Exponential decay of reconstruction error from binary measurements of sparse signals

Richard Baraniuk[r], Simon Foucart[g], Deanna Needell[c], Yaniv Plan[b], Mary Wootters[m*†]

[r]Department of Electrical and Computer Engineering, Rice University,
6100 Main Street, Houston, TX 77005 USA. Email: richb@rice.edu

[g]Department of Mathematics, University of Georgia
321C Boyd Building, Athens, GA 30602 USA. Email: foucart@math.uga.edu

[c]Department of Mathematical Sciences, Claremont McKenna College,
850 Columbia Ave., Claremont, CA 91711, USA. Email: dneedell@cmc.edu

[b]Department of Mathematics, University of British Columbia
1984 Mathematics Road, Vancouver, B.C. Canada V6T 1Z2. Email: yaniv@math.ubc.ca

[m]Department of Mathematics, University of Michigan
530 Church Street, Ann Arbor, MI 48109, USA. Email: wootters@umich.edu

July 30, 2014

Abstract

Binary measurements arise naturally in a variety of statistical and engineering applications. They may be inherent to the problem—e.g., in determining the relationship between genetics and the presence or absence of a disease—or they may be a result of extreme quantization. A recent influx of literature has suggested that using prior signal information can greatly improve the ability to reconstruct a signal from binary measurements. This is exemplified by *one-bit compressed sensing*, which takes the compressed sensing model but assumes that only the sign of each measurement is retained. It has recently been shown that the number of one-bit measurements required for signal estimation mirrors that of unquantized compressed sensing. Indeed, s-sparse signals in $\mathbb{R}^n$ can be estimated (up to normalization) from $\Omega(s \log(n/s))$ one-bit measurements. Nevertheless, controlling the precise accuracy of the error estimate remains an open challenge. In this paper, we focus on optimizing the decay of the error as a function of the oversampling factor $\lambda := m/(s \log(n/s))$, where m is the number of measurements. It is known that the error in reconstructing sparse signals from standard one-bit measurements is bounded below by $\Omega(\lambda^{-1})$. Without adjusting the measurement procedure, reducing this polynomial error decay rate is impossible. However, we show that an adaptive choice of the thresholds used for quantization may lower the error rate to $e^{-\Omega(\lambda)}$. This improves upon guarantees for other methods of adaptive thresholding as proposed in Sigma-Delta quantization. We develop

[*] Authors are listed in alphabetical order.

[†] This research was performed as part of the AIM SQuaRE program. In addition, Baraniuk is partially supported by NSF grant number CCF-0926127 and ARO MURI grant number W911NF-09-1-0383, Foucart by NSF grant number DMS–1120622, Needell by Simons Foundation Collaboration grant number 274305, Alfred P. Sloan Fellowship and NSF Career grant number 1348721, Plan by an NSF Postdoctoral Research Fellowship grant number 1103909, Wootters by a Rackham Predoctoral Fellowship and by the Simons Institute for the Theory of Computing.

a general recursive strategy to achieve this exponential decay and two specific polynomial-time algorithms which fall into this framework, one based on convex programming and one on hard thresholding. This work is inspired by the one-bit compressed sensing model, in which the engineer controls the measurement procedure. Nevertheless, the principle is extendable to signal reconstruction problems in a variety of binary statistical models as well as statistical estimation problems like logistic regression.

Keywords. compressed sensing, quantization, one-bit compressed sensing, convex optimization, iterative thresholding, binary regression

1 Introduction

Many practical acquisition devices in signal processing and algorithms in machine learning use a small number of linear measurements to represent a high-dimensional signal. Compressed sensing is a technology which takes advantage of the fact that, for some interesting classes of signals, one can use far fewer measurements than dictated by traditional Nyquist sampling paradigm. In this setting, one obtains m linear measurements of a signal $\boldsymbol{x} \in \mathbb{R}^n$ of the form

$$y_i = \langle \boldsymbol{a}_i, \boldsymbol{x} \rangle \,, \qquad i = 1, \ldots, m.$$

Written concisely, one obtains the measurement vector $\boldsymbol{y} = \mathbf{A}\boldsymbol{x}$, where $\mathbf{A} \in \mathbb{R}^{m \times n}$ is the matrix with rows $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m$. From these (or even from corrupted measurements $\boldsymbol{y} = \mathbf{A}\boldsymbol{x} + \boldsymbol{e}$), one wishes to recover the signal $\boldsymbol{x}$. To make this problem well-posed, one must exploit a priori information on the signal $\boldsymbol{x}$, for example that it is *s-sparse*, i.e.,

$$\|\boldsymbol{x}\|_0 \overset{\text{def}}{=} |\text{supp}(\boldsymbol{x})| = s \ll n,$$

or is well-approximated by an s-sparse signal. After a great deal of research activity in the past decade (see the website [DSP] or the references in the monographs [EK12, FR13]), it is now well known that when $\mathbf{A}$ consists of, say, independent standard normal entries, one can, with high probability, recover all s-sparse vectors $\boldsymbol{x}$ from the $m \approx s \log(n/s)$ linear measurements $y_i = \langle \boldsymbol{a}_i, \boldsymbol{x} \rangle$, $i = 1, \ldots, m$.

However, in practice, the compressive measurements $\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle$ must be quantized: one actually observes $\boldsymbol{y} = Q(\mathbf{A}\boldsymbol{x})$, where the map $Q : \mathbb{R}^m \to \mathcal{A}^m$ is a quantizer that acts entrywise by mapping each real-valued measurement to a discrete quantization alphabet $\mathcal{A}$. This type of quantization with an alphabet $\mathcal{A}$ consisting of only two elements was introduced in the compressed sensing setting by [BB08] and dubbed *one-bit compressed sensing* . In this work, we focus on this one-bit approach and seek quantization schemes Q and reconstruction algorithms Δ so that $\hat{\boldsymbol{x}} = \Delta(Q(\mathbf{A}\boldsymbol{x}))$ is a good approximation to $\boldsymbol{x}$. In particular, we are interested in the trade-off between the error of the approximation and the *oversampling factor*

$$\lambda \overset{\text{def}}{=} \frac{m}{s \log(n/s)}.$$

1.1 Motivation and previous work

The most natural quantization method is *Memoryless Scalar Quantization* (MSQ), where each entry of $\boldsymbol{y} = \mathbf{A}\boldsymbol{x}$ is rounded to the nearest element of some quantization alphabet $\mathcal{A}$. If $\mathcal{A} = \delta\mathbb{Z}$ for

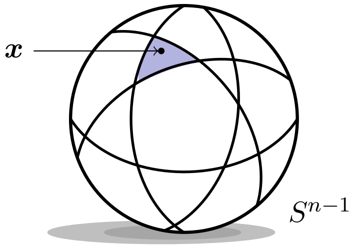

Figure 1: *Geometric interpretation of one-bit compressed sensing.* Each quantized measurement reveals which side of a hyperplane (or great circle, when restricted to the sphere) the signal $\boldsymbol{x}$ lies on. After several measurements, we know that $\boldsymbol{x}$ lies in one unique region. However, if the measurements are non-adaptive, then as the region of interest becomes smaller, it becomes less and less likely that the next measurement yields any new information about $\boldsymbol{x}$.

some suitably small $\delta > 0$, then this rounding error can be modeled as an additive measurement error [DPM09], and the recovery algorithm can be fine-tuned to this particular situation [JHF11]. In the one-bit case, however, the quantization alphabet is $\mathcal{A} = \{\pm 1\}$ and the quantized measurements take the form $\boldsymbol{y} = \operatorname{sign}(\boldsymbol{A}\boldsymbol{x})$, meaning that sign[1] acts entrywise as

$$y_i = Q_{\mathrm{MSQ}}(\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle) = \operatorname{sign}(\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle), \qquad i = 1, \ldots, m.$$

One-bit compressed sensing was introduced in [BB08], and it has generated a considerable amount of work since then, see [DSP] for a growing list of literature in this area. Several efficient recovery algorithms have been proposed, based on linear programming [PV13a, PV13b, GNJN13] and on modifications of iterative hard thresholding [JLBB13, JDDV13]. As shown in [JLBB13], with high probability one can perform the reconstruction from one-bit measurements with error

$$\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2 \lesssim \frac{1}{\lambda} \qquad \text{for all } \boldsymbol{x} \in \Sigma'_s := \{\boldsymbol{v} \in \mathbb{R}^n \ : \ \|\boldsymbol{v}\|_0 \le s, \|\boldsymbol{v}\|_2 = 1\}.$$

In other words, a uniform ℓ_2-reconstruction error of at most $\gamma > 0$ can be achieved with $m \asymp \gamma^{-1} s \log(n/s)$ one-bit measurements.

Despite the dimension reduction from n to $s\log(n/s)$, MSQ presents substantial limitations [JLBB13, GVT98]. Precisely, according to [GVT98], even if the support of $\boldsymbol{x} \in \Sigma'_s$ is known, the best recovery algorithm Δ_{opt} must obey

$$\|\boldsymbol{x} - \Delta_{\mathrm{opt}}(Q_{\mathrm{MSQ}}(\mathbf{A}\boldsymbol{x}))\|_2 \gtrsim \frac{1}{\lambda} \tag{1}$$

up to a logarithmic factor. An intuition for the limited accuracy of MSQ is given in Figure 1.

Alternative quantization schemes have been developed to overcome this drawback. For a specific signal model and reconstruction algorithm, [SG09] obtained the optimal quantization scheme, but more general quantization schemes remain open.

Recently, Sigma-Delta quantization schemes have also been proposed as a more general quantization model [GLP+10, KSY14]. These works show that, with high probability on measurement

[1] We define $\operatorname{sign}(0) = 1$.

matrices with independent subgaussian entries, r-th order Sigma-Delta quantization can be applied to the standard compressed sensing problem to achieve, for any $\alpha \in (0,1)$, the reconstruction error

$$\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2 \lesssim_r \lambda^{-\alpha(r-1/2)} \tag{2}$$

with a number of measurements

$$m \approx s\left(\log(n/s)\right)^{1/(1-\alpha)}.$$

For suitable choices of α and r, the guarantee (2) overcomes the limitation (1), but it is still polynomial in λ. This leads us to ask whether an exponential dependence can be achieved.

1.2 Our contributions

In this work, we focus on improving the trade-off between the error $\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2$ and the oversampling factor λ. To the best of our knowledge, all quantized compressed sensing schemes obtain guarantees of the form

$$\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2 \lesssim \lambda^{-c} \qquad \text{for all } \boldsymbol{x} \in \Sigma_s' \tag{3}$$

with some constant $c > 0$. We develop one-bit quantizers $Q : \mathbb{R}^m \to \{\pm 1\}$, coupled with two efficient recovery algorithms $\Delta : \{\pm 1\} \to \mathbb{R}^m$ that yield the reconstruction guarantee

$$\|\boldsymbol{x} - \Delta(Q(\mathbf{A}\boldsymbol{x}))\|_2 \leq \exp(-\Omega(\lambda)) \qquad \text{for all } \boldsymbol{x} \in \Sigma_s'. \tag{4}$$

It is not hard to see that the dependence on λ in (4) is optimal, since any method of quantizing measurements that provides the reconstruction guarantee $\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2 \leq \gamma$ must use at least $\log_2 \mathcal{N}(\Sigma_s', \gamma) \geq s \log_2(1/\gamma)$ bits, where $\mathcal{N}(\cdot)$ denotes the covering number.

1.2.1 Adaptive measurement model

A key element of our approach is that the quantizers are *adaptive* to previous measurements of the signal in a manner similar to Sigma-Delta quantization [GLP$^+$10]. In particular, the measurement matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ is assumed to have independent standard normal entries and the quantized measurements take the form of thresholded signs, i.e.,

$$y_i = \operatorname{sign}(\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle - \tau_i) = \begin{cases} 1 & \text{if} \quad \langle \boldsymbol{a}_i, \boldsymbol{x} \rangle \geq \tau_i, \\ -1 & \text{if} \quad \langle \boldsymbol{a}_i, \boldsymbol{x} \rangle < \tau_i. \end{cases} \tag{5}$$

Such measurements are readily implementable in hardware, and they retain the simplicity and storage benefits of the one-bit compressed sensing model. However, as we will show, this model is much more powerful in the sense that it permits optimal guarantees of the form (4), which are impossible with standard MSQ one-bit quantization. As in the Sigma-Delta quantization approach, we allow the quantizer to be adaptive, meaning that the quantization threshold τ_i of the ith entry may depend on the 1st, 2nd, …, $(i-1)$st quantized measurements. In the context of (5), this means that the thresholds τ_i will be chosen adaptively, resulting in a feedback loop as depicted in Figure 2. The thresholds τ_i can also be interpreted as an additive *dither*, which is oft-used in the theory and practice of analog-to-digital conversion.

In contrast to Sigma-Delta quantization, the feedback loop involves the calculation of the quantization threshold. This is the concession made to arrive at exponentially decaying error rates. It is an interesting open problem to determine low-memory quantization methods with such error rates that do not require such a calculation.

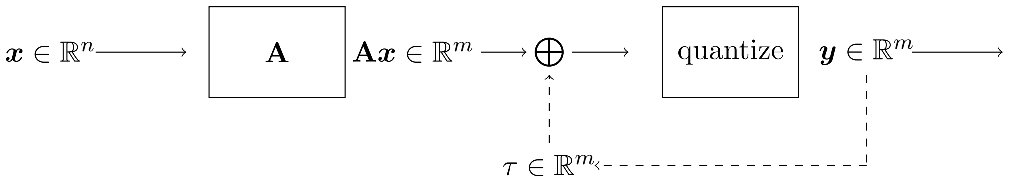
$\boldsymbol{x} \in \mathbb{R}^n$
A
$\mathbf{A}\boldsymbol{x} \in \mathbb{R}^m$
$\bigoplus$
quantize
$\boldsymbol{y} \in \mathbb{R}^m$
$\tau \in \mathbb{R}^m$

Figure 2: Our closed-loop feedback system for binary measurements.

1.2.2 Overview of our main result

Our main result is that there is a recovery algorithm using measurements of the form (5) and providing a guarantee of the form (4). For clarity of exposition, we overview a simplified version of our main result below. The full result is stated in Section 3.

Theorem 1 (Main theorem, simplified version)**.** *Let Q and Δ be the quantization and recovery algorithms given below in Algorithms 1 and 2, respectively. Suppose that $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and $\tau \in \mathbb{R}^m$ have independent standard normal entries. Then, with probability at least $C\lambda \exp(-cs\log(n/s))$ over the choice of $\boldsymbol{A}$ and τ, for all $\boldsymbol{x} \in B_2^n$ with $\|\boldsymbol{x}\|_0 \le s$,*

$$\|\boldsymbol{x} - \Delta(Q(\boldsymbol{A}\boldsymbol{x}, \boldsymbol{A}, \tau))\|_2 \le \exp(-\Omega(\lambda)), \qquad \textit{where} \qquad \lambda = \frac{m}{s\log(n/s)}.$$

The quantization algorithm works iteratively. First, a small batch of measurements are quantized in a memoryless fashion. From this first batch, one gains a very rough estimate of $\boldsymbol{x}$ (called $\boldsymbol{x}_1$). The next batch of measurements are quantized with a focus on encoding the difference between $\boldsymbol{x}$ and $\boldsymbol{x}_1$, and so on. Thus, the trap depicted in Figure 1 is avoided; each hyperplane is translated with an appropriate dither, with the aim of cutting the size of the feasible region. The recovery algorithm also works iteratively and its iterates are in fact intertwined with the iterates of the quantization algorithm. We artificially separate the two algorithms below.

Note that we present Algorithms 1 and 2 at this point mainly because they are the simplest to state. Below we will provide a more general framework for algorithms that satisfy the guarantees of Theorem 1 and develop a second set of algorithms with computational advantages.

1.2.3 Robustness

Our algorithms are robust to two different kinds of measurement corruption. First, they allow for perturbed linear measurements of the form $\langle \boldsymbol{a}_i, \boldsymbol{x}\rangle + e_i$ for an error vector $\boldsymbol{e} \in \mathbb{R}^m$ with bounded ℓ_∞-norm. Second they allow for post-quantization sign flips, recorded as a vector $\mathbf{f} \in \{\pm 1\}^m$.

Formally, the measurements take the form

$$y_i = f_i \operatorname{sign}(\langle \boldsymbol{a}_i, \boldsymbol{x}\rangle - \tau_i + e_i), \qquad i = 1, \ldots, m. \tag{6}$$

It is known that for inaccurate measurements with pre-quantization noise on the same order of magnitude as the signal, even unquantized compressed sensing algorithms must obey a lower bound of the form (1) [CD13]. Our algorithms respect this reality and exhibit exponentially fast convergence until the estimate hits the "noise floor"—that is, until the error $\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2$ is on the order of $\|\boldsymbol{e}\|_\infty$.

Table 1 summarizes the various noise models, adaptive threshold calculations, and algorithms we develop and study below.

Algorithm 1: Adaptive quantization

```
**Input**: Linear measurements $\mathbf{A}\boldsymbol{x} \in \mathbb{R}^m$; measurement matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$; sparsity parameter $s$; thresholds $\tau \in \mathbb{R}^m$; parameter $q \geq Cs\log(n/s)$ for the size of batches.
**Output**: Quantized measurements $\boldsymbol{y} \in \{\pm 1\}^m$.
$T \leftarrow \left\lfloor \frac{m}{q} \right\rfloor$
Partition $\boldsymbol{A}$ and $\tau$ into $T$ blocks $\boldsymbol{A}^{(1)}, \ldots, \boldsymbol{A}^{(T)} \in \mathbb{R}^{q\times n}$ and $\tau^{(1)}, \ldots .\tau^{(T)} \in \mathbb{R}^q$.
$\boldsymbol{x}_0 \leftarrow \mathbf{0}$
**for** $t = 1, \ldots, T$ **do**
    $\sigma^{(t)} \leftarrow \mathbf{A}^{(t)}\boldsymbol{x}_{t-1}$
    $\boldsymbol{y}^{(t)} \leftarrow \mathrm{sign}(\boldsymbol{A}^{(t)}\boldsymbol{x} - 2^{2-t}\tau^{(t)} - \sigma^{(t)})$
    $\boldsymbol{z}_t \leftarrow \mathrm{argmin}\|\boldsymbol{z}\|_1$   subject to   $\|\boldsymbol{z}\|_2 \leq 2^{2-t},\ y_i^{(t)}\left(\left\langle \boldsymbol{a}_i^{(t)}, \boldsymbol{z}\right\rangle - 2^{2-t}\tau_i^{(t)}\right) \geq 0$   for all $i$
    //$\boldsymbol{z}_t$ is an approximation of $\boldsymbol{x} - \boldsymbol{x}_{t-1}$
    $\boldsymbol{x}_t \leftarrow H_s(\boldsymbol{x}_{t-1} + \boldsymbol{z}_t)$
    //$H_s$ keeps $s$ largest (in magnitude) entries and zeroes out the rest
**return** $\boldsymbol{y}^{(t)}$ for $t = 1, \ldots, T$
//Notice that we discard $\sigma^{(t)}$
```

Table 1: Summary of the noise models, adaptive threshold calculations, and algorithms considered. See Section 2 for further discussion of the trade-offs between the two algorithms.

Noise model	**Threshold algorithm**	**Recovery algorithm**
Additive error e_i in (6)	Algorithm 7, instantiated by Algorithm 3	Convex programming: Algorithm 8, instantiated by Algorithm 4
Additive error e_i and sign flips f_i in (6)	Algorithm 7, instantiated by Algorithm 5	Iterative hard thresholding: Algorithm 8, instantiated by Algorithm 6

1.2.4 Relationship to binary regression

Our one-bit adaptive quantization and reconstruction algorithms are more broadly applicable to a certain kind of statistical classification problem related to sparse binary regression, and in particular sparse logistic and probit regression. These techniques are often used to explain statistical data in which the response variable is binary. In regression, it is common to assume that the data $\{z_i\}$ is generated according to the *generalized linear model*, where $z_i \in \{0, 1\}$ is a Bernoulli random variable satisfying

$$\mathbb{E}\left[z_i\right] = f(\langle \boldsymbol{a}_i, \boldsymbol{x}\rangle) \tag{7}$$

for some function $f : \mathbb{R} \to [0, 1]$. The generalized linear model is equivalent to the noisy one-bit compressed sensing model when the measurements $y_i = 2z_i - 1 \in \{\pm 1\}$ and

$$P(y_i = 1) =: f(\langle \boldsymbol{a}_i, \boldsymbol{x}\rangle),$$

Algorithm 2: Recovery

```
Input: Quantized measurements $\boldsymbol{y} \in \{\pm 1\}^m$; measurement matrix $\mathbf{A}$; sparsity parameter
       $s$; thresholds $\tau \in \mathbb{R}^m$; size of batches $q$.
Output: Approximation $\hat{\boldsymbol{x}} \in \mathbb{R}^n$.
$T \leftarrow \left\lfloor \frac{m}{q} \right\rfloor$
Partition $\boldsymbol{A}$ and $\tau$ into $T$ blocks $\boldsymbol{A}^{(1)}, \dots, \boldsymbol{A}^{(T)} \in \mathbb{R}^{q \times n}$ and $\tau^{(1)}, \dots . \tau^{(T)} \in \mathbb{R}^q$.
$x_0 \leftarrow \mathbf{0}$
for $t = 1, \dots, T$ do
    $\boldsymbol{z}_t \leftarrow \operatorname{argmin} \|\boldsymbol{z}\|_1$   subject to   $\|\boldsymbol{z}\|_2 \leq 2^{2-t},\ y_i^{(t)} \left( \left\langle \boldsymbol{a}_i^{(t)}, \boldsymbol{z} \right\rangle - 2^{2-t} \tau_i^{(t)} \right) \geq 0$   for all $i$
    $\boldsymbol{x}_t = H_s(\boldsymbol{x}_{t-1} + \boldsymbol{z}_t)$
return $\boldsymbol{x}_T$
```

or equivalently, when

$$y_i = \operatorname{sign}(\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle + e_i)$$

with $f(t) := P(e_i \geq -t)$. In summary, one-bit compressed sensing is equivalent to binary regression as long as f is the cumulative distribution function (CDF) of the noise variable e_i. The most commonly used CDFs in binary regression are the inverse logistic link function $f(t) = \frac{1}{1+e^t}$ in logistic regression and the inverse probit link function $f(t) = \int_{-\infty}^{t} \mathcal{N}(s) \mathrm{d}s$ in probit regression. These cases correspond to the noise variable e_i being logistic and Gaussian distributed, respectively.

The new twist here is that the quantization thresholds are selected adaptively; see Section 6.1 for some examples. Specifically, our adaptive threshold measurement model is equivalent to the adaptive binary regression model

$$y_i = \operatorname{sign}(\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle + e_i - \tau_i)$$

with

$$P(y_i = 1) = P(e_i - \tau_i >= -t) = f(t - \tau_i).$$

The effect of τ_i in this adaptive binary regression is equivalent to an offset term added to all measurements y_i. Standard binary regression corresponds to the special case with $\tau_i = 0$.

1.3 Organization

In Section 2, we introduce two methods to recover not only the direction, but also the magnitude, of a signal from one-bit compressed sensing measurements of the form (6). These methods may be of independent interest (in one-bit compressed sensing, only the direction can be recovered), but they do not exhibit the exponential decay in the error that we seek. In Section 3, we will show how to use these schemes as building blocks to obtain (4). The proofs of all of our results are given in Section 4. In Section 5, we present some numerical results for the new algorithms. We conclude in Section 6 with a brief summary.

1.4 Notation

Throughout the paper, we use the standard notation $\|\boldsymbol{v}\|_2 = \sqrt{\sum_i v_i^2}$ for the ℓ_2-norm of a vector $\boldsymbol{v} \in \mathbb{R}^n$, $\|\boldsymbol{v}\|_1 = \sum_i |v_i|$ for its ℓ_1-norm, and $\|\boldsymbol{v}\|_0$ for its number of nonzero entries. A vector $\boldsymbol{v}$ is called s-sparse if $\|\boldsymbol{v}\|_0 \leq s$ and effectively s-sparse if $\|\boldsymbol{v}\|_1 \leq \sqrt{s}\|\boldsymbol{v}\|_2$. We write $H_s(\boldsymbol{v})$ to represent the vector in $\mathbb{R}^n$ agreeing with $\boldsymbol{v}$ on the index set of largest s entries of $\boldsymbol{v}$ (in magnitude) and with the zero vector elsewhere. We use a prime to indicate ℓ_2-normalization, so that $H_s'(\boldsymbol{v})$ is defined as $H_s'(\boldsymbol{v}) := H_s(\boldsymbol{v}) / \left\|H_s(\boldsymbol{v})\right\|_2$. The set $\Sigma_s := \{\boldsymbol{v} \in \mathbb{R}^n : \|\boldsymbol{v}\|_0 \leq s\}$ of s-sparse vectors is accompanied by the set $\Sigma_s' := \{\boldsymbol{v} \in \mathbb{R}^n : \|\boldsymbol{v}\|_0 \leq s, \|\boldsymbol{v}\|_2 = 1\}$ of ℓ_2-normalized s-sparse vectors. For $R > 0$, we write $R\Sigma_s'$ to mean the set $\{\boldsymbol{v} \in \mathbb{R}^n : \|\boldsymbol{v}\|_0 \leq s, \|\boldsymbol{v}\|_2 = R\}$. We also write $B_2^n = \{\boldsymbol{v} \in \mathbb{R}^n : \|\boldsymbol{v}\|_2 \leq 1\}$ for the ℓ_2-ball in $\mathbb{R}^n$ and RB_2^n for the appropriately scaled version. We consider the task of recovering $\boldsymbol{x} \in \Sigma_s$ from measurements of the form (5) or (6) for $i = 1, \ldots, m$. These measurements are organized as a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ with rows $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m$ and a vector $\tau \in \mathbb{R}^m$ of thresholds. Matching the Sigma-Delta quantization model, the $\boldsymbol{a}_i \in \mathbb{R}^n$ may be random but are non-adaptive, while the $\tau_i \in \mathbb{R}$ may be chosen adaptively, in either a random or deterministic fashion. The Hamming distance between sign vectors $\boldsymbol{y}, \tilde{\boldsymbol{y}} \in \{\pm 1\}^m$ is defined as $d_H(\boldsymbol{y}, \tilde{\boldsymbol{y}}) = \sum_i \mathbf{1}_{\{y_i \neq \tilde{y}_i\}}$.

2 Magnitude recovery

Given an s-sparse vector $\boldsymbol{x} \in \mathbb{R}^n$, several convex programs are provably able to extract an accurate estimate of the direction of $\boldsymbol{x}$ from $\operatorname{sign}(\mathbf{A}\boldsymbol{x})$ or $\operatorname{sign}(\mathbf{A}\boldsymbol{x}+\boldsymbol{e})$ [PV13b, PV13a]. However, recovery of the magnitude of $\boldsymbol{x}$ is challenging in this setting [KSW14]. Indeed, all magnitude information about $\boldsymbol{x}$ is lost in measurements of the form $\operatorname{sign}(\mathbf{A}\boldsymbol{x})$. Fortunately, if random (non-adaptive) dither is added before quantization, then magnitude recovery becomes possible, i.e., noise can actually help with signal reconstruction. This observation has also been made in the concurrently written paper [KSW14] and also in the literature on binary regression in statistics [DPvdBW14].

Our main result will show that both the magnitude and direction of $\boldsymbol{x}$ can be estimated with exponentially small error bounds. In this section, we first lay the groundwork for our main result by developing two methods for one-bit signal acquisition and reconstruction that provide accurate reconstruction of both the magnitude and direction of $\boldsymbol{x}$ with polynomially decaying error bounds.

We propose two different order-one recovery schemes. The first is based on second-order cone programming and is simpler but more computationally intensive. The second is based on hard thresholding, is faster, and is able to handle a more general noise model (in particular, random sign flips of the measurements) but requires an adaptive dither. Recall Table 1.

2.1 Second-order cone programming

The size of the appropriate dither/threshold depends on the magnitude of $\boldsymbol{x}$. Thus, let $R > 0$ satisfy $\left\|\boldsymbol{x}\right\|_2 \leq R$. We take measurements of the form

$$y_i = \operatorname{sign}(\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle - \tau_i + e_i), \qquad i = 1, \ldots, q, \tag{8}$$

where $\tau_1, \ldots, \tau_q \sim N(0, 4R^2)$ are known independent normally distributed dithers that are also independent of the rows $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_q$ of the matrix $\mathbf{A}$ and $e_1, \ldots, e_q$ are small deterministic errors (possibly adversarial) satisfying $|e_i| \leq cR$ for an absolute constant c. The following second-order cone program

$$\operatorname{argmin} \|\boldsymbol{z}\|_1 \qquad \text{subject to} \quad \left\|\boldsymbol{z}\right\|_2 \leq 2R, \quad y_i(\langle \boldsymbol{a}_i, \boldsymbol{z} \rangle - \tau_i) \geq 0 \quad \text{for all } i = 1, \ldots, q \tag{9}$$

provides a good estimate of $\boldsymbol{x}$, as formally stated below.

Algorithm 3: T_0: Threshold production for second-order cone programming

```
    Input: Bound $R$ on $\|\boldsymbol{x}\|_2$
    Output: Thresholds $\tau \in \mathbb{R}^q$
    return $\tau \sim N(0, R^2 I_q)$
```

Algorithm 4: Δ_0: Recovery procedure for second-order cone programming

```
    Input: Quantized measurements $\boldsymbol{y} \in \{\pm 1\}^q$; measurement matrix $\boldsymbol{A} \in \mathbb{R}^{q \times n}$; bound $R$ on
           $\|\boldsymbol{x}\|_2$; thresholds $\tau \in \mathbb{R}^q$.
    Output: Approximation $\hat{\boldsymbol{x}}$
    return
    $\operatorname{argmin} \|\boldsymbol{z}\|_1 \quad \text{subject to} \quad \|\boldsymbol{z}\|_2 \le 2R, \quad y_i(\langle \boldsymbol{a}_i, \boldsymbol{z} \rangle - \tau_i) \ge 0 \quad \text{for all } i = 1, \ldots, q.$
```

Theorem 2. *Let $1 \ge \delta > 0$, let $\boldsymbol{A} \in \mathbb{R}^{q \times n}$ have independent standard normal entries, and let $\tau_1, \ldots, \tau_q \in \mathbb{R}$ be independent normal variables with variance $4R^2$. Suppose that $n \ge 2q$ and*

$$q \ge C' \delta^{-4} s \log(n/s).$$

Then, with probability at least $1 - 3\exp(-c_0 \delta^4 q)$ over the choice of $\boldsymbol{A}$ and the dithers $\tau_1, \ldots, \tau_q$, the following holds for all $\boldsymbol{x} \in R B_2^n \cap \Sigma_s$ and $\boldsymbol{e} \in \mathbb{R}^q$ satisfying $\|\boldsymbol{e}\|_\infty \le c\delta^3 R$: for $\boldsymbol{y}$ obeying the measurement model (8)*, the solution $\hat{\boldsymbol{x}}$ to* (9) *satisfies*

$$\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2 \le \delta R.$$

The positive constants C', c and c_0 above are absolute constants.

Remark 1. *The choice of the constraint $\|\boldsymbol{z}\|_2 \le 2R$ and the variance $4R^2$ for the τ_i's allows for the above theoretical guarantees in the presence of pre-quantization error $\boldsymbol{e} \ne \boldsymbol{0}$. However, in the ideal case $\boldsymbol{e} = \boldsymbol{0}$, the guarantees also hold if we impose $\|\boldsymbol{z}\|_2 \le R$ and take a variance of R^2. This more natural choice seems to give better results in practice, even in the presence of pre-quantization error (as R was already an overestimation for $\|\boldsymbol{x}\|_2$). This is the route followed in the numerical experiments of Section 5. It only requires changing 2^{2-t} to 2^{1-t} in Algorithms 1 and 2.*

To fit into our general framework for exponential error decay, it is helpful to think of the program (9) as two separate algorithms: an algorithm T_0 that produces thresholds and an algorithm Δ_0 that performs the recovery. These are formally described in Algorithms 3 and 4.

2.2 Hard thresholding

The convex programming approach is attractive in many respects; in particular, the thresholds/dithers τ_i are non-adaptive, which makes them especially easy to apply in hardware. However, the recovery algorithm Δ_0 in Algorithm 4 can be costly. Further, while the convex programming approach can handle additive pre-quantization error, it cannot necessarily handle post-quantization error (sign flips). In this section, we present an alternative scheme for estimating magnitude, based on iterative hard thresholding that addresses these challenges. The only downside is that the thresholds/dithers τ_i become *adaptive* within the order-one recovery scheme.

Given an s-sparse vector $\boldsymbol{x} \in \mathbb{R}^n$, one can easily extract from $\operatorname{sign}(\mathbf{A}\boldsymbol{x})$ a good estimate for the direction of $\boldsymbol{x}$. For example, we will see that $H_s'(\mathbf{A}^* \operatorname{sign}(\mathbf{A}\boldsymbol{x}))$ is a good approximation of $\boldsymbol{x}/\|\boldsymbol{x}\|_2$.

Algorithm 5: T_0: Threshold production for hard thresholding

```
Input: Measurements $\boldsymbol{Ax} \in \mathbb{R}^q$; measurement matrix $\boldsymbol{A} \in \mathbb{R}^{q\times n}$; sparsity parameter $s$;
       bound $R$ on $\|\boldsymbol{x}\|_2$.
Output: Thresholds $\tau \in \mathbb{R}^q$
Partition $\boldsymbol{Ax}$ into $\boldsymbol{A}_1\boldsymbol{x}$, $\boldsymbol{A}_2\boldsymbol{x} \in \mathbb{R}^{q/2}$.
$\boldsymbol{u} \leftarrow H_s'(\boldsymbol{A}_1^*\mathrm{sign}(\boldsymbol{A}_1\boldsymbol{x}))$
$\boldsymbol{v} \leftarrow V(\boldsymbol{u})$
$\boldsymbol{w} \leftarrow 2R \cdot (\boldsymbol{u}+\boldsymbol{v})$
return $\mathbf{0} \in \mathbb{R}^{q/2}$, $\boldsymbol{A}_2\boldsymbol{w} \in \mathbb{R}^{q/2}$
```

Algorithm 6: Δ_0: Recovery procedure for hard thresholding

```
Input: Quantized measurements $\boldsymbol{y} \in \{\pm 1\}^q$; measurement matrix $\boldsymbol{A} \in \mathbb{R}^{q\times n}$; sparsity
       parameter $s$; bound $R$ on $\|\boldsymbol{x}\|_2$.
Output: Approximation $\hat{\boldsymbol{x}}$
Partition $\boldsymbol{y}$ into $\boldsymbol{y}_1$, $\boldsymbol{y}_2 \in \mathbb{R}^{q/2}$.
$\boldsymbol{u} \leftarrow H_s'(\boldsymbol{A}_1^*\boldsymbol{y}_1)$
$\boldsymbol{v} \leftarrow V(\boldsymbol{u})$
$\boldsymbol{t} \leftarrow -H_s'(\boldsymbol{A}_2^*\boldsymbol{y}_2)$
return $2Rf(\langle \boldsymbol{t}, \boldsymbol{v}\rangle)\cdot \boldsymbol{u}$, where $f(\xi) = 1 - \frac{\sqrt{1-\xi^2}}{\xi}$
```

However, as mentioned earlier, there is no hope of recovering the magnitude $\|\boldsymbol{x}\|_2$ of the signal from $\mathrm{sign}(\mathbf{A}\boldsymbol{x})$. To get around this, we use a second estimator, this time for the direction of $\boldsymbol{x}-\boldsymbol{z}$ for a well-chosen vector $\boldsymbol{z} \in \mathbb{R}^n$ obtained by computing $H_s'(\mathbf{A}^*\mathrm{sign}(\mathbf{A}(\boldsymbol{x}-\boldsymbol{z})))$. This allows us to estimate both the direction and the magnitude of $\boldsymbol{x}$.

As above, we break the measurement/recovery process into two separate algorithms. The first is an algorithm T_0 describing how to generate the thresholds τ_i. The second is a recovery algorithm Δ_0 that describes how to recover an approximation $\hat{\boldsymbol{x}}$ to $\boldsymbol{x}$ based on measurements of the form (6), using the τ_i as thresholds. These are formally described in Algorithms 5 and 6. In the algorithm statements, V denotes any fixed rule associating to a vector $\boldsymbol{u}$ an ℓ_2-normalized vector $V(\boldsymbol{u})$ that is both orthogonal to $\boldsymbol{u}$ and has the same support.

The analysis for T_0 and Δ_0 relies on the following theorems.

Theorem 3. *Let $1 \geq \delta > 0$ and let $\boldsymbol{A} \in \mathbb{R}^{q\times n}$ have independent standard normal entries. Suppose that $n \geq 2q$ and $q \geq c_1\delta^{-7}s\log(n/s)$. Then, with probability at least $1 - c_2\exp(-c_3\delta^2 q)$ over the choice of $\boldsymbol{A}$, the following holds for all s-sparse $\boldsymbol{x} \in \mathbb{R}^n$, all $\boldsymbol{e} \in \mathbb{R}^q$ with $\|\boldsymbol{e}\|_2 \leq c_6\sqrt{q}\,\|\boldsymbol{x}\|_2$, and all $\boldsymbol{y} \in \{\pm1\}^q$:*

$$\left\| \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2} - H_s'\left(\mathbf{A}^*\boldsymbol{y}\right)\right\|_2 \leq \delta + c_4\frac{\|\boldsymbol{e}\|_2}{\sqrt{q}\,\|\boldsymbol{x}\|_2} + c_5\sqrt{\frac{d_H(\boldsymbol{y}, \mathrm{sign}\left(\mathbf{A}\boldsymbol{x}+\boldsymbol{e}\right))}{q}} \qquad (10)$$

The positive constants c_1, c_2, c_3, c_4, c_5, and c_6 above are absolute constants.

The proof of Theorem 3 is given in Section 4. Once Theorem 3 is shown, we will be able to establish the following results when the threshold production and recovery procedures T_0 and Δ_0 are given by Algorithms 5 and 6.

Theorem 4. *Let $1 \geq \delta > 0$, let $\boldsymbol{A} \in \mathbb{R}^{q \times n}$ have independent standard normal entries, and let T_0 and Δ_0 be as in Algorithms 5 and 6. Suppose that $n \geq 2q$ and*

$$q \geq c_1 \delta^{-7} s \log(n/s).$$

Further assume that whenever a signal $\boldsymbol{z}$ is measured, the corruption errors satisfy $\|\mathbf{e}\|_\infty \leq c\delta\|\boldsymbol{z}\|_2$ and $|\{i : f_i = -1\}| \leq c'\delta q$. Then, with probablity at least $1 - c_7 \exp(-c_8 \delta^2 q)$ over the choice of $\boldsymbol{A}$, the following holds for all $\boldsymbol{x} \in RB_2^n \cap \Sigma_s$: for $\boldsymbol{y}$ obeying the measurement model (6) with $\tau = T_0(\boldsymbol{Ax}, \boldsymbol{A}, s, R)$, the vector $\hat{\boldsymbol{x}} = \Delta_0(\boldsymbol{y}, \boldsymbol{A}, s, R)$ satisfies

$$\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2 \leq \delta R.$$

The positive constants c_1, c, c', c_7, and c_8 above are absolute constants.

Having proposed two methods for recovering both the direction and magnitude of a sparse vector from binary measurements, we now turn to our main result.

3 Exponential decay: General framework

In the previous section, we developed two methods for approximately recovering $\boldsymbol{x}$ from binary measurements. Unfortunately, these methods exhibit polynomial error decay in the oversampling factor, and our goal is to obtain an exponential decay. We can achieve this goal by applying the rough estimation methods iteratively, in batches, with *adaptive thresholds/dithers*. As we show below, this leads to an extremely accurate recovery scheme. To make this framework precise, we first define an *order-one recovery scheme* (T_0, Δ_0).

Definition 5 (Order-one recovery scheme)**.** *An order-one recovery scheme with sparsity parameter s, measurement complexity q, and noise resilience (η, b) is a pair of algorithms (T_0, Δ_0) such that:*

- *The* thresholding algorithm T_0 *takes a parameter R and, optionally, a set of linear measurements $\boldsymbol{Ax} \in \mathbb{R}^q$ and the measurement matrix $\boldsymbol{A} \in \mathbb{R}^{q \times n}$. It outputs a set of thresholds $\tau \in \mathbb{R}^q$.*
- *The* recovery algorithm Δ_0 *takes q corrupted quantized measurements of the form (6), i.e.,*

  $$y_i = f_i \operatorname{sign}(\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle - \tau_i + e_i),$$

  where $\boldsymbol{e} \in \mathbb{R}^q$ is a pre-quantization error and $\mathbf{f} \in \{\pm 1\}^q$ is a post-quantization error. It also takes as input the measurement matrix $\boldsymbol{A} \in \mathbb{R}^{q \times n}$, a parameter R, and, optionally, a sparsity parameter s and the thresholds τ returned by T_0. It outputs a vector $\hat{\boldsymbol{x}} \in \mathbb{R}^n$.
- *With probability at least $1 - C\exp(-cq)$ over the choice of $\boldsymbol{A} \in \mathbb{R}^{q \times n}$ and the randomness of T_0, the following holds: for all $\boldsymbol{x} \in RB_2^n \cap \Sigma_s$, all $\boldsymbol{e} \in \mathbb{R}^q$ with $\|\boldsymbol{e}\|_\infty \leq \eta\|\boldsymbol{x}\|_2$, and all $\mathbf{f} \in \{\pm 1\}^q$ with at most b sign flips, the estimate $\hat{\boldsymbol{x}} = \Delta_0(\boldsymbol{y}, \boldsymbol{A}, R, s, \tau)$ satisfies*

  $$\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2 \leq \frac{R}{4}.$$

Algorithm 7: Q: Quantization

```
Input: Linear measurements $\boldsymbol{Ax} \in \mathbb{R}^m$; measurement matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$; sparsity parameter $s$; bound $R$ on $\|\boldsymbol{x}\|_2$; parameter $q \geq Cs\log(n/s)$ for the size of batches.
Output: Quantized measurements $\boldsymbol{y} \in \{\pm 1\}^m$ and thresholds $\tau \in \mathbb{R}^m$
$T \leftarrow \left\lfloor \frac{m}{q} \right\rfloor$
Partition $\boldsymbol{A}$ into $T$ blocks $\boldsymbol{A}^{(1)}, \dots, \boldsymbol{A}^{(T)} \in \mathbb{R}^{q\times m}$
$\boldsymbol{x}_0 \leftarrow \boldsymbol{0}$
for $t = 1, \dots, T$ do
    $R_t = 2^{-t+1}$
    $\tau^{(t)} \leftarrow T_0(\boldsymbol{A}^{(t)}(\boldsymbol{x} - \boldsymbol{x}_{t-1}), \boldsymbol{A}^{(t)}, R_t)$
    $\sigma^{(t)} \leftarrow \boldsymbol{A}^{(t)}\boldsymbol{x}_{t-1}$
    $\boldsymbol{y}^{(t)} \leftarrow \mathbf{f}^{(t)} \odot \operatorname{sign}(\boldsymbol{A}^{(t)}\boldsymbol{x} - \tau^{(t)} - \sigma^{(t)} + \boldsymbol{e}^{(t)})$
    $\boldsymbol{x}_t \leftarrow H_s(\boldsymbol{x}_{t-1} + \Delta_0(\boldsymbol{y}^{(t)}, \boldsymbol{A}^{(t)}, R_t, \tau^{(t)}))$
return $\boldsymbol{y}^{(t)}, \tau^{(t)}$ for $t = 1, \dots, T$
```

Algorithm 8: Δ: Recovery

```
Input: Quantized measurements $\boldsymbol{y} \in \{\pm 1\}^m$; measurement matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$; sparsity parameter $s$; bound $R$ on $\|\boldsymbol{x}\|_2$; thresholds $\tau \in \mathbb{R}^m$; size of batches $q$.
Output: Approximation $\hat{\boldsymbol{x}} \in \mathbb{R}^n$
$T \leftarrow \left\lfloor \frac{m}{q} \right\rfloor$
Partition $\boldsymbol{A}$ into $T$ blocks $\boldsymbol{A}^{(1)}, \dots, \boldsymbol{A}^{(T)} \in \mathbb{R}^{q\times m}$
$\boldsymbol{x}_0 \leftarrow \boldsymbol{0}$
for $t = 1, \dots, T$ do
    $$\boldsymbol{x}_t \leftarrow H_s(\boldsymbol{x}_{t-1} + \Delta_0(\boldsymbol{y}^{(t)}, \boldsymbol{A}^{(t)}, R2^{-t+1}, \tau^{(t)})) \qquad (11)$$
return $\boldsymbol{x}_T$
```

We saw two examples of order-one recovery schemes in Section 2. The scheme based on second-order cone programming is an order-one recovery scheme with sparsity parameter s, measurement complexity $q = C_0 s\log(n/s)$, and noise resilience $\eta = c_0 R$ and $b = 0$. The scheme based on iterated hard thresholding is an order-one recovery scheme with sparsity parameter s, measurement complexity $q = C_1 s\log(n/s)$, and noise resilience $\eta = c_1 R$ and $b = c_2 q$. Above, $c_0, c_1, c_2, C_0, C_1 > 0$ are absolute constants.

We use an order-one recovery scheme to build a pair of one-bit quantization and recovery algorithms for sparse vectors that exhibits extremely fast convergence. Our quantization and recovery algorithms Q and Δ are given in Algorithms 7 and 8, respectively. They are in reality intertwined, but again we separate them for expositional clarity.

The intuition motivating Step (11) is that $\Delta_0(\boldsymbol{y}^{(t)}, \boldsymbol{A}^{(t)}, R_t, \tau^{(t)},)$ estimates $\boldsymbol{x} - \boldsymbol{x}_{t-1}$; hence $\boldsymbol{x}_t$ approximates $\boldsymbol{x}$ better than $\boldsymbol{x}_{t-1}$ does. Note the similarity to the intuition motivating iterative hard thresholding, with the key difference being that the quantization is also performed iteratively.

Remark 2 (Computational and storage considerations)**.** *Let us analyze the storage requirements* *and computational complexity of Q and Δ, both during and after quantization.*

We begin by considering the approach based on convex programming. In this case, the final storage requirements of the quantizer Q are similar to those in standard one-bit compressed sensing. The "algorithm" T_0 is straightforward: it simply draws random thresholds/dithers. In particular, we may treat these thresholds as predetermined independent normal random variables in the same way as we treat $\boldsymbol{A}$. If $\boldsymbol{A}$ and τ are generated by a short seed, then all that needs to be stored after quantization are the binary measurements $\boldsymbol{y} \in \{\pm 1\}^q$. During quantization, the algorithm Q needs to store $\boldsymbol{x}_t$. However, this requires small memory since $\boldsymbol{x}_t$ is s-sparse.

While the convex programming approach is designed to ease storage burdens, the order-one recovery scheme based on hard thresholding is built for speed. In this case, the threshold algorithm T_0 (Algorithm 5) is more complicated, and the adaptive thresholds τ need to be stored. On the other hand, the computation of $\boldsymbol{x}_t$ is much faster, and both the quantization and recovery algorithms are very efficient.

Given an order-one recovery scheme (T_0, Δ_0), the quantizer Q given in Algorithm 7 and the recovery algorithm Δ given in Algorithm 8 have the desired exponential convergence rate. This is formally stated in the theorem below and proved in Section 4.

Theorem 6. *Let (T_0, Δ_0) be an order-one recovery scheme with sparsity parameter $2s$, measurement complexity q, and noise resilience (η, b). Fix $R > 0$ and recall that $T := \lfloor m/q \rfloor$. With probability at least $1 - CT\exp(-cq)$ over the choice of $\boldsymbol{A}$ and the randomness of T_0, the following holds for all $\boldsymbol{x} \in RB_2^n \cap \Sigma_s$, all $\boldsymbol{e} \in \mathbb{R}^m$ with $\|\boldsymbol{e}\|_\infty \le \eta 2^{-T}\|\boldsymbol{x}\|_2$, and all $\mathbf{f} \in \{\pm 1\}^m$ with $|\{i \ : \ f_i = -1\}| \le b$ in the measurement model* (6)*:*

for $\boldsymbol{y} \in \{\pm 1\}^m$ and $\tau = Q(\boldsymbol{Ax}, \boldsymbol{A}, s, R, q) \in \mathbb{R}^m$, the output $\hat{\boldsymbol{x}}$ of $\Delta(\boldsymbol{y}, \boldsymbol{A}, s, R, \tau, q)$ satisfies

$$\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2 \le R\, 2^{-T}. \tag{12}$$

The positive constants η, b, c, and C above are absolute constants.

Our two order-one recovery schemes each have measurement complexity $q = Cs\log(n/s)$. This implies the announced exponential decay in the error rate.

Corollary 7. *Let Q, Δ be as in Algorithms 7 and 8 with one-bit recovery schemes (T_0, Δ_0) given either by Algorithms (3,4) or (5,6). Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ have independent standard normal entries. Fix $R > 0$ and recall that $\lambda = m/(slog(n/s))$. With probability at least $1 - C\lambda\exp(-cs\log(n/s))$ over the choice of $\boldsymbol{A}$ and the randomness of T_0, the following holds for all $\boldsymbol{x} \in RB_2^n \cap \Sigma_s$, all $\boldsymbol{e} \in \mathbb{R}^m$ with $\|\boldsymbol{e}\|_\infty \le \eta 2^{-T}\|\boldsymbol{x}\|_2$, and all $\mathbf{f} \in \{\pm 1\}^m$ with $|\{i \ : \ f_i = -1\}| \le b$ in the measurement model* (6) *($b = 0$ if (T_0, Δ_0) is based on convex programming or $b = cs\log(n/s)$ if (T_0, Δ_0) is based on hard thresholding):*

for $\boldsymbol{y} \in \{\pm 1\}^m$ $\tau = Q(\boldsymbol{Ax}, \boldsymbol{A}, s, R, q) \in \mathbb{R}^m$, the output $\hat{\boldsymbol{x}}$ of $\Delta(\boldsymbol{y}, \boldsymbol{A}, s, R, \tau, q)$ satisfies

$$\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2 \le R\, 2^{-c\lambda}. \tag{13}$$

The positive constants η, c', c, and C above are absolute constants.

4 Proofs

4.1 Exponentially decaying error rate from order-one recovery schemes

First, we prove Theorem 6 which states that, given an appropriate order-one recovery scheme, the recovery algorithm Δ in Algorithm 8 converges with exponentially small reconstruction error when the measurements are obtained by the quantizer Q of Algorithm 7.

Proof of Theorem 6. For $\boldsymbol{x} \in RB_2^n \cap \Sigma_s$, we verify by induction on $t \in \{0, 1, \ldots, T\}$ that

$$\|\boldsymbol{x} - \boldsymbol{x}_t\|_2 \leq R2^{-t}.$$

This induction hypothesis holds for $t = 0$. Now, suppose that it holds for $t - 1$, $t \in \{1, \ldots, T\}$. Consider $\Delta_0(\boldsymbol{y}^{(t)}, \boldsymbol{A}^{(t)}, R_t, \tau^{(t)})$, the estimate returned by the order-one recovery scheme in (11). By definition, the thresholds $\tau^{(t)}$ were obtained in step t by running T_0 on $A^{(t)}(\boldsymbol{x} - \boldsymbol{x}_{t-1})$. Similarly, the quantized measurements $\boldsymbol{y}^{(t)}$ are formed by quantizing (with noise) the affine measurements

$$\boldsymbol{A}^{(t)}\boldsymbol{x} - \sigma^{(t)} - \tau^{(t)} = \boldsymbol{A}^{(t)}(\boldsymbol{x} - \boldsymbol{x}_{t-1}) - \tau^{(t)}.$$

Thus, we have effectively run the order-one recovery scheme on the $2s$-sparse vector $\boldsymbol{x} - \boldsymbol{x}_t$. By the guarantee of the order-one recovery algorithm, with probability at least $1 - C\exp(-cq)$,

$$\left\|(\boldsymbol{x} - \boldsymbol{x}_{t-1}) - \Delta_0(\boldsymbol{y}^{(t)}, \boldsymbol{A}^{(t)}, R_t, \tau^{(t)})\right\|_2 \leq R_t/4 = R2^{-t+1}/4.$$

Suppose that this occurs. Let

$$\boldsymbol{z} = \boldsymbol{x}_{t-1} + \Delta_0(\boldsymbol{y}^{(t)}, \boldsymbol{A}^{(t)}, R_t, \tau^{(t)}),$$

so $\|\boldsymbol{x} - \boldsymbol{z}\|_2 \leq R2^{-t+1}/4$. Since $\boldsymbol{x}_t = H_s(\boldsymbol{z})$ is the best s-term approximation to $\boldsymbol{z}$, it follows that

$$\|\boldsymbol{x} - \boldsymbol{x}_t\|_2 = \|\boldsymbol{x} - H_s(\boldsymbol{z})\|_2 \leq \|\boldsymbol{x} - \boldsymbol{z}\|_2 + \|H_s(\boldsymbol{z}) - \boldsymbol{z}\|_2 \leq 2\|\boldsymbol{x} - \boldsymbol{z}\|_2 \leq R2^{-t}.$$

Thus, the induction hypothesis holds for t. A union bound over the T iterations completes the proof, since the announced result is the inductive hypothesis in the case that $t = T$. □

4.2 Hard-thresholding-based order-one recovery scheme

The proof of Theorem 3 relies on three properties of random matrices $\boldsymbol{A} \in \mathbb{R}^{q \times n}$ with independent standard normal entries. In their descriptions below, the positive constants c, C, and d are absolute constants.

- The *restricted isometry property* of order s ([FR13, Theorems 9.6 and 9.27]): for any $\delta > 0$, with failure probability at most $2\exp(-c\delta^2 q)$, the estimates

$$(1 - \delta)\|\boldsymbol{x}\|_2^2 \leq \frac{1}{q}\|\mathbf{A}\boldsymbol{x}\|_2^2 \leq (1 + \delta)\|\boldsymbol{x}\|_2^2 \tag{14}$$

  hold for all s-sparse $\boldsymbol{x} \in \mathbb{R}^n$ provided $q \geq C\delta^{-2}s\log(n/s)$.

- The *sign product embedding property* of order s ([JDDV13, PV13b]): for any $\delta > 0$, with failure probability at most $8\exp(-c\delta^2 q)$, the estimates
$$\left| \frac{\sqrt{\pi/2}}{q} \left\langle \mathbf{A}\boldsymbol{w}, \operatorname{sign}\left(\mathbf{A}\boldsymbol{x}\right)\right\rangle - \left\langle \boldsymbol{w}, \boldsymbol{x}\right\rangle \right| \leq \delta \tag{15}$$
hold for all effectively s-sparse $\boldsymbol{w}, \boldsymbol{x} \in \mathbb{R}^n$ with $\|\boldsymbol{w}\|_2 = \|\boldsymbol{x}\|_2 = 1$ provided $q \geq C\delta^{-6} s\log(n/s)$.
- The ℓ_1*-quotient property* ([Woj09] or [FR13, Theorem 11.19]): if $n \geq 2q$, then with failure probability at most $\exp(-cq)$, every $\boldsymbol{e} \in \mathbb{R}^q$ can be written as
$$\boldsymbol{e} = \mathbf{A}\boldsymbol{u} \qquad \text{with} \quad \|\boldsymbol{u}\|_1 \leq d\sqrt{s_*}\,\|\boldsymbol{e}\|_2/\sqrt{q} \quad \text{where } s_* := \frac{q}{\log(n/q)}. \tag{16}$$

Combining the ℓ_1-quotient property and the restricted isometry property (of order $2s$ for a fixed $\delta \in (0, 1/2)$, say) yields the *simultaneous* (ℓ_2, ℓ_1)*-quotient property* (use, for instance, [FR13, Theorem 6.13 and Lemma 11.16]); that is, there are absolute constants $d, d' > 0$ such that every $\boldsymbol{e} \in \mathbb{R}^q$ can be written as
$$\boldsymbol{e} = \mathbf{A}\boldsymbol{u} \qquad \text{with} \quad \begin{cases} \|\boldsymbol{u}\|_2 & \leq & d\,\|\boldsymbol{e}\|_2/\sqrt{q}, \\ \|\boldsymbol{u}\|_1 & \leq & d'\sqrt{s_*}\,\|\boldsymbol{e}\|_2/\sqrt{q}. \end{cases} \tag{17}$$

Proof of Theorem 3. We target the inequalities
$$\left\| \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2} - \frac{\sqrt{\pi/2}}{q} H_s\left(\mathbf{A}^*\boldsymbol{y}\right)\right\|_2 \leq \delta + c_4 \frac{\|\boldsymbol{e}\|_2}{\sqrt{q}\,\|\boldsymbol{x}\|_2} + c_5\sqrt{\frac{d_H(\boldsymbol{y}, \operatorname{sign}\left(\mathbf{A}\boldsymbol{x} + \boldsymbol{e}\right))}{q}}. \tag{18}$$

The desired inequalities (10) then follows modulo a change of constants, because $H_s'\left(\mathbf{A}^*\boldsymbol{y}\right)$ is the best unit-norm approximation to $\sqrt{\pi/2}\,q^{-1} H_s\left(\mathbf{A}^*\boldsymbol{y}\right)$, so that
$$\begin{aligned} \left\| \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2} - H_s'\left(\mathbf{A}^*\boldsymbol{y}\right)\right\|_2 &\leq \left\| \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2} - \frac{\sqrt{\pi/2}}{q} H_s\left(\mathbf{A}^*\boldsymbol{y}\right)\right\|_2 + \left\| H_s'\left(\mathbf{A}^*\boldsymbol{y}\right) - \frac{\sqrt{\pi/2}}{q} H_s\left(\mathbf{A}^*\boldsymbol{y}\right)\right\|_2 \\ &\leq 2\left\| \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2} - \frac{\sqrt{\pi/2}}{q} H_s\left(\mathbf{A}^*\boldsymbol{y}\right)\right\|_2. \end{aligned}$$

With $s_* = q/\log(n/q)$ as in (16), we remark that it is enough to consider the case $s = cs_*$, $c := c_1^{-1}\delta^7$. Indeed, the inequality $q \geq c_1\delta^{-7} s\log(n/s)$ yields $q \geq c^{-1} s\log(n/q)$, i.e., $s \leq cs_*$. Then (18) for s follows from (18) for cs_* modulo a change of constants because $H_s(\mathbf{A}^*\boldsymbol{y})$ is the best s-term approximation to $H_{cs_*}(\mathbf{A}^*\boldsymbol{y})$, so that
$$\begin{aligned} &\left\| \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2} - \frac{\sqrt{\pi/2}}{q} H_s\left(\mathbf{A}^*\boldsymbol{y}\right)\right\|_2 \\ &\quad \leq \left\| \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2} - \frac{\sqrt{\pi/2}}{q} H_{cs_*}\left(\mathbf{A}^*\boldsymbol{y}\right)\right\|_2 + \left\| \frac{\sqrt{\pi/2}}{q} H_s\left(\mathbf{A}^*\boldsymbol{y}\right) - \frac{\sqrt{\pi/2}}{q} H_{cs_*}\left(\mathbf{A}^*\boldsymbol{y}\right)\right\|_2 \\ &\quad \leq 2\left\| \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2} - \frac{\sqrt{\pi/2}}{q} H_{cs_*}\left(\mathbf{A}^*\boldsymbol{y}\right)\right\|_2. \end{aligned}$$

We now assume that $s = cs_*$. This reads $q = c_1\delta^{-7} s \log(n/q)$ and arguments similar to [FR13, Lemma C.6(c)] lead to $q \geq (c_1\delta^{-7}/\log(ec_1\delta^{-7}))s\log(n/s)$. Thus, if c_1 is chosen large enough at the start, we have $q \geq C\delta^{-6} s \log(n/s)$. This ensures that the sign product embedding property (15) of order $2s$ with constant $\delta/2$ holds with high probability. Likewise, the restricted isometry property (14) of order $2s$ with constant $9/16$, say, holds with high probability. In turn, the simultaneous (ℓ_2, ℓ_1)-quotient property (17) holds with high probability.

We place ourselves in the situation where all three properties hold simultaneously, which occurs with failure probability at most $c_2 \exp(-c_3\delta^2 q)$ for some absolute constants $c_2, c_3 > 0$. Then, writing $S = \mathrm{supp}\,(\boldsymbol{x})$ and $T = \mathrm{supp}\,(H_s\,(\mathbf{A}^*\boldsymbol{y}))$, we remark that $H_s\,(\mathbf{A}^*\boldsymbol{y})$ is the best s-term approximation to $\mathbf{A}^*_{S\cup T}\boldsymbol{y}$, so that

$$\begin{aligned}\left\|\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2} - \frac{\sqrt{\pi/2}}{q} H_s\,(\mathbf{A}^*\boldsymbol{y})\right\|_2 &\leq \left\|\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2} - \frac{\sqrt{\pi/2}}{q}\mathbf{A}^*_{S\cup T}\boldsymbol{y}\right\|_2 + \left\|\frac{\sqrt{\pi/2}}{q} H_s\,(\mathbf{A}^*\boldsymbol{y}) - \frac{\sqrt{\pi/2}}{q}\mathbf{A}^*_{S\cup T}\boldsymbol{y}\right\|_2 \\ &\leq 2\left\|\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2} - \frac{\sqrt{\pi/2}}{q}\mathbf{A}^*_{S\cup T}\boldsymbol{y}\right\|_2. \end{aligned} \tag{19}$$

We continue with the fact that

$$\begin{aligned}&\left\|\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2} - \frac{\sqrt{\pi/2}}{q}\mathbf{A}^*_{S\cup T}\boldsymbol{y}\right\|_2 \\ &\qquad\leq \left\|\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2} - \frac{\sqrt{\pi/2}}{q}\mathbf{A}^*_{S\cup T}\mathrm{sign}\,(\mathbf{A}\boldsymbol{x}+\boldsymbol{e})\right\|_2 + \frac{\sqrt{\pi/2}}{q}\left\|\mathbf{A}^*_{S\cup T}\,(\boldsymbol{y} - \mathrm{sign}\,(\mathbf{A}\boldsymbol{x}+\boldsymbol{e}))\right\|_2. \end{aligned} \tag{20}$$

The second term on the right-hand side of (20) can be bounded with the help of the restricted isometry property (14) as

$$\begin{aligned}\|\mathbf{A}^*_{S\cup T}\,(\boldsymbol{y} - \mathrm{sign}\,(\mathbf{A}\boldsymbol{x}+\boldsymbol{e}))\|_2^2 &= \langle \mathbf{A}_{S\cup T}\mathbf{A}^*_{S\cup T}\,(\boldsymbol{y} - \mathrm{sign}\,(\mathbf{A}\boldsymbol{x}+\boldsymbol{e}))\,, \boldsymbol{y} - \mathrm{sign}\,(\mathbf{A}\boldsymbol{x}+\boldsymbol{e})\rangle \\ &\leq \|\mathbf{A}_{S\cup T}\mathbf{A}^*_{S\cup T}\,(\boldsymbol{y} - \mathrm{sign}\,(\mathbf{A}\boldsymbol{x}+\boldsymbol{e}))\|_2\,\|\boldsymbol{y} - \mathrm{sign}\,(\mathbf{A}\boldsymbol{x}+\boldsymbol{e})\|_2 \\ &\leq \sqrt{1+\frac{9}{16}}\sqrt{q}\,\|\mathbf{A}^*_{S\cup T}\,(\boldsymbol{y} - \mathrm{sign}\,(\mathbf{A}\boldsymbol{x}+\boldsymbol{e}))\|_2\,\|\boldsymbol{y} - \mathrm{sign}\,(\mathbf{A}\boldsymbol{x}+\boldsymbol{e})\|_2\,. \end{aligned}$$

Simplifying by $\|\mathbf{A}^*_{S\cup T}\,(\boldsymbol{y} - \mathrm{sign}\,(\mathbf{A}\boldsymbol{x}+\boldsymbol{e}))\|_2$, we obtain

$$\|\mathbf{A}^*_{S\cup T}\,(\boldsymbol{y} - \mathrm{sign}\,(\mathbf{A}\boldsymbol{x}+\boldsymbol{e}))\|_2 \leq \frac{5}{4}\sqrt{q}\,\|\boldsymbol{y} - \mathrm{sign}\,(\mathbf{A}\boldsymbol{x}+\boldsymbol{e})\|_2 = \frac{5}{2}\sqrt{q}\sqrt{d_H\,(\boldsymbol{y}, \mathrm{sign}(\mathbf{A}\boldsymbol{x}+\boldsymbol{e}))}. \tag{21}$$

The first term on the right-hand side of (20) can be bounded with the help of the simultaneous (ℓ_2, ℓ_1)-quotient property (17) and of the sign product embedding property (15). We start by writing $\mathbf{A}\boldsymbol{x}+\boldsymbol{e}$ as $\mathbf{A}\,(\boldsymbol{x}+\boldsymbol{u})$ for some $\boldsymbol{u} \in \mathbb{R}^n$ as in (17). We then notice that

$$\begin{aligned}&\|\boldsymbol{x}+\boldsymbol{u}\|_2 \geq \|\boldsymbol{x}\|_2 - \|\boldsymbol{u}\|_2 \geq \|\boldsymbol{x}\|_2 - d\,\|\boldsymbol{e}\|_2/\sqrt{q} \geq (1-dc_6)\,\|\boldsymbol{x}\|_2\,, \\ &\|\boldsymbol{x}+\boldsymbol{u}\|_1 \leq \|\boldsymbol{x}\|_1 + \|\boldsymbol{u}\|_1 \leq \sqrt{s}\,\|\boldsymbol{x}\|_2 + d'\sqrt{s_*}\,\|\boldsymbol{e}\|_2/\sqrt{q} \leq \left(\frac{1}{\sqrt{2}} + \frac{d'c_6}{\sqrt{2c}}\right)\sqrt{2s}\,\|\boldsymbol{x}\|_2\,. \end{aligned}$$

Hence, if c_6 is chosen small enough at the start, then we have $\|\boldsymbol{x}+\boldsymbol{u}\|_1 \leq \sqrt{2s}\,\|\boldsymbol{x}+\boldsymbol{u}\|_2$, i.e., $\boldsymbol{x}+\boldsymbol{u}$ is effectively $(2s)$-sparse. The sign product embedding property (15) of order $2s$ then implies that

$$\left|\left\langle \boldsymbol{w}, \frac{\boldsymbol{x}+\boldsymbol{u}}{\|\boldsymbol{x}+\boldsymbol{u}\|_2} - \frac{\sqrt{\pi/2}}{q}\mathbf{A}^*_{S\cup T}\mathrm{sign}\left(\mathbf{A}\boldsymbol{x}+\boldsymbol{e}\right)\right\rangle\right|$$
$$= \left|\left\langle \boldsymbol{w}, \frac{\boldsymbol{x}+\boldsymbol{u}}{\|\boldsymbol{x}+\boldsymbol{u}\|_2}\right\rangle - \frac{\sqrt{\pi/2}}{q}\left\langle \mathbf{A}\boldsymbol{w}, \mathrm{sign}\left(\mathbf{A}\left(\boldsymbol{x}+\boldsymbol{u}\right)\right)\right\rangle\right| \leq \frac{\delta}{2}$$

for all unit-normed $\boldsymbol{w} \in \mathbb{R}^n$ supported on $S \cup T$. This gives

$$\left\|\frac{\boldsymbol{x}+\boldsymbol{u}}{\|\boldsymbol{x}+\boldsymbol{u}\|_2} - \frac{\sqrt{\pi/2}}{q}\mathbf{A}^*_{S\cup T}\mathrm{sign}\left(\mathbf{A}\boldsymbol{x}+\boldsymbol{e}\right)\right\|_2 \leq \frac{\delta}{2},$$

and in turn

$$\begin{aligned}\left\|\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2} - \frac{\sqrt{\pi/2}}{q}\mathbf{A}^*_{S\cup T}\mathrm{sign}\left(\mathbf{A}\boldsymbol{x}+\boldsymbol{e}\right)\right\|_2 &\leq \frac{\delta}{2} + \left\|\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2} - \frac{\boldsymbol{x}+\boldsymbol{u}}{\|\boldsymbol{x}+\boldsymbol{u}\|_2}\right\|_2 \\ &\leq \frac{\delta}{2} + \left\|\left(\frac{1}{\|\boldsymbol{x}\|_2} - \frac{1}{\|\boldsymbol{x}+\boldsymbol{u}\|_2}\right)\boldsymbol{x}\right\|_2 + \left\|\frac{\boldsymbol{u}}{\|\boldsymbol{x}+\boldsymbol{u}\|_2}\right\|_2 \\ &\leq \frac{\delta}{2} + \frac{|\,\|\boldsymbol{x}+\boldsymbol{u}\|_2 - \|\boldsymbol{x}\|_2\,|}{\|\boldsymbol{x}+\boldsymbol{u}\|_2} + \frac{\|\boldsymbol{u}\|_2}{\|\boldsymbol{x}+\boldsymbol{u}\|_2} \leq \frac{\delta}{2} + \frac{2\,\|\boldsymbol{u}\|_2}{\|\boldsymbol{x}+\boldsymbol{u}\|_2}.\end{aligned}$$

From $\|\boldsymbol{u}\|_2 \leq d\,\|\boldsymbol{e}\|_2/\sqrt{q}$ and $\|\boldsymbol{x}+\boldsymbol{u}\|_2 \geq (1-dc_6)\,\|\boldsymbol{x}\|_2 \geq \|\boldsymbol{x}\|_2/2$ for c_6 is small enough, we derive that

$$\left\|\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2} - \frac{\sqrt{\pi/2}}{q}\mathbf{A}^*_{S\cup T}\left(\mathrm{sign}\left(\mathbf{A}\boldsymbol{x}+\boldsymbol{e}\right)\right)\right\|_2 \leq \frac{\delta}{2} + \frac{4d\,\|\boldsymbol{e}\|_2}{\sqrt{q}\,\|\boldsymbol{x}\|_2}. \tag{22}$$

Substituting (21) and (22) into (20) enables us to derive the desired result (18) from (19). □

The proof of Theorem 4 presented next follows from Theorem 3.

Proof of Theorem 4. For later purposes, we introduce the constant

$$C := \max_{\xi\in\left[\frac{1}{\sqrt{2}}-\frac{1}{20},\frac{2}{\sqrt{5}}+\frac{1}{20}\right]} \left|f'(\xi)\right| \geq 2, \qquad f(\xi) := 1 - \frac{\sqrt{1-\xi^2}}{\xi}.$$

Given $\boldsymbol{x} \in RB_2^n \cap \Sigma_s$, we acquire a corrupted version $\boldsymbol{y}_1 \in \{\pm 1\}^{q/2}$ of the quantized measurements $\mathrm{sign}(\mathbf{A}_1\boldsymbol{x})$. Since the number of rows of the matrix $\mathbf{A}_1 \in \mathbb{R}^{(q/2)\times n}$ is large enough for Theorem 3 to hold with $\delta_0 = \delta/(4(1+2C))$ instead of δ, we obtain

$$\left\|\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2} - \boldsymbol{u}\right\|_2 \leq \delta_0 + c_4 c\delta + c_5 c'\delta \leq 2\delta_0, \qquad \boldsymbol{u} := H_s'(\mathbf{A}_1^*\boldsymbol{y}_1),$$

provided that the constants c and c' are small enough. With $\boldsymbol{x}^\sharp$ denoting the orthogonal projection of $\boldsymbol{x}$ onto the line spanned by $\boldsymbol{u}$, we have

$$\left\|\boldsymbol{x} - \boldsymbol{x}^\sharp\right\|_2 \leq \|\boldsymbol{x} - \|\boldsymbol{x}\|_2\,\boldsymbol{u}\|_2 \leq 2\delta_0\,\|\boldsymbol{x}\|_2\,.$$

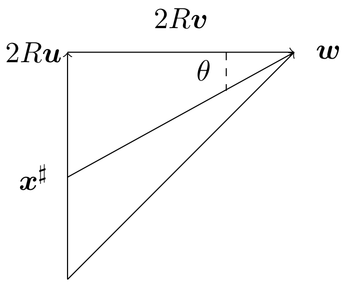

Figure 3: The situation in the plane spanned by $\boldsymbol{u}$ and $\boldsymbol{v}$.

We now consider a unit-norm vector $\boldsymbol{v} \in \mathbb{R}^n$ supported on $\mathrm{supp}(\boldsymbol{u})$ and orthogonal to $\boldsymbol{u}$. The situation in the plane spanned by $\boldsymbol{u}$ and $\boldsymbol{v}$ is summarized in Figure 3.

We point out that $\|\boldsymbol{x}^\sharp\| \le \|\boldsymbol{x}\| \le R$ gave $\|\boldsymbol{x}^\sharp\|_2 \le 2R$, but that $2R$ was just an arbitrary choice to ensure that $\cos(\theta)$ stays away from 1—here, $\cos(\theta) \in [1/\sqrt{2}, 2/\sqrt{5}]$. Forming the s-sparse vector $\boldsymbol{w} := 2R \cdot (\boldsymbol{u} + \boldsymbol{v})$, we now acquire a corrupted version $\boldsymbol{y}_2 \in \{\pm 1\}^{q/2}$ of the quantized measurements $\mathrm{sign}(\mathbf{A}_2(\boldsymbol{x} - \boldsymbol{w}))$ on the $2s$-sparse vector $\boldsymbol{x} - \boldsymbol{w}$. Since the number of rows of the matrix $\mathbf{A}_2 \in \mathbb{R}^{(q/2)\times n}$ is large enough for Theorem 3 to hold with $\delta_0 = \delta/(4(1+2C))$ instead of δ and $2s$ instead of s, we obtain

$$\left\| \frac{\boldsymbol{w} - \boldsymbol{x}}{\|\boldsymbol{w} - \boldsymbol{x}\|_2} - \boldsymbol{t} \right\|_2 \le \delta_0 + c_4 c \delta + c_5 c' \delta \le 2\delta_0, \qquad \boldsymbol{t} = -H_s'(\mathbf{A}_2^* \boldsymbol{y}_2).$$

We deduce that $\boldsymbol{t}$ also approximates $(\boldsymbol{w} - \boldsymbol{x}^\sharp)/\left\|\boldsymbol{w} - \boldsymbol{x}^\sharp\right\|_2$ with error

$$\begin{aligned}
&\left\| \frac{\boldsymbol{w} - \boldsymbol{x}^\sharp}{\|\boldsymbol{w} - \boldsymbol{x}^\sharp\|} - \boldsymbol{t} \right\|_2 \\
&\le \left\| \frac{\boldsymbol{w} - \boldsymbol{x}^\sharp}{\|\boldsymbol{w} - \boldsymbol{x}^\sharp\|_2} - \frac{\boldsymbol{w} - \boldsymbol{x}}{\|\boldsymbol{w} - \boldsymbol{x}^\sharp\|_2} \right\|_2 + \left\| \left(\frac{1}{\|\boldsymbol{w} - \boldsymbol{x}^\sharp\|_2} - \frac{1}{\|\boldsymbol{w} - \boldsymbol{x}\|_2} \right) (\boldsymbol{w} - \boldsymbol{x}) \right\|_2 + \left\| \frac{\boldsymbol{w} - \boldsymbol{x}}{\|\boldsymbol{w} - \boldsymbol{x}\|_2} - \boldsymbol{t} \right\|_2 \\
&\le \frac{\left\|\boldsymbol{x} - \boldsymbol{x}^\sharp\right\|_2}{\|\boldsymbol{w} - \boldsymbol{x}^\sharp\|_2} + \frac{\left| \|\boldsymbol{w} - \boldsymbol{x}\|_2 - \|\boldsymbol{w} - \boldsymbol{x}^\sharp\|_2 \right|}{\|\boldsymbol{w} - \boldsymbol{x}^\sharp\|_2} + 2\delta_0 \le 2 \frac{\left\|\boldsymbol{x} - \boldsymbol{x}^\sharp\right\|_2}{\|\boldsymbol{w} - \boldsymbol{x}^\sharp\|_2} + 2\delta_0 \le 2 \frac{2\delta_0 \|\boldsymbol{x}\|_2}{2R} + 2\delta_0 \\
&\le 4\delta_0.
\end{aligned}$$

It follows that $\langle \boldsymbol{t}, \boldsymbol{v} \rangle$ approximates $\left\langle (\boldsymbol{w} - \boldsymbol{x}^\sharp)/\|\boldsymbol{w} - \boldsymbol{x}^\sharp\|, \boldsymbol{v} \right\rangle = \cos(\theta)$ with error

$$|\cos(\theta) - \langle \boldsymbol{t}, \boldsymbol{v} \rangle| = \left| \left\langle \frac{\boldsymbol{w} - \boldsymbol{x}^\sharp}{\|\boldsymbol{w} - \boldsymbol{x}^\sharp\|_2} - \boldsymbol{t}, \boldsymbol{v} \right\rangle \right| \le \left\| \frac{\boldsymbol{w} - \boldsymbol{x}^\sharp}{\|\boldsymbol{w} - \boldsymbol{x}^\sharp\|_2} - \boldsymbol{t} \right\|_2 \|\boldsymbol{v}\|_2 \le 4\delta_0.$$

We then notice that

$$\left\|\boldsymbol{x}^\sharp\right\|_2 = 2R - 2R\tan(\theta) = 2Rf(\cos(\theta)),$$

so that $2Rf(\langle \boldsymbol{t}, \boldsymbol{v} \rangle)$ approximates $\left\|\boldsymbol{x}^\sharp\right\|_2$ with error

$$\left| \left\|\boldsymbol{x}^\sharp\right\|_2 - 2Rf(\langle \boldsymbol{t}, \boldsymbol{v} \rangle) \right| = 2R|f(\cos(\theta)) - f(\langle \boldsymbol{t}, \boldsymbol{v} \rangle)| \le 2RC\,|\cos(\theta) - \langle \boldsymbol{t}, \boldsymbol{v} \rangle\,| \le 2RC\,4\delta_0 = 8C\delta_0 R.$$

Here, we used the facts that $\cos(\theta) \in [1/\sqrt{2}, 2/\sqrt{5}]$ and that $\langle \boldsymbol{t}, \boldsymbol{v} \rangle \in [1/\sqrt{2} - 4\delta_0, 2/\sqrt{5} + 4\delta_0] \subseteq [1/\sqrt{2} - 1/20, 2/\sqrt{5} + 1/20]$. We derive that

$$\begin{aligned}\left| \left\| \boldsymbol{x} \right\|_2 - 2Rf(\langle \boldsymbol{t}, \boldsymbol{v} \rangle) \right| &\leq \left| \left\| \boldsymbol{x} \right\|_2 - \left\| \boldsymbol{x}^\sharp \right\|_2 \right| + \left| \left\| \boldsymbol{x}^\sharp \right\|_2 - 2Rf(\langle \boldsymbol{t}, \boldsymbol{v} \rangle) \right| \\ &\leq \left\| \boldsymbol{x} - \boldsymbol{x}^\sharp \right\|_2 + \left| \left\| \boldsymbol{x}^\sharp \right\|_2 - 2Rf(\langle \boldsymbol{t}, \boldsymbol{v} \rangle) \right| \\ &\leq 2\delta_0 \left\| \boldsymbol{x} \right\|_2 + 8C\,\delta_0 R \leq 2(1+4C)\delta_0 R.\end{aligned}$$

Finally, with the estimate $\hat{\boldsymbol{x}}$ for $\boldsymbol{x}$ being defined as

$$\hat{\boldsymbol{x}} := 2Rf(\langle \boldsymbol{t}, \boldsymbol{v} \rangle)\,\boldsymbol{u},$$

the previous considerations lead to the error estimate

$$\begin{aligned}\left\| \boldsymbol{x} - \hat{\boldsymbol{x}} \right\|_2 &\leq \left\| \boldsymbol{x} - \left\| \boldsymbol{x} \right\|_2 \boldsymbol{u} \right\|_2 + \left| \left\| \boldsymbol{x} \right\|_2 - 2Rf(\langle \boldsymbol{t}, \boldsymbol{v} \rangle) \right| \left\| \boldsymbol{u} \right\|_2 \leq 2\delta_0 \left\| \boldsymbol{x} \right\|_2 + 2(1+4C)\delta_0 R \\ &\leq 4(1+2C)\delta_0 R.\end{aligned}$$

Our initial choice of $\delta_0 = \delta/(4(1+2C))$ enables us to conclude that $\left\| \boldsymbol{x} - \hat{\boldsymbol{x}} \right\|_2 \leq \delta R$. □

4.3 Second-order-cone-programming-based order-one recovery scheme

Proof of Theorem 2. Without loss of generality, we assume that $R = 1/2$. The general argument follows from a rescaling. We begin by considering the exact case in which $\boldsymbol{e} = \boldsymbol{0}$. Observe that, by the Cauchy–Schwarz inequality,

$$\left\| \boldsymbol{x} \right\|_1 \leq \sqrt{\left\| \boldsymbol{x} \right\|_0} \cdot \left\| \boldsymbol{x} \right\|_2 \leq \sqrt{s}.$$

Since $\boldsymbol{x}$ is feasible for program (9), we also have $\left\| \hat{\boldsymbol{x}} \right\|_1 \leq \sqrt{s}$. The result will follow from the following two observations:

- $\boldsymbol{x}, \hat{\boldsymbol{x}} \in \sqrt{s}B_1^n \cap B_2^n$
- $\operatorname{sign}(\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle - \tau_i) = \operatorname{sign}(\langle \boldsymbol{a}_i, \hat{\boldsymbol{x}} \rangle - \tau_i), \qquad i = 1, \ldots, q.$

Each equation $\langle \boldsymbol{a}_i, \boldsymbol{z} \rangle - \tau_i = 0$ defines a hyperplane perpendicular to $\boldsymbol{a}_i$ and translated proportionally to τ_i; further, $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$ are on the same side of the hyperplane. To visualize this, imagine $\sqrt{s}B_1^n \cap B_2^n$ as an oddly shaped apple that we are trying to dice. Each hyperplane randomly slices the apple, eventually cutting it into small sections. The vectors $\hat{\boldsymbol{x}}$ and $\boldsymbol{x}$ belong to the same section. Thus, we ask: *how many random slices are needed for all sections to have small diameter?* Similar questions have been addressed in a broad context in [PV14]. We give a self-contained proof that $O(s\log(n/s))$ slices suffice based on the following result [PV14, Theorem 3.1].

Theorem 8 (Random hyperplane tessellations of $\sqrt{s}B_1^n \cap S^{n-1}$)**.** *Let $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_q \in \mathbb{R}^n$ be independent standard normal vectors. If*

$$q \geq C\delta^{-4} s \log(n/s),$$

then, with probability at least $1 - 2\exp(-c\delta^4 q)$, all $\boldsymbol{x}, \boldsymbol{x}' \in \sqrt{s}B_1^n \cap S^{n-1}$ with

$$\operatorname{sign}\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle = \operatorname{sign}\langle \boldsymbol{a}_i, \boldsymbol{x}' \rangle, \qquad i = 1, \ldots, q,$$

satisfy

$$\|\boldsymbol{x} - \boldsymbol{x}'\|_2 \leq \frac{\delta}{8}.$$

The positive constants c and C are absolute constants.

We translate the above result into a tessellation of $\sqrt{s}B_1^n \cap B_2^n$ in the following corollary.

Corollary 9 (Random hyperplane tessellations of $\sqrt{s}B_1^n \cap B_2^n$)**.** *Let $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_q \in \mathbb{R}^n$ be independent standard normal vectors and let $\tau_1, \tau_2, \ldots, \tau_q$ be independent standard normal random variables. If*

$$q \geq C\delta^{-4} s \log(n/s),$$

then, with probability at least $1 - 2\exp(-c\delta^4 q)$, all $\boldsymbol{x}, \boldsymbol{x}' \in \sqrt{s}B_1^n \cap B_2^n$ with

$$\operatorname{sign}(\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle - \tau_i) = \operatorname{sign}(\langle \boldsymbol{a}_i, \boldsymbol{x}' \rangle - \tau_i), \qquad i = 1, \ldots, q,$$

satisfy

$$\|\boldsymbol{x} - \boldsymbol{x}'\|_2 \leq \frac{\delta}{4}.$$

The positive constants c and C are absolute constants.

Proof. For any $\boldsymbol{z} \in \sqrt{s}B_1^n \cap B_2^n$, we notice that $\operatorname{sign}(\langle \boldsymbol{a}_i, \boldsymbol{z} \rangle - \tau_i) = \operatorname{sign}(\langle [\boldsymbol{a}_i, -\tau_i], [\boldsymbol{z}, 1] \rangle)$, where the augmented vectors $[\boldsymbol{a}_i, -\tau_i] \in \mathbb{R}^{n+1}$ and $[\boldsymbol{z}, 1] \in \mathbb{R}^{n+1}$ are the concatenations of $\boldsymbol{a}_i$ with $-\tau_i$ and $\boldsymbol{z}$ with 1, respectively. Thus, we have moved to the ditherless setup by only increasing the dimension by one. Since

$$\|[\boldsymbol{z}, 1]\|_2 \geq 1 \quad \text{and} \quad \|[\boldsymbol{z}, 1]\|_1 = \|\boldsymbol{z}\|_1 + 1 \leq \sqrt{s} + 1 \leq \sqrt{4s},$$

we may apply Theorem 8 after projecting on S^n to derive

$$\left\| \frac{[\boldsymbol{x}, 1]}{\|[\boldsymbol{x}, 1]\|_2} - \frac{[\boldsymbol{x}', 1]}{\|[\boldsymbol{x}', 1]\|_2} \right\|_2 \leq \frac{\delta}{8}. \tag{23}$$

with probability at least $1 - 2\exp(c\delta^4 q)$. We now show that the inequality (23) implies that $\|\boldsymbol{x} - \boldsymbol{x}'\|_2 \leq \delta/4$.

First note that

$$\|\boldsymbol{x} - \boldsymbol{x}'\|_2 \leq \sqrt{2} \left\| \frac{\boldsymbol{x}}{\|[\boldsymbol{x}, 1]\|_2} - \frac{\boldsymbol{x}'}{\|[\boldsymbol{x}, 1]\|_2} \right\|_2$$

since $\|\boldsymbol{x}\|_2 \leq 1$. Subtract and add $\boldsymbol{x}' / \|[\boldsymbol{x}', 1]\|_2$ inside the norm and apply triangle inequality to obtain

$$\|\boldsymbol{x} - \boldsymbol{x}'\|_2 \leq \sqrt{2} \left(\left\| \frac{\boldsymbol{x}}{\|[\boldsymbol{x}, 1]\|_2} - \frac{\boldsymbol{x}'}{\|[\boldsymbol{x}', 1]\|_2} \right\|_2 + \|\boldsymbol{x}'\|_2 \cdot \left| \frac{1}{\|[\boldsymbol{x}, 1]\|_2} - \frac{1}{\|[\boldsymbol{x}', 1]\|_2} \right| \right).$$

Since $\|\boldsymbol{x}'\|_2 \leq 1$, we may remove $\|\boldsymbol{x}'\|_2$ from in front of the second term in parenthesis. Next, use the inequality $a + b \leq \sqrt{2} \cdot \sqrt{a^2 + b^2}$ on the two terms in parenthesis. This bounds the right-hand side by precisely

$$2 \left\| \frac{[\boldsymbol{x}, 1]}{\|[\boldsymbol{x}, 1]\|_2} - \frac{[\boldsymbol{x}', 1]}{\|[\boldsymbol{x}', 1]\|_2} \right\|_2,$$

which is bounded by $\delta/4$ according to (23). □

This corollary immediately completes the proof of Theorem 2 in the case $\boldsymbol{e} = \mathbf{0}$. We now turn to the general problem where $\|\boldsymbol{e}\|_\infty \leq c\delta^3$ and thus $\|\boldsymbol{e}\|_2 \leq c\delta^3\sqrt{q}$. We reduce to the exact problem using the simultaneous (ℓ_1, ℓ_2)-quotient property (17), which guarantees that the error can be represented by a signal with small ℓ_1-norm. In particular, (17) implies that, with probability at least $1 - \exp(-cq)$, there exists a vector $\boldsymbol{u}$ satisfying

$$\boldsymbol{e} = \mathbf{A}\boldsymbol{u} \qquad \text{with} \quad \begin{cases} \|\boldsymbol{u}\|_2 & \leq & \delta/4, \\ \|\boldsymbol{u}\|_1 & \leq & c_1\delta^3\sqrt{q/\log(n/q)} \end{cases} \tag{24}$$

where c_1 is an absolute constant which we may choose as small as we need. We may now replace $\boldsymbol{x}$ with $\tilde{\boldsymbol{x}} = \boldsymbol{x} + \boldsymbol{u}$ and proceed as in the proof in the noiseless case. Reconstruction of $\tilde{\boldsymbol{x}}$ to accuracy $\delta/4$ yields reconstruction of $\boldsymbol{x}$ to accuracy $\delta/2$, as desired. By replacing $\boldsymbol{x}$ with $\tilde{\boldsymbol{x}}$, we have (mildly) increased the bound on the ℓ_1-norm and the ℓ_2-norm. Fortunately, $\|\tilde{\boldsymbol{x}}\|_2 \leq \|\boldsymbol{x}\|_2 + \|\boldsymbol{u}\|_2 \leq 1$ and thus $\tilde{\boldsymbol{x}}$ remains feasible for the program (9). Further, $\tilde{\boldsymbol{x}}$ is approximately sparse in the sense that $\|\tilde{\boldsymbol{x}}\|_1 \leq \|\boldsymbol{x}\|_1 + \|\boldsymbol{u}\|_1 \leq \sqrt{s} + c_1\delta^3\sqrt{q/\log(n/q)} =: \sqrt{\tilde{s}}$. To conclude the proof, we must show that the requirement of Theorem 2, namely $q \geq C'\delta^{-4}s\log(n/s)$, implies that the required condition of Corollary 9, namely $q \geq C\delta^{-4}\tilde{s}\log(n/\tilde{s})$, is still satisfied. The result follows from massaging the equations, as sketched below.

If $s \geq c_1^2\delta^6 q/\log(n/q)$, then $\sqrt{\tilde{s}} \leq 2\sqrt{s}$ and the desired result follows quickly. Suppose then that $s < c_1^2\delta^6 q/\log(n/q)$ and thus $\tilde{s} \leq c_2\delta^6 q/\log(n/q)$. To conclude, note that

$$C\delta^{-4}\tilde{s}\log(n/\tilde{s}) \leq q \cdot C \cdot c_2\frac{\delta^2}{\log(n/q)} \cdot (\log(n/q) + \log(1/c_2) + 6\log(1/\delta) + \log(\log(n/q)) \leq q,$$

where the first inequality follows since $s\log(n/s)$ is increasing in s and thus $\tilde{s}$ may be replaced by its upper bound, $c_2\delta^6 q/\log(n/q)$. The last inequality follows by taking c_2 small enough. This concludes the proof. □

5 Numerical Results

This brief section provides several experimental validations of the theory developed above. The computations, performed in MATLAB, are reproducible and can be downloaded from the second author's webpage. The random measurements $\boldsymbol{a}_i$ were always generated as vectors with independent standard normal entries. As for the random sparse vectors $\boldsymbol{x}$, after a random choice of their supports, their nonzero entries also consisted of independent standard normal variables.

Our first experiment (results not displayed here) verified on a single sparse vector that both its direction and magnitude can be accurately estimated via order-one recovery schemes, while only its direction could be accurately estimated using convex programs [PV13a, PV13b], ℓ_1-regularized logistic regression, or binary iterative hard thresholding [JLBB13]. We also noted the reduction of the reconstruction error by several orders of magnitude from the same number m of quantized measurements when Algorithms 7-8 are used instead of the above methods. We remark in passing that this number m is significantly larger than the number of measurements in classical compressed sensing with real-valued measurements, as intuitively expected.

Our second experiment corroborates the exponential decay of the error rate. The results are summarized in Figure 4, whose logarithmic scale on the vertical axis confirms the behavior $\log(\|\boldsymbol{x} - \boldsymbol{x}^*\|_2/\|\boldsymbol{x}\|_2) \leq -c\lambda$ for the relative reconstruction error as a function of the oversampling factor

$\lambda = m/\log(n/s)$. The tests were conducted on four sparsity levels s at a fixed dimension n for an oversampling ratio λ varying through the increase of the number m of measurements. The number T of iterations in Algorithms 7 and 8 was fixed throughout the experiment based on hard thresholding and throughout the experiment based on second-order cone programming. The values of all these parameters are reported directly in Figure 4. We point out that we could carry out a more exhaustive experiment for the faster hard-thresholding-based version than for the slower second-order-cone-programming-based version, both in terms of problem scale and of number of tests.

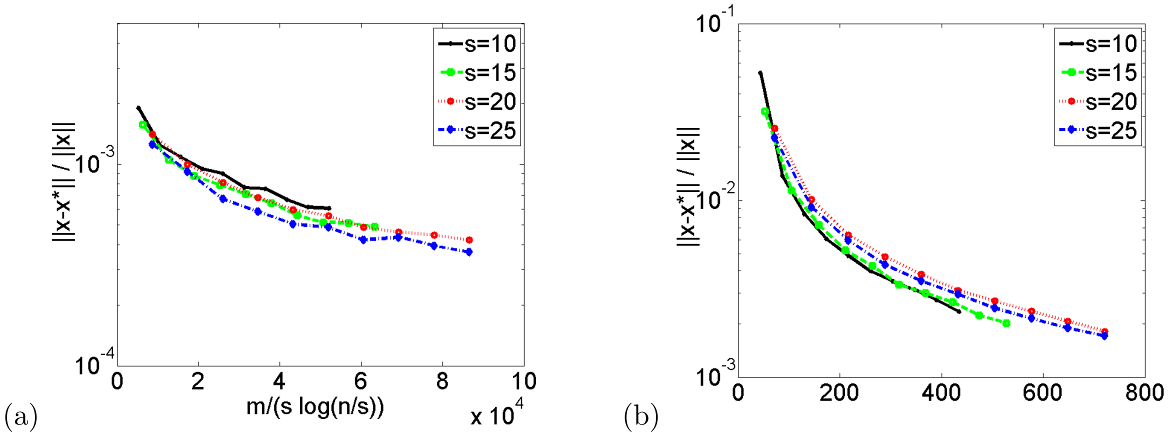

Figure 4: Averaged relative error for the reconstruction of sparse vectors ($n = 100$) by the outputs of Algorithms 7-8 based on (a) hard thresholding and (b) second-order cone programming as a function of the oversampling ratio.

Our third experiment examines the effect of measurement errors on the reconstruction via Algorithms 7 and 8. Once again, the problem scale was much larger when relying on hard thresholding than on second-order cone programming. The values of the size parameters are reported on Figure 5. This figure shows how the reconstruction error decreases as the iteration count t increases in Algorithms 7 and 8. For the hard-thresholding-based version, see Figure 5(a), we observe an error decreasing by a constant factor at each iteration when the measurements are totally accurate. Introducing a pre-quantization noise $\boldsymbol{e} \sim N(0, \sigma^2 \mathbf{I})$ in $\boldsymbol{y} = \mathrm{sign}(\mathbf{A}\boldsymbol{x} + \boldsymbol{e})$ does not affect this behavior too much until the "noise floor" is reached. Flipping a small fraction of the bits sign $\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle$ by multiplying them with $f_i = \pm 1$, most of which being equal to $+1$, seems to have an even smaller effect on the reconstruction. However, these bit flips prevent the use of the second-order-cone-programming-based version, as the constraints of the optimization problems become infeasible. But we still remark that the pre-quantization noise is not very damaging in this case either, see Figure 5(b), where the results of an experiment using ℓ_1-regularized logistic regression in Algorithms 7 and 8 are also displayed.

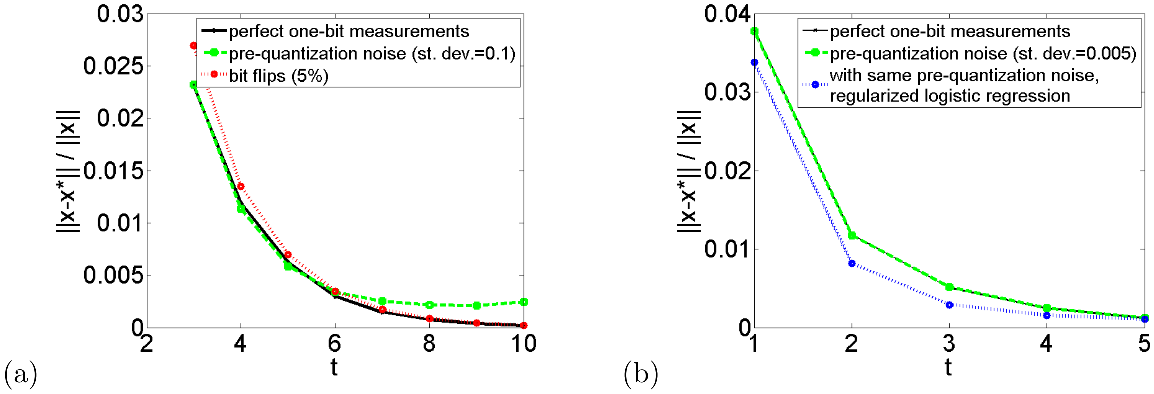

Figure 5: Averaged relative error for the reconstruction of sparse vectors ($n = 100$) by the outputs of Algorithms 7-8 based on (a) hard thresholding ($s = 15$, $m = 10^5$) and second-order cone programming and (b) ℓ_1-regularized logistic regression ($s = 10$, $m = 2 \cdot 10^4$) as a function of the iteration count when measurement error is present.

6 Discussion

6.1 Related work

The one-bit compressed sensing framework developed by Boufounos and Baraniuk [BB08] is a relatively new line of work, with theoretical backing only recently being developed. Empirical evidence and convergence analysis of algorithms for quantized measurements appear in the works of Boufounos et al. and others [Bou09, BB08, LWYB11, ZBC10]. Theoretical bounds on recovery error have only recently been studied, outside from results which model the one-bit setting as classical compressed sensing with specialized additive measurement error [DPM09, JHF11, SG09]. Other settings analyze quantized measurements where the number of bits used depends on signal parameters like sparsity level or the dynamic range [ACS09, GLP$^+$10, GLP$^+$13]. Boufounos develops hierarchical and scalar quantization with modified quantization regions which aim to balance the rate-distortion trade-off [Bou11, Bou12]. These results motivate our work but do not directly apply to the compressed sensing setting.

Theoretical guarantees more in line with the objectives of this paper began with Jacques et al. [JLBB13] who proved robust recovery from approximately $s \log n$ one-bit measurements. However, the program used has constraints which require sparsity estimation, making it NP-Hard in general. Gupta et al. offers a computationally feasible method via a scheme which either depends on the dynamic range of the signal or is adaptive [GNR10]. Plan and Vershynin analyze a tractable non-adaptive convex program which provides accurate recovery without these types of dependencies [PV13a, PV13b, ALPV14]. Other methods have also been proposed, many of which are largely motivated by classical compressed sensing methods (see e.g. [Bou09, MPD12, YYO12, MBN13, JDDV13]).

In order to break the bound (3) and obtain an exponential rather than polynomial dependence on the oversampling factor, one cannot take traditional non-adaptive measurements. Several schemes have employed adaptive samples including the work of Kamilov et. al. which utilizes a generalized approximate message passing algorithm (GAMP) for recovery, and the adaptive thresholds are selected in line with this recovery method. Adaptivity is also considered in [GNR10] which allows for a constant factor improvement in the number of measurements required. However, to our best knowledge our work is the first to break the bound given by (3).

Regarding the link between our methods and sparse binary regression, there is a number of related theoretical results focusing on sparse logistic regression [NRWY12, Bun08, VDG08, Bac10, RWL10, MVDGB08, KSST10], but these are necessarily constrained by the same limited accuracy of the one-bit compressed sensing model discussed in Section 1.

We also point to the closely related threshold group testing literature, see e.g., [Che13]. In many cases, the statistician has some control over the threshold beyond which the measurement maps to a one. For example, the wording of a binary survey may be adjusted to only ask for a positive answer in an extreme case; a study of the relationship of heart attacks to various factors may test whether certain subjects have heart attacks in a short window of time and other subjects have heart attacks in a long window of time. The main message of this paper is that by carefully choosing this threshold the accuracy of reconstruction of the parameter vector $\boldsymbol{x}$ can be greatly increased.

6.2 Conclusions

We have proposed a recursive framework for adaptive thresholding quantization in the setting of compressed sensing. We have developed both a second-order-cone-programming-based method and a hard-thresholding-based method for signal recovery from these type of quantized measurements. Both of our methods feature a bound on the recovery error of the form $e^{-\Omega(\lambda)}$, an exponential dependence on the oversampling factor λ. To our best knowledge, this is the first result of this kind, and it improves upon the best possible dependence of $\Omega(1/\lambda)$ for non-adaptively quantized measurements.

Acknowledgements

We would like to thank the AIM SQuaRE program for hosting our initial collaboration and also Mr. Lan for discussions around the relationship of our work to logistic regression.

References

[ACS09] E. Ardestanizadeh, M. Cheraghchi, and A. Shokrollahi. Bit precision analysis for compressed sensing. In *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2009.

[ALPV14] A. Ai, A. Lapanowski, Y. Plan, and R. Vershynin. One-bit compressed sensing with non-Gaussian measurements. *Linear Algebra and its Applications*, 441:222–239, 2014.

[Bac10] F. Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4:384–414, 2010.

[BB08] P. T. Boufounos and R. G. Baraniuk. 1-bit compressive sensing. In *Proceedings of the 42nd Annual Conference on Information Sciences and Systems (CISS)*, pages 16–21. IEEE, 2008.

[Bou09] P. T. Boufounos. Greedy sparse signal reconstruction from sign measurements. In *Asilomar Conference on Signals, Systems and Computers*, November 2009.

[Bou11] P. T. Boufounos. Hierarchical distributed scalar quantization. In *Proceedings of the 9th International Conference on Sampling Theory and Applications (SampTA)*, 2011.

[Bou12] P. T. Boufounos. Universal rate-efficient scalar quantization. *IEEE Transactions on Information Theory*, 58(3):1861–1872, 2012.

[Bun08] F. Bunea. Honest variable selection in linear and logistic regression models via ℓ_1 and $\ell_1 + \ell_2$ penalization. *Electronic Journal of Statistics*, 2:1153–1194, 2008.

[CD13] E. J. Candès and M. A. Davenport. How well can we estimate a sparse vector? *Applied and Computational Harmonic Analysis*, 34(2):317–323, 2013.

[Che13] M. Cheraghchi. Improved constructions for non-adaptive threshold group testing. *Algorithmica*, 67(3):384–417, 2013.

[DPM09] W. Dai, H. V. Pham, and O. Milenkovic. A comparative study of quantized compressive sensing schemes. In *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2009.

[DPvdBW14] M. A. Davenport, Y. Plan, E. van den Berg, and M. Wootters. 1-bit matrix completion. *Information and Inference*, 2014.

[DSP] Compressive sensing webpage. http://dsp.rice.edu/cs.

[EK12] Y. C. Eldar and G. Kutyniok. *Compressed sensing: theory and applications*. Cambridge University Press, 2012.

[FR13] S. Foucart and H. Rauhut. *A mathematical introduction to compressive sensing*. Birkhäuser, 2013.

[GLP$^+$10] C. S. Güntürk, M. Lammers, A. M. Powell, R. Saab, and Ö. Yılmaz. Sigma-Delta quantization for compressed sensing. In *Proceedings of the 44th Annual Conference on Information Sciences and Systems (CISS)*. IEEE, 2010.

[GLP$^+$13] C. S. Güntürk, M. Lammers, A. M. Powell, R. Saab, and Ö. Yılmaz. Sobolev duals for random frames and Sigma-Delta quantization of compressed sensing measurements. *Foundations of Computational Mathematics*, 13(1):1–36, 2013.

[GNJN13] S. Gopi, P. Netrapalli, P. Jain, and A. Nori. One-bit compressed sensing: Provable support and vector recovery. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, pages 154–162, 2013.

[GNR10] A. Gupta, R. Nowak, and B. Recht. Sample complexity for 1-bit compressed sensing and sparse classification. In *Proceedings of the International Symposium on Information Theory (ISIT)*. IEEE, 2010.

[GVT98] V. K. Goyal, M. Vetterli, and N. T. Thao. Quantized overcomplete expansions in $\mathbb{R}^N$: analysis, synthesis, and algorithms. *IEEE Transactions on Information Theory*, 44(1):16–31, 1998.

[JDDV13] L. Jacques, K. Degraux, and C. De Vleeschouwer. Quantized iterative hard thresholding: Bridging 1-bit and high-resolution quantized compressed sensing. In *Proceedings of the 10th International Conference on Sampling Theory and Applications (SampTA)*, pages 105–108, 2013.

[JHF11] L. Jacques, D. Hammond, and J. Fadili. Dequantizing compressed sensing: When oversampling and non-gaussian constraints combine. *IEEE Transactions on Information Theory*, 57(1):559–571, 2011.

[JLBB13] L. Jacques, J. N. Laska, P. T. Boufounos, and R. G. Baraniuk. Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors. *IEEE Transactions on Information Theory*, 59(4):2082–2102, April 2013.

[KSST10] S. Kakade, O. Shamir, K. Sridharan, and A. Tewari. Learning exponential families in high-dimensions: Strong convexity and sparsity. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*. JMLR, 2010.

[KSW14] K. Knudson, R. Saab, and R. Ward. One-bit compressive sensing with norm estimation. *arXiv preprint arXiv:1404.6853*, 2014.

[KSY14] F. Krahmer, R. Saab, and Ö. Yılmaz. Sigma-Delta quantization of sub-Gaussian frame expansions and its application to compressed sensing. *Information and Inference*, 2014.

[LWYB11] J. N. Laska, Z. Wen, W. Yin, and R. G. Baraniuk. Trust, but verify: Fast and accurate signal recovery from 1-bit compressive measurements. *IEEE Transactions on Signal Processing*, 59(11):5289–5301, 2011.

[MBN13] Y. Ma, D. Baron, and D. Needell. Two-part reconstruction in compressed sensing. In *Proceedings of the IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 1041–1044, 2013.

[MPD12] A. Movahed, A. Panahi, and G. Durisi. A robust rfpi-based 1-bit compressive sensing reconstruction algorithm. In *Proceedings of the IEEE Information Theory Workshop (ITW)*, pages 567–571. IEEE, 2012.

[MVDGB08] L. Meier, S. Van De Geer, and P. Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71, 2008.

[NRWY12] S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

[PV13a] Y. Plan and R. Vershynin. One-bit compressed sensing by linear programming. *Communications on Pure and Applied Mathematics*, 66(8):1275–1297, 2013.

[PV13b] Y. Plan and R. Vershynin. Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *IEEE Transactions on Information Theory*, 59(1):482–494, 2013.

[PV14] Y. Plan and R. Vershynin. Dimension reduction by random hyperplane tessellations. *Discrete & Computational Geometry*, 51(2):438–461, 2014.

[RWL10] P. Ravikumar, M. J. Wainwright, and J. D. Lafferty. High-dimensional Ising model selection using $\ell 1$-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319, 2010.

[SG09] J. Sun and V. Goyal. Optimal quantization of random measurements in compressed sensing. In *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2009.

[VDG08] S. Van De Geer. High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36(2):614–645, 2008.

[Woj09] P. Wojtaszczyk. Stability and instance optimality for Gaussian measurements in compressed sensing. *Foundations of Computational Mathematics*, 10(1):1–13, April 2009.

[YYO12] M. Yan, Y. Yang, and S. Osher. Robust 1-bit compressive sensing using adaptive outlier pursuit. *IEEE Transactions on Signal Processing*, 60(7):3868–3875, 2012.

[ZBC10] A. Zymnis, S. Boyd, and E. Candès. Compressed sensing with quantized measurements. *IEEE Signal Processing Letters*, 17(2):149–152, February 2010.