

## Quantization and Compressive Sensing

Boufounos, P.T.; Jacques, L.; Krahmer, F.; Saab, R.

TR2015-008 January 2015

### Abstract

Quantization is an essential step in digitizing signals, and, therefore, an indispensable component of any modern acquisition system. This chapter explores the interaction of quantization and compressive sensing and examines practical quantization strategies for compressive acquisition systems. Specifically, we first provide a brief overview of quantization and examine fundamental performance bounds applicable to any quantization approach. Next, we consider several forms of scalar quantizers, namely uniform, non-uniform, and 1-bit. We provide performance bounds and fundamental analysis, as well as practical quantizer designs and reconstruction algorithms that account for quantization. Furthermore, we provide an overview of Sigma-Delta ( $\Sigma\Delta$ ) quantization in the compressed sensing context, and also discuss implementation issues, recovery algorithms and performance bounds. As we demonstrate, proper accounting for quantization and careful quantizer design has significant impact in the performance of a compressive acquisition system.

*Springer/Birkhaeuser Book, Compressed Sensing and its Applications*

This work may not be copied or reproduced in whole or in part for any commercial purpose. Permission to copy in whole or in part without payment of fee is granted for nonprofit educational and research purposes provided that all such whole or partial copies include the following: a notice that such copying is by permission of Mitsubishi Electric Research Laboratories, Inc.; an acknowledgment of the authors and individual contributions to the work; and all applicable portions of the copyright notice. Copying, reproduction, or republishing for any other purpose shall require a license with payment of fee to Mitsubishi Electric Research Laboratories, Inc. All rights reserved.



# Quantization and Compressive Sensing

Petros T. Boufounos, Laurent Jacques, Felix Krahmer, and Rayan Saab

**Abstract** Quantization is an essential step in digitizing signals, and, therefore, an indispensable component of any modern acquisition system. This chapter explores the interaction of quantization and compressive sensing and examines practical quantization strategies for compressive acquisition systems. Specifically, we first provide a brief overview of quantization and examine fundamental performance bounds applicable to any quantization approach. Next, we consider several forms of scalar quantizers, namely uniform, non-uniform, and 1-bit. We provide performance bounds and fundamental analysis, as well as practical quantizer designs and reconstruction algorithms that account for quantization. Furthermore, we provide an overview of Sigma-Delta ( $\Sigma\Delta$ ) quantization in the compressed sensing context, and also discuss implementation issues, recovery algorithms and performance bounds. As we demonstrate, proper accounting for quantization and careful quantizer design has significant impact in the performance of a compressive acquisition system.

---

Petros T. Boufounos

Mitsubishi Electric Research Laboratories, 201 Broadway, Cambridge, MA, USA, e-mail: petrosb@merl.com

Laurent Jacques

ISPGROUP, ICTEAM/ELEN, Université catholique de Louvain, Place du Levant 2, PO box L5.04.04, B1348 Louvain-la-Neuve, Belgium e-mail: laurent.jacques@uclouvain.be

Felix Krahmer

Georg-August-Universität Göttingen, Lotzestraße 16-18, 37083 Göttingen, Germany e-mail: f.krahmer@math.uni-goettingen.de

Rayan Saab

University of California, San Diego, 9500 Gilman Drive #0112, La Jolla, CA 92093-0112, USA e-mail: rsaab@ucsd.edu

## 1 Introduction

In order to store and manipulate signals using modern devices, it is necessary to digitize them. This involves two steps: sampling (or measurement) and quantization. The compressed sensing theory and practice described in the remainder of this book provides a novel understanding of the measurement process, enabling new technology and approaches to reduce the sampling burden. This chapter explores the very interesting interaction of compressed sensing with quantization.

Sampling maps a signal to a set of coefficients, typically using linear measurements. This map can often be designed to be lossless, *i.e.*, to perfectly represent all signals in a certain class, as well as robust to noise and signal modeling errors. The Nyquist theorem, as well as more recent compressive sampling theorems are examples of such sampling approaches [25, 9, 75].

The guarantees in sampling theorems are typically stated in terms of the critical measurement rate, *i.e.*, the number of measurements necessary to perfectly represent signals in a given class. Oversampling, compared to that minimum rate, typically provides robustness to errors in the representation, noise in the acquisition, and mismatches in signal models. The latter is especially important in compressive sensing systems as they provide perfect reconstruction guarantees for exactly sparse signals; in practice, the acquired signal is almost never exactly sparse.

Quantization, on the other hand, is the process of mapping the representation coefficients—which potentially belong to an uncountably infinite set—to elements in a finite set, and representing them using a finite number of bits. Due to the many-to-one nature of such a map, the quantized representation is in general lossy, *i.e.*, distorts the representation and, therefore, the signal. This distortion occurs even if the measurement process is lossless.

The interaction of quantization with sampling introduces interesting trade-offs in the acquisition process. A system designed to sample signals at (or slightly above) the critical rate may be less robust to errors introduced by quantization. Consequently, it requires a sophisticated quantizer design that ensures very small quantization errors. On the other hand, a simpler quantizer architecture (e.g., with fewer bits per measurement) could introduce significant error to the representation and require some oversampling to compensate. Practical systems designs navigate this trade-off, for example, according to the complexity of the corresponding hardware.

Compressive acquisition systems amplify the importance of the trade-off between quantizer complexity and oversampling. The sampling rate is significantly reduced in such systems, at the expense of increased sensitivity to noise and signal model mismatch. Thus, loss of information due to quantization can be detrimental, especially when not properly handled. One may revert to oversampling here as well, however the incoherent and often randomized nature of compressive measurements poses challenges. Thus, powerful oversampling based quantization approaches, such as Sigma-Delta quantization can be applied, but only after careful consideration.

Nevertheless, the sparse signal models and the computational methods developed for compressed sensing can alleviate a number of performance bottlenecks due to quantization in conventional systems. Using computational approaches originating

in frame theory and oversampling, it is possible to significantly reduce the distortion due to quantization, to significantly improve the performance due to saturation, and to enable reconstruction from measurements quantized as coarsely as 1 bit. The theory and practice for such methods are described in Sec. 3.

It might seem counter-intuitive that compressed sensing attempts to remove sampling redundancy, yet successful reconstruction approaches employ tools developed for oversampled representations. In fact there is a strong connection between compressed sensing and oversampling, which we explore in various points in this chapter. Furthermore, with sufficient care, this connection can be exposed and exploited to implement Sigma-Delta quantization in CS-based acquisition systems, and significantly improve performance over scalar quantization. The details are discussed in Sec. 4.

The next section presents general principles of quantization, including a brief background on vector, scalar, and Sigma-Delta quantization for general acquisition systems. It is not an exhaustive survey of the topic. For this we refer the reader to [39, 32, 77]. Instead, it serves to establish notation and as quick reference for the subsequent discussion. Sec. 3 and Sec. 4 examine the interaction of compressive sensing and quantization in significant detail. Sec. 5 concludes with some discussion of the literature, promising directions and open problems.

**Notation:** In addition to the notational conventions defined in Chapter 1, this chapter also uses the following general notations. The logarithm in base  $a > 0$  is noted  $\log_a$  and whenever the base is not specified,  $\log$  refers to the natural logarithm. Note that in some cases, such as asymptotic results, the logarithm base is not important. This chapter also uses the following non-asymptotic orderings: For two functions  $f$  and  $g$ , we write  $f \lesssim g$  if there exists a constant  $C > 0$  independent of the function arguments such that  $f \leq Cg$ , with a similar definition for  $f \gtrsim g$ . Moreover,  $f \asymp g$  if we have both  $f \lesssim g$  and  $f \gtrsim g$ . Occasionally, we also rely on the well-established big- $O$  and big- $\Omega$  asymptotic notation to concisely explain asymptotic behavior when necessary. More specific notation is defined at first occurrence.

## 2 Fundamentals of Quantization

For the purposes of this section, a quantizer operates on signals  $x$ , viewed as vectors in a bounded set  $V \subset \mathbb{R}^n$ . The goal of a quantizer  $Q(\cdot)$  is to represent those signals as accurately as possible using a rate of  $R$  bits, *i.e.*, using a quantization point  $q = Q(x)$  chosen from a set of  $2^R$  possible ones often referred to as codebook. Of course, when  $V$  contains an infinite number of signals, signals will be distorted through this representation.

In this section, we first define common quantization performance metrics and determine fundamental bounds on the performance of a quantizer. Then, in preparation for the next sections, we examine common approaches to quantization, namely

scalar and Sigma-Delta quantization, which are very useful in compressive sensing applications.

## 2.1 Quantization Performance Bounds

To measure the accuracy of the quantizer we consider the distortion, i.e., the  $\ell_2$  distance of a quantization point from its original signal  $\|x - Q(x)\|_2$ . The overall performance of the quantizer is typically evaluated either using the average distortion over all the signals—often computed using a probability measure on the signal space  $V$ —or using the worst case distortion over all signals in  $V$ . In this chapter, in the spirit of most of the compressed sensing literature, we quantify the performance of the quantizer using the worst case distortion on any signal, i.e.,

$$\varepsilon = \sup_{x \in V} \|x - Q(x)\|_2. \quad (1)$$

This choice enables very strong guarantees, irrespective of the accuracy of any probabilistic assumption on the signal space.

A lower bound on the distortion of any quantizer can be derived by constructing a covering of the set  $V$ . A covering of radius  $r$  is a set of points  $q$  such that each element in  $V$  has distance at most  $r$  from its closest point in the covering. If we can construct a covering using  $P$  points, then we can also define a quantizer that uses  $R = \lceil \log_2 P \rceil$  bits and has worst case distortion  $\varepsilon = r$  as each signal is quantized to the closest point in the covering.

To determine a lower bound for the number of points in such a covering, we consider balls of radius  $r$  centered at  $q$ , defined as

$$\mathcal{B}_r(q) = \{x \in \mathbb{R}^n \mid \|q - x\|_2 \leq r\}. \quad (2)$$

Since each signal in  $V$  is at most  $r$  away from some point in the covering, if we place a ball of radius  $r$  at the center of each point of the covering, then the union of those balls covers  $V$ . Thus, the total volume of the balls should be at least as large as the volume of the set, denoted  $\text{vol}(V)$ . Since the volume of a ball of radius  $r$  in  $n$  dimensions is  $\text{vol}(\mathcal{B}_r(q)) = r^n \pi^{n/2} / \Gamma(1 + n/2)$ , where  $\Gamma(\cdot)$  is the Gamma function, the best possible error given the rate  $R$  can be derived using

$$\text{vol}(V) \leq \frac{\pi^{n/2}}{\Gamma(1 + \frac{n}{2})} 2^R r^n \Rightarrow r \gtrsim 2^{-\frac{R}{n}}. \quad (3)$$

In other words, the worst-case error associated with an optimal quantizer can, at best, decay exponentially as the bit rate increases. Moreover, the decay rate depends on the ambient dimension of the signal. In short,

$$\varepsilon \gtrsim 2^{-\frac{R}{n}}. \quad (4)$$

The smallest achievable worst case distortion for a set is also known as the  $(R + 1)$ -dyadic entropy number of the set, whereas the number of bits necessary to achieve a covering with worst-case distortion equal to  $\varepsilon$  is known as the Kolmogorov  $\varepsilon$ -entropy or metric entropy of the set.

For the models commonly assumed in compressive sensing, these quantities are not straightforward to calculate and depend on the sparsity model assumed. For example, compressible signals are commonly modeled as being drawn from a unit  $\ell_p$  ball, where  $0 < p < 1$  (cf. Chapter 1 for a discussion on compressibility). In this case, the worst case distortion is bounded by

$$\varepsilon \gtrsim \begin{cases} 1 & \text{if } 1 \leq R \leq \log_2 n \\ \left(\frac{1}{R} \log_2 \left(\frac{n}{R} + 1\right)\right)^{\frac{1}{p} - \frac{1}{2}} & \text{if } \log_2 n \leq R \leq n \\ 2^{-\frac{R}{n}} n^{\frac{1}{2} - \frac{1}{p}} & \text{if } R \geq n, \end{cases} \quad (5)$$

where the constant implicit in our notation is independent of  $R$  and  $n$  [88, 36, 64, 24].

In the case of exactly  $k$ -sparse signals, the volume of the union of subspaces they occupy has measure zero in the  $n$ -dimensional ambient space. However, by considering the  $\binom{n}{k}$   $k$ -dimensional subspaces and coverings of their unit balls, a lower bound on the error can be derived [18], namely

$$\varepsilon \gtrsim \frac{2^{-\frac{R}{k}} n}{k}. \quad (6)$$

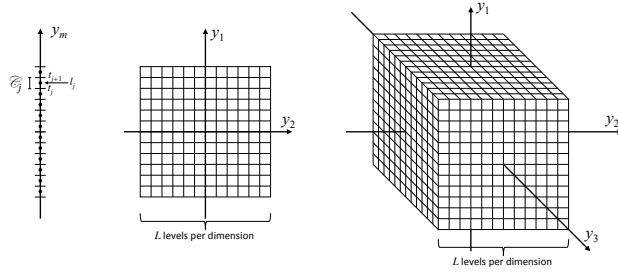
Note that this lower bound can be achieved in principle using standard transform coding (TC), *i.e.*, by first representing the signal using its sparsity basis, using  $\log_2 \binom{n}{k} \lesssim k \log_2(n/k)$  bits to represent the support of the non-zero coefficients and using the remaining bits to represent the signal in the  $k$ -dimensional subspace at its Kolmogorov entropy

$$\varepsilon_{\text{TC}} \lesssim 2^{-\frac{R - k \log_2(n/k)}{k}} = \frac{2^{-\frac{R}{k}} n}{k}. \quad (7)$$

Unfortunately, compressive sensing systems do not have direct access to the sparse vectors. They can only access the measurements,  $y = Ax$ , which must be quantized upon acquisition—in practice using analog circuitry. Thus, transform coding is not possible. Instead, we must devise simple quantization algorithms that act directly on the measurements in such a way that permits accurate reconstruction.

## 2.2 Scalar Quantization

The simplest approach to quantization is known as *scalar quantization* and often referred to as *pulse code modulation* (PCM), or *memoryless scalar quantization* (MSQ). Scalar quantization directly quantizes each measurement of the signal, with-



**Fig. 1** A finite uniform scalar quantizer and the uniform grid it generates in 2 and 3 dimensions

out taking other measurements into account. In other words a 1-dimensional, *i.e.*, scalar, quantizer is applied separately to each measurement of the signal.

### 2.2.1 Measurement and Scalar Quantization

A scalar quantizer can be defined using a set of levels,  $\mathcal{Q} = \{l_i \in \mathbb{R} : l_j < l_{j+1}\}$ , comprising the quantization codebook, and a set of thresholds  $\mathcal{T} = \{t_i \in \mathbb{R} : t_j < t_{j+1}\}$ , implicitly defining the quantization intervals  $\mathcal{C}_j = [t_j, t_{j+1})$ . Assuming no measurement noise, the quantizer is applied element-wise to the measurement coefficients,  $y = Ax$ , to produce the quantized measurements  $q = Q(y)$ ,  $q_i = Q(y_i)$ . Using a rate of  $B$  bits per coefficient, *i.e.*,  $R = mB$  total bits, the quantizer represents  $L = 2^B$  total levels per coefficient. A scalar value  $y_i$  quantizes to the quantization level corresponding to the quantization interval in which the coefficient lies.

$$Q(y_i) = l_j \Leftrightarrow y_i \in \mathcal{C}_j. \quad (8)$$

A scalar quantizer is designed by specifying the quantization levels and the corresponding thresholds. Given a source signal with measurements modeled as a continuous random variable  $X$ , a (*distortion*) *optimal* scalar quantizer minimizes the error

$$\mathbb{E}|X - Q(X)|^2. \quad (9)$$

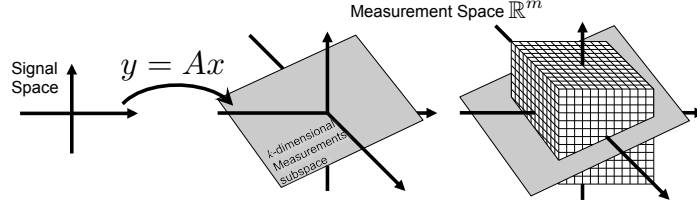
Such an optimal quantizer necessarily satisfies the Lloyd-Max conditions [71, 74]

$$l_j = \mathbb{E}\{X | X \in \mathcal{C}_j\}, \quad t_j = \frac{1}{2}(l_j + l_{j+1}), \quad (10)$$

which define a fixed point equation for levels and thresholds and the corresponding fixed-point iteration—known as the Lloyd-Max algorithm—to compute them.

Alternatively, a simpler design approach is the uniform scalar quantizer, which often performs almost as well as an optimal scalar quantizer design. It is significantly less complex and can be shown to approach optimality as the bit-rate increases [39]. The thresholds of a uniform scalar quantizer are defined to be equispaced, *i.e.*,  $t_{j+1} - t_j = \Delta$ , where  $\Delta$  is referred to as the quantization bin width





**Fig. 2** A  $k$ -dimensional space measured using  $m$  measurements spans a  $k$ -dimensional subspace of  $\mathbb{R}^m$  and intersects only a few of the  $L^m$  available quantization cells.

or *resolution*. The levels are typically set to the mid-point  $l_j = \frac{1}{2}(t_j + t_{j+1})$  of the quantization bin  $\mathcal{C}_j$ . Thus, the quantization error introduced to each coefficient is bounded by  $\Delta/2$ . A uniform quantizer defines a uniform grid in the  $m$ -dimensional measurement space, as shown in Fig. 1.

In practical systems, the scalar quantizer has finite range, *i.e.*, it saturates if the signal exceeds a saturation level  $S$ . In particular, a uniform finite-range scalar quantizer using  $B$  bits per coefficient has quantization interval  $\Delta = S2^{-B+1}$ . If a coefficient exceeds  $S$ , the quantizer maps the coefficient to the largest quantization level, *i.e.*, it saturates. Depending on the magnitude of the coefficient, this may introduce significant error. However, it is often convenient in theoretical analysis to assume an infinite quantizer that does not saturate. This assumption is often justified, as  $S$  in practice is set large enough to avoid saturation given a signal class. As described in Sec. 3.3, this is often suboptimal in compressive sensing applications.

Compared to classical systems, optimal scalar quantizer designs for compressive sensing measurements require extra care. An optimal design with respect to the measurement error is not necessarily optimal for the signal, due to the non-linear reconstruction inherent in compressed sensing. While specific designs have been derived for very specific reconstruction and probabilistic signal models, *e.g.*, [89, 56], a general optimal design remains an open problem. Thus the literature has focused mostly, but not exclusively, on uniform scalar quantizers.

## 2.2.2 Scalar Quantization and Oversampling

When a signal is oversampled, a scalar quantizer makes suboptimal use of the bit-rate. The  $k$ -dimensional signal space mapped through the measurement operator to an  $m$ -dimensional measurement space, where  $m > k$ , spans, at most, a  $k$ -dimensional subspace of  $\mathbb{R}^m$ , as shown in Fig. 2. As evident from the figure, this subspace intersects only a few of the available quantization cells and, therefore, does not use the available bits effectively. For an  $L$ -level quantizer, the number of quantization cells intersected  $I_{k,m,L}$  is bounded by [90, 38, 14]

$$I_{k,m,L} \lesssim \left(\frac{Lm}{k}\right)^k \quad (11)$$

Using a simple covering argument as in Sec. 2.1, it is thus possible to derive a lower bound on the error performance as a function of the number of measurements  $m$

$$\epsilon \gtrsim \left( \frac{2^{-Bk}}{m} \right) \quad (12)$$

The bounds hold for any scalar quantizer design, not just uniform ones.

Linear reconstruction, *i.e.*, reconstruction using a linear operator acting on the scalar quantized measurements, does not achieve the bound (12) [91, 38]. The quantization error using linear reconstruction can only decay as fast as

$$\epsilon \gtrsim \frac{2^{-Bk}}{\sqrt{m}}. \quad (13)$$

Instead, *consistent* reconstruction achieves the optimal bound in a number of cases. Consistent reconstruction treats the quantization regions as reconstruction constraints and ensures that the reconstructed signal  $\hat{x}$  quantizes to the same quantization points when measured using the same system. Thus in the oversampled setting where  $A$  is an  $m \times k$  matrix with  $m > k$ , and where  $q = Q(Ax)$  one solves the problem:

$$\text{find any } \hat{x} \text{ s.t. } q = Q(A\hat{x}). \quad (14)$$

If the measurement operator  $A$  is a tight frame formed by an oversampled Discrete Fourier Transform (DFT), the root mean square error (RMSE) of such a reconstruction (with respect to a random signal model) decays as  $O(1/m)$  [91, 38], *i.e.*, as (12). In the case of random frames with frame vectors drawn independently from a Gaussian distribution [50] or from a suitable distribution on the  $(m-1)$ -sphere [84], the reconstruction method in (14) also displays RMSE and worst case reconstruction error decreasing as  $O(1/m)$  and  $O((\log m)/m)$ , respectively.

The constraints imposed by consistent reconstruction are convex and can be imposed on any convex optimization algorithm. This makes them particularly suitable for a number of reconstruction algorithms already used in compressive sensing systems, as we explore in Sec. 3.

The bounds (12) and (13)—which can be achieved with proper design of the measurement process and the reconstruction algorithm—demonstrate that the most efficient use of the rate  $R = mB$  is in refining each measurement using more bits per measurement,  $B$ , rather than in increasing the number of measurements,  $m$ . They suggest that in terms of error performance, by doubling the oversampling it is possible to save 0.5 bits per coefficient if linear reconstruction is used and 1 bit per coefficient if consistent reconstruction is used. This means that a doubling of the rate by doubling the oversampling factor, is equivalent to a linear increase in the rate by  $m/2$  or  $m$  through an increase in  $B$ , for linear and consistent reconstruction, respectively. So in principle, if rate-efficiency is the objective, the acquisition system should only use a sufficient number of measurements to reconstruct the signal and no more. All the rate should be devoted to refining the quantizer. However, these

bounds ignore the practical advantages in oversampling a signal, such as robustness to erasures, robustness to measurement noise and implementation complexