delta/2$. Thus, we observe that the total quantization error is less than the error of a signal where each element takes the value of the maximum saturated measurement. If we define $\overline{G} = G - \Delta/2$ then we have that

$$\mathrm{SQNR}(\beta x) \geq \frac{\beta^2 \left\|x\right\|_2^2}{N(\beta\|x\|_\infty - \overline{G})^2}. \tag{A.3}$$

By design, we have that $\beta\|x\|_\infty > G$, and hence

$$\begin{aligned}\left(\beta\|x\|_\infty - \overline{G}\right)^2 &= \beta^2\|x\|_\infty^2 - 2\overline{G}\beta\|x\|_\infty + \overline{G}^2 \\ &\leq \beta^2\|x\|_\infty^2 - 2\overline{G}G + \overline{G}^2 = \beta^2\|x\|_\infty^2 - G^2 + (\Delta/2)^2.\end{aligned}$$

From this we observe that

$$\frac{\beta^2 \left\|x\right\|_2^2}{N(\beta\|x\|_\infty - \overline{G})^2} \geq \frac{\beta^2 \left\|x\right\|_2^2}{N(\beta^2\|x\|_\infty^2 - G^2 + (\Delta/2)^2)},$$

and so from (A.3) we have that if

$$\frac{\beta^2 \left\|x\right\|_2^2}{N(\beta^2\|x\|_\infty^2 - G^2 + (\Delta/2)^2)} > C$$

then $\mathrm{SQNR}(\beta x) > C$. By rearranging, we see that this will occur provided that

$$\beta^2 < \frac{CN}{\|x\|_2^2}\left(\frac{G^2 - (\Delta/2)^2}{C\gamma(x)^2 - 1}\right).$$

Thus we can set

$$\beta_C^{\max}(x)^2 = \frac{CN}{\|x\|_2^2}\left(\frac{G^2 - (\Delta/2)^2}{C\gamma(x)^2 - 1}\right).$$

Combining our expressions for $\beta_C^{\min}(x)$ and $\beta_C^{\max}(x)$ we obtain

$$\mathrm{DR}_C(x) \geq \left(\frac{\beta_C^{\max}(x)}{\beta_C^{\min}(x)}\right)^2 = \frac{1}{C\gamma(x)^2 - 1}\left(\left(\frac{2G}{\Delta}\right)^2 - 1\right),$$

which simplifies to establish (2.9). □

APPENDIX B

Binary Stable Embeddings

B.1 Lemma 5: Intersections of Orthants by Subspaces

In this section, we demonstrate that while there are 2^M available quantization points provided by 1-bit measurements, a K sparse signal will not use all of them. To understand how effectively the quantization bits are used, we first need to investigate how the K-dimensional subspaces projected from the N-dimensional K-sparse signal spaces intersect orthants in the M-dimensional measurement space.

An orthant in M dimensions is a set of points in $\mathbb{R}^M$ that all have the same sign pattern:

$$\mathcal{O}_s = \{x \mid \operatorname{sign} x = s\},$$

where s is a vector of ± 1. Each orthant has M boundaries of dimension $M-1$, defined as the subspace with a coordinate set to 0:

$$\mathfrak{B}_i = \{x \mid (x)_i = 0\}.$$

We split each boundary into 2^{M-1} faces, defined as the set

$$\mathcal{F}_{i,s} = \{x \mid (x)_i = 0 \text{ and } \operatorname{sign}(x)_j = (s)_j \text{ for all } i \neq j\},$$

where s is the sign vector of a bordering orthant, and i is the boundary in which the face lies. Each face borders two orthants. Note that the faces are $M-1$ dimensional orthants in the $M-1$ dimensional boundary subspace. The geometry of the problem in $\mathbb{R}^3$ is summarized in Figure B.1(a).

We use $I(M,K)$ to denote the maximum number of orthants in M dimensions intersected by a K-dimensional subspaces (with $I(M,1) = 2$). We upper bound $I(M,K)$ using an inductive argument that relies on the following two lemmas:

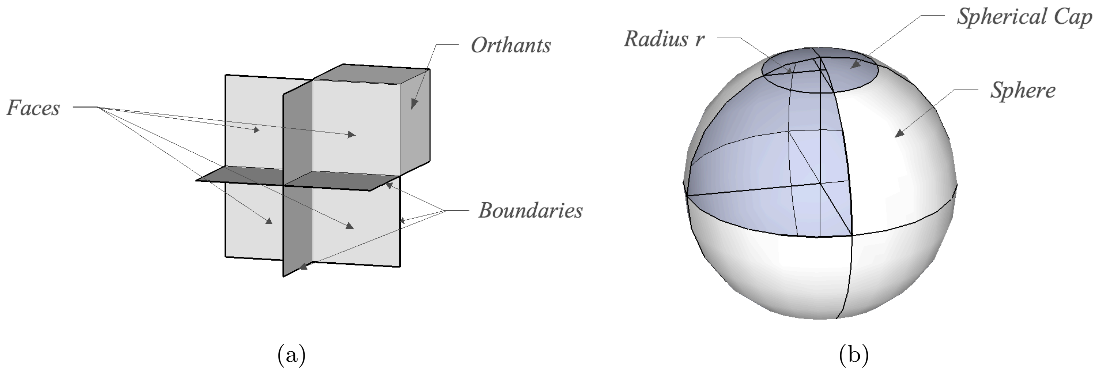


Figure B.1: (a) The geometry of orthants in $\mathbb{R}^3$. (b) The geometry of spherical caps.

Lemma 12. *If a K-dimensional subspace $\mathcal{S} \subset \mathbb{R}^M$ is not the subset of a boundary $\mathfrak{B}_i$, then the subspace and boundary do intersect and their intersection is a $K-1$ dimensional subspace of $\mathfrak{B}_i$.*

Proof. We count the dimensions of the relevant spaces. If $\mathcal{S}$ is not a subset of $\mathfrak{B}_i$, then it equals the direct sum $\mathcal{S} = (\mathcal{S} \cap \mathfrak{B}_i) \oplus \mathcal{W}$, where $\mathcal{W} \subset \mathbb{R}^M$ is also not a subspace of $\mathfrak{B}_i$. Since $\dim \mathfrak{B}_i = M-1$, $\dim \mathcal{W} \leq 1$, and $\dim \mathcal{S} \cap \mathfrak{B}_i = K-1$ follows. □

Lemma 13. *For $K > 1$, a K-dimensional subspace that intersects an orthant also non-trivially intersects at least K faces bordering that orthant.*

Proof. Consider a K-subspace $\mathcal{S}$, a point $p \in \mathcal{S}$ interior to the orthant $\mathcal{O}_{\operatorname{sign} p}$, and a vector $x_1 \in \mathcal{S}$ non-parallel to p. The following iterative procedure can be used to prove the result:

1. Starting from 0, grow a until the set $p \pm a x_l$ intersects a boundary $\mathfrak{B}_i$, say at $a = a_l$. It is straightforward to show that as a grows, a boundary will be intersected. The point of intersection is in the face $\mathcal{F}_{i,\operatorname{sign} p}$. The set $\{p \pm a x_l | a \in (0, a_l)\}$ is in the orthant $\mathcal{O}_{\operatorname{sign} p}$.
2. Determine a vector $x_{l+1} \in \mathcal{S}$ parallel to all the boundaries already intersected and not parallel to p, set $l = l+1$ and iterate from step 1.

A vector can always be found in step 2 for the first K iterations since $\mathcal{S}$ is K dimensional. The vector is parallel to all the boundaries intersected in the previous iterations and therefore $p \pm a x_l$ always intersects a boundary not intersected before. Therefore, at least K distinct faces are intersected. □

Lemmas 12 and 13 lead to the main result in this section. Lemma 5 in Section 3.4.1 follows trivially.

Lemma 14. *The number of orthants intersected by a K-dimensional subspace $\mathcal{S}$ in an M dimensional space $\mathcal{V}$ is upper bounded by*

$$I(M, K) \leq \binom{M}{K} 2^K.$$

Proof. The main intuition is that since the faces on each boundary are equivalent to orthants in the lower dimensional subspace of the boundary, the maximum number of faces intersected at each boundary is a problem of dimension $I(M-1, K-1)$.

If $\mathcal{S}$ is contained in one of the boundaries in $\mathcal{V}$, the number of orthants of $\mathcal{V}$ intersected is at most $I(M-1, K)$. Since $I(M, K)$ is non-decreasing in M and K, we can ignore this case in determining the upper bound.

If $\mathcal{S}$ is not contained in one of the boundaries then Lemma 12 shows that the intersection of $\mathcal{S}$ with any boundary $\mathfrak{B}_i$ is a $K-1$ dimensional subspace in $\mathfrak{B}_i$. To count the faces of $\mathfrak{B}_i$ intersected by $\mathcal{S}$ we use the observation in the definition of faces above, that each face is also an orthant of $\mathfrak{B}_i$. Therefore, the maximum number of faces of $\mathfrak{B}_i$ intersected is a recursion of the same problem in lower dimensions, i.e., is upper bounded by $I(M-1, K-1)$. Since there are M boundaries in $\mathcal{V}$, it follows that the number of faces in $\mathcal{V}$ intersected by $\mathcal{S}$ is upper bounded by $M \cdot I(M-1, K-1)$.

Using Lemma 13 we know that for an orthant to be intersected, at least K faces adjacent to it should be intersected. Since each face is adjacent to two orthants, the total number of orthants intersected cannot be greater than twice the number of faces intersected divided by K:

$$I(M, K) \leq \frac{2M \cdot I(M-1, K-1)}{K}. \tag{B.1}$$

The result follows by induction. □

A tighter achievable bound is also known [139, 140]:

$$\begin{aligned} I(M, K) &\leq 2\sum_{l=0}^{K-1} \binom{M-1}{l} \qquad \text{(B.2)} \\ &\leq 2K\binom{M-1}{K-1} \text{ if } K \leq \frac{M-1}{2}. \end{aligned}$$

Although (B.2) is tighter and achieved with a subspace in a general configuration, it leads to expressions on the same asymptotical order of our main results. We use (B.1) for the remainder of this chapter because of its simpler form.

B.2 Theorem 4: Distributing Signals to Quantization Points

To prove Theorem 4 we consider how the available quantization points optimally cover the set of signals of interest. We consider unit norm signals that belong in a union of L subspaces, each of dimension K. Thus the set of interest is the union of L unit spheres of K dimensions.

First we need to understand how to measure the sets of signals of interest. We denote the unit sphere in K dimensions—which is the surface of the K-dimensional unit ball—using S^{K-1}, and the rotationally invariant area measure on the sphere using $\sigma(\cdot)$. Thus the area of the whole sphere is equal to $\sigma(S^{K-1})$. If subspaces intersect, the area of the sphere inside the intersection has measure zero. Therefore, the total surface area of all L spheres is LS^{K-1}.

The most efficient cover of this area is achieved if every point covers a spherical cap of radius r, denoted using $C(r)$. The geometry of the problem is demonstrated in Figure B.1(b). From [141] the surface area of a spherical cap of radius r satisfies

$$\sigma(C(r)) \leq r^K \sigma(S^{K-1}),$$

For $L\binom{M}{K}2^K$ points to cover the area $L\sigma(S^{K-1})$ we require

$$L\binom{M}{K}2^K\sigma(C(r)) \geq L\sigma(S^{K-1}) \Rightarrow \left(\frac{Me2r}{K}\right)^K \geq 1$$
$$\Rightarrow r \geq \frac{K}{2eM} = \Omega\left(K/M\right),$$

using the bound $\binom{M}{K} \leq (eM/K)^K$. Incidentally, this proof gives an obvious solution to a Grassmanian covering problem of 1-dimensional subspaces in K dimensional spaces. Although Grassmanian packing problems have been examined in the literature (e.g., in the context of frame theory [142]), to our knowledge, the Grassmanian covering problem has not been posed or attempted.

B.3 Theorem 5: Optimal Performance via Gaussian Projections

To prove Theorem 5, we follow the procedure given in [89, Theorem 3.3]. We begin by restricting our analysis to the support set $T \subset \{1, \cdots, N\}$ with $|T| \leq D \leq N$, and thus we consider vectors that lie on the (sub) sphere $\Sigma^*(T) = \{x : \operatorname{supp} x \subset T, \|x\|_2 = 1\} \subset \mathbb{R}^N$. We remind the reader that $B_\delta(x) := \{a \in S^{N-1} : \|x - a\|_2 < \delta\}$ is the ball of unit norm vectors of Euclidean distance δ around x, and we write $B^*_\delta(x) = B_\delta(x) \cap \Sigma^*(T)$ as in Section 3.5.2.

Given a vector $\varphi \sim \mathcal{N}^{N\times 1}(0,1)$ and two distinct points p and q in Q_δ, we have that

$$\mathbb{P}\Big[\forall\, u \in B^*_\delta(p), \forall v \in B^*_\delta(q) : \operatorname{sign}\varphi^T u \neq \operatorname{sign}\varphi^T v\Big] \;\geq\; d_S(p,q) \;-\; \sqrt{\tfrac{\pi}{2}D}\,\delta,$$

from Lemma 15 (given in Section B.4). When $\epsilon_o > 2\delta$, we have the relationship

$$\pi\, d_S(p,q) \geq 2\sin(\frac{\pi}{2}\, d_S(p,q)) = \|p-q\|_2 \geq \|u-v\|_2 - 2\delta > \epsilon_o - 2\delta,$$

and thus

$$\mathbb{P}\Big[\forall\, u \in B^*_\delta(p), \forall v \in B^*_\delta(q) : \operatorname{sign}\varphi^T u \neq \operatorname{sign}\varphi^T v \mid \|u-v\|_2 > \epsilon_o\Big] \;\geq\; \tfrac{\epsilon_o}{\pi} \;-\; (\tfrac{2}{\pi} + \sqrt{\tfrac{\pi}{2}D})\,\delta.$$

By setting $\delta = \pi\epsilon_o/(4 + \pi\sqrt{2\pi D})$ (and reversing the inequality), we obtain

$$\mathbb{P}\Big[\exists\, u \in B^*_\delta(p), \exists v \in B^*_\delta(q) : \operatorname{sign}(\varphi^T u) = \operatorname{sign}(\varphi^T v) \mid \|u-v\|_2 > \epsilon_o\Big] \;\leq\; 1 - \tfrac{\epsilon_o}{2}.$$

Thus, for M different random vectors φ_i arranged in $\Phi = (\varphi_1, \cdots, \varphi_M)^T \sim \mathcal{N}^{M\times N}(0,1)$, and for the associated mapping A defined in (3.1), we have that

$$\mathbb{P}\Big[\exists\, u \in B^*_\delta(p), \exists v \in B^*_\delta(q) : A(u) = A(v) \mid \|u-v\|_2 > \epsilon_o\Big] \leq (1 - \tfrac{\epsilon_o}{2})^M.$$

In words, we have found a bound on the probability that two vectors' measurements are consistent, even if their euclidean distance is greater than ϵ_o, but only for vectors in the restricted (sub) sphere $\Sigma^*(T)$. Now we seek to cover the rest of the space Σ^*_K (unit norm K-sparse signals).

Given a radius $\delta > 0$, the sphere $\Sigma^*(T)$ can be covered with a finite set $Q_\delta \subset \Sigma^*(T)$ of no more than $(3/\delta)^D$ points such that, for any $w \in \Sigma^*(T)$, there exists a $q \in Q_\delta$ with $w \in B^*_\delta(q)$ [45]. Since there are no more than $\binom{|Q_\delta|}{2} \leq (|Q_\delta|)^2 \leq (3/\delta)^{2D}$ pairs of distinct points in Q_δ, we find

$$\mathbb{P}\Big[\exists\, u, v \in \Sigma^*(T) : d_H(A(u), A(v)) = 0 \mid \|u - v\|_2 > \epsilon_o\Big] \;\leq\; \Big(\tfrac{1}{\pi\epsilon_o}(12 + 3\pi\sqrt{2\pi D})\Big)^{2D}\; (1 - \tfrac{\epsilon_o}{2})^M.$$

To obtain the final bound, we observe that any pair of unit K-sparse vectors x and s in Σ^*_K belongs to some $\Sigma^*(T)$ with $T = \operatorname{supp} x \cup \operatorname{supp} s$ and $|T| \leq 2K$. There are no more than $\binom{N}{2K} \leq (N/2K)^{2K}$ of such sets T, and thus setting $D = 2K$ above yields

$$\begin{aligned}
&\mathbb{P}\Big[\exists\, u, v \in \Sigma^*_K : d_H(A(u), A(v)) = 0 \mid \|u - v\|_2 > \epsilon_o\Big] \\
&\leq\; (\tfrac{N}{2K})^{2K}\; (\tfrac{1}{\pi\epsilon_o}(12 + 6\pi\sqrt{\pi K}))^{4K}\; (1 - \tfrac{\epsilon_o}{2})^M \\
&\leq \exp\Big[2K \log(\tfrac{N}{2K}) + 4K \log(\tfrac{1}{\pi\epsilon_o}(12 + 6\pi\sqrt{\pi K})) - M\tfrac{\epsilon_o}{2}\Big],
\end{aligned}$$

where the second inequality follows from $1 - \frac{\epsilon_o}{2} \leq \exp\frac{\epsilon_o}{2}$. By upper bounding this probability by η and solving for M, we obtain

$$M \geq \tfrac{1}{\epsilon_o}\Big(2K\, \log\tfrac{N}{2K} + 4K \log(\tfrac{1}{\pi\epsilon_o}(12 + 6\pi\sqrt{\pi K})) + \log\tfrac{1}{\eta}\Big).$$

Since $K \geq 1$, we have that $\frac{1}{\pi}(12 + 6\pi\sqrt{\pi K}) < 12\sqrt{\pi K} < 16\sqrt{2K}$, and thus the previous relation is then satisfied when

$$\begin{aligned}
M \;&\geq\; \tfrac{1}{\epsilon_o}\Big(2K\, \log\tfrac{N}{2K} + 4K \log(\tfrac{1}{\epsilon_o}(16\sqrt{2K})) + \log\tfrac{1}{\eta}\Big) \\
&=\; \tfrac{1}{\epsilon_o}\Big(2K\, \log\tfrac{N}{2} + 4K \log(\tfrac{1}{\epsilon_o}(16\sqrt{2})) + \log\tfrac{1}{\eta}\Big) \\
&=\; \tfrac{1}{\epsilon_o}\Big(2K\, \log N + 4K \log(\tfrac{16}{\epsilon_o}) + \log\tfrac{1}{\eta}\Big).
\end{aligned}$$

B.4 Lemma 8: Concentration of Measure for δ-Balls

Proving Lemma 8 amounts to showing that, for some fixed $\epsilon > 0$ and $0 \leq \delta \leq 1$, given a Gaussian matrix $\Phi \in \mathbb{R}^{M\times D}$, the mapping $A : \mathbb{R}^D \to \mathcal{B}^M$ defined as $A(u) = \operatorname{sign}(\Phi u)$, and for some $x, s \in S^{D-1}$, we have

$$\mathbb{P}\Big(\;\Big|d_H\big(A(u), A(v)\big) \;-\; d_S(x, s)\Big| \;\leq\; \epsilon + \sqrt{\tfrac{\pi}{2}D}\,\delta\Big) \;\geq\; 1 - 2\,e^{-2\epsilon^2 M}, \quad \forall u \in B^*_\delta(x),\ \forall v \in B^*_\delta(s),$$

where the balls B_δ are also restricted to $\mathbb{R}^D$.

Given $u \in B^*_\delta(x)$ and $v \in B^*_\delta(s)$, the quantity $M d_H\big(A(u), A(v)\big)$ is the sum $\sum_i A_i(u) \oplus A_i(v)$, where $A_i(u)$ stands for the i^{th} component of $A(u)$. For one index $1 \leq i \leq M$

$$\begin{aligned}
A_i(u) \oplus A_i(v) \;&\leq\; Z_i^+ := \max\Big\{A_i(p) \oplus A_i(q) :\; p \in B^*_\delta(x), q \in B^*_\delta(s)\Big\}, \\
A_i(u) \oplus A_i(v) \;&\geq\; Z_i^- := \min\Big\{A_i(p) \oplus A_i(q) :\; p \in B^*_\delta(x), q \in B^*_\delta(s)\Big\},
\end{aligned}$$

and therefore

$$Z^- \ := \sum_{i=1}^{M} Z_i^- \ \le \ M\, d_H\big(A(u), A(v)\big) \ \le \ \sum_{i=1}^{M} Z_i^+ \ =: \ Z^+.$$

Of course, the occurrence of $Z_i^+ = 0$ ($Z_i^- = 1$) means that all vector pairs taken separately in $B^*_\delta(x)$ and $B^*_\delta(s)$ have consistent (or respectively, inconsistent) measurements on the i^{th} sensing component A_i. More precisely, since $\varphi_i \sim \mathcal{N}^{N\times 1}(0,1)$, $Z_i^\pm$ are binary random variables such that $\mathbb{P}[Z_i^+ = 1] = 1 - p_0$ and $\mathbb{P}[Z_i^- = 1] = p_1$ independently of i, where the probabilities p_0 and p_1 are defined by

$$\begin{aligned} p_0(d_S(x,s),\delta) \ &= \ \mathbb{P}[Z_i^+ = 0] \ = \ \mathbb{P}\Big[\forall p \in B^*_\delta(x), \forall q \in B^*_\delta(s),\ A_i(u) = A_i(v)\Big],\\ p_1(d_S(x,s),\delta) \ &= \ \mathbb{P}\Big[\forall p \in B^*_\delta(x), \forall q \in B^*_\delta(s),\ A_i(u) \neq A_i(v)\Big]. \end{aligned}$$

In summary, Z^+ and Z^- are binomially distributed with M trials and probability of success $1-p_0$ and p_1, respectively. Furthermore, we have that $\mathbb{E}Z^+ = M\,(1-p_0)$ and $\mathbb{E}Z^- = M\,p_1$, thus by the Chernoff-Hoeffding inequality,

$$\begin{aligned} \mathbb{P}\Big[\, Z^+ \ &> \ M\,(1-p_0) + M\epsilon\,\Big] \ \le \ e^{-2M\epsilon^2},\\ \mathbb{P}\Big[\, Z^- \ &< \ M\,p_1 - M\epsilon\,\Big] \ \le \ e^{-2M\epsilon^2}. \end{aligned}$$

This indicates that with a probability higher than $1 - 2e^{-2M\epsilon^2}$, we have

$$p_1 - \epsilon \ \le \ d_H\big(A(u), A(v)\big) \ \le \ (1-p_0) + \epsilon.$$

The final result follows by lower bounding p_0 and p_1 as in Lemma 15.

Lemma 15. *Given $0 \le \delta < 1$ and two unit vectors $x, s \in S^{D-1}$, we have*

$$p_0 = \mathbb{P}\Big[\forall u \in B_\delta(x),\ \forall v \in B_\delta(s), \operatorname{sign}\langle\varphi, u\rangle \ = \ \operatorname{sign}\langle\varphi, v\rangle\Big] \ \ge \ 1 \ - d_S(x,s) \ - \ \sqrt{\tfrac{\pi}{2}D}\,\delta, \qquad (B.3)$$

$$p_1 = \mathbb{P}\Big[\forall u \in B_\delta(x),\ \forall v \in B_\delta(s), \operatorname{sign}\langle\varphi, u\rangle \ \neq \ \operatorname{sign}\langle\varphi, v\rangle\Big] \ \ge \ d_S(x,s) \ - \ \sqrt{\tfrac{\pi}{2}D}\,\delta. \qquad (B.4)$$

Proof of Lemma 15. We begin by introducing some useful properties of Gaussian vector distribution. If $\varphi \sim \mathcal{N}^{D\times 1}(0,1)$, the probability that $\varphi \in \mathcal{A} \subset \mathbb{R}^D$ is simply the measure μ of $\mathcal{A}$ with respect to the standard Gaussian density $\gamma(\varphi) = \frac{1}{(2\pi)^{D/2}}\, e^{-\|\varphi\|^2/2}$, i.e.,

$$\mathbb{P}[\,\varphi \in \mathcal{A}\,] = \mu(\mathcal{A}) = \int_{\mathcal{A}} \mathrm{d}^D\varphi\ \gamma(\varphi),$$

with $\mu(\mathbb{R}^D) = 1$. It may be easier to perform this integration over a hyper-spherical set of coordinates. Specifically, we let any vector φ be represented by the values $(r, \phi_1, \cdots, \phi_{D-1})$ where $r \in \mathbb{R}_+$ stands for the vector length, $\phi_1, \cdots, \phi_{D-2} \in [0,\pi]$ corresponds to the vector angles in each dimension, and $\phi_{D-1} \in [0, 2\pi]$ being the angle of φ in the "xs" plane. This is possible since γ is rotionally invariant and thus we may assume the "xs" plane is spanned by the canonical vectors

$e_D = x$ and e_{D-1} in the canonical basis $\{e_1, \cdots, e_D\}$ of $\mathbb{R}^D$, with $e_1 = (x \wedge s) \,/\, \|x \wedge s\|_2$ and $e_{D-1} = e_D \wedge e_1$.

The change of coordinates is then defined as $\varphi_1 = r\cos\phi_1$, $\varphi_2 = r\sin\phi_1\cos\phi_2$, ..., $\varphi_{D-1} = r\,\sin\phi_1 \cdots \sin\phi_{D-2}\cos\phi_{D-1}$, and $\varphi_D = r\,\sin\phi_1 \cdots \sin\phi_{D-2}\sin\phi_{D-1}$, while, conversely, $r = \|\varphi\|_2$, $\tan\phi_1 = (\varphi_D^2 + \cdots + \varphi_2^2)^{1/2}/\varphi_1$, ..., $\tan\phi_{D-2} = (\varphi_D^2 + \varphi_{D-1}^2)^{1/2} \,/\, \varphi_{D-2}$, and $\tan\phi_{D-1} = \varphi_D \,/\, \varphi_{D-1}$.[1]

We now seek a lower bound on p_1. Computing this probability amounts to estimating

$$p_1 = \mathbb{P}[\,\forall u \in B_\delta(x),\ \forall v \in B_\delta(s),\ \langle\varphi, u\rangle\langle\varphi, v\rangle \le 0\,] \;=\; \mu(\mathcal{W}_\delta),$$

where $\mathcal{W}_\delta = \{\varphi : \langle\varphi, u\rangle\langle\varphi, v\rangle \le 0,\ \forall u \in B_\delta(x),\ \forall v \in B_\delta(s)\}$ is the set of all vectors φ such that its inner product with u and v result in different signs. Note that if $B_\delta(x) \cap B_\delta(s)$ is not empty, then we have $p_1 = 0$ since for $w \in B_\delta(x) \cap B_\delta(s)$, we have $\langle\varphi, w\rangle^2$. This term cannot be negative and thus $\mathcal{W}_\delta = \{\varphi : \langle\varphi, w\rangle = 0\}$, which has measure zero with respect to μ. In order to avoid this trouble, we must choose $d_S(x,s) \ge \frac{4}{\pi}\arcsin\delta/2$. Furthermore, since $\arcsin\lambda \le \frac{\pi}{2}\lambda$ for any $0 \le \lambda \le 1$, this occurs if $d_S(x,s) \ge \delta$.

The remainder of the proof is devoted to finding an appropriate way to integrate the set $\mathcal{W}_\delta$. To this end, we begin by demonstrating that estimating p_1 can be simplified with the following equivalence (proved just after the completion of the proof of Lemma 15).

Lemma 16. *The set $\mathcal{W}_\delta \subset \mathbb{R}^D$ is equal to the set*

$$\mathcal{V}_\delta^- \;=\; \{\varphi : \langle\varphi, x\rangle\langle\varphi, s\rangle \le 0, \|x \,-\, \mathcal{P}_{\Pi(\varphi)}\,x\| \ge \delta, \|s \,-\, \mathcal{P}_{\Pi(\varphi)}\,s\| \ge \delta\},$$

where $\mathcal{P}_{\Pi(\varphi)}$ is the orthogonal projection on the plane $\Pi(\varphi) = \{u \in \mathbb{R}^D : \langle\varphi, u\rangle = 0\}$.

Using the hyper spherical coordinate system developed earlier, membership in $\mathcal{V}_\delta^-$ can be expressed as

$$\varphi = (r, \phi_1, \cdots, \phi_{D-1}) \in \mathcal{V}_\delta^- \;\Leftrightarrow\; \begin{cases} \tan\phi_{D-1} \;\in\; [0, \tan\theta], & \text{(R1)} \\ \sin\phi_1 \,\cdots\, \sin\phi_{D-2}\,|\sin\phi_{D-1}| \;\ge\; \delta, & \text{(R2)} \\ \sin\phi_1 \,\cdots\, \sin\phi_{D-2}\,|\sin(\phi_{D-1} - \theta)| \;\ge\; \delta. & \text{(R3)} \end{cases}$$

Indeed, requirement (R1) enforces $\langle\varphi, x\rangle\langle\varphi, s\rangle \le 0$, while (R2) and (R3) are direct translations of the requirements that $\|x \,-\, \mathcal{P}_{\Pi(\varphi)}\,x\| \;=\; |\langle\widehat{\varphi}, x \;=\; e_D\rangle| \;\ge\; \delta$ and $\|s \,-\, \mathcal{P}_{\Pi(\varphi)}\,s\| \;=\; |\langle\widehat{\varphi}, y = -\sin\theta\, e_D + \cos\theta\, e_{D-1}\rangle| \ge \delta$, with $\widehat{\varphi} = \frac{1}{\|\varphi\|}\varphi$.

We are now ready to integrate to find p_1:

$$p_1 = \mu(\mathcal{V}_\delta^-) \;=\; \tfrac{1}{(2\pi)^{D/2}} \int_{\mathbb{R}_+} \mathrm{d}r\; r^{D-1} e^{-r^2/2} \,\Big[\Big(\int_0^\pi \mathrm{d}\phi_1\, \sin^{D-2}\phi_1\Big) \cdots \Big(\int_0^\pi \mathrm{d}\phi_{D-2}\, \sin\phi_{D-2}\Big)\Big] \cdots$$

$$\Big[\int_{[0,\theta]\,\cup\,[\pi,\pi+\theta]} \mathrm{d}\phi_{D-1}\; \chi_{g(\delta,\varphi)}(\phi_{D-1})\, \chi_{g(\delta,\varphi)}(\phi_{D-1} - \theta)\Big],$$

[1] This change of coordinates can be very convenient. For instance, the proof of Lemma 7 relies on the computation $\mathbb{P}[A_i(x) \neq A_i(s)] = \mu(\mathcal{A} = \{\varphi : \phi_{D-1} \in [0, \pi\, d_S(x,s)] \cup [\pi, \pi + \pi\, d_S(x,s)]\}) = d_S(x,s)$, since for (almost) all $\varphi \in \mathcal{A}$, x and s live in the two different subvolumes determined by the plane $\{u : \langle\varphi, u\rangle = 0\}$ [96, 97].

with $\chi_\lambda(\phi) = 1$ if $|\sin\phi\,| \geq \lambda$ and 0 else, for some $\lambda \in [0,1]$, and $g(\delta,\varphi) = \delta/(\sin\phi_1\,\cdots\,\sin\phi_{D-2})$.

However,

$$\int_{[0,\theta]\,\cup\,[\pi,\pi+\theta]} \mathrm{d}\phi\;\chi_\lambda(\phi)\,\chi_\lambda(\phi-\theta) \;=\; 2\theta - 4\arcsin\lambda \;\geq\; 2\theta - 2\pi\lambda,$$

since $\lambda \leq \arcsin\lambda \leq \frac{\pi}{2}\lambda$ for any $\lambda \in [0,1]$. Consequently,

$$\begin{aligned}\mu(\mathcal{V}_\delta^-) \geq\; & \tfrac{1}{(2\pi)^{D/2}} \int_{\mathbb{R}_+} \mathrm{d}r\; r^{D-1} e^{-r^2/2}\;\cdots \\ & \Big[\Big(\int_0^\pi \mathrm{d}\phi_1\,\sin^{D-2}\phi_1\Big)\cdots\Big(\int_0^\pi \mathrm{d}\phi_{D-2}\,\sin\phi_{D-2}\Big)\Big]\big(2\theta - \tfrac{2\pi\delta}{(\sin\phi_1\,\cdots\,\sin\phi_{D-2})}\big) \\ =\; & \frac{\theta}{\pi} \;-\; \frac{\pi\,\delta\;I_{D-3}\,I_{D-4}\cdots I_0}{I_{D-2}\,I_{D-3}\,I_{D-4}\cdots I_0} \;=\; \frac{\theta}{\pi} \;-\; \frac{\pi\,\delta}{I_{D-2}},\end{aligned}$$

with $I_n = \int_0^\pi \mathrm{d}\phi\,\sin^n\phi$ and knowing that $(2\pi)^{D/2} = 2(I_{D-2}\,\cdots I_0)\int_{\mathbb{R}_+}\mathrm{d}r\; r^{D-1}e^{-r^2/2}$.

Using the fact that $I_n = \sqrt{\pi}\,\Gamma(\frac{n+1}{2})/\Gamma(\frac{n}{2}+1) \geq \sqrt{\pi}/\sqrt{\frac{n}{2}+\frac{1}{4}}$, we obtain $I_{D-2} \geq \frac{\sqrt{\pi}}{\sqrt{\frac{D}{2}-\frac{3}{4}}} \geq \sqrt{\frac{2\pi}{D}}$, and thus

$$p_1 \;\geq\; d_S(x,s) \;-\; \sqrt{\tfrac{\pi}{2}D}\,\delta.$$

If we want a meaningful bound for $p_1 \geq 0$, then we must have $d_S(x,s) \geq \sqrt{\frac{\pi}{2}d}\,\delta \geq \delta$. Therefore, as soon as the lower bound is positive, the aforementioned condition $d_S(x,s) \geq \delta$ always holds.

The lower bound for p_0 is obtained similarly. It is straightforward to show that $p_0 = \mu(\mathcal{V}_\delta^+)$, with $\mathcal{V}_\delta^+ = \{\varphi : \langle\varphi,x\rangle\langle\varphi,s\rangle > 0, \|x - \mathcal{P}_{\Pi(\varphi)}\,x\| \geq \delta, \|y - \mathcal{P}_{\Pi(\varphi)}\,s\| \geq \delta\}$. Lower bounding $\mu(\mathcal{V}_\delta^+)$ as for $\mu(\mathcal{V}_\delta^+)$, the only difference occurring with the integral on ϕ_{D-2} given by

$$\begin{aligned}\int_{[\theta,\pi]\,\cup\,[\pi+\theta,2\pi]} \mathrm{d}\phi_{D-1}\;\chi_{g(\delta,\varphi)}(\phi_{D-1})\,\chi_{g(\delta,\varphi)}(\phi_{D-1}-\theta)\;\cdots \\ =\; 2\pi - 2\theta - 4\arcsin g(\delta,\varphi) \;\geq\; 2(\pi-\theta) - 2\pi g(\delta,\varphi).\end{aligned}$$

Therefore, the lower bound of p_0 amounts to change $\theta \to \pi - \theta$ in the one of p_1, which provides the result. □

Proof of Lemma 16. If $\delta = 0$, there is nothing to prove. Therefore $\delta > 0$ and if φ^* belongs to either $\mathcal{V}_\delta$ or $\mathcal{W}_\delta$, we must have $\langle\varphi,x\rangle\langle\varphi,s\rangle < 0$. It is also sufficient to work on the restriction of $\mathcal{V}_\delta$ and $\mathcal{W}_\delta$ to unit vectors.

(i) $\mathcal{V}_\delta \subset \mathcal{W}_\delta$: By contradiction, let us assume that $\varphi^* \in \mathcal{V}_\delta$ but $\varphi^* \notin \mathcal{W}_\delta$. Without any loss of generality, $\langle\varphi^*,x\rangle > 0$ and $\langle\varphi^*,s\rangle < 0$. Since $\varphi^* \notin \mathcal{W}_\delta$, there exist two vectors $u^* \in B_\delta(x)$ and $v^* \in B_\delta(y)$ such that $\langle\varphi^*,u^*\rangle\langle\varphi^*,v^*\rangle > 0$. If $\langle\varphi^*,u^*\rangle > 0$ and $\langle\varphi^*,v^*\rangle > 0$, then, since $\langle\varphi^*,s\rangle < 0$ and by continuity of the inner product, there exist a $\lambda \in (0,1)$ such that $\langle\varphi^*,s(\lambda)\rangle = 0$ with $s(\lambda) = y + \lambda(v^* - s)$. Therefore, $s(\lambda) \in \Pi(\varphi)$ and, by definition of the orthogonal projection, $\|s - \mathcal{P}_{\Pi\varphi}\,s\| \leq \|s - s(\lambda)\| \leq \lambda\delta < \delta$ which is a contradiction. If $\langle\varphi^*,u^*\rangle < 0$ and $\langle\varphi^*,v^*\rangle < 0$, we apply the same reasoning on x and u^*. Therefore, $\mathcal{V}_\delta \subset \mathcal{W}_\delta$.

(ii) $\mathcal{W}_\delta \subset \mathcal{V}_\delta$: If $\varphi^* \in \mathcal{W}_\delta$ with $\varphi^* \notin \mathcal{V}_\delta$, we have either $\|x - \mathcal{P}_{\Pi(\varphi^*)}\, x\| < \delta$ or $\|s - \mathcal{P}_{\Pi(\varphi^*)}\, s\| < \delta$. Let us say that $\|x - \mathcal{P}_{\Pi(\varphi^*)}\, x\| < \delta$. Then, for $w = x + \delta\,(\mathcal{P}_{\Pi(\varphi^*)}\, x - x)/\|\mathcal{P}_{\Pi(\varphi^*)}\, x - x\| \in B^*_\delta(x)$, $\langle\varphi^*, x\rangle\langle\varphi^*, w\rangle = (\langle\varphi^*, x\rangle)^2\big(1 - \delta/\|\mathcal{P}_{\Pi(\varphi^*)}\, x - x\|\big) + \delta\,\langle\varphi^*, \mathcal{P}_{\Pi(\varphi^*)}\, x\rangle < 0$. However, $\varphi^* \in \mathcal{W}_\delta$ and $\langle\varphi^*, x\rangle\langle\varphi^*, s\rangle < 0$, leading to $\langle\varphi^*, w\rangle\langle\varphi^*, s\rangle > 0$, which is a contradiction. □

B.5 Theorem 6: Gaussian Matrices Provide BϵSEs

The strategy for proving Theorem 6 will be to count the number of pairs of K-sparse signals that are Euclidean distance δ apart. We will then apply the concentration results of Lemma 8 to demonstrate that the angles between these pairs are approximately preserved. We specifically proceed by focusing on a single K-dimensional subspace (intersected with the unit sphere) and then by applying a union bound to account for all possible subspaces.

Let $T \subset \{1, \dots, N\}$ be an index set of size $|T| = K$, $\Sigma^*(T) = \{w \in \mathbb{R}^N : \operatorname{supp} w \subset T,\ \|w\|_2 = 1\}$ be the sphere of unit vectors with support T. We first use again the fact that the sphere $\Sigma^*(T)$ can be δ-covered by a finite set of points $Q_{T,\delta}$. That is, for any $w \in \Sigma^*(T)$, there exists a $q \in Q_{T,\delta}$ such that $w \in B^*_\delta(q) = B_\delta(q) \cap \Sigma^*_T = \{w' \in \Sigma^*_T : \|w' - q\|_2 \le \delta\}$ [45]. Note that the size of $Q_{T,\delta}$ is bounded by $|Q_{T,\delta}| \le C_\delta = (3/\delta)^K$.

Let Φ_T be the matrix formed by the columns of Φ indexed by T and note that $\Phi_T w = \Phi w$. Since $\epsilon \ge 0$ is given, then for all pairs of points $x, y \in Q_{T,\delta}$, we have

$$\mathbb{P}\Big(\Big|\, d_H\big(A(p), A(q)\big) \;-\; d_S(x,y) \Big| \;\le\; \epsilon + \sqrt{\tfrac{\pi}{2}K}\,\delta \Big) \;\ge\; 1 \;-\; 2\,(\tfrac{3}{\delta})^{2K}\, e^{-2\epsilon^2 M}, \tag{B.5}$$

for all $p \in B^*_\delta(x)$ and $q \in B^*_\delta(y)$. This follows from Lemma 8 with $D = K$, since Φ_T is a Gaussian matrix and by invoking the union bound, since there are $\binom{C_\delta}{2} \le C^2_\delta = (3/\delta)^{2K}$ such pairs x, y.

The bound (B.5) can be extended to all possible index sets T of size K via the union bound. Specifically, for all $T \subset \{1, \cdots, N\}$ and all pairs of points $x, y \in Q_{T,\delta}$, we have

$$\mathbb{P}\Big(\Big| d_H\big(A(p), A(q)\big) \;-\; d_S(x,y) \Big| \;\le\; \epsilon + \sqrt{\tfrac{\pi}{2}K}\,\delta \Big) \;\ge\; 1 \;-\; 2\,(\tfrac{eN}{K})^{K}\,(\tfrac{3}{\delta})^{2K}\, e^{-2\epsilon^2 M} \tag{B.6}$$

for all $p \in B^*_\delta(x)$ and $q \in B^*_\delta(y)$, since there are no more than $\binom{N}{K} \le (eN/K)^K$ possible T.

To summarize, for any points on the sphere $u, v \in S^{N-1}$ with $|\operatorname{supp} u \cup \operatorname{supp} v| \le K$, there exists an index set T of size K such that $u, v \in \Sigma^*(T)$ and from (B.6) there exists two points $x, y \in Q_{T,\delta}$ such that $u \in B^*_\delta(x)$ and $v \in B^*_\delta(y)$ with a probability exceeding $1 - 2\,(\frac{eN}{K})^K\,(\frac{3}{\delta})^{2K}\, e^{-2\epsilon^2 M}$. Furthermore, when this occurs we have

$$\Big| d_H\big(A(u), A(v)\big) \;-\; d_S(x,y) \Big| \;\le\; \epsilon + \sqrt{\tfrac{\pi}{2}K}\,\delta. \tag{B.7}$$

To obtain our final bound, consider that $u \in B^*_\delta(x)$ implies that $\pi\, d_S(u,x) \le 2 \arcsin \delta/2 \le \pi\delta/2$, and $d_S(v,y)$ can be similarly bounded. Thus, $d_S(u,v) \ge d_S(x,y) - \delta$ and $d_S(u,v) \le d_S(x,y) + \delta$, and (B.7) becomes

$$\Big| d_H\big(A(u), A(v)\big) \;-\; d_S(u,v) \Big| \;\le\; \epsilon + (1 + \sqrt{\tfrac{\pi}{2}K})\,\delta. \tag{B.8}$$

By bounding the probability of failure as $2\,(\frac{eN}{K})^K\,(\frac{3}{\delta})^{2K}\,e^{-2\epsilon^2 M} \leq \eta$, where $0<\eta<1$, and setting $\epsilon = (1+\sqrt{\frac{\pi}{2}K})\,\delta$, solving for M, we obtain

$$M \;\geq\; \tfrac{4}{\epsilon^2}\Big(K\,\log(\tfrac{9eN}{K}) + 2K\,\log(\tfrac{2(1+\sqrt{2\pi K})}{\epsilon}) + \log(\tfrac{2}{\eta})\Big).$$

Since $K \geq 1$, we have that $2(1+\sqrt{2\pi K}) < 4\sqrt{2\pi K}$, and thus the previous relation is satisfied if

$$\begin{aligned} M \;&\geq\; \tfrac{4}{\epsilon^2}\Big(K\,\log(\tfrac{9eN}{K}) + 2K\,\log(\tfrac{4\sqrt{2\pi K}}{\epsilon}) + \log(\tfrac{2}{\eta})\Big), \\ &=\; \tfrac{4}{\epsilon^2}\Big(K\,\log(9eN) + 2K\,\log(\tfrac{4\sqrt{2\pi}}{\epsilon}) + \log(\tfrac{2}{\eta})\Big), \\ &=\; \tfrac{4}{\epsilon^2}\Big(K\,\log(N) + 2K\,\log(\tfrac{12\sqrt{2\pi e}}{\epsilon}) + \log(\tfrac{2}{\eta})\Big), \end{aligned}$$

which can be further simplified to $M \;\geq\; \frac{4}{\epsilon^2}\Big(K\,\log(N) + 2K\,\log(\frac{50}{\epsilon}) + \log(\frac{2}{\eta})\Big)$.

B.6 Lemma 9: Stability with Measurement Noise

In Lemma 9, since $\Phi \sim \mathcal{N}^{M\times N}(0,1)$, each $y_i = (\Phi x)_i$ follows a Gaussian distribution $\mathcal{N}(0,\|x\|_2^2)$, and furthermore, since we have independent additive noise, $z_i = y_i + n_i = (\Phi x)_i + n_i$ follows the Gaussian distribution $\mathcal{N}(0,\|x\|_2^2+\sigma^2)$.

We begin by bounding the probability that any noisy measurement z_i has a different sign than the original corresponding measurement y_i, i.e., we bound $p_0 = \mathbb{P}(z_i y_i < 0)$. This quantity is interesting since $M\,d_H\big(A_n(x), A(x)\big)$ follows a Binomial distribution with M trials and probability of success p_0 and thus we also have $\mathbb{E}\Big(d_H\big(A_n(x), A(x)\big)\Big) = p_0$.

To solve for the bound, we compute

$$p_0 = \int_{\mathbb{R}} \mathrm{d}u\;\mathbb{P}(z_i y_i < 0\,|\,y_i = u\,)\,f_{y_i}(u) \;=\; \int_{\mathbb{R}} \mathrm{d}u\;\mathbb{P}(u^2 + u n_i < 0)\,g(u;\|x\|_2),$$

with the pdf $f_{y_i}(t) = g(t;\sigma') = \frac{1}{\sqrt{2\pi t}}\exp(-t^2/2\sigma'^2)$. This leads to

$$\begin{aligned} p_0 \;&=\int_0^\infty \mathrm{d}u\;\mathbb{P}(n_i < -u)\,g(u;\|x\|_2) \;+\; \int_{-\infty}^0 \mathrm{d}u\;\mathbb{P}(n_i > -u)\,g(u;\|x\|_2) \\ &= \int_0^\infty \mathrm{d}u\;2\,Q(u/\sigma)\,g(u;\|x\|_2) \;\leq\; \int_0^\infty \mathrm{d}u\;e^{-\frac{u^2}{2\sigma^2}}\,g(u;\|x\|_2) \\ &=\; \tfrac{1}{\sqrt{2\pi}\|x\|_2}\int_0^\infty \mathrm{d}u\;e^{-\frac{1}{2}(\frac{(\|x\|_2^2+\sigma^2)\,u^2}{\sigma^2\|x\|_2^2})} \;=\; \tfrac{1}{2}\frac{\sigma}{\sqrt{\|x\|_2^2+\sigma^2}}, \end{aligned}$$

where $Q(u) = \int_u^\infty \mathrm{d}t\,g(t;1)$ denotes the tail integral of the standard Gaussian distribution which is bounded by $Q(t) \leq \frac{1}{2}e^{-t^2/2}$ for $t \geq 0$ (see for instance [143, Eq. (13.48)]).

Thus, we have $p_0 \leq e(\sigma, \|x\|_2) = \frac{1}{2}\frac{\sigma}{\sqrt{\|x\|_2^2+\sigma^2}}$ and, by applying the Chernoff-Hoeffding inequality to the distribution of $d_H\big(A_n(x), A(x)\big)$,

$$\begin{aligned} \mathbb{P}\big[M\, d_H\big(A_n(x), A(x)\big) > \ & M\, e(\sigma, \|x\|_2) + M\epsilon\,\big] \\ \leq \ & \mathbb{P}\big[M\, d_H\big(A_n(x), A(x)\big) > \ M\, p_0 + M\epsilon\,\big] \\ \leq \ & e^{-2M\epsilon^2}, \end{aligned}$$

which proves the lemma.

B.7 Corollary 3: Stability with Compressible Signals

The proof of Corollary 3 is as follows. Since $x = x_K + (x - x_K)$ then $\Phi x = \Phi x_K + n$ where $n = \Phi(x - x_K)$ is a random Gaussian vector. Thus $A(x) = A_n(x_K)$ where A_n is defined as in Lemma 9. The vector n is also independent of Φx_K since the supports of x_K and $(x - x_K)$ are disjoint. Finally, the variance σ of each i.i.d. component n_i of n is $\|x - x_K\|_2^2$, thus the result follows from Lemma 9 with the bound $e(\sigma, \|x_K\|_2) = \frac{1}{2}\frac{\sigma}{\sqrt{\|x_K\|_2^2+\sigma^2}} = \frac{\|x-x_K\|_2}{2\|x\|_2}$ since $\|x\|_2^2 = \|x_K\|_2^2 + \|x - x_K\|_2^2$.

APPENDIX C

The RSS Algorithm

C.1 Quadratic Penalty Framework

The quadratic penalty framework can also make use of the RSS-inner subroutine to solve (3.17). This framework has been is used by FPC and *gradient projection for sparse reconstruction* (GPSR) to solve conventional CS reconstruction problems [27]. This approach proceeds by iteratively minimizing a sequence of penalty functions:

$$\min_{x \in \mathbb{R}^N} \ \|x\|_1 + \frac{\mu^s}{2} \| \min\{Ax - b, 0\}\|_2^2 \text{ s.t. } \ \|x\|_2 = 1, \tag{C.1}$$

where $\mu^s > 0$ is the penalty parameter, and we increase $\mu^s \to +\infty$ by setting $\mu^{s+1} := \kappa \mu^s$ with $\kappa > 1$. In fact, (C.1) often only needs to be solved once for some values of $\mu^s = \mu$, as is done in practice with FPC and GPSR.

It is then straightforward to see that (C.1) is of the form (3.18) and can be solved by the RSS-inner subroutine.

C.2 Convergence Proof of Algorithm 4

Before proving Lemma 10 from Section 3.7.1-B, we introduce an additional Lemma 17 that provides bounds on the reduction of the first-order approximation $m_s(x)$. We then present a proof that demonstrates both Lemmas.

Lemma 17. *Suppose that x^s is not a stationary point of* (3.18)*. Denote by $d^s := z^s - x^s$ the search direction computed at $x^s \in \mathbb{R}^N$. Then the predicted reduction satisfies $\delta(x^s, z^s) \geq \frac{\tau^s}{2}\|d^s\|_2^2$.*

In particular, if $\mathcal{S}(\tau^s x^s - \mu g^s) \neq 0$, the reduction of the objective function of the subproblem (3.23) *satisfies*

$$m_s(x^s) - m_s(z^s) \geq \frac{1}{2}\|\mathcal{S}^s\|_2 \|d^s\|_2^2. \tag{C.2}$$

Proof of Lemmas 10 and 17: The corresponding first-order optimality conditions of (3.23) are

$$p + \mu g^s + \tau^s(z^s - x^s) - \lambda z^s = 0, \quad \lambda \in \mathbb{R}, \quad \|z^s\|_2 = 1, \tag{C.3}$$

where $p \in \partial \|z^s\|_1$. Given any feasible solution x with $\|x\|_2 = 1$, we have

$$\begin{aligned}
& m_s(x) - m_s(z^s) \\
= \;& \|x\|_1 - \|z^s\|_1 + \left(\mu g^s + \tau^s(z^s - x)\right)^\top (x - z^s) \\
& + \frac{\tau^s}{2}\|x - z^s\|_2^2 \\
= \;& \|x\|_1 - \|z^s\|_1 + (\lambda z^s - p)^\top (x - z^s) + \frac{\tau^s}{2}\|x - z^s\|_2^2 \\
= \;& \|x\|_1 - p^\top x + (\lambda z^s)^\top x - \lambda + \frac{\tau^s}{2}\|x - z^s\|_2^2 \\
= \;& \|x\|_1 - p^\top x + \frac{\tau^s - \lambda}{2}\|x - z^s\|_2^2 \qquad \text{(C.4)} \\
\geq \;& \frac{\tau^s - \lambda}{2}\|x - z^s\|_2^2, \qquad \text{(C.5)}
\end{aligned}$$

where the first equality follows from the Taylor expansion of the smooth terms of $m_s(x)$, the second equality comes from (C.3), the third equality uses $p^\top z^s = \|z^s\|_1$ and $\|z^s\|_2 = 1$, and the fact $\|x\|_1 = \max_{q \in [-1,1]} q^\top x$ gives the last inequality.

It follows from (C.3) that $(\tau^s - \lambda) z^s := \tau^s x^s - \mu g^s - p$. We now discuss the following cases:

1. $\tau^s - \lambda > 0$ and $\mathcal{S}(\tau^s x^s - \mu g^s) \neq 0$. It can be verified that $\tau^s - \lambda = \|\mathcal{S}\|_2$, and $z^s = \frac{\mathcal{S}^s}{\|\mathcal{S}^s\|_2}$ is a global minimizer. Substituting $x = x^s$ into (C.5) gives (C.2).

2. $\tau^s - \lambda > 0$ and $|\tau^s x^s - \mu g^s| \leq 1$. Without loss of generality, we can assume that there exists a component $z_i^s > 0$. Then $p_i = 1$ and $\tau^s x^s - \mu g^s - 1 \leq 0$ which contradicts $(\tau^s - \lambda) z_i^s > 0$.

3. $\tau^s - \lambda < 0$. Suppose that z^s has at least two nonzero components. Without loss of generality, we can assume that there exits a component $z_i^s > 0$. Let $\bar{x}_i = z_i^s + \epsilon$ with $\epsilon > 0$ and $\bar{x}_j = z_j^s$ for all other $j \neq i$. It is obvious that $x = \frac{\bar{x}}{\|\bar{x}\|_2}$ is feasible, $x \neq z^s$ and $p \in \partial \|x\|_1$. Hence $\|x\|_1 - p^\top x = 0$ and (C.4) implies that $m_s(x) < m_s(z^s)$, which contradicts the fact $m_s(x) \geq m_s(z^s)$. Therefore, the solution z^s only has one nonzero element and its value must be either -1 or 1. Note that

$$m_s(x) = \|x\|_1 + (\mu g^s - \tau^s x^s)^\top x + \text{constant}.$$

It can be verified that $z_i^s = \operatorname{sgn}(\tau_i^s x^s - \mu g_i^s)$, where $i = \arg\max_{k=1,\ldots,n} |\tau^s x_k^s - \mu g_k^s|$ (select only one i if there are multiple solutions); otherwise $z_i^s = 0$. In fact, the set $\{i \mid |\tau^s x_i^s - \mu g_i^s| = 1\}$ is empty.

4. $\tau^s - \lambda = 0$. Then we must have $|\tau^s x^s - \mu g^s| \leq 1$, $p \notin \partial\|x^s\|_1$, the set $\{i \mid |\tau^s x_i^s - \mu g_i^s| = 1\}$ is not empty and the closed-form solution of the subproblem (3.23) is given by (3.26).

□

The next lemma shows that iteration s will be successful for a sufficient large τ^s, hence, the number of unsuccessful iterations between two successful iterations cannot be infinity.

Lemma 18. *Suppose that $\|d^s\|_2 > 0$ and $\tau^s \geq \bar{\tau} := \frac{2\mu L}{1-\eta_2}$. Then the s-th iteration is a very successful iteration which satisfies $\tau^{s+1} \leq \tau^s$.*

Proof: Using the definition of r_s, Lemma 17 and Assumption 1, we obtain

$$\begin{aligned} |r_s - 1| &= \left|\frac{\zeta_\mu(x^k) - \zeta_\mu(z^s) - \delta(x^s, z^s)}{\delta(x^s, z^s)}\right| \\ &= \left|\frac{\mu f(x^s) + \mu(g^s)^\top d^s - \mu f(z^s)}{\delta(x^s, z^s)}\right| \\ &\leq \frac{2\mu\|g(x^s + \xi d^s) - g(x^s)\|_2\,\|d^s\|_2}{\tau^s\|d^s\|_2^2}, (\xi \in (0,1)) \\ &\leq \frac{2\mu L}{\tau^s} \leq 1 - \eta_2. \end{aligned}$$

Therefore, $r_s \geq \eta_2$ and the s-th iteration is very successful. The rule (3.28) ensures that $\tau^{s+1} \leq \tau^s$.

□

The following lemma gives a useful alternative characterization of stationarity.

Lemma 19. *For any successful iteration k with $\tau_s < +\infty$, the point x^s is a stationary point of* (3.18) *if only if $d^s = 0$.*

Proof: Suppose that $d^s \neq 0$. Since iteration k is successful, Lemma 17 and the ratio (3.27) testing show that the function value at z^s is smaller than that of x^s, implying that x^s is not a stationary point. Conversely, if $d^s = 0$, then it follows from (C.3) and $x^s = z^s$ that

$$p + \mu g^s - \lambda x^s = 0, \quad \lambda \in \mathbb{R}, \quad \|x^s\|_2 = 1,$$

which are the first-order optimality conditions of (3.18). □

We are now ready to prove Theorem 7.

C.2.1 Proof of Theorem 7

If Algorithm 4 has finitely many successful iterations, then for sufficiently large s, the iteration is unsuccessful. Thus, the sequence $\{\tau^s\}$ converges to $+\infty$. Suppose that s_0 is the index of the last successful iteration and $\|d^s\|_2 > 0$ for $s > s_0$. It follows from Lemma 18 that there must exist a very successful iteration of index s larger than s_0, which is a contradiction to the assumption.

Suppose that Algorithm 4 has infinitely many successful iterations. Since an unsuccessful iterate in the sequence $\{x^s\}$ remains the same and makes no progress, it can be substituted by the same successful iterate. The substituted sequence which only consists of different successful iterates is still denoted by the same notation $\{x^s\}$. Since the sequence satisfying $\|x^s\| = 1$ lies in a compact set, there exists at least one cluster point x^* such that $\|x^*\| = 1$.

Suppose that the cluster point x^* is not a stationary point. According to Lemma 18, there exits a constant $\tilde{\tau}$ such that $\tau^s \leq \tilde{\tau} < +\infty$ for all s. Hence, there exists a subsequence $\{x^{s_i}\}$ approaches x^* and $\lim\limits_{s_i \to \infty} \tau^{s_i} = t^* \geq 0$. Since x^* is not a stationary point, by Lemma 19, $d^* \neq 0$ and

$$\delta(x^*, x^* + d^*) = \|x^*\|_1 - \|x^* + d^*\|_1 - \mu(g^*)^\top d^* = \epsilon > 0.$$

Using the fact that the shrinkage operator is non-expansive, i.e.,

$$\|\mathcal{S}(x) - \mathcal{S}(y)\|_2 \leq \|x - y\|_2,$$

we obtain

$$\begin{array}{ll} \multicolumn{2}{l}{\|\mathcal{S}^{s_i} - \mathcal{S}^*\|_2} \\ = & \|\mathcal{S}(\tau^{s_i} x^{s_i} - \mu g(x^{k_i})) - \mathcal{S}(\tau^* x^* - \mu g(x^*))\|_2 \\ \leq & \|\tau^{s_i} x^{s_i} - \tau^* x^* - \mu(g(x^{s_i} - g(x^*))\|_2 \\ \leq & |\tau^*| \|x^{s_i} - x^*\|_2 + |\tau^{s_i} - \tau^*| \|x^{s_i}\|_2 + \mu M \|x^{s_i} - x^*\|_2, \end{array}$$

which implies that $\lim\limits_{s_i \to \infty} \mathcal{S}^{s_i} = \mathcal{S}^*$ and $\lim\limits_{s_i \to \infty} d^{s_i} = d^*$. Note that $g(x)$ and $\|x\|_1$ are continuous. For s_i large enough, we have therefore that

$$\delta(x^{s_i}, x^{s_i} + d^{s_i}) = \|x^{s_i}\|_1 - \|x^{s_i} + d^{s_i}\|_1 - \mu(g^{s_i})^\top d^{s_i} \geq \frac{\epsilon}{2}.$$

It follows from the acceptance rule for successful iterations (3.28) that

$$\zeta_\mu(x^{s_i}) - \zeta_\mu(x^{s_i+1}) \geq \eta_1 \delta(x^s, x^{s_i} + d^{s_i}) \geq \frac{\eta_1 \epsilon}{2}. \tag{C.6}$$

However, since the series with positive terms

$$\begin{array}{rcl} \sum\limits_{i=1}^{\infty} \zeta_\mu(x^{s_i}) - \zeta_\mu(x^{s_i+1}) & \leq & \sum\limits_{s=s_1}^{\infty} \zeta_\mu(x^{s_i}) - \zeta_\mu(x^{s_{i+1}}) \\ & = & \zeta_\mu(x^{s_1}) - \zeta_\mu(x^*) < +\infty \end{array}$$

is convergent, we have

$$\lim_{s_i \to \infty} \zeta_\mu(x^{s_i}) - \zeta_\mu(x^{s_i+1}) = 0,$$

which contradicts (C.6) and completes the proof. □

APPENDIX D

Regime Change: Proof of Theorem 8

We first extend the upper bound of Theorem 4.1 in [38] on the oracle-assisted reconstruction error to account for correlated measurement noise.

Lemma 20. *Suppose that $y = \Phi x + z$, where $z \in \mathbb{R}^M$ is a zero-mean, random vector with covariance matrix $\Sigma = \mathbb{E}(zz^T)$, and that x is K-sparse. Furthermore, suppose that Φ satisfies the RIP of order K with constant δ. Then the estimate $\widehat{x}$ provided by the oracle-assisted reconstruction algorithm (1.10) satisfies*

$$\mathbb{E}\left(\|x - \widehat{x}\|_2^2\right) \leq \frac{K}{1-\delta}\lambda_{\max}(\Sigma), \tag{D.1}$$

where $\lambda_{\max}(\Sigma)$ is the largest eigenvalue of Σ.

Proof. For a fixed support set $\Omega \in \{1, \ldots, N\}$ with $|\Omega| = K$, the RIP ensures that Φ_Ω is full rank, and thus the oracle estimate satisfies

$$\widehat{x}|_\Omega = x|_\Omega + \Phi_\Omega^\dagger z. \tag{D.2}$$

We seek to estimate $\mathbb{E}\left(\|\Phi_\Omega^\dagger z\|_2^2\right)$.

For any $K \times M$ matrix A we have that

$$\begin{aligned} \mathbb{E}\left(\|Az\|_2^2\right) &= \mathbb{E}(\mathrm{Tr}(Az(Az)^T)) = \mathbb{E}(\mathrm{Tr}(Azz^T A^T)) \\ &= \mathrm{Tr}(A\mathbb{E}(zz^T)A^T) = \mathrm{Tr}(A\Sigma A^T) \\ &= \sum_{j=1}^{K} \lambda_j(A\Sigma A^T), \end{aligned} \tag{D.3}$$

where $\lambda_j(A\Sigma A^T)$ denotes the j-th eigenvalue of $A\Sigma A^T$, and (D.3) follows since $A\Sigma A^T$ is a $K \times K$ matrix. Lemma 8.2 of [38] explains that the eigenvalues of this matrix can be upper bounded as

$$\begin{aligned} \lambda_{\max}(A\Sigma A^T) &\leq \lambda_{\max}(AA^T)\lambda_{\max}(\Sigma) \\ &\leq s_{\max}(A)^2\lambda_{\max}(\Sigma), \end{aligned} \tag{D.4}$$

where $s_{\max}(A)$ denotes the maximum singular value of A.

Thus, to obtain the final bound, we combine (D.3) with (D.4) and substitute $A = \Phi_\Omega^\dagger$, yielding

$$\begin{aligned} \mathbb{E}\left(\|\Phi_\Omega^\dagger z\|_2^2\right) &\leq K s_{\max}(\Phi_\Omega^\dagger)^2\lambda_{\max}(\Sigma) \\ &\leq \frac{K}{1-\delta}\lambda_{\max}(\Sigma), \end{aligned} \tag{D.5}$$

since we have that $s_{\max}(\Phi_\Omega^\dagger)^2 \leq \frac{1}{1-\delta}$ from Lemma 8.1 of [38]. □

We next demonstrate that, by choosing a signal model with random values and supports, the noiseless measurements Φx are identically distributed and uncorrelated.

Lemma 21. *Let $x \in \mathbb{R}^N$ be a sparse signal with support $\Omega \in \{1, \ldots, N\}$ and $|\Omega| = K$, where the elements Ω are chosen uniformly at random and the amplitudes of the non-zero coefficients are drawn according to $x_j \in \Omega \sim \mathcal{N}(0, \sigma_x^2)$. Furthermore, let the $M \times N$ matrix Φ satisfy $\Phi\Phi^T = \frac{N}{M} I_M$. Then the vector Φx is distributed as a mixture of Gaussians with*

$$\mathbb{E}((\Phi x)_i) = 0, \qquad \mathbb{E}((\Phi x)(\Phi x)^T) = \frac{K}{M}\sigma_x^2 I_M, \tag{D.6}$$

i.e., the elements $(\Phi x)_i$ of Φx are zero-mean uncorrelated variables.

Proof. For a fixed support Ω, each element $(\Phi x)_i$ is Gaussian distributed with mean zero since it is the sum of K zero-mean Gaussian variables. Furthermore, the distribution of $(\Phi x)_i$ over all possible supports is the sum of the distribution for each fixed support, scaled by the probability that they occur. Thus, $(\Phi x)_i$ is a mixture of Gaussians with $\mathbb{E}((\Phi x)_i) = 0$.

To derive the variance of the elements and also show that they are uncorrelated, we first examine $\mathbb{E}(xx^T)$. The off-diagonal elements are zero, i.e., $\mathbb{E}(x_i x_j)_{i \neq j} = 0$, since the elements of x are uncorrelated, by definition. Furthermore, the variance of the diagonal elements can be computed as

$$\mathbb{E}(x_i^2) = \sigma_x^2 \mathbb{P}(i \in \Omega) = \frac{K}{N}\sigma_x^2,$$

since the K non-zero support locations are chosen uniformly, any location j is chosen with probability K/N. Thus, $\mathbb{E}(xx^T) = \frac{K}{N}\sigma_x^2 I_N$. We next compute the correlation of the measurements Φx to obtain

$$\begin{aligned} \mathbb{E}(\Phi x(\Phi x)^T) &= \Phi\mathbb{E}(xx^T)\Phi^T \\ &= \frac{K}{N}\sigma_x^2\Phi\Phi^T = \frac{K}{M}\sigma_x^2 I_M, \end{aligned} \tag{D.7}$$

which concludes the proof. □

Proof of Theorem 8. Denote the error between the noiseless ideal measurements and y_Q by

$$z := \Phi x - \mathcal{Q}_B(\Phi x + \Phi n). \tag{D.8}$$

Our goal is to determine a bound on the variance $\sigma_{z_i}^2$ of each element z_i of z. We begin by rewriting the norm squared of z as

$$\begin{aligned} z_i^2 &= [(\Phi x)_i - \mathcal{Q}_B(\Phi x + \Phi n)_i)]^2 \\ &= [(\Phi x + \Phi n)_i - \mathcal{Q}_B(\Phi x + \Phi n)_i - (\Phi n)_i]^2 \\ &\leq 2[(\Phi x + \Phi n)_i - \mathcal{Q}_B(\Phi x + \Phi n)_i]^2 + 2(\Phi n)_i^2, \end{aligned} \tag{D.9}$$

where the index i denotes individual elements of the respective vector.

We now seek an upper bound on the expected value of each of the quantities in (D.9). We begin with the second term in (D.9). From the definition of Φ, we have that the elements of Φn have variance

$$\sigma_{\Phi n}^2 = \mathbb{E}((\Phi n)_i^2) = \frac{N}{M}\sigma_n^2, \tag{D.10}$$

and furthermore are uncorrelated, as was reviewed in Section 1.2.

To bound the first term in (D.9), we note that the optimal scalar quantizer of rate B for a Gaussian variable g with variance σ^2 has MSE given by $\mathbb{E}(g - \mathcal{Q}_B(g))^2 = \sigma^2 2^{-2B}$. Furthermore, the MSE of an optimal quantizer of rate B for any variable with variance σ^2 is upper bounded by that of a Gaussian variable. Our goal is to apply this quantization bound to $(\Phi x + \Phi n)_i$. Since $(\Phi x)_i$ and $(\Phi n)_i$ are zero mean and independent of each other, then we immediately have that $\mathbb{E}\left((\Phi x + \Phi n)_i^2\right) = \frac{K}{M}\sigma_x^2 + \frac{N}{M}\sigma_n^2$, where the first term follows from Lemma 21, and the second term follows from (D.10). Thus, we can bound the first term in (D.9) as

$$\begin{aligned} \mathbb{E}\left([(\Phi x + \Phi n)_i - \mathcal{Q}_B(\Phi x + \Phi n)_i]^2\right) &\leq \mathbb{E}\left((\Phi x + \Phi n)_i^2\right) 2^{-2B} \\ &\leq \frac{K}{M}\sigma_x^2 2^{-2B} + \frac{N}{M}\sigma_n^2 2^{-2B}. \end{aligned} \tag{D.11}$$

Combining (D.10) and (D.11) as in (D.9) yields

$$\sigma_{z_i}^2 \leq 2\frac{K}{M}\sigma_x^2 2^{-2B} + 2\frac{N}{M}\sigma_n^2\left(1 + 2^{-2B}\right). \tag{D.12}$$

We have thus far established an upper bound on the variance $\sigma_{z_i}^2$ of the error z_i of each measurement. We next obtain a bound on the eigenvalues of the covariance matrix $\Sigma = \mathbb{E}(zz^T)$. The off-diagonal elements of Σ can be written as

$$\begin{aligned} \mathbb{E}(z_i z_j)_{i \neq j} &= \mathbb{E}((\Phi x)_i(\Phi x)_j)) - \mathbb{E}((\Phi x)_i \mathcal{Q}_B(\Phi x + \Phi n)_j) \\ &\quad -\mathbb{E}((\Phi x)_j \mathcal{Q}_B(\Phi x + \Phi n)_i) + \mathbb{E}(\mathcal{Q}_B(\Phi x + \Phi n)_i \mathcal{Q}_B(\Phi x + \Phi n)_j) \\ &= -\mathbb{E}(\mathcal{Q}_B(\Phi x + \Phi n)_i \mathcal{Q}_B(\Phi x + \Phi n)_j), \end{aligned} \tag{D.13}$$

since $\mathbb{E}((\Phi x)_i(\Phi x)_j)) = 0$ by design and, for an optimal scalar quantizer, we have that $\mathbb{E}(\mathcal{Q}_B(\Phi x + \Phi n)_i \mathcal{Q}_B(\Phi x + \Phi n)_j) = \mathbb{E}((\Phi x)_j \mathcal{Q}_B(\Phi x + \Phi n)_i)$ [118]. Thus, the matrix Σ has $\sigma_{z_i}^2$ along its diagonal and $\mathfrak{S}$ for all other entries. We next apply Gershgorin's circle theorem, which explains that any eigenvalue is upper bounded by the diagonal entry plus the sum of the magnitudes of the off-diagonal entries of each row of Σ. Thus, we have

$$\lambda_{\max}(\Sigma) \leq \sigma_{z_i}^2 + (M-1)\mathfrak{S}, \tag{D.14}$$

where $\mathfrak{S} = \max_{i \neq j} |\mathbb{E}(z_i z_j)|$.

To obtain the final bound, we combine to (D.12) with (D.14) and apply the upper bound in Lemma 20. We express the bound with the substitution $M = \mathfrak{B}/B$. □

Bibliography

[1] D. Healy, “Analog-to-information,” 2005, BAA #05-35.

[2] C. Shannon, “Communication in the presence of noise,” *Proc. Institute of Radio Engineers*, vol. 37, no. 1, pp. 10–21, 1949.

[3] H. Nyquist, “Certain topics in telegraph transmission theory,” *Trans. AIEE*, vol. 47, pp. 617–644, 1928.

[4] E. Whittaker, “On the functions which are represented by the expansions of the interpolation theory,” *Proc. Royal Soc. Edinburgh, Sec. A*, vol. 35, pp. 181–194, 1915.

[5] V. Kotelnikov, “On the carrying capacity of the ether and wire in telecommunications,” in *Izd. Red. Upr. Svyazi RKKA*, Moscow, Russia, 1933.

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

[7] B. Le, T. W. Rondeau, J. H. Reed, and C. W. Bostian, “Analog-to-digital converters,” *IEEE Signal Processing Mag.*, Nov. 2005.

[8] D. Donoho, “Denoising by soft-thresholding,” *IEEE Trans. Inform. Theory*, vol. 41, no. 3, pp. 613–627, 1995.

[9] R. DeVore, B. Jawerth, and V. Popov, “Compression of wavelet decompositions,” *American J. of Math.*, vol. 114, no. 4, pp. 737–785, 1992.

[10] J. Bazerque and G. Giannakis, “Distributed spectrum sensing for cognitive radio networks by exploiting sparsity,” *IEEE Trans. Signal Processing*, vol. 58, no. 3, pp. 1847–1862, 2010.

[11] J. Mitola III, “Cognitive radio for flexible mobile multimedia communications,” *Mobile Networks and Applications*, vol. 6, no. 5, pp. 435–441, 2001.

[12] W. Pennebaker and J. Mitchell, *JPEG Still Image Data Compression Standard*, Van Nostrand Reinhold, 1993.

[13] D. Taubman and M. Marcellin, *JPEG 2000: Image Compression Fundamentals, Standards and Practice*, Kluwer, 2001.

[14] E. Candès, “Compressive sampling,” in *Proc. Int. Congress Math.*, Madrid, Spain, Aug. 2006.

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

[16] J. Tropp and A. Gilbert, “Signal recovery from partial information via orthogonal matching pursuit,” *IEEE Trans. Inform. Theory*, vol. 53, no. 12, pp. 4655–4666, 2007.

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

[18] J. Tropp, J. Laska, M. Duarte, J. Romberg, and R. Baraniuk, “Beyond Nyquist: Efficient sampling of sparse, bandlimited signals,” *IEEE Trans. Inform. Theory*, vol. 56, no. 1, pp. 520–544, 2010.

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

[20] J. Slavinsky, J. Laska, M. Davenport, and R. Baraniuk, “The compressive multiplexer for multi-band signal acquistion,” in *International Conf. on Acoustics, Speech, and Signal Proc. (ICASSP)*, Prague, May 2011.

[21] W. Bajwa, J. Haupt, G. Raz, S. Wright, and R. Nowak, “Toeplitz-structured compressed sensing matrices,” in *Proc. IEEE Work. Stat. Signal Processing*, Madison, WI, Aug. 2007.

[22] M. Duarte and R. Baraniuk, “Spectral compressive sensing,” Preprint, 2010.

[23] C. Hegde, M. Duarte, and V. Cevher, “Compressive sensing recovery of spike trains using a structured sparsity model,” in *Proc. Work. Struc. Parc. Rep. Adap. Signaux (SPARS)*, Saint-Malo, France, Apr. 2009.

[24] R. Baraniuk, V. Cevher, M. Duarte, and C. Hegde, “Model-based compressive sensing,” *IEEE Trans. Inform. Theory*, vol. 56, no. 4, pp. 1982–2001, 2010.

[25] E. Hale, W. Yin, and Y. Zhang, “A fixed-point continuation method for ℓ_1-regularized minimization with applications to compressed sensing,” Tech. Rep. TR07-07, Rice Univ., CAAM Dept., 2007.

[26] W. Yin, S. Osher, D. Goldfarb, and J. Darbon, “Bregman iterative algorithms for ℓ_1-minimization with applications to compressed sensing,” *SIAM J. Imaging Sci.*, vol. 1, no. 1, pp. 143–168, 2008.

[27] M. A. T. Figueiredo, R. D. Nowak, and S. J. Wright, “Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems,” *IEEE J. of Selected Topics in Signal Processing*, Sept. 2007.

[28] E. van den Berg and M. Friedlander, “Probing the Pareto frontier for basis pursuit solutions,” *SIAM J. on Sci. Comp.*, vol. 31, no. 2, pp. 890–912, 2008.

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

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

[31] D. Donoho, A. Maleki, and A. Montanari, “Message passing algorithms for compressed sensing,” *Proc. Natl. Acad. Sci.*, vol. 106, no. 45, pp. 18914–18919, 2009.

[32] J. Treichler, “Personal communication,” Oct. 2009.

[33] P. Boufounos, “Reconstruction of sparse signals from distorted randomized measurements,” in *Proc. Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Dallas, TX, Mar. 2010.

[34] J. Laska, P. Boufounos, M. Davenport, and R. Baraniuk, “Democracy in action: Quantization, saturation, and compressive sensing,” *App. Comp. and Harm. Anal.*, vol. 31, no. 3, pp. 429–443, 2011.

[35] M. Davenport, J. Laska, P. Boufounos, and R. Baraniuk, “A simple proof that random matrices are democratic,” *Rice University Technical Report TREE-0906*, Nov. 2009.

[36] E. Candès and M. A. Davenport, “How well can we estimate a sparse vector?,” Preprint, 2011.

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

[38] M. Davenport, J. Laska, J. Treichler, and R. Baraniuk, “The pros and cons of compressive sensing: Noise folding and dynamic range,” Preprint, 2011.

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

[40] M. Yuan and Y. Lin, “Model selection and estimation in regression with grouped variables,” *J. Roy. Stat. Soc. Ser. B*, vol. 68, no. 1, pp. 49–67, 2006.

[41] M. Duarte, W. Bajwa, and R. Calderbank, “Regression performance of group lasso for arbitrary design matrices,” in *Sampling Theory and Applications (SampTA)*, Singapore, May 2011.

[42] E. J. Candès and Y. Plan, “A probabilistic and RIPless theory of compressed sensing,” *IEEE Trans. Inform. Theory*, vol. 57, no. 11, pp. 7235–7254, 2010.

[43] W. Bajwa, R. Calderbank, and S. Jafarpour, “Why Gabor frames? Two fundamental measures of coherence and their role in model selection,” *J. Commun. Netw.*, vol. 12, no. 4, pp. 289–307, Aug. 2010.

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

[45] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, “A simple proof of the restricted isometry property for random matrices,” *Const. Approx.*, vol. 28, no. 3, pp. 253–263, 2008.

[46] M. Davenport, *Random observations on random observations: Sparse signal acquisition and processing*, Ph.d. thesis, Rice University, Aug. 2010.

[47] J. Romberg, “Compressive sensing by random convolution,” *SIAM J. on IMaging Science*, vol. 2, no. 4, pp. 1098–1128, Nov. 2009.

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

[49] M. Mishali and Y. Eldar, “From theory to practice: Sub-Nyquist sampling of sparse wideband analog signals,” *IEEE J. of Selected Topics on Signal Proc.*, vol. 4, no. 2, pp. 375–391, 2010.

[50] L. Jacques, P. Vandergheynst, A. Bibet, V. Majidzadeh, A. Schmid, and Y. Leblebici, “Cmos compressed imaging by random convolution,” in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, 2009, pp. 1113–1116, doi:10.1109/ICASSP.2009.4959783.

[51] B.K. Natarajan, “Sparse Approximate Solutions to Linear Systems,” *SIAM Journal on Computing*, vol. 24, pp. 227, 1995.

[52] S. S. Chen, D.L. Donoho, and M.A. Saunders, “Atomic decomposition by basis pursuit,” *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 33–61, 1998.

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

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

[55] S. Osher, Y. Mao, B. Dong, and W. Yin, “Fast linearized Bregman iterations for compressive sensing and sparse denoising,” *Comm. in Math. Sciences*, vol. 8, no. 1, pp. 93–111, 2010.

[56] R. Tibshirani, “Regression shrinkage and selection via the Lasso,” *J. Royal Statist. Soc B*, vol. 58, no. 1, pp. 267–288, 1996.

[57] P. Bickel, Y. Ritov, and A. Tsybakov, “Simultaneous analysis of Lasso and Dantzig selector,” *Ann. Stat.*, vol. 37, no. 4, pp. 1705–1732, 2009.

[58] F. Ye and C. Zhang, “Rate minimaxity of the Lasso and Dantzig selector for the ℓ_q loss in ℓ_r balls,” *J. Machine Learning Research*, vol. 11, pp. 3519–3540, 2010.

[59] I. Drori and D. Donoho, “Solution of ℓ_1 minimization problems by LARS / homotopy methods,” in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing (ICASSP)*, Toulouse, France, May 2006.

[60] J. Haupt and R. Nowak, “Signal reconstruction from noisy random projections,” *IEEE Trans. Inform. Theory*, vol. 52, no. 9, pp. 4036–4048, Sept. 2006.

[61] J. Tropp, “Greed is good: Algorithmic results for sparse approximation,” *IEEE Trans. Inform. Theory*, vol. 50, no. 10, pp. 2231–2242, 2004.

[62] Wenyu Sun and Ya-Xiang Yuan, *Optimization Theory and Methods*, vol. 1 of *Springer Optimization and Its Applications*, Springer, New York, 2006, Nonlinear programming.

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

[64] L. Jacques, D. Hammond, and M. Fadili, “Dequantizing compressed sensing: When oversampling and non-gaussian contraints combine,” *IEEE Trans. Inform. Theory*, vol. 57, no. 1, pp. 559–571, 2011.

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

[66] A. Zymnis, S. Boyd, and E. Candès, “Compressed sensing with quantized measurements,” *IEEE Signal Processing Letters*, vol. 17, no. 2, Feb. 2010.

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

[68] U. Kamilov, V. Goyal, and S. Rangan, “Message-passing estimation from quantized samples,” Preprint, 2011.

[69] U. Kamilov, V. Goyal, and S. Rangan, “Optimal quantization for compressive sensing under message passing reconstruction,” in *IEEE Int. Symp. on Inform. Thy. (ISIT)*, St. Petersburg, Jul. 2011, pp. 459–463.

[70] R. Lyons, *Understanding digital signal processing*, Pearson Education, Inc., Upper Saddle River, NJ, 2004.

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

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

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

[74] L. Jacques, J. Laska, P. Boufounos, and R. Baraniuk, “Robust 1-bit compressive sensing via binary stable embeddings,” *Preprint*, 2011.

[75] J. Laska, Z. Wen, W. Yin, and R. Baraniuk, “Trust, but verify: Fast and accurate signal recovery from 1-bit compressive measurements,” *Submitted*, 2010.

[76] P. Boufounos, “Greedy sparse signal reconstruction from sign measurements,” in *Proc. Asilomar Conf. on Signals Systems and Comput.*, Asilomar, California, Nov. 2009.

