Quantized Compressed Sensing: A Survey

Sjoerd Dirksen

Abstract The field of quantized compressed sensing investigates how to jointly design a measurement matrix, quantizer, and reconstruction algorithm in order to accurately reconstruct low-complexity signals from a minimal number of measurements that are quantized to a finite number of bits. In this short survey, we give an overview of the state-of-the-art rigorous reconstruction results that have been obtained for three popular quantization models: one-bit quantization, uniform scalar quantization, and noise-shaping methods.

1 Introduction

In the last 15 years, compressed sensing [8, 9, 23, 29] has matured into a new paradigm in signal processing. This theory predicts that high-dimensional signals can be accurately reconstructed from a small number of measurements provided that the signal has low complexity. Whereas compressed sensing initially focused on the recovery of signals that can be approximately sparsely represented, many rigorous reconstruction results have been obtained for other low-complexity models, such as low-rank matrices and tensors, structured sparse signals, and signals located in a low-dimensional manifold, see e.g., [2, 15, 17, 24, 29, 50, 56] and the references therein.

In the standard compressed sensing model, one assumes that one has direct access to noisy analog linear measurements of the unknown signal x of the form $y = Ax + \nu$. In reality, these analog measurements need to be quantized to a finite number of bits before they can be transmitted, stored, and processed. This operation can be

S. Dirksen (✉)
Utrecht University, Mathematical Institute, P.O. Box 80010, 3508 Utrecht, TA, The Netherlands
e-mail: s.dirksen@uu.nl

H. Boche et al. (eds.), *Compressed Sensing and Its Applications*,
Applied and Numerical Harmonic Analysis,
https://doi.org/10.1007/978-3-319-73074-5_2

modeled by the application of a quantizer map $Q : \mathbb{R}^m \to \mathcal{Q}^m$, where $\mathcal{Q}$ is a finite (or sometimes, countable) alphabet. Accordingly, one has access to

$$q = Q(Ax + \nu). \tag{1}$$

Early works on compressed sensing assumed implicitly that the impact of quantization is negligible in the sense that the error due to the quantization step, i.e., $\eta = Q(Ax + \nu) - (Ax + \nu)$, is small in ℓ_2-norm, say. With this perspective, recovering x from (1) is simply a "usual" noisy compressed sensing problem and one can use standard methods, e.g., basis pursuit denoising, to recover the signal. This approach to recovery from quantized measurements, which we will call the *agnostic* approach, has two downsides. To ensure that the error η is small, one needs to use a very high-resolution quantizer, which may not be realistic or inefficient in practice, and even if this is possible, the estimates on the reconstruction error are pessimistic: the error will not decay beyond the noise floor, in particular not beyond the quantization error.

The area of *quantized compressed sensing* has shown that one can substantially improve over the agnostic approach by designing the triple $(A, Q, \mathscr{A})$ of measurement matrix A, quantizer Q and reconstruction algorithm $\mathscr{A}$ *in unison*. In the last few years, many fascinating results have been obtained in this area. The purpose of this survey is to give an introduction to the main emerging ideas. We do not intend to give an exhaustive overview of the area, but rather focus on rigorous reconstruction guarantees that have been obtained for three popular models in quantized compressed sensing: one-bit compressed sensing, uniform scalar quantization, and noise-shaping methods.

1.1 Notation

Throughout we will use the following notation. We reserve m for the number of measurements, n for the signal dimension, and ρ for the target reconstruction error. For any $N \in \mathbb{N}$ we write $[N] = \{1, \ldots, N\}$. We let $|S|$ denote the cardinality of a set S. We use $\|x\|_p$ to denote the ℓ_p-norm of a vector and $B_p^n = \{x \in \mathbb{R}^n : \|x\|_p \leq 1\}$. We write $\|x\|_0 = |\{i \in [n] : x_i \neq 0\}|$. We use S^{n-1} to denote the Euclidean unit sphere. d_H is the (unnormalized) Hamming distance on the discrete cube. For a random variable ξ we let $\|\xi\|_{L_p}$ denote its L_p-norm. We call ξ L-subgaussian if

$$\sup_{p\geq 1} \frac{\|\xi\|_{L^p}}{\sqrt{p}\|\xi\|_{L^2}} \leq L.$$

is finite. For a given measurement matrix $A \in \mathbb{R}^{m\times n}$ we let $a_1, \ldots, a_m$ denote its rows and refer to them as measurement vectors. We use $A^* \in \mathbb{R}^{n\times m}$ to denote the transpose of A. For a given $T \subset \mathbb{R}^n$ and $1 \leq p, q \leq \infty$, a matrix $A \in \mathbb{R}^{m\times n}$ is said to satisfy $\mathrm{RIP}_{p,q}(T,\varepsilon)$ if

$$(1-\varepsilon)\|x\|_q \leq \|Ax\|_p \leq (1+\varepsilon)\|x\|_q, \qquad \text{for all } x \in T. \tag{2}$$

We call a matrix $A \in \mathbb{R}^{m\times n}$ standard Gaussian if all its entries are i.i.d. standard Gaussian, Bernoulli if its entries are i.i.d. symmetric Bernoulli, or (L-)subgaussian if its entries are independent, mean-zero, unit variance, and (L-)subgaussian. For any $x \in \mathbb{R}^n$ we let $\Gamma_x \in \mathbb{R}^{n\times n}$ be the circulant matrix generated by x, i.e., $(\Gamma_x)_{i,j} = x_{(i-j) \bmod n}$. A circulant matrix implements the discrete circular convolution with x, i.e., $\Gamma_x z = x * z$ for all $z \in \mathbb{R}^n$. If ξ is a vector with independent, mean-zero, unit variance, (L-)subgaussian entries, then we call Γ_ξ an (L-)subgaussian circulant matrix. If the ξ_i are i.i.d. standard Gaussian or symmetric Bernoulli, then we call Γ_ξ a standard Gaussian or Bernoulli circulant matrix. A subsampled partial circulant matrix is obtained by selecting m rows from a circulant matrix. In the literature three different random selection models are considered, which we will give an explicit name here in order to distinguish between them. In the *row picking model*, one selects m rows independently of each other. Each row is picked uniformly at random from the set of $[n]$ rows of Γ_ξ. In the *uniformly at random model*, one selects a subset I uniformly at random from the set of all subsets of $[n]$ of cardinality m. One then considers the measurement matrix $R_I\Gamma_\xi$, where $R_I : \mathbb{R}^n \to \mathbb{R}^{|I|}$ is the operator defined by $R_I z = (z_i)_{i\in I}$. Finally, in the *selector model* one picks a vector $\theta \in \mathbb{R}^n$ of i.i.d. random selectors with mean m/n, sets $I = \{i \in [n] \,:\, \theta_i = 1\}$ and considers the measurement matrix $R_I\Gamma_\xi$. Note that $\mathbb{E}|I| = m$, so m corresponds to the expected number of measurements in this model.

If T is a closed set, then we let P_T be the ℓ_2-projection operator, which assigns to an element $x \in \mathbb{R}^n$ a certain solution of the optimization problem $\min_{z\in T} \|x-z\|_2$. In general, there is not a unique solution unless T is convex. For instance, if T is the set $\Sigma_s = \{x \in \mathbb{R}^n \,:\, \|x\|_0 \leq s\}$ of all s-sparse vectors, then $T = H_s$ is the hard thresholding operator. Finally, c and C denote absolute constants and their value many change from line to line. We use c_α or $c(\alpha)$ to denote a constant that only depends on the parameter α. We write $a \lesssim_\alpha b$ if $a \leq c_\alpha b$, and $a \simeq_\alpha b$ means that both $a \lesssim_\alpha b$ and $a \gtrsim_\alpha b$ hold.

2 Key Concepts

Before investigating the three different quantization models, we first introduce some important general concepts in quantized compressed sensing. We start by specifying the signals that we try to recover and the measurement matrices that we wish to analyze.

- **Low-complexity signal sets**. Any compressed sensing-type scheme exploits the fact that, even though the signal x that we would like to recover may be high-dimensional, it is a priori known to belong to a set of low *intrinsic dimension* or *complexity*. For instance, it is known empirically that many signals are (approximately) sparse in terms of a suitable basis, e.g., natural images can often be approximately sparsely represented in terms of wavelets. Accordingly, the number of measurements that need to be collected to ensure accurate reconstruction is governed by certain parameters that measure the complexity of the signal set. For our purposes, a suitable complexity measure is the *Gaussian width* of a bounded signal set $T \subset \mathbb{R}^n$, which is defined by

$$w(T) = \mathbb{E} \sup_{x \in T} \langle g, x \rangle,$$

where $g \in \mathbb{R}^n$ is standard Gaussian. Another measure that we will use is the ε-covering number $N(T, \varepsilon)$ of T, the minimal number of Euclidean balls of radius ε needed to cover T. The Gaussian width and covering numbers are closely related by Sudakov's and Dudley's inequality, which are the lower and upper bounds, respectively, in

$$\sup_{\varepsilon > 0} \varepsilon \sqrt{\log N(T, \varepsilon)} \lesssim w(T) \lesssim \int_0^\infty \sqrt{\log N(T, \varepsilon)}\, d\varepsilon.$$

Neither of the two bounds is sharp in general, see e.g., [62] for more details. Several of the results that we discuss below state rigorous reconstruction guarantees for a general signal set T and give a bound on the sufficient number of measurements for recovery in terms of the Gaussian width and covering numbers. Other results only concern sparse recovery. To allow for easy comparison, let us recall the following. If $\Sigma_s = \{x \in \mathbb{R}^n \ : \ \|x\|_0 \leq s\}$ is the set of sparse signals, then $w^2(\Sigma_s \cap B_2^n) \simeq s \log(en/s)$ and $\log N(\Sigma_s \cap B_2^n, \rho) \lesssim s \log(en/(s\rho))$. As a model for approximate sparsity, we also consider the larger set of *s-effectively sparse* signals $\Sigma_s^{\text{eff}} = \{x \in \mathbb{R}^n \ : \ \|x\|_1 \leq \sqrt{s}\|x\|_2\}$. If x is s-effectively sparse and $\|x\|_2 \leq 1$, then x belongs to the set of *s-compressible* signals $\sqrt{s}B_1^n \cap B_2^n$. The latter set is essentially the convex hull of the set of s-sparse vectors in the unit ball (see [53, Lemma 3.1]):

$$\operatorname{conv}(\Sigma_s \cap B_2^n) \subset \sqrt{s}B_1^n \cap B_2^n \subset 2\operatorname{conv}(\Sigma_s \cap B_2^n). \tag{3}$$

Since the Gaussian width is invariant under taking convex hulls, one finds $w^2(\sqrt{s}B_1^n$
$\cap B_2^n) \simeq s \log(en/s)$.

- **Random matrices**. Similarly to the situation in "unquantized" compressed sensing, the best-known recovery guarantees in quantized compressed sensing have been obtained for *random* measurement matrices. In particular, in quantized compressed sensing, optimal results have been obtained for standard Gaussian measurement matrices, i.e., matrices with independent standard Gaussian entries. These results are mostly of theoretical interest, as these matrices are difficult to realize in a practical measurement setup. On the other hand, it has proven very challenging to establish recovery guarantees for deterministic measurement matrices involving a number of measurements that is close to optimal. As a compromise between completely random matrices and deterministic ones, it is of interest to study *structured random matrices*, which arise when introducing randomness in (more) realistic measurement models. Two particularly popular classes of matrices, which can be considered as the “fruitflies” of compressed sensing with structured matrices, are *partial random circulant matrices* and *randomly subsampled bounded orthonormal systems*. The former model is connected to SAR radar imaging, Fourier optical imaging, and channel estimation (see e.g., [58] and the references therein). The latter model is relevant to many applications, for instance, models in compressive magnetic resonance imaging [47]. In standard compressed sensing it has been shown that stable and robust sparse recovery can be achieved with a near-optimal (i.e., up to logarithmic factors) number of measurements, see [7, 35, 44, 49, 59] for the best known bounds for the two respective classes of matrices. Recently, substantial progress has been made on quantized compressed sensing with structured random matrices. We will mostly restrict our discussion to (sub)gaussian matrices and circulant matrices, as results for these matrices have been obtained for all three quantization models that we consider in this survey.

Let us now discuss some terminology regarding quantization.

- **Memoryless versus adaptive schemes**. The quantizer $Q : \mathbb{R}^m \to \mathcal{Q}^m$ is called *memoryless* if it quantizes each entry of its input vector independently of the others. In contrast, an *adaptive* quantizer quantizes the i-th measurement using knowledge of previous analog measurements, their quantizations, and in some cases, even reconstructions of the signal based on the previous $i-1$ quantized measurements. As we will discuss below, adaptive methods can achieve a fundamentally better error decay rate. Whereas the reconstruction error cannot decay faster than linear (i.e., as $O(1/m)$) in terms of the number of measurements if a memoryless scalar quantization scheme is used, adaptive schemes can achieve a polynomial or even an optimal exponential error decay rate. This improved rate comes at a price: the implementation of adaptive schemes generally requires hardware that is more complicated and consumes more energy in operation. In addition, since by their very nature adaptive methods require measurements to be acquired sequentially, their implementation may be difficult or impossible in some sensing scenarios, e.g., in distributed sensing with a sensor network.
- **Dithering**. In the engineering literature on quantization, it has been known for a long time (at least since the work [57], see also [31, 32]) that it is potentially helpful to add random noise to the analog measurements before quantizing. This operation is called *dithering*. Note that the term “random noise” is somewhat misleading, since at least we have the freedom to *design* the distribution of the dithering vector. Indeed, as we will see below, it was recently shown rigorously that dithering with *well-chosen* distributions can substantially improve reconstruction guarantees in quantized compressed sensing.

Finally, we formalize some concepts regarding recovery methods.

- **Uniform versus non-uniform recovery**. The reconstruction results in quantized compressed sensing involving random matrices or dithering are guarantees to reconstruct a signal x or a class of signals with "high probability", which typically means that recovery will only fail with a probability that decays exponentially in terms of the number of measurements. These results can either be *uniform*, meaning that a high probability event exists upon which one can reconstruct any signal $x \in T$ (e.g., the set of all sparse vectors with unit norm), or *non-uniform*, meaning that the high probability event depends on the specific signal x which is to be recovered. Accordingly, a uniform guarantee is sometimes informally called a "for all" guarantee, whereas a non-uniform one is called a "for one" guarantee. To understand the difference between the two from a practical point of view, suppose that $A = R_I U$ is a randomly subsampled unitary matrix and suppose that T is the set of all s-sparse vectors on the unit sphere. A uniform recovery guarantee means that when we draw a random sample of the rows of U then, with high probability, we can recover any unit norm s-sparse vector from $Q(Ax + \nu)$. Thus, with high probability, a *single* random draw of the rows will yield a matrix that can be used for compressed sensing of *any* signal from the set T. A non-uniform guarantee is much weaker: only for a *fixed* signal x one shows that with high probability one can draw a random subset of the rows so that x can be recovered from its measurements. Hence, in this setting, we only guarantee good reconstruction performance with high probability if we draw a new random subset of the rows of U each time that we measure a new signal.
- **Quantization consistency**. A vector $x^{\#}$ is called *quantization consistent* with the true signal x if, when we were to measure and quantize $x^{\#}$, we would reproduce the observed quantized measurements. For instance, if we observe $q = Q(Ax)$, then $x^{\#}$ is quantization consistent if $q = Q(Ax^{\#})$. Several successful reconstruction methods that will be introduced below search for a quantization consistent vector.
- **Stability and robustness**. A triple $(A, Q, \mathscr{A})$ can only be expected to perform satisfactorily if it is *stable* and *robust*. We say that it is stable if the reconstruction performance does not deteriorate sharply if the signal lies "slightly outside of" the low-complexity set T. For instance, in the context of sparse recovery it is desirable to be able to accurately recover vectors that are not exactly sparse, but only effectively sparse or compressible. In addition, we would like to ensure that $(A, Q, \mathscr{A})$ is robust with respect to both *pre-quantization noise*, i.e., the noise ν on the analog measurements, as well as *post-quantization noise*, i.e., bit corruptions occurring during the quantization process.

3 Two Fundamental Limits

To set benchmarks for the reconstruction results for the three different quantization models, let us first formulate two fundamental lower bounds for the recovery error. The first concerns a lower bound for (uniform) recovery of signals from a set T in terms of its covering numbers. Suppose that we wish to quantify how many bits we need to collect to ensure that the worst case ℓ_2-reconstruction error of a reconstruction map $\mathscr{A}$ over the set T, i.e.,

$$\sup_{x \in T} \|x - \mathscr{A}(Q(Ax + \nu))\|_2,$$

is at most ρ. If this is fulfilled, then the set of Euclidean balls with radii ρ and centers in the image set $\mathscr{A}(Q(Ax + \nu))$ form a covering of T. If our quantization scheme Q encodes any analog measurement vector $Ax + \nu$ into at most B bits, then this cover has at most 2^B elements. Thus, the minimal total number of bits required to attain worst case error ρ over T satisfies

$$B \geq \log_2 N(T, \rho).$$

In particular, if we collect L bits per measurement, then at least $m \gtrsim \log_2 N(T, \rho)/L$ measurements are necessary. As an example, $\log_2 N(T, \rho) \simeq s \log_2(1/\rho)$ if T is the intersection of the Euclidean unit sphere with an s-dimensional subspace, so the worst case reconstruction error cannot decay faster than exponential in terms of the number of measurements in this case. In particular, one cannot obtain a better worst case error decay rate for the set of s-sparse vectors on the sphere.

The second fundamental lower bound concerns non-uniform recovery of sparse vectors.

Theorem 1 ([21, Theorem 1.3]) *Suppose that ν contains i.i.d. centered Gaussian random variables with variance σ^2. Let A be a (random) measurement matrix that satisfies, with probability at least* 0.95*,*

$$\|Ax\|_2 \leq \kappa\sqrt{m}\|x\|_2, \qquad \text{for all } x \in \Sigma_s \cap B_2^n. \tag{4}$$

Let Ψ be any recovery procedure such that, for every fixed $x \in \Sigma_s \cap B_2^n$, when receiving as data the measurement matrix A and the noisy linear measurements $Ax + \nu$, Ψ returns $x^\sharp$ that satisfies $\|x^\sharp - x\|_2 \leq \rho$ with probability 0.9*. Then*

$$m \geq c\kappa^{-2}\sigma^2 \frac{s \log(en/s)}{\rho^2}.$$

Note that the condition (4) is satisfied by many popular random measurement matrices if $m \gtrsim s \log^\alpha(n)$, in particular by subgaussian matrices, partial subgaussian circulant matrices and randomly subsampled bounded orthonormal systems. For these matrices the sample size required for recovery with accuracy ρ is at least $\sigma^2 s \log(en/s)/\rho^2$, even if one receives the noisy analog linear measurements *prior to quantization*, and is then free to use those measurements as one sees fit. In particular, in a high noise setting one cannot hope to achieve a better error decay rate than $O(1/\sqrt{m})$.

4 One-Bit Compressed Sensing

We start by discussing *one-bit compressed sensing*, which studies the extreme case where each measurement is quantized to a *single* bit. Specifically, we consider the map $Q_\tau : \mathbb{R}^m \to \{-1, 1\}^m$ defined by $Q_\tau(z) = \mathrm{sign}(z + \tau)$, where sign is the signum function applied element-wise and $\tau \in \mathbb{R}^m$ is a vector of quantization thresholds. This quantizer is memoryless if τ is a fixed or a randomly generated vector. In this case, the one-bit quantizer can be easily implemented by voltage comparison to fixed thresholds (τ deterministic) combined with dithering (τ random). Due to the efficiency of the memoryless one-bit quantizer, one-bit compressed sensing is one of the most popular quantized compressed sensing models. For a memoryless one-bit quantizer we cannot expect better than linear decay of the reconstruction error [6, 30, 42]. However, as we will see in Sect. 4.4, optimal error decay can be achieved by choosing the thresholds adaptively.

In the context of one-bit compressed sensing, post-quantization noise takes the form of "bit flips": the quantizer erroneously produces the bit $-q_i$ rather than $q_i = \mathrm{sign}(\langle a_i, x\rangle + \tau_i)$. One can either assume that bit corruptions occur in a random fashion, i.e., one observes a vector $q_c \in \{-1, 1\}^m$ satisfying $(q_c)_i = f_i q_i$, where the f_i are independent random variables satisfying $\mathbb{P}(f_i = -1) = 1 - \mathbb{P}(f_i = 1) = p$, i.e., a bit is corrupted with probability p. Alternatively, one can assume that a small fraction β of the bits are arbitrarily corrupted, i.e., one observes a vector $q_c \in \{-1, 1\}^m$ satisfying $d_H(q, q_c) \le \beta m$. Clearly, the second noise model is more challenging to analyze, as bit corruptions can in principle occur in an adversarial fashion.

4.1 Memoryless One-Bit Compressed Sensing: Zero Thresholds

One-bit compressed sensing was first considered by Boufounos and Baraniuk [5] in the completely noiseless case (i.e., neither pre- nor post-quantization noise) and $\tau = 0$. In this case, one simply observes $q = \mathrm{sign}(Ax)$. Since the sign function is invariant under positive scaling, the energy $\|x\|_2$ of the signal x is lost during quantization and one can only hope to recover its direction $x/\|x\|_2$. For this reason, it is standard in this original one-bit compressed sensing model to assume that $\|x\|_2 = 1$. From a geometric perspective, the vector q is a rough encoding of the position of x on S^{n-1}. To see this, note that each measurement vector a_i (i.e., the i-th row of A) determines a hyperplane $H_{a_i} = \{z \in \mathbb{R}^n : \langle a_i, z\rangle = 0\}$ passing through the origin. The corresponding quantized measurement $\mathrm{sign}(\langle a_i, x\rangle)$ indicates on which side of the hyperplane x is located. By taking m measurements, the space $\mathbb{R}^n$ is tessellated into (at most) 2^m cells, and the bit sequence $q = \mathrm{sign}(Ax) = (\mathrm{sign}(\langle a_i, x\rangle))_{i=1}^m \in \{-1, 1\}^m$ encodes in which cell x is located.

The original paper [5] considered recovery of a sparse vector from its one-bit measurements and proposed to reconstruct the signal via

$$\min_{z\in\mathbb{R}^n} \|z\|_0 \qquad \text{s.t.} \qquad q = \text{sign}(Az), \ \ \|z\|_2 = 1. \tag{5}$$

The linear constraint $q = \text{sign}(Az)$ forces any solution $x^\#$ to (5) to be quantization consistent. Geometrically, a vector z is quantization consistent with x precisely when it is located in the same cell of the hyperplane tessellation induced by the quantized measurements. To show that one can recover any $x \in \Sigma_s \cap S^{n-1}$ via (5) up to error ρ, one therefore needs to ensure that the measurement vectors tessellate $\Sigma_s \cap S^{n-1}$ into cells with diameter at most ρ. It was shown in [42, Theorem 2] that standard Gaussian vectors have this property: if $A \in \mathbb{R}^{m\times n}$ is standard Gaussian and $m \gtrsim \rho^{-1}s\log(n/\rho)$ then, with high probability, any s-sparse x, x' with $\|x\|_2 = \|x'\|_2 = 1$ and $\text{sign}(Ax) = \text{sign}(Ax')$ satisfy $\|x - x'\|_2 \leq \rho$. In particular, any solution $x^\#$ to (5) satisfies $\|x^\# - x\|_2 \leq \rho$. The number of measurements needed for this reconstruction is essentially optimal: in fact, the reconstruction $x^\#$ of an s-sparse vector produced by *any* method using $\text{sign}(Ax)$ as its input must satisfy the lower bound $\|x^\# - x\|_2 \gtrsim s/(m + s^{3/2})$ [42, Theorem 1]. Hence, the reconstruction error cannot decay faster than linear (i.e., than $O(1/m)$). This linear decay bottleneck is common to all memoryless scalar quantization methods, see Sect. 5.

Even though the error of the reconstruction produced by (5) decays essentially optimally if A is standard Gaussian, this program is hard to solve. Although one can convexify the objective of (5) by replacing $\|z\|_0$ by $\|z\|_1$, the constraint $\|z\|_2 = 1$ is problematic (note that the relaxation $\|z\|_2 \leq 1$ leads to a trivial program). A solution to this problem was proposed by Plan and Vershynin [53]: the simple, yet effective, idea is to observe that if A is standard Gaussian, then for any $z \in \mathbb{R}^n$,

$$\frac{1}{m}\mathbb{E}\|Az\|_1 = \sqrt{\frac{2}{\pi}}\|z\|_2.$$

This suggests to use the reconstruction program

$$\min_{z\in\mathbb{R}^n} \|z\|_1 \qquad \text{s.t.} \qquad q = \text{sign}(Az), \ \ \|Az\|_1 = m\sqrt{\frac{2}{\pi}}, \tag{6}$$

which is a linear program. Plan and Vershynin showed that using $m \gtrsim \rho^{-5}s\log^2(n/s)$ standard Gaussian measurements one can, with high probability, recover every $x \in \mathbb{R}^n$ with $\|x\|_1 \leq \sqrt{s}$ and $\|x\|_2 = 1$ via (6) up to reconstruction error ρ. This was the first uniform reconstruction result for stable recovery of sparse vectors from their one-bit measurements via a tractable program. Still, the program (6) has a weakness, which is common to any recovery program that enforces quantization consistency: the program can easily fail in the presence of post-quantization noise. Indeed, already a single bit corruption can cause (6) to be infeasible: there will simply be no vector

z which is consistent with the observed corrupted quantized measurements (see [20] for a detailed discussion).

In order to handle post-quantization noise, Plan and Vershynin introduced a different program in [54], which can be used to robustly reconstruct signals from an arbitrary set $T \subset S^{n-1}$, namely

$$\max_{z\in\mathbb{R}^n} \langle q_c, Az\rangle \qquad \text{s.t.} \qquad z \in T. \tag{7}$$

That is, we search for a vector that maximizes the correlation between the linear and observed corrupted quantized measurements. This program is convex if T is convex and therefore [54] suggested to use this program with $T = \text{conv}(\Sigma_s \cap B_2^n)$ for stable sparse recovery. By (3), this leads to the tractable program

$$\max_{z\in\mathbb{R}^n} \langle q_c, Az\rangle \qquad \text{s.t.} \qquad \|z\|_1 \leq \sqrt{s}, \ \|z\|_2 \leq 1.$$

In a non-uniform recovery setting, Plan and Vershynin showed that $m \gtrsim \rho^{-4} w^2(T)$ measurements suffice to reconstruct a fixed signal in T with high probability up to error ρ, even if pre-quantization noise is present and quantization bits are randomly flipped with a probability that is allowed to be arbitrarily close to $1/2$. A much deeper result is the following uniform recovery theorem, which proves robustness of (7) to adversarial post-quantization noise.

Theorem 2 ([54, Theorem 1.3]) *Fix $0 < \rho, \beta \leq 1$, let $T \subset B_2^n$ and let $A \in \mathbb{R}^{m\times n}$ be standard Gaussian. Suppose that*

$$m \geq c_2 \frac{\log^3(e/\rho)}{\rho^{12}} w^2(T), \qquad \beta\sqrt{\log(e/\beta)} = c_3\rho^2.$$

Then with probability at least $1 - e^{-c_1 m\rho^4/\log(e/\rho)}$ the following holds for any $x \in T$ with $\|x\|_2 = 1$. If we observe $q_c \in \{-1, 1\}^m$ with $d_H(q_c, \text{sign}(Ax)) \leq \beta m$, then any solution $x^{\#}$ to (7) satisfies $\|x^{\#} - x\|_2 \leq \rho$.

The results mentioned so far all concern standard Gaussian measurement matrices. For other measurement matrices, signal recovery from the one-bit measurements $q = \text{sign}(Ax)$ can very easily fail, even if the measurement matrix enjoys optimal recovery guarantees in "unquantized" compressed sensing. For instance, it was pointed out in [1] that if $A \in \mathbb{R}^{m\times n}$ is a matrix with entries in $\{-1, 1\}$ (e.g., a Bernoulli matrix), then there are already two-sparse vectors that cannot be accurately recovered. For instance, for any $0 < \lambda < 1$, the vectors

$$x_{+\lambda} = (1+\lambda^2)^{-1/2}(1, \lambda, 0, \ldots, 0), \qquad x_{-\lambda} = (1+\lambda^2)^{-1/2}(1, -\lambda, 0, \ldots, 0) \tag{8}$$

produce identical one-bit measurements $\text{sign}(Ax_{+\lambda}) = \text{sign}(Ax_{-\lambda})$, irrespective of the draw of A and the number of measurements. Hence, there is no hope to accurately recover these vectors. Nevertheless, in [1] some non-uniform recovery results from [54] were generalized to subgaussian matrices by imposing additional restrictions. For a fixed $x \in T \subset S^{n-1}$ they showed that $m \gtrsim \rho^{-4} w^2(T)$ suffice to reconstruct x up to error ρ via (7) with high probability provided that either $\|x\|_\infty \leq \rho^4$ (since $\|x\|_2 = 1$, this means that the energy of the signal must be sufficiently spread out over its coordinates) or the total variation distance between the subgaussian distribution of the entries of A and the standard Gaussian distribution is at most ρ^{16}.

Even though one-bit compressed sensing generally fails for subgaussian matrices, Foucart [27] identified a different class of matrices for which accurate one-bit compressed sensing is possible. He showed that one can accurately recover signals from one-bit measurements if the measurement matrix satisfies an appropriate RIP-type property of the form (2).

Theorem 3 ([27, Theorem 8]) *If A satisfies $RIP_{1,2}(\Sigma_{2s}, \varepsilon)$, then for every $x \in \mathbb{R}^n$ with $\|x\|_0 \leq s$ and $\|x\|_2 = 1$, the hard thresholding reconstruction $x^\#_{HT} = H_s(A^*q)$ satisfies $\|x - x^\#_{HT}\|_2 \leq 2\sqrt{5\varepsilon}$.*

Let $\varepsilon \leq 1/5$. If A satisfies $RIP_{1,2}(\Sigma^{eff}_{9s}, \varepsilon)$, then for every $x \in \mathbb{R}^n$ with $\|x\|_1 \leq \sqrt{s}$ and $\|x\|_2 = 1$, any solution $x^\#_{LP}$ to (6) satisfies $\|x - x^\#_{LP}\|_2 \leq 2\sqrt{5\varepsilon}$.

A special case of a result of Schechtman [61] shows that if B is standard Gaussian and $A = \frac{1}{m}\sqrt{\frac{\pi}{2}}B$, then A satisfies $\mathrm{RIP}_{1,2}(T, \varepsilon)$ with probability at least $1 - 2e^{-m\varepsilon^2/2}$ if $m \gtrsim \varepsilon^{-2} w^2(T_n)$, where $T_n = \{x/\|x\|_2 : x \in T\}$ (see also [55, Lemma 2.1] for a short proof of this special case). In particular, for $T = \Sigma_{2s}$ or $T = \Sigma^{\mathrm{eff}}_{9s}$ this is satisfied if $m \gtrsim \varepsilon^{-2} s \log(en/s)$. Hence, the first statement of Theorem 3 shows that in this case the hard thresholding reconstruction $x^\#_{\mathrm{HT}}$ achieves error ρ if $m \gtrsim \rho^{-4} s \log(en/s)$, which is slightly better than [41, Propositions 1 and 2]. The second statement shows that any solution to the linear program (6) achieves reconstruction error ρ if $m \gtrsim \rho^{-4} s \log(en/s)$, which is a small improvement of the condition originally obtained in [53].

Theorem 3 can be made robust to a small amount of pre-quantization noise: if we observe $q = \mathrm{sign}(Ax + \nu)$, then the first statement holds with error bound $\|x - x^\#_{\mathrm{HT}}\|_2 \lesssim \sqrt{\varepsilon + \|\nu\|_1}$. A similar error bound can be obtained for solutions to an augmented version of the linear program (6), which accounts for the noise. In addition, one can prove a result analogous to Theorem 3 for recovery of low-rank matrices via hard thresholding or a semidefinite program (in the noiseless case, the latter arises by replacing the objective $\|z\|_1$ in (6) by the nuclear norm). We refer to [28] for these extensions and resulting recovery results of low-rank matrices from one-bit standard Gaussian measurements.

In [18], Theorem 3 was used to derive uniform recovery guarantees for randomly subsampled standard Gaussian circulant matrices under a *small sparsity assumption*. For a target reconstruction accuracy $0 < \rho \leq 1$, it is assumed that the sparsity s is small enough, i.e.,

$$s \lesssim \rho^2 \sqrt{n/\log(n)}. \tag{9}$$

If $0 < \rho \leq (\log^2(s)\log(n))^{-1/4}$ and

$$m \gtrsim \rho^{-4} s \log(en/(s\rho^4))$$

then, with high probability, for any $x \in \mathbb{R}^n$ with $\|x\|_0 \leq s$ and $\|x\|_2 = 1$ the hard thresholding reconstruction $x^{\#}_{\mathrm{HT}}$ satisfies $\|x - x^{\#}_{\mathrm{HT}}\|_2 \leq \rho$. Under slightly stronger conditions a similar uniform reconstruction result can be obtained for effectively sparse vectors on the unit sphere via (6). It is conjectured that a small sparsity assumption is not necessary for these results.

4.2 Memoryless One-Bit Compressed Sensing With Dithering

Memoryless one-bit quantization with zero thresholds suffers from two downsides. First, one can only recover signals located on the unit sphere or, viewed differently, only the direction of signals. Second, it is easy to find measurement matrices that perform optimally in "unquantized" compressed sensing for which one-bit compressed sensing fails. These two issues can be resolved by introducing dithering in the quantization process. Let $Q_\tau : \mathbb{R}^m \to \{-1, 1\}^m$ again denote the map $Q_\tau(z) = \mathrm{sign}(z + \tau)$ and consider the measurements $q = Q_\tau(Ax)$. We can interpret this measurement vector geometrically in a similar way as before, except that each measurement now determines a hyperplane $H_{a_i,\tau_i} = \{z \in \mathbb{R}^n \ : \ \langle a_i, z\rangle + \tau_i = 0\}$, which is a parallel shift of the hyperplane H_{a_i}. This immediately explains why dithering can be helpful to recover signals outside of the unit sphere: whereas two signals lying on a straight line cannot be separated by a hyperplane through the origin (and are therefore located in the same cell of the tessellation if $\tau = 0$), they can be separated by shifted hyperplanes. Later we will see that dithering can also greatly extend the class of measurement matrices for which accurate recovery from one-bit measurements can be achieved.

In the setting of Gaussian measurement matrices, recovery results for sparse vectors in the unit ball were first obtained in [4, 43]. In particular, [43] used Gaussian thresholds τ_i and used a slight modification of the linear program (6) for recovery. We will discuss a similar result that was obtained in [4] for the second- order cone program

$$\min_{z\in\mathbb{R}^n} \|z\|_1 \qquad \text{s.t.} \qquad q = \mathrm{sign}(Az + \tau), \ \|z\|_2 \leq R, \tag{10}$$

with $q = \mathrm{sign}(Ax + \tau)$. The rough idea behind the results in [4, 43] is a reduction to the 'standard' one-bit compressed sensing model of Sect. 4.1: we view the dithered measurements $\mathrm{sign}(Ax + \tau)$ as zero-threshold one-bit measurements $\mathrm{sign}([A\ \frac{\tau}{R}]\bar{x})$ of the unit norm vector $\bar{x} = [x, R]/\|[x, R]\|_2 \in S^{n+1}$, where the vector $[x, R] \in \mathbb{R}^{n+1}$ is obtained from x by appending the scalar R as an extra entry. To find an approximant of x, it suffices to find an approximant of $\bar{x}$ of the form $\bar{z}$: by the argument in the proof of [4, Corollary 9] one finds $\|x - z\|_2 \leq 2R\|\bar{x} - \bar{z}\|_2$ for any two vectors $x, z \in RB_{\ell_2^n}$. If A is standard Gaussian, then a small amount of adversarial pre-quantization noise can be handled in a similar fashion by using that A satisfies a *simultaneous* (ℓ_2, ℓ_1)*-quotient* property: with probability at least $1 - e^{-cm}$ any

$\nu \in \mathbb{R}^m$ can be written as $\nu = Au$ for some $u \in \mathbb{R}^n$ with $\|u\|_2 \leq c_1\|\nu\|_2/\sqrt{m}$ and $\|u\|_1 \leq c_1\|\nu\|_2/\sqrt{\log(n/m)}$.

Based on the above reasoning and the binary embedding result (16) stated below, the following was shown.

Theorem 4 ([4, Theorem 2]) *There exist absolute constants c_0, c_1, c_2 such that the following holds. Suppose that $A \in \mathbb{R}^{m\times n}$ is standard Gaussian, $\tau_1, \ldots, \tau_m$ are independent $\mathcal{N}(0, 4R^2)$-distributed. If*

$$m \geq c_0\rho^{-4}s\log(n/s),$$

then the following holds with probability at least $1 - 3e^{-c_1 m\rho^4}$: for any $x \in \mathbb{R}^n$ with $\|x\|_0 \leq s$ and $\|x\|_2 \leq R$ and $q = sign(Ax + \nu + \tau)$ with $\|\nu\|_\infty \leq c_2R\rho^3$, any solution $x^\#$ to (10) satisfies $\|x - x^\#\|_2 \leq R\rho$.

The linear programming result of [43] and Theorem 4 were extended further to recovery of (effectively) dictionary sparse signals in [3].

Similarly to Theorem 3, uniform recovery via (10) can be ensured via an appropriate RIP$_{1,2}$-property. Suppose that $\nu = 0$ and consider

$$\min_{z\in\mathbb{R}^n} \|z\|_1 \quad \text{s.t.} \quad \text{sign}(C[z, R]) = \text{sign}(C[x, R]), \ \|z\|_2 \leq R, \tag{11}$$

then (10) is obtained by taking $C = [A\ \frac{\tau}{R}]$. It was shown in [18] that if $\varepsilon < 1/5$ and C satisfies RIP$_{1,2}(\Sigma^{\text{eff}}_{36(\sqrt{s}+1)^2}, \varepsilon)$, then for any $x \in \mathbb{R}^n$ satisfying $\|x\|_1 \leq \sqrt{s}\|x\|_2$ and $\|x\|_2 \leq R$, any solution $x^\#$ to (11) satisfies $\|x - x^\#\|_2 \leq 2R\sqrt{\varepsilon}$. To connect this to Theorem 4, note that if τ contains i.i.d. $\mathcal{N}(0, R)$-distributed entries, then $C = [A\ \frac{\tau}{R}]$ is standard Gaussian. By Schechtman's result, $\frac{1}{m}\sqrt{\frac{\pi}{2}}C$ satisfies RIP$_{1,2}(\Sigma^{\text{eff}}_{36(\sqrt{s}+1)^2}, \varepsilon)$ if $m \gtrsim \varepsilon^{-2}s\log(en/s)$ and this immediately implies Theorem 4 (in the case $\nu = 0$). In [18] it was shown that if A is a random partial standard Gaussian circulant matrix, then $\frac{1}{m}\sqrt{\frac{\pi}{2}}C$ with high probability satisfies the same RIP property if $m \gtrsim \varepsilon^{-4}s\log(en/s) + s\log^2 s\log^2 n$ and a certain small sparsity assumption (similar to (9)) is satisfied. Thus, the conclusion of Theorem 4 (for $\nu = 0$) remains valid in this case if $m \gtrsim \rho^{-8}s\log(en/s) + s\log^2 s\log^2 n$.

The program (10) (as well as the linear program in [43]) reconstruct by enforcing quantization consistency. For this reason, this program can easily fail in the case of post-quantization noise, as has been discussed in Sect. 4.1. In addition, since the approaches in [4, 18, 43] essentially reduce to the standard one-bit compressed sensing model, the type of measurement matrices for which results can be obtained is relatively limited: so far only reconstruction results are known for standard Gaussian and, under additional restrictions, randomly subsampled standard Gaussian circulant matrices and subgaussian matrices. These limitations were overcome in [20, 21] by using uniform dithering, as we will now discuss.

In [20], recovery results were obtained for matrices with i.i.d. subgaussian or heavy-tailed rows, which are stable and robust with respect to both pre- and post-quantization noise. Suppose that we observe a vector $q_c \in \{-1, 1\}^m$ satisfying

$$d_H(q_c, \operatorname{sign}(Ax + \nu + \tau)) \leq \beta m,$$

i.e., at most a fraction β of the bits are arbitrarily corrupted during quantization. Consider

$$\min_{z \in \mathbb{R}^n} d_H(q_c, \operatorname{sign}(Az + \tau)) \qquad \text{s.t.} \qquad z \in T. \tag{12}$$

This (non-convex) program selects an $x^{\#}$ whose noiseless one-bit measurements minimize the Hamming distance to the corrupted vector of quantized noisy measurements. The following recovery theorem applies to subgaussian random matrices. A more general version of this result can be proved for *heavy-tailed* measurement vectors, see [20].

Theorem 5 ([20, Theorem 1.5]) *There exist constants $c_0, \ldots, c_4 > 0$ depending only on L such that the following holds. Suppose that $A \in \mathbb{R}^{m \times n}$ has i.i.d. symmetric, isotropic, L-subgaussian rows, ν has i.i.d. L-subgaussian entries with variance σ^2, and τ has i.i.d. entries which are uniformly distributed on $[-\lambda, \lambda]$. Let $T \subset RB_2^n$, set $\lambda \geq c_0(R + \sigma) + \rho$, put $r = c_1\rho/\sqrt{\log(e\lambda/\rho)}$, and let $T_r = (T - T) \cap rB_2^n$. Assume that*

$$m \geq c_2 \lambda \left(\frac{w^2(T_r)}{\rho^3} + \frac{\log \mathcal{N}(T, r)}{\rho} \right), \tag{13}$$

and that $|\mathbb{E}\nu_1| \leq c_3\rho$, $\sigma \leq c_3\rho/\sqrt{\log(e\lambda/\rho)}$ and $\beta \leq c_3\rho/\lambda$. Then, with probability at least $1 - 10\exp(-c_4 m\rho/\lambda)$, for every $x \in T$, any solution $x^{\#}$ of (12) satisfies $\|x^{\#} - x\|_2 \leq \rho$.

If $T \subset B_2^n$ and $\sigma \leq 1$ then λ is a constant that depends only on L. In this case (see [20] for details) (13) holds if

$$m = c(L) \frac{\log(e/\rho)}{\rho^3} w^2(T).$$

In the special case $T = \Sigma_s \cap B_2^n$, a much better estimate is possible:

$$m = c'(L)\rho^{-1} s \log\left(\frac{en}{s\rho}\right).$$

The latter is optimal in terms of s and n and optimal up to the log-factor in terms of ρ.

The result in Theorem 5 is still rather sensitive to pre-quantization noise: the mean and variance of the noise should be of the order of ρ. In addition to this sensitivity to pre-quantization noise, the program (12) is computationally hard to solve. To resolve these two issues a different program, which is essentially obtained by convexifying the objective of (12), was introduced in [20]: for $\lambda > 0$ consider

$$\max_{z \in \mathbb{R}^n} \frac{1}{m}\langle q_c, Az \rangle - \frac{1}{2\lambda}\|z\|_2^2 \qquad \text{s.t.} \qquad z \in U, \tag{14}$$

where either $U = T$ or $U = \operatorname{conv}(T)$. In the first case, we can view (14) as a regularized version of (7). As is pointed out in [21], (14) is equivalent to

$$\min \left\| z - \frac{\lambda}{m} A^* q_c \right\|_2 \qquad \text{s.t.} \qquad z \in U, \tag{15}$$

i.e., it computes an ℓ_2-projection $P_U(\frac{\lambda}{m} A^* q_c)$ of $\frac{\lambda}{m} A^* q_c$ onto U. If $U = \operatorname{conv}(T)$, then (14) is convex, has a unique solution and can be expected to be stable. On the other hand, if T is "simple", then it may be advantageous to take $U = T$. For instance, if $U = T = \Sigma_s \cap B_2^n$, then (14) has a closed-form solution

$$x^{\#} = \min \left\{ \frac{\lambda}{m}, \frac{1}{\|H_s(A^* q_c)\|_2} \right\} H_s(A^* q_c),$$

where H_s is the hard thresholding operator. The following result is stated for $U = \operatorname{conv}(T)$ in [20, Theorem 1.7], the case $U = T$ is immediate from the proof.

Theorem 6 ([20, Theorem 1.7]) *There exist constants $c_0, \ldots, c_4$ that depend only on L for which the following holds. Suppose that either $U = T$ and $T - T$ is star-shaped or $U = \operatorname{conv}(T)$. Suppose that A has i.i.d. symmetric, isotropic, L-subgaussian rows, ν has i.i.d. mean-zero, L-subgaussian entries with variance σ, and τ has i.i.d. entries which are uniformly distributed on $[-\lambda, \lambda]$. Let $T \subset RB_2^n$, fix $\rho > 0$, set $U_\rho = (U - U) \cap \rho B_2^n$,*

$$\lambda \geq c_0(\sigma + R)\sqrt{\log(c_0(\sigma + R)/\rho)}$$

and let $r = c_1\rho/\log(e\lambda/\rho)$. If m and β satisfy

$$m \geq c_2 \left(\left(\frac{\lambda w(U_\rho)}{\rho^2} \right)^2 + \lambda^2 \frac{\log \mathcal{N}(T, r)}{\rho^2} \right), \qquad \beta\sqrt{\log(e/\beta)} = c_3 \frac{\rho}{\lambda},$$

then, with probability at least $1 - 8\exp(-c_4 m\rho^2/\lambda^2)$, for any $x \in T$ any solution $x^{\#}$ of (14) satisfies $\|x^{\#} - x\|_2 \leq \rho$.

If T is the set of sparse or compressible vectors in RB_2^n, then Theorem 6 can be extended to randomly subsampled subgaussian circulant matrices (with rows selected according to the selector model). The only difference is that some additional logarithmic factors appear in the result. We refer to [21, Theorem 1.1] for details.

If $T = \Sigma_s \cap B_2^n$ and $\sigma \geq 1$, then we can take $\lambda = c(L)\sigma\sqrt{\log(c(L)\sigma/\rho)}$ and

$$m = c'(L)\frac{\sigma^2}{\rho^2} s \log\left(\frac{\sigma}{\rho}\right)\left(\log\left(\frac{en}{s\rho}\right) + \log\log\left(\frac{e\sigma}{\rho}\right)\right),$$

which is optimal up to logarithmic factors by Theorem 1.

4.3 Relation to Binary Embeddings

The robust recovery result in Theorem 2 relies on a beautiful geometric result due to Plan and Vershynin [55]. They showed that if $T \subset S^{n-1}$, $m \gtrsim \rho^{-6}w^2(T)$, and $A \in \mathbb{R}^{m\times n}$ is a standard Gaussian matrix then, with probability at least $1 - 2e^{-cm\rho^2}$, for all $x, y \in T$,

$$d_{S^{n-1}}(x, y) - \rho \le \frac{1}{m} d_H(\operatorname{sign}(Ax), \operatorname{sign}(Ay)) \le d_{S^{n-1}}(x, y) + \rho. \tag{16}$$

In other words, if x and y are "separated enough", then the fraction of the random Gaussian hyperplanes $H_{a_i} = \{z \in \mathbb{R}^n : \langle a_i, z\rangle = 0\}$ that separate x and y approximates their geodesic distance in a very sharp way. It was later shown in [52] that (16) remains true if $m \gtrsim \rho^{-4}w^2(T)$. Moreover, for certain "simple" sets (e.g., if T is the set of unit norm sparse vectors) it is known that $m \gtrsim \rho^{-2}w^2(T)$ suffices for (16) (see [42, 52, 55] for examples).

In a similar way, the reconstruction results in Theorems 5 and 6 are connected to "isomorphic" versions of (16). To give a concrete example from [20], suppose that A has i.i.d. symmetric, isotropic, L-subgaussian rows and that the entries of τ are i.i.d. uniformly distributed on $[-\lambda, \lambda]$. If $T \subset RB_2^n$, $\lambda = c_0 R$ and

$$m \ge c_1 \frac{R \log(eR/\rho)}{\rho^3} w^2(T),$$

then with probability at least $1 - 8\exp(-c_2 m\rho/R)$, for any $x, y \in \operatorname{conv}(T)$ such that $\|x - y\|_2 \ge \rho$, one has

$$c_3 \frac{\|x - y\|_2}{R} \le \frac{1}{m} d_H(\operatorname{sign}(Ax + \tau), \operatorname{sign}(Ay + \tau)) \le c_4 \sqrt{\log(eR/\rho)} \cdot \frac{\|x - y\|_2}{R}, \tag{17}$$

where $c_0, \dots, c_4$ depend only on L. Hence, if x and y are separated enough, then the fraction of the hyperplanes $H_{a_i,\tau_i} = \{v \in \mathbb{R}^n : \langle a_i, v\rangle + \tau_i = 0\}$ that separate x and y accurately approximates their Euclidean distance.

4.4 Exponential Error Decay Via Adaptive Thresholds

Let us now briefly discuss how one can achieve optimal, exponential error decay in terms of the number of measurements by using adaptive thresholds, following the idea put forward in [4]. Interestingly, this scheme completely integrates the analog measurement, quantization, and reconstruction procedures. Our presentation follows [22].

To sketch the high-level idea, recall that in memoryless one-bit compressed sensing, by taking measurements we geometrically produce hyperplanes through the origin (if $\tau = 0$) or shifted versions thereof ($\tau \neq 0$). In both cases, the origin is our "reference point" for producing hyperplanes. Intuitively, this is a good strategy to locate x if x happens to lie close to the origin, but relatively ineffective if x is far away. This is reflected by the appearance of the radius R of the signal set in the reconstruction results discussed in Sect. 4.2. To improve the error decay, we can proceed as follows: we first take a small batch of memoryless quantized measurements and run a reconstruction algorithm to obtain a rough estimate $\hat{x}$ of the location of x. In the next round, we use $\hat{x}$ as a new reference point to produce hyperplanes. Continuing in this fashion, we "move in" on the target signal x and are able to produce more informative measurements in each round.

Formally, fix a closed signal set $K \subset \mathbb{R}^n$ with $0 \in K$ and let P_K be the ℓ_2-projection onto this set. We set $I_i = \{(i-1)m/B + 1, \ldots, im/B\}$ and divide the measurement matrix A into the submatrices $A_{(i)} = R_{I_i}A$, $1 \leq i \leq B$, each containing m/B consecutive rows of A. We let $\nu_{(i)} = R_{I_i}\nu$, $\tau_{(i)} = R_{I_i}\tau$ and $R_i = 2^{-i}R$. Suppose that we have an algorithm $\mathscr{A}_i$ which, with probability at least $1-\eta$, satisfies the following for any $w \in K - K$ with $\|w\|_2 \leq R_{i-1}$: based on the input $(A_{(i)}, \tau_{(i)}, (q_c)_{(i)}, R_{i-1})$, with $(q_c)_{(i)} \in \{-1, 1\}^{m/B}$ satisfying for a certain $\bar{\tau}_{(i)} = \bar{\tau}_{(i)}(A_{(i)}, \tau_{(i)}, R_{i-1}) \in \mathbb{R}^{m/B}$

$$d_H((q_c)_{(i)}, \operatorname{sign}(A_{(i)}w + \nu_{(i)} + \bar{\tau}_{(i)})) \leq \beta m/B, \tag{18}$$

$\mathscr{A}_i$ produces a $w^\# \in \mathbb{R}^n$ so that $\|w - w^\#\|_2 \leq R_{i-1}/4$. We can then produce partial reconstructions $(\bar{x}_{(i)})_{i=1}^B$ of x iteratively as follows. Suppose that we produced an $\bar{x}_{(i-1)} \in K$ satisfying $\|x - \bar{x}_{(i-1)}\|_2 \leq R_{i-1}$. We acquire corrupted measurements $(q_c)_{(i)}$ satisfying (18) for $w = x - \bar{x}_{(i-1)}$. Since

$$\operatorname{sign}(A_{(i)}w + \nu_{(i)} + \bar{\tau}_{(i)}) = \operatorname{sign}(A_{(i)}x + \nu_{(i)} + \mu_{(i)} + \bar{\tau}_{(i)}),$$

with $\mu_{(i)} = -A_{(i)}\bar{x}_{(i-1)}$, the desired $(q_c)_{(i)}$ can be acquired by measuring x with $A_{(i)}$ and using $Q_{(\mu_{(i)}+\bar{\tau}_{(i)})}$ as a quantizer.

We now input $(A_{(i)}, \tau_{(i)}, (q_c)_{(i)}, R_{i-1})$ into the algorithm $\mathscr{A}_i$ and let $x^\#_{(i)}$ be its output. Define $\bar{x}_{(i)} = P_K(\bar{x}_{(i-1)} + x^\#_{(i)})$. Clearly, since $x \in K$,

$$\begin{aligned}\|x - \bar{x}_{(i)}\|_2 &\leq \|x - \bar{x}_{(i-1)} - x^\#_{(i)}\|_2 + \|\bar{x}_{(i-1)} + x^\#_{(i)} - P_K(\bar{x}_{(i-1)} + x^\#_{(i)})\|_2 \\ &\leq 2\|x - \bar{x}_{(i-1)} - x^\#_{(i)}\|_2 \leq 2\frac{R_{i-1}}{4} = R_i.\end{aligned}$$

Hence, if $\|x\|_2 \leq R$ and we set $\bar{x}_{(0)} = 0$, then by induction we find $\|x - \bar{x}_{(i)}\|_2 \leq R2^{-i}$ for all $1 \leq i \leq B$. In summary, if we set $B = \log_2(R/\rho)$ then, with probability at least $1 - B\eta$, $\|x - \bar{x}_{(B)}\|_2 \leq \rho$ for any $x \in K$.

In the original paper [4], recovery results with exponential error decay were obtained via the above scheme for s-sparse vectors and standard Gaussian measurement matrices using either hard thresholding operations or Gaussian dithering and the second-order cone program (10) to produce partial reconstructions. In [28], these results were extended to recovery of low-rank matrices, using either hard thresholding or a semidefinite program. In [21, 22], an exponential decay scheme was derived for sparse vectors and randomly subsampled subgaussian circulant matrices using uniform dithering and hard thresholding for partial reconstruction.

As a variation of the result in [21, 22], we will derive a general result valid for any signal set K which is a closed cone, any $A \in \mathbb{R}^{m\times n}$ with i.i.d. symmetric, isotropic, L-subgaussian rows, and uniform dithering. We only need to specify the "base algorithms" $\mathscr{A}_i$. We consider a $w \in (K-K)\cap R_{i-1}B_2^n$ and acquire measurements $(q_c)_{(i)}$ satisfying

$$d_H((q_c)_{(i)}, \operatorname{sign}(A_{(i)}w + \nu_{(i)} + \bar{\tau}_{(i)})) \leq \beta m/B$$

with $A_{(i)} = R_{I_i}A$, $\nu_{(i)} = R_{I_i}\nu$, $\tau_{(i)} = R_{I_i}\tau$, and $\bar{\tau}_{(i)} = R_{i-1}\tau_{(i)}$, where τ has i.i.d. entries which are uniformly distributed on $[-\lambda, \lambda]$. Clearly,

$$\begin{aligned} &d_H((q_c)_{(i)}, \operatorname{sign}(A_{(i)}w + \nu_{(i)} + \bar{\tau}_{(i)})) \\ &\quad = d_H((q_c)_{(i)}, \operatorname{sign}(A_{(i)}(w/R_{i-1}) + \nu_{(i)}/R_{i-1} + \tau_{(i)})). \end{aligned}$$

Define $\tilde{w} = P_{(K-K)\cap B_2^n}(\frac{\lambda}{m}A_{(i)}^*(q_c)_{(i)})$. Since K is a cone, $w/R_{i-1} \in (K-K)\cap B_2^n$. Hence, Theorem 6 (applied with $T = (K-K)\cap B_2^n$, $\rho = 1/4$, and $R = 1$) shows that if we assume that ν contains i.i.d. mean-zero, L-subgaussian entries with variance $\sigma^2 \leq \rho^2 \leq R_{i-1}^2$ and set

$$m/B \geq c_1 w^2((K-K)\cap B_2^n), \qquad \lambda = c_2, \qquad \beta\sqrt{\log(e/\beta)} = c_3,$$

then, with probability at least $1 - 8\exp(-c_4 m/B)$, for all $w \in (K-K)\cap R_{i-1}B_2^n$ the vector $\tilde{w}$ satisfies $\|\frac{w}{R_{i-1}} - \tilde{w}\|_2 \leq 1/4$. Hence, the vector $w^\# = R_{i-1}\tilde{w}$ has the desired properties.

Our considerations lead to the following algorithm and result.

Algorithm 1: exponentially decaying scheme

```
**Input**: $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{N}$, $\tau \in \mathbb{R}^m$, $R > 0$
**Initialization:** $\bar{x}_{(0)} = 0$.
**for** $i=1,\ldots,B$ **do**
    $A_{(i)} = R_{I_i}A$
    $\mu_{(i)} = -A_{(i)}\bar{x}_{(i-1)}$
    $\nu_{(i)} = R_{I_i}\nu$
    $\tau_{(i)} = \mu_{(i)} + R2^{-(i-1)}R_{I_i}\tau$
    Produce corrupted quantized measurements $(q_c)_{(i)} \in \{-1,1\}^{m/B}$ with
        $d_H((q_c)_{(i)}, \operatorname{sign}(A_{(i)}x + \nu_{(i)} + \tau_{(i)})) \leq \beta m/B$
    $x_{(i)}^\# = R2^{-(i-1)}P_{(K-K)\cap B_2^n}\left(\frac{\lambda}{m}A_{(i)}^*(q_c)_{(i)}\right)$
    $\bar{x}_{(i)} = P_K(\bar{x}_{(i-1)} + x_{(i)}^\#)$
**end**
**Output**: $x^\# = \bar{x}_{(B)}$
```

Theorem 7 *There exist constants c_1, c_2, c_3, c_4 depending only on L such that the following holds. Let $K \subset \mathbb{R}^n$ be a closed cone, fix $0 < \rho \leq 1$ and $R > 0$, set $B = \log_2(R/\rho)$, $\lambda = c_1$, $m \geq c_2 B w^2((K - K) \cap B_2^n)$, $\beta = c_3$. Suppose that A has i.i.d. symmetric, isotropic, L-subgaussian rows, ν has i.i.d. mean-zero, L-subgaussian entries with variance $\sigma \leq \rho$, τ has i.i.d. entries which are uniformly distributed on $[-\lambda, \lambda]$, and A, ν, τ are independent. Then with probability at least $1 - Be^{-c_4 m/B}$ the following holds: for any $x \in K$ with $\|x\|_2 \leq R$, the output $x^\#$ of Algorithm 1 satisfies $\|x - x^\#\|_2 \leq \rho$.*

The decay of the reconstruction error in Theorem 7 is clearly superior to the error decay in Theorem 6. The total number of measurements generated in Algorithm 1 is

$$m \sim \log(R/\rho) w^2((K - K) \cap B_2^n),$$

so the reconstruction error decays exponentially in terms of the number of measurements, which is optimal (see the discussion in Sect. 2). In addition, the total number of adversarial bit corruptions is βm, a constant fraction of m.

The price to pay for this superior scheme is more complicated hardware and higher energy consumption in operation. The quantizer needs to be equipped with memory and the capability to compute and set new thresholds in each round.

5 Memoryless Multi-bit Compressed Sensing

Let us now consider memoryless multi-bit quantization schemes. A *memoryless scalar quantizer* is defined by fixing a quantization alphabet $\mathcal{Q} \subset \mathbb{R}$ and setting, for a given $z \in \mathbb{R}^m$ and $i \in [m]$,

$$Q_{\mathrm{MSC}}(z)_i = \min\{\mathrm{argmin}_{t \in \mathcal{Q}} |z_i - t|\}.$$

For example, by taking the alphabet $\mathcal{Q} = \{-1, 1\}$ we find the one-bit quantizer with zero thresholds studied in Sect. 4.1. Before discussing specific recovery algorithms, let us first point out that the best reconstruction error decay in terms of the number of measurements that *any* reconstruction algorithm can achieve when receiving memoryless scalar quantized measurements as input is, in general, *linear*. Specifically, it was shown in [6, 30] that if $A \in \mathbb{R}^{m \times n}$ and $E \subset \mathbb{R}^n$ is a k-dimensional subspace, then $\sup_{x \in E} \|x - \mathscr{A}(Q_{\mathrm{MSC}}(Ax))\|_2 \geq c\frac{k}{m}$ for any reconstruction map $\mathscr{A} : \mathbb{R}^m \to \mathbb{R}^n$.

The most studied memoryless multi-bit compressed sensing model involves the memoryless scalar quantizer with alphabet $\mathcal{Q} = \delta\mathbb{Z}$, i.e., the quantizer $Q_\delta : \mathbb{R}^m \to (\delta\mathbb{Z})^m$ defined by

$$Q_\delta(z) = \left(\delta \lfloor z_i/\delta \rfloor\right)_{i=1}^m.$$

For brevity, we will call this map the *uniform scalar quantizer*. Geometrically, Q_δ divides $\mathbb{R}^m$ into half-open cubes with side lengths equal to δ and maps any vector

$z \in \mathbb{R}^m$ to the corner of the cube in which it is located. From a practical point of view, this quantizer is somewhat idealized: in a realistic implementation the range of the quantizer is limited and measurements $\langle a_i, x\rangle$ which exceed the quantizer's range incur a potentially unbounded quantization error. One calls such measurements *saturated*. The work [46] analyzes some strategies to deal with saturated measurements. We will restrict ourselves to the idealized uniform scalar quantizer.

Let us first consider the "agnostic" approach to reconstruct x from uniformly scalar quantized measurements $q = Q_\delta(Ax)$, i.e., we simply treat the error due to quantization as additive noise. Note that the ℓ_∞-distance of Ax to the center of its quantization cell, i.e., $q + (\delta/2)\mathbb{1}$ where $\mathbb{1} \in \mathbb{R}^m$ is the vector which has all entries equal to 1, is at most $\delta/2$. Hence, we can reconstruct the signal x via the linear program

$$\min_{z\in\mathbb{R}^n} \|z\|_1 \qquad \text{s.t.} \qquad \|Az - (q + (\delta/2)\mathbb{1})\|_\infty \leq \delta/2. \tag{19}$$

Note that this method is very close to minimizing the ℓ_1-norm under a quantization consistency constraint, i.e., to solving

$$\min \|z\|_1 \qquad \text{s.t.} \qquad y = Q_\delta(Az). \tag{20}$$

Indeed, whereas z is feasible for (20) if and only if Az lies in the same quantization cell as Ax, z is feasible for (19) precisely when it lies in the closure of that cell.

From the standard theory of compressed sensing, it is easy to extract (see [18, Theorem A.1]) that if $A \in \mathbb{R}^{m\times n}$ is such that $\frac{1}{\sqrt{m}}A$ satisfies $\text{RIP}_{2,2}(\Sigma_s, c)$ with constant $c < 4/\sqrt{41}$, then for any $x \in \mathbb{R}^n$ and $y = Q_\delta(Ax)$ any solution $x^\#$ to (19) satisfies

$$\|x - x^\#\|_2 \lesssim \delta + s^{-1/2} \inf_{z\in\Sigma_s} \|x - z\|_1. \tag{21}$$

In particular, this applies to partial subgaussian circulant matrices (with deterministically selected rows) if $m \gtrsim s \log^2 s \log^2 n$ [44] and randomly subsampled discrete bounded orthonormal systems if $m \gtrsim s \log^2 s \log n$ [35]. A different argument, which relies on Mendelson's small ball method [48] instead of an RIP-based analysis, shows that even for a variety of heavy-tailed random matrices the reconstruction guarantee (21) holds in the optimal regime $m \gtrsim s \log(en/s)$ (see [19, Section V] for several results).

Although these results exhibit the same dependence of m on s and n as in "unquantized" compressed sensing, they have a clear downside: by treating the quantization error as noise, the reconstruction error does not decay beyond the resolution δ of the quantizer, which corresponds to the noise floor. Intuitively, one could hope to be able to decrease the reconstruction error even beyond the resolution δ by taking more measurements. In a series of works by L. Jacques and co-authors [37, 39, 40, 63], this is shown to be possible if one introduces appropriate dithering at the quantizer. Let us denote by $Q_{\delta,\tau} = Q_\delta(\cdot + \tau)$ the uniform scalar quantizer with dithering vector $\tau \in \mathbb{R}^m$. It was first observed in [37] (see also [63, Appendix A]) that if the entries

τ_i of τ are i.i.d. uniformly distributed on $[0, \delta]$, then for any $y \in \mathbb{R}^m$, $\mathbb{E}Q_{\delta,\tau}(y) = y$. Hence, at least in expectation, dithering that matches the resolution can "cancel out" the error caused by the uniform scalar quantizer. This fact can be exploited to prove recovery results for general signals sets and a large class of measurement matrices. We start by describing a result from [63]. Let $T \subset \mathbb{R}^n$ be a closed set of signals. For $x \in T$ consider its quantized measurements $q = Q_{\delta,\tau}(Ax)$ and define

$$x^{\#}_{\mathrm{PBP}} = P_T\Big(\frac{1}{m}A^*q\Big).$$

Since A^*q is usually called the back projection of q, this reconstruction is coined the *projected back projection* in [63]. If $T = \Sigma_s$, then the projected back projection is up to scaling the same as the hard thresholding map in Theorem 3. To give the flavor of the recovery results in [63], we state a recovery result if T is a union of subspaces. Further results are obtained for low-rank matrices and star-shaped convex sets.

Theorem 8 ([63]) *Let $T = \cup_{i=1}^N T_i \subset \mathbb{R}^n$ be a union of subspaces. Suppose that the entries of τ are i.i.d. uniformly distributed on $[0, \delta]$. Let $A \in \mathbb{R}^{m\times n}$ be a random matrix that, for any fixed $0 < \varepsilon < 1$, satisfies*

$$\Big|\frac{1}{m}\|Az\|_2^2 - \|z\|_2^2\Big| \leq \varepsilon, \qquad \textit{for all } z \in T \cap B_2^n$$

with probability at least $1 - \eta$ if

$$m \gtrsim \varepsilon^{-2} w^2(T \cap B_2^n)\,\mathrm{polylog}(m, n, 1/\eta).$$

Let $T^{(4)} = \sum_{i=1}^4 T$. If $m \gtrsim \rho^{-2}(1+\delta)^2 w^2(T^{(4)} \cap B_2^n)\,\mathrm{polylog}(m, n, \delta, 1/\rho, 1/\eta)$, then with probability at least $1-\eta$, for any $x \in T \cap B_2^n$ the projected back projection $x^{\#}_{PBP}$ satisfies $\|x - x^{\#}\|_2 \leq \rho$.

In the special case $T = \Sigma_s$, the assumption of Theorem 8 is e.g. satisfied if A is subgaussian, a partial subgaussian circulant matrix or a randomly subsampled discrete bounded orthonormal system. Hence, for these matrices, one can uniformly recover all s-sparse vectors from their projected back projections if $m \gtrsim \rho^{-2}(1+\delta)^2 s \log(en/s)\,\mathrm{polylog}(m, n, \delta, 1/\rho, 1/\eta)$.

The reconstruction error in Theorem 8 does not decrease to zero as the bin width δ goes to zero, as e.g. in (21). In fact, this cannot be expected as $x^{\#}_{\mathrm{PBP}}$ will, loosely speaking, start behaving as $H_s(\frac{1}{m}A^*Ax)$ as $\delta \to 0$, i.e., as the first step of the iterative hard thresholding algorithm in "unquantized" compressed sensing. Therefore, it is of interest to derive a "best of both worlds" result that exhibits both a decaying reconstruction error in terms of the number of measurements and, at the same time, a reconstruction error decaying to zero if $\delta \to 0$ once m exceeds the threshold of $Cs\log(en/s)$ measurements, which are needed for uniform recovery from unquantized measurements. One can get very close to such a result by using a relation between uniform scalar quantization and so-called *quantized Johnson-Lindenstrauss* *embeddings*. This relation is analogous to the connection between one-bit compressed sensing and binary embeddings sketched in Sect. 4.3. For concreteness, we consider the following embedding result.

Theorem 9 ([40, Proposition 1]) *If $m \gtrsim \varepsilon^{-2}\log N(T,\delta\varepsilon^2)$ and $\frac{1}{m}A \in \mathbb{R}^{m\times n}$ satisfies RIP$_{1,2}(T-T,\theta)$, then for certain absolute constants $c, C > 0$, with probability at least $1-Ce^{-cm\varepsilon^2}$ the map $f(x) = Q_{\delta,\tau}(Ax)$ satisfies*

$$(1-\theta)\|x-y\|_2 - c\delta\varepsilon \le \frac{1}{m}\|f(x)-f(y)\|_1 \le (1+\theta)\|x-y\|_2 + c\delta\varepsilon \qquad (22)$$

for all $x, y \in T$.

By the lower bound in (22), for any given signal $x \in T$, any $x^{\#} \in T$ that is quantization consistent with x satisfies $\|x-x^{\#}\|_2 \le c\delta\varepsilon/(1-\theta)$. Thus, under the conditions of Theorem 9 we can recover x via a program that finds a quantization consistent vector in T. In particular, if $T = \Sigma_s \cap B_2^n$ then we can use the non-convex program

$$\min \|z\|_0 \qquad \text{s.t.} \qquad q = Q_{\delta,\tau}(Az), \qquad \|z\|_2 \le 1. \qquad (23)$$

If B is standard Gaussian and $A = \sqrt{\frac{\pi}{2}}B$, then $\frac{1}{m}A$ satisfies RIP$_{1,2}(\Sigma_{2s},\theta)$ with probability at least $1-2e^{-cm\theta^2}$ if $m \gtrsim \theta^{-2}s\log(en/s)$. Combining this fact with Theorem 9 and the estimate $\log N(\Sigma_s \cap B_2^n, \delta\varepsilon^2) \lesssim s\log(en/(s\delta\varepsilon^2))$, we find that if $m \gtrsim \varepsilon^{-2}s\log(en/(s\delta\varepsilon^2))$, then with probability at least $1-Ce^{-cm\varepsilon^2}$, for any $x \in \Sigma_s \cap B_2^n$, any solution $x^{\#}$ to (23) satisfies $\|x-x^{\#}\|_2 \le \delta\varepsilon$.

This result can still be improved, since to derive a recovery result it suffices to prove a much weaker property than (22). In [38, 39] a direct analysis was made of the required property

$$Q_{\delta,\tau}(Az) = Q_{\delta,\tau}(Ax) \Rightarrow \|x-z\|_2 \le \theta, \qquad \text{for all } x, z \in T. \qquad (24)$$

If (24) holds for $T = \Sigma_s \cap B_2^n$ and $\theta = \delta\varepsilon$, then for any $x \in \Sigma_s \cap B_2^n$ any solution $x^{\#}$ to (23) satisfies $\|x^{\#}-x\|_2 \le \delta\varepsilon$. It was shown in [38, Theorem 2] that a standard Gaussian matrix $A \in \mathbb{R}^{m\times n}$ satisfies this property with high probability if $m \gtrsim \varepsilon^{-1}s\log(en/(\sqrt{s}\delta\varepsilon))$. Since for a fixed δ the reconstruction error cannot decay faster than linear in m, the dependence of m on ε is near-optimal in this result.

We refer to [39, 40] for further results on quantized Johnson-Lindenstrauss embeddings, in particular versions involving RIP$_{2,2}$-matrices and subgaussian matrices, and to [38, 39] for further results concerning the property (24). The latter results are used in [51] to derive reconstruction guarantees for generalizations of (23) in which $\|z\|_0$ is replaced by an atomic norm.

In [18], Theorem 9 was used to prove a uniform recovery result for effectively s-sparse vectors in the unit ball from randomly subsampled Gaussian circulant measurements (with rows selected according to the selector model) via a convex program that enforces quantization consistency. Loosely speaking, [18, Theorem 6.2] shows that with high probability one can achieve a reconstruction error $\varepsilon\delta^{2/3}$ using roughly

$m \sim \varepsilon^{-6} s \log(en/s)$ measurements, provided that a small sparsity condition is satisfied. Interestingly, this result uses a combination of Gaussian and uniform dithering in the quantizer.

6 Noise-Shaping Methods

Finally, we discuss quantized compressed sensing with a family of adaptive quantization methods called *noise-shaping methods*. The most prominent example in this family are $\Sigma\Delta$-quantization methods, which are very popular in practice. Noise-shaping quantizers were first studied mathematically in the context of analog-to-digital conversion of bandlimited functions (see e.g., [14, 33]) and afterwards have been successfully extended to the frameworks of finite frames and compressed sensing (see e.g., the survey [13] and the references therein). In the setting of compressed sensing, the first reconstruction results for exactly sparse signals were obtained via a two-stage approach [25, 34, 45]. First, one estimates only the support of the original sparse signal via a traditional compressed sensing method for noisy measurements. Once the support is known, one can use reconstruction methods developed in the framework of finite frames to fully reconstruct the signal, e.g., by using an appropriate Sobolev dual frame. For the sake of brevity, we will not discuss this approach and refer to the survey [13] for details. We will only discuss a recent one-stage recovery approach via a convex program, which was developed in [10, 13, 26, 36, 60]. In contrast to the two-stage approach sketched above, this method is proven to be stable with respect to approximate sparsity, robust with respect to (a small amount of) pre-quantization noise and has been successfully applied to structured random measurement matrices [26, 36].

A *noise-shaping quantizer* $Q : \mathbb{R}^m \to \mathcal{Q}^m$ associated with a *noise transfer operator* H, is defined so that for each $y \in \mathbb{R}^m$ the quantization $q = Q(y)$ satisfies the *noise-shaping relation*

$$y - q = Hu \tag{25}$$

where $u = u(y, Q) \in \mathbb{R}^m$ is an auxiliary vector called the *internal state vector*. The matrix $H \in \mathbb{R}^{m \times m}$ is chosen to be a lower triangular Toeplitz matrix with unit diagonal, so that the quantization scheme can be implemented via a recursion. The noise-shaping quantizer is called *stable* if, for all $y \in \mathbb{R}^m$ with $\|y\|_\infty \le \mu$, $\|u\|_\infty \le C_{Q,\mu}$, where $C_{Q,\mu}$ is a constant independent of m called the *stability constant*. The most important examples of noise-shaping quantizers are *$\Sigma\Delta$-quantizers*, which compute a solution to the noise-shaping relation (25) for $H = D^r$, where $D \in \mathbb{R}^m$ is the first-order difference matrix defined by

$$D_{ij} = \begin{cases} 1 & \text{if } i = j \\ -1 & \text{if } i = j + 1 \\ 0 & \text{else.} \end{cases}$$

We call r the *order* of the scheme. The construction of a stable r-th order $\Sigma\Delta$-scheme is non-trivial. It was shown in [16] that for any $L \in \mathbb{N}$ and $\delta > 0$ there exists a stable r-th order $\Sigma\Delta$-scheme with a fixed alphabet $\mathcal{Q}_{\delta,L} = \{\pm(2\ell - 1)\delta \ : \ 1 \leq \ell \leq L\}$ and constant

$$C_{\mathcal{Q},\mu} \leq C\delta \left(\frac{er}{\pi} \left\lceil \frac{\pi^2}{(\cosh^{-1}(2L - \frac{\mu}{\delta}))^2} \right\rceil \right)^r .$$

In particular, taking $L = 1$, $\delta = 1$, we find an r-th order scheme with the one-bit alphabet $\mathcal{Q} = \{-1, 1\}$ which is stable in the sense that $\|u\|_\infty \leq Cc_\mu^r r^r$ whenever $\|y\|_\infty \leq \mu < 1$.

Let us now turn to the compressed sensing scenario, where $y = Ax$ and the noise-shaping relation is

$$Ax - q = Hu.$$

To see how we could recover x, multiply both sides by a designed preconditioning matrix $V \in \mathbb{R}^{p \times m}$ to obtain

$$VAx - Vq = VHu.$$

Since we observe Vq, we can interpret this equation as a linear measurement equation $Vq = VAx + e$, where VA is the measurement matrix and $e = -VHu$ is the noise on the measurements. To recover x, we can then use methods for recovery from "unquantized" noisy measurements. For instance, we can use basis pursuit denoising

$$\min_{z \in \mathbb{R}^n} \|z\|_1 \qquad \text{s.t.} \qquad \|VAz - Vq\|_2 \leq \eta. \tag{26}$$

By a standard result in compressed sensing, one can recover any s-sparse x via (26) if VA satisfies $\mathrm{RIP}_{2,2}(\Sigma_s, c)$ for c a small enough absolute constant and $\|e\|_2 \leq \eta$ (see e.g., [29, Chapter 9]). To satisfy the latter condition, if we assume that the quantization scheme is stable and $\|Ax\|_\infty \leq \mu$, it suffices to ensure that $\|VH\|_{\ell_\infty \to \ell_2}$ is small.

In the presence of pre-quantization noise, the noise-shaping relation changes to

$$V(Ax + \nu) - Vq = VHu.$$

It was suggested in [60] to replace the program (26) by

$$\min_{(z,w) \in \mathbb{R}^{n+m}} \|z\|_1 \qquad \text{s.t.} \qquad \|V(Az + w) - Vq\|_2 \leq \eta, \ \|w\|_2 \leq \kappa. \tag{27}$$

The following result summarizes two reconstruction results for subgaussian [60] and randomly subsampled subgaussian circulant matrices [26].

Theorem 10 ([60, Theorem 9]) and [26, Theorem 5]) *Let Q be the stable r-th order $\Sigma\Delta$-scheme with the one-bit alphabet $\mathcal{Q} = \{-1, 1\}$ as above and let $C_{Q,\mu}$ be its stability constant. Let $A \in \mathbb{R}^{m\times n}$ be a subgaussian matrix. Suppose that*

$$m \geq p \geq Cs\log(en/s).$$

Then the following holds with probability at least $1 - e^{-cp}$. For any $x \in \mathbb{R}^n$ satisfying $\|Ax\|_\infty \leq \mu < 1$ and $q = Q(Ax + \nu)$ with $\|\nu\|_\infty \leq \varepsilon < 1 - \mu$, any solution $x^\#$ to (26) with $V = D^{-r}$, $\eta = C_{Q,\mu}\sqrt{m}$, $\kappa = \varepsilon\sqrt{m}$ satisfies

$$\|x^\# - x\|_2 \lesssim_{\mu,r} \left(\frac{p}{m}\right)^{r-\frac{1}{2}} + \frac{\sigma_s(x)_1}{\sqrt{s}} + \sqrt{\frac{m}{p}}\varepsilon, \tag{28}$$

where $\sigma_s(x)_1 = \min_{z\in\Sigma_s} \|x - z\|_1$.

If A is a randomly subsampled subgaussian circulant matrix (with rows selected according to the uniformly at random model), then the same result holds with probability at least $1 - e^{-t}$ provided that, for some $0 \leq \alpha < 1/2$,

$$m \gtrsim t^{1/(1-2\alpha)} s \log^{2/(1-2\alpha)}(s) \log^{2/(1-2\alpha)}(n)$$

and $p = m(\frac{s}{m})^\alpha$.

The result in Theorem 10 essentially relies on proving that the matrix $D^{-r}A$ satisfies $\mathrm{RIP}_{2,2}(\Sigma_s, c)$, which has proven to be difficult for structured random matrices. To overcome this problem, [36] constructed a different preconditioner V for $\Sigma\Delta$-schemes as follows. For $p < m$ let $\lambda = m/p$. For simplicity, we assume that $\lambda \in \mathbb{N}$ and that there is a $\tilde{\lambda} \in \mathbb{N}$ such that $\lambda = r\tilde{\lambda} - r + 1$. Suppose that $u \in \mathbb{R}^\lambda$ contains the coefficients of the polynomial $(1 + z + \ldots + z^{\tilde{\lambda}-1})^r$. Define $V \in \mathbb{R}^{p\times m}$ by

$$V_{\Sigma\Delta} = \frac{1}{\sqrt{p}\|u\|_2} I_p \otimes u^T = \frac{1}{\sqrt{p}\|u\|_2} \begin{bmatrix} u^T & 0 & \cdots & 0 \\ 0 & u^T & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & u^T \end{bmatrix}. \tag{29}$$

Using this construction, [36] obtained the following result for partial Bernoulli circulant matrices with randomized row signs. It can be easily modified in the case of pre-quantization noise to produce an error bound similar to (28). In addition, a similar result was obtained for randomly subsampled discrete bounded orthonormal systems (again with randomized row signs).

Theorem 11 ([36, Theorem 6.1]) *Let Q be the stable r-th order $\Sigma\Delta$-scheme with the one-bit alphabet $\mathcal{Q} = \{-1, 1\}$ as above. Let B be a partial Bernoulli circulant matrix (with rows selected according to the row picking model), let D_ξ be a diagonal* *matrix with i.i.d. symmetric Bernoulli random variables on its diagonal which are independent of* B *and let* $A = D_\xi B$. *Fix* $\theta > 0$, $s \in [n]$ *and suppose that*

$$m \geq p \geq Cs \log^4 n.$$

Then the following holds with probability at least $1 - e^{-cp^2/(sm)}$. *For any* $x \in \mathbb{R}^n$ *satisfying* $\|Ax\|_\infty \leq \mu < 1$ *and* $q = Q(Ax)$, *any solution to* (26) *with* $V = V_{\Sigma\Delta}$ *satisfies*

$$\|x^\# - x\|_2 \lesssim_{\mu,r} \left(\frac{p}{m}\right)^{r-\frac{1}{2}} + \frac{\sigma_s(x)_1}{\sqrt{s}}.$$

The reconstruction error in Theorems 10 and 11 decays polynomially in terms of the number of measurements. If x is s-sparse ($\sigma_s(x)_1 = 0$) and there is no prequantization noise ($\varepsilon = 0$), then one can optimize the bound (28) (including the implicit constant depending on r) in terms of r. This yields an r depending on s and m for which the reconstructions error decays root-exponentially, i.e., as $e^{-\sqrt{m}}$, in terms of the number of measurements (see e.g., [60, Corollary 11]). Exponential error decay can be achieved by using a different noise-shaping method, called distributed noise-shaping quantization [11, 12]. For such recovery results with partial Bernoulli circulant matrices and randomly subsampled discrete bounded orthonormal systems (both with randomized row signs), see [36].

Acknowledgements It is a pleasure to thank the anonymous reviewer, Rayan Saab, and especially Laurent Jacques for many comments that improved this book chapter. This work was supported by the DFG through the project Quantized Compressive Spectrum Sensing (QuaCoSS), which is part of the Priority Program SPP 1798 Compressive Sensing in Information Processing (COSIP).

References

1. A. Ai, A. Lapanowski, Y. Plan, R. Vershynin, One-bit compressed sensing with non-Gaussian measurements. Linear Algebr. Appl. **441**, 222–239 (2014)
2. U. Ayaz, S. Dirksen, H. Rauhut, Uniform recovery of fusion frame structured sparse signals. Appl. Comput. Harmon. Anal. **41**(2), 341–361 (2016)
3. R. Baraniuk, S. Foucart, D. Needell, Y. Plan, M. Wootters, One-bit compressive sensing of dictionary-sparse signals. Inf. Inference: A J. IMA **7**(1), 83–104 (2017)
4. R.G. Baraniuk, S. Foucart, D. Needell, Y. Plan, M. Wootters, Exponential decay of reconstruction error from binary measurements of sparse signals. IEEE Trans. Inf. Theory **63**(6), 3368–3385 (2017)
5. P.T. Boufounos, R.G. Baraniuk, 1-bit compressive sensing, in *2008 42nd Annual Conference on Information Sciences and Systems* (IEEE 2008), pp. 16–21
6. P. T. Boufounos, L. Jacques, F. Krahmer, R. Saab, Quantization and compressive sensing, in *Compressed Sensing and its Applications* (Springer, 2015), pp. 193–237

7. J. Bourgain, An improved estimate in the restricted isometry problem, in *Geometric Aspects of Functional Analysis*, ed. B. Klartag, E. Milman, volume 2116 of *Lecture Notes in Mathematics* (Springer International Publishing, 2014), pp. 65–70
8. E.J. Candès, J., T. Tao, J.K. Romberg, Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. IEEE Trans. Inform. Theory **52**(2), 489–509 (2006)
9. E.J. Candès, J.K. Romberg, T. Tao, Stable signal recovery from incomplete and inaccurate measurements. Comm. Pure Appl. Math. **59**(8), 1207–1223 (2006)
10. E. Chou, *Beta-duals of frames and applications to problems in quantization*. PhD thesis, New York University (2013)
11. E. Chou, C.S. Güntürk, Distributed noise-shaping quantization: I. Beta duals of finite frames and near-optimal quantization of random measurements. Constr. Approx. **44**(1), 1–22 (2016)
12. E. Chou, C. S. Güntürk, Distributed noise-shaping quantization: II. Classical frames, in *Excursions in Harmonic Analysis, Volume 5* (Springer, 2017), pp. 179–198
13. E. Chou, C. S. Güntürk, F. Krahmer, R. Saab, Ö. Yılmaz, Noise-shaping quantization methods for frame-based and compressive sampling systems, in *Sampling Theory, a Renaissance* (Springer, 2015), pp. 157–184
14. I. Daubechies, R. DeVore, Approximating a bandlimited function using very coarsely quantized data: A family of stable sigma-delta modulators of arbitrary order. Ann. Math. **158**(2), 679–710 (2003)
15. M.A. Davenport, J. Romberg, An overview of low-rank matrix recovery from incomplete observations. IEEE J. Sel. Top. Signal Process. **10**(4), 608–622 (2016)
16. P. Deift, F. Krahmer, C.S. Güntürk, An optimal family of exponentially accurate one-bit sigma-delta quantization schemes. Commun. Pure Appl. Math. **64**(7), 883–919 (2011)
17. S. Dirksen, Dimensionality reduction with subgaussian matrices: a unified theory. Found. Comput. Math. **16**(5), 1367–1396 (2016)
18. S. Dirksen, H.C. Jung, H. Rauhut, One-bit compressed sensing with Gaussian circulant matrices. arXiv:1710.03287 (2017)
19. S. Dirksen, G. Lecué, H. Rauhut, On the gap between restricted isometry properties and sparse recovery conditions. IEEE Trans. Inform. Theory **64**(8), 5478–5487 (2018)
20. S. Dirksen, S. Mendelson, Non-gaussian hyperplane tessellations and robust one-bit compressed sensing. arXiv:1805.09409
21. S. Dirksen, S. Mendelson, Robust one-bit compressed sensing with partial circulant matrices. arXiv:1812.06719
22. S. Dirksen, S. Mendelson. Unpublished manuscript
23. D.L. Donoho, Compressed sensing. IEEE Trans. Inform. Theory **52**(4), 1289–1306 (2006)
24. A. Eftekhari, M.B. Wakin, New analysis of manifold embeddings and signal recovery from compressive measurements. Appl. Comput. Harmon. Anal. **39**(1), 67–109 (2015)
25. J.-M. Feng, F. Krahmer, An RIP-based approach to $\Sigma\Delta$ quantization for compressed sensing. IEEE Signal Process. Lett. **21**(11), 1351–1355 (2014)
26. J.-M. Feng, F. Krahmer, R. Saab, Quantized compressed sensing for partial random circulant matrices. arXiv:1702.04711 (2017)
27. S. Foucart, *Flavors of Compressive Sensing* (Springer International Publishing, Cham, 2017), pp. 61–104
28. S. Foucart, R. Lynch, Recovering low-rank matrices from binary measurements. Preprint (2018)
29. S. Foucart, H. Rauhut, *A Mathematical Introduction to Compressive Sensing* (Applied and Numerical Harmonic Analysis. Birkhäuser/Springer, New York, 2013)
30. V.K. Goyal, M. Vetterli, N.T. Thao, Quantized overcomplete expansions in $\mathbb{R}^N$ analysis, synthesis, and algorithms. IEEE Trans. Inform. Theory **44**(1), 16–31 (1998)
31. R.M. Gray, D.L. Neuhoff, Quantization. IEEE Trans. Inf. Theory **44**(6), 2325–2383 (1998)
32. R.M. Gray, T.G. Stockham, Dithered quantizers. IEEE Trans. Inf. Theory **39**(3), 805–812 (1993)
33. C.S. Güntürk, One-bit sigma-delta quantization with exponential accuracy. Commun. Pure Appl. Math. **56**(11), 1608–1630 (2003)

34. C.S. Güntürk, M. Lammers, A.M. Powell, R. Saab, Ö. Yılmaz, Sobolev duals for random frames and $\Sigma\Delta$ quantization of compressed sensing measurements. Found. Comput. Math. **13**(1), 1–36 (2013)
35. I. Haviv, O. Regev, The restricted isometry property of subsampled Fourier matrices, in *SODA '16* (Philadelphia, PA, USA, 2016), pp. 288–297
36. T. Huynh, R. Saab, Fast binary embeddings, and quantized compressed sensing with structured matrices. arXiv:1801.08639 (2018)
37. L. Jacques, A quantized Johnson-Lindenstrauss lemma: the finding of Buffon's needle. IEEE Trans. Inf. Theory **61**(9), 5012–5027 (2015)
38. L. Jacques, Error decay of (almost) consistent signal estimations from quantized gaussian random projections. IEEE Trans. Inf. Theory **62**(8), 4696–4709 (2016)
39. L. Jacques, Small width, low distortions: quantized random embeddings of low-complexity sets. IEEE Trans. Inf. Theory **63**(9), 5477–5495 (2017)
40. L. Jacques, V. Cambareri, Time for dithering: fast and quantized random embeddings via the restricted isometry property. Inf. Inference: A J. IMA **6**(4), 441–476 (2017)
41. L. Jacques, K. Degraux, C. De Vleeschouwer, Quantized iterative hard thresholding: Bridging 1-bit and high-resolution quantized compressed sensing. arXiv:1305.1786 (2013)
42. L. Jacques, J.N. Laska, P.T. Boufounos, R.G. Baraniuk, Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors. IEEE Trans. Inform. Theory **59**(4), 2082–2102 (2013)
43. K. Knudson, R. Saab, R. Ward, One-bit compressive sensing with norm estimation. IEEE Trans. Inform. Theory **62**(5), 2748–2758 (2016)
44. F. Krahmer, S. Mendelson, H. Rauhut, Suprema of chaos processes and the restricted isometry property. Comm. Pure Appl. Math. **67**(11), 1877–1904 (2014)
45. F. Krahmer, R. Saab, Ö. Yilmaz, Sigma-delta quantization of sub-gaussian frame expansions and its application to compressed sensing. Inf. Inference **3**(1), 40–58 (2014)
46. J.N. Laska, P.T. Boufounos, M.A. Davenport, R.G. Baraniuk, Democracy in action: quantization, saturation, and compressive sensing. Appl. Comput. Harmonic Anal. **31**(3), 429–443 (2011)
47. M. Lustig, D.L. Donoho, J.M. Santos, J.M. Pauly, Compressed sensing MRI. IEEE Signal Process. Mag. **25**(2), 72–82 (2008)
48. S. Mendelson, Learning without concentration. J. ACM **62**(3), Art. 21, 25 (2015)
49. S. Mendelson, H. Rauhut, R. Ward, Improved bounds for sparse recovery from subsampled random convolutions. Ann. Appl. Probab. **28**(6), 3491–3527 (2018)
50. A. Montanari, N. Sun, Spectral algorithms for tensor completion. Comm. Pure Appl. Math. **71**(11), 2381–2425 (2018)
51. A. Moshtaghpour, L. Jacques, V. Cambareri, K. Degraux, C. De Vleeschouwer, Consistent basis pursuit for signal and matrix estimates in quantized compressed sensing. IEEE Signal Process. Lett. **23**(1), 25–29 (2016)
52. S. Oymak, B. Recht, Near-optimal bounds for binary embeddings of arbitrary sets. arXiv:1512.04433 (2015)
53. Y. Plan, R. Vershynin, One-bit compressed sensing by linear programming. Comm. Pure Appl. Math. **66**(8), 1275–1297 (2013)
54. Y. Plan, R. Vershynin, Robust 1-bit compressed sensing and sparse logistic regression: a convex programming approach. IEEE Trans. Inform. Theory **59**(1), 482–494 (2013)
55. Y. Plan, R. Vershynin, Dimension reduction by random hyperplane tessellations. Discrete Comput. Geom. **51**(2), 438–461 (2014)
56. H. Rauhut, R. Schneider, Z. Stojanac, Low rank tensor recovery via iterative hard thresholding. Linear Algebra Appl. **523**, 220–262 (2017)
57. L. Roberts, Picture coding using pseudo-random noise. IRE Trans. Inf. Theory **8**(2), 145–154 (1962)
58. J. Romberg, Compressive sensing by random convolution. SIAM J. Imaging Sci. **2**(4), 1098–1128 (2009)

59. M. Rudelson, R. Vershynin, On sparse reconstruction from Fourier and Gaussian measurements. Comm. Pure Appl. Math. **61**(8), 1025–1045 (2008)
60. R. Saab, R. Wang, Ö. Yılmaz, Quantization of compressive samples with stable and robust recovery. Appl. Comput. Harmonic Anal. **44**(1), 123–143 (2018)
61. G. Schechtman, Two observations regarding embedding subsets of Euclidean spaces in normed spaces. Adv. Math. **200**(1), 125–135 (2006)
62. R. Vershynin, *High-Dimensional Probability* (Cambridge University Press, 2018)
63. C. Xu, L. Jacques, Quantized compressive sensing with RIP matrices: the benefit of dithering. arXiv:1801.05870 (2018)