[77] P. Boufounos, *Quantization and erasures in frame representations*, Ph.D. thesis, MIT EECS, Cambridge, MA, Jan. 2006.

[78] P. Boufounos and R. Baraniuk, “Quantization of sparse representations,” in *Proc. data compression conference (DCC)*, Mar. 2007, p. 378.

[79] J. Z. Sun and V. K. Goyal, “Optimal quantization of random measurements in compressed sensing,” in *IEEE Int. Symp. on Inform. Thy. (ISIT)*, June 2009.

[80] N. Thao and M. Vetterli, “Reduction of the mse in R-times oversampled A/D conversion $O(1/R)$ to $O(1/R^2)$,” *IEEE Trans. Signal Processing*, vol. 42, no. 1, pp. 200–203, 2994.

[81] V. K. Goyal, M. Vetterli, and N. Thao, “Quantized overcomplete expansions in $\mathbb{R}^n$: Analysis, synthesis, and algorithms,” *IEEE Trans. Inform. Theory*, vol. 44, no. 1, pp. 16–31, 1998.

[82] P. M. Aziz, H. V. Sorensen, and J. Van Der Spiegel, “An overview of Sigma-Delta converters,” *IEEE Signal Processing Magazine*, vol. 13, no. 1, pp. 61–84, Jan. 1996.

[83] N. T. Thao, “Vector quantization analysis of $\Sigma\Delta$ modulation,” *IEEE Trans. Signal Processing*, vol. 44, no. 4, pp. 808–817, Apr. 1996.

[84] J. C. Candy and G. C. Temes, Eds., *Oversampling Delta-Sigma Converters*, IEEE Press, 1992.

[85] J. J. Benedetto, A. M. Powell, and O. Yilmaz, “Sigma-delta quantization and finite frames,” *IEEE Trans. Info. Theory*, vol. 52, no. 5, pp. 1990–2005, May 2006.

[86] P. Boufounos and A. V. Oppenheim, “Quantization noise shaping on arbitrary frame expansions,” *EURASIP Journal on Applied Signal Processing, Special issue on Frames and Overcomplete Representations in Signal Processing, Communications, and Information Theory*, vol. 2006, pp. Article ID 53807, 12 pages, DOI:10.1155/ASP/2006/53807, 2006.

[87] P. Boufounos and R. Baraniuk, “Sigma Delta Quantization for Compressive Sensing,” in *Proc. SPIE Waveletts XII. Proc. of SPIE Vol. 6701, 670104*, San Diego, CA, Aug. 2007.

[88] S. Güntürk an A. Powell, R. Saab, and Ö. Yilmaz, “Sobolev Duals for Random Frames and Sigma-Delta Quantization of Compressed Sensing Measurements,” Preprint, 2010.

[89] P. Boufounos, “Universal rate-efficient scalar quantization,” *to appear in IEEE Trans. Inform. Theory*, 2011.

[90] H. Q. Nguyen, V.K. Goyal, and L.R. Varshney, “Frame Permutation Quantization,” *Applied and Computational Harmonic Analysis (ACHA)*, Nov. 2010.

[91] P. Indyk and R. Motwani, “Approximate nearest neighbors: towards removing the curse of dimensionality,” in *Proceedings of the thirtieth annual ACM symposium on Theory of computing.* ACM, 1998, pp. 604–613.

[92] A. Andoni, M. Datar, N. Immorlica, P. Indyk, and V. Mirrokni, “Locality-sensitive hashing scheme based on p-stable distributions,” *Nearest neighbor methods in learning and vision: Theory and practice (book)*, 2006.

[93] A. Andoni and P. Indyk, “Near-optimal hashing algorithms for approximate nearest nieghbor in high dimensions,” *Commun. ACM*, vol. 51, no. 1, pp. 117–122, 2008.

[94] M. Raginsky and S. Lazebnik, “Locality-sensitive binary codes from shift-invariant kernels,” *The Neural Information Processing Systems*, vol. 22, 2009.

[95] M. Bridson, “Geometric and combinatorial group theory,” in *The Princeton companion to mathematics*, T. Gowers, J. Barrow-Green, and I. Leader, Eds., chapter IV.10, pp. 431–447. Princeton Univ. Press, 2008.

[96] M. Goemans and D. Williamson, “Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming,” *J. ACM*, vol. 42, no. 6, pp. 1145, 1995.

[97] S. Shariati, L. Jacques, F. X. Standaert, B. Macq, M. A. Salhi, and P. Antoine, “Randomly driven fuzzy key extraction of unclonable images,” in *IEEE Int. conf. on image proc. (ICIP)*, 2010.

[98] A. Gupta, B. Recht, and R. Nowak, “Sample complexity for 1-bit compressed sensing and sparse classification,” in *Proc. Intl. Symp. on Information Theory (ISIT)*, 2010.

[99] H. Nguywn, V. Goyal, and L. Varshney, “Frame permutation quantization,” *Applied and Computational Harmonic Analysis*, 2010.

[100] J. Barzilai and J. Borwein, “Two-point step size gradient methods,” *IMA J. Numer. Anal.*, vol. 8, no. 1, pp. 141–148, 1988.

[101] A. Conn, N. Gould, and P. Toint, *Trust-region methods*, MPS/SIAM Series on Optimization. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2000.

[102] J. Nocedal and S. Wright, *Numerical Optimization*, Springer Ser. in Operations Res. and Financial Eng. Springer, second edition, 2006.

[103] N. S. Aybat and G. Iyengar, “A first-order augmented Lagrangian method for compressed sensing,” Preprint, 2010.

[104] C. Li, “An efficient algorithm for total variation regularization with applications to the signle pixel camera and compressive sensing,” M.S. Thesis, Rice University, 2009.

[105] T. Blumensath, “Compressed sensing with nonlinear observations,” Preprint, 2010.

[106] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer Series in Statistics, 2nd edition, Feb. 2009.

[107] P. J. Huber, “Robust regression: Asymptotics, conjectures, and Monte Carlo,” *Statistics*, vol. 1, pp. 799–821, 1973.

[108] Y. Plan and R. Vershynin, “One-bit compressed sensing by linear programming,” Preprint, 2011.

[109] L. Rosasco, E. De Vito, A. Caponnetto, M. Piana, and A. Verri, “Are loss functions all the same?,” *Neural Comp.*, vol. 16, no. 5, pp. 1063–1076, Mar. 2004.

[110] N. Srebro, K. Sridharan, and A. Tewari, “Smoothness, low-noise, and fast rates,” in *Advances in Neural Information Processing Systems (NIPS)*, Dec. 2010.

[111] O. Bartlett, Y. Freund, W. S. Lee, and R. E. Schapire, “Boosting the margin: a new explanation for the effectiveness of voting methods,” *Annals of Statistics*, vol. 26, no. 5, pp. 1651–1686, 1998.

[112] G. Ratsch and M. Warmuth, “Efficient margin maximizing with boosting,” *The J. of Machine Learning Research*, vol. 6, Dec. 2005.

[113] A. Blum, *On-Line Algorithms in Machine Learning*, vol. 1442/1998 of *Lecture Notes in Computer Science*, Springer, 1998.

[114] A. Bourquard, F. Aguet, and M. Unser, “Optical imaging using binary sensors,” *Optics Express*, vol. 18, no. 5, Mar. 2010.

[115] T. Cover and J. Thomas, *Elements of Information Theory*, Wiley-Interscience, New York, NY, 1991.

[116] S. Sarvotham, D. Baron, and R. Baraniuk, “Measurements vs. bits: Compressed sensing meets information theory,” in *Proc. Allerton Conf. Communication, Control, and Computing*, Monticello, IL, Sept. 2006.

[117] J. Laska and R. Baraniuk, “Regime change: Bit-depth vs. measurement rate in compressive sensing,” Tech. Rep., Preprint, 2011.

[118] R. M. Gray and D. L. Neuhoff, “Quantization,” *IEEE Trans. Info. Theory*, vol. 44, no. 6, pp. 2325–2383, 1998.

[119] Z. Cvetković and I. Daubechies, “Single-bit oversampled A/D conversion with exponential accurace in the bit-rate,” *IEEE Trans. Info. Theory*, vol. 53, no. 11, pp. 3979–3989, 2007.

[120] S. Hoyos, B. Sadler, and G. Arce, “Monobit digital receivers for ultrawideband communications,” *IEEE Trans. on Wireless Comm.*, vol. 4, no. 4, pp. 1337–1344, July 2005.

[121] R. Feynman, “There’s plenty of room at the bottom,” in *Caltech Eng. and Sci.*, American Physical Society, Ed., 1960, vol. 23, pp. 22–36.

[122] E. Candès and J. Romberg, “Sparsity and incoherence in compressive sampling,” *Inverse Problems*, vol. 23, no. 3, pp. 969–985, 2007.

[123] J. Haupt, R. Castro, R. Nowak, G. Fudge, and A. Yeh, “Compressive sampling for signal classification,” in *Proc. Asilomar Conf. Signals, Systems, and Computers*, Pacific Grove, CA, Nov. 2006.

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

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

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

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

[128] I. Maravić and M. Vetterli, “Sampling and reconstruction of signals with finite innovation in the presence of noise,” *IEEE Trans. Signal Processing*, vol. 53, no. 8, pp. 2788–2805, 2005.

[129] J. Laska, J.P. Slavinsky, and R. Baraniuk, “The polyphase random demodulator for wideband compressive sensing,” in *Proc. Asilomar Conf. on Signals Systems and Comput.*, 2011.

[130] Z. Yu, S. Hoyos, and B. Sadler, “Mixed-signal parallel compressed sensing and reception for cognitive radio,” in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing (ICASSP)*, Las Vegas, NV, Apr. 2008, pp. 3861–3864.

[131] M. A. Lexa, M. E. Davies, and J. S. Thompson, "Reconciling compressive sampling schemes for spectrally-sparse continuous-time signals," *to appear in IEEE Trans. Sig. Proc.*, 2011.

[132] E. Candès, "Some applications and hardware implementations of compressive sensing," LMS Invited Lecturer Series, Center for Mathematical Sciences, Mar. 2011.

[133] A. Harms, W. Bajwa, and R. Calderbank, "Beating Nyquist through correlations: A constrained random demodulator for sampling of sparse bandlimited signals," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing (ICASSP)*, 2011.

[134] J. Romberg, "Multiple channel estimation using spectrally random probes," in *SPIE Wavelets XIII*, Aug. 2009.

[135] J. Romberg and R. Neelamani, "Sparse channel separation using random probes," *submitted to Inverse Problems*, Feb. 2010.

[136] S. Sardy, A. Bruce, and P. Tseng, "Block coordinate relxation methods for nonparametric wavelet denoising," *J. of Comp. and Graph. Stat.*, vol. 9, no. 2, pp. 361–379, Jun. 2000.

[137] J. Bobin, Y. Moudden, and J. Starck, "Morphological diversity and sparsity for multichannel data restoration," *J. Math Imaging Vis*, 2008.

[138] E. Candès, Y. Eldar, and D. Needell, "Compressed sensing with coherent and redundant dictionaries," *submitted*, 2010.

[139] T. M. Cover, "Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition," *IEEE Trans. on Electronic Comp.*, vol. 3, pp. 326–334, Jun. 1965.

[140] L. Flatto, "A new proof of the transposition theorem," *Proc. American Mathematical Society*, vol. 24, no. 1, pp. 29–31, jan 1970.

[141] K. Ball, "An elementary introduction to modern convex geometry," in *Flavors of Geometry*, Silvio Levy, Ed., vol. 31 of *MSRI Publications*, pp. 1–58. Cambridge University Press, Cambridge, UK, 1997.

[142] T. Strohmer and R. W. Heath, "Grassmannian frames with applications to coding and communication," *Applied and Computational Harmonic Analysis*, vol. 14, no. 3, pp. 257–275, 2003.

[143] N. Johnson, S. Kotz, and N. Balakrishnan, *Continuous Univariate Distributions, Volume 1*, Wiley, 1994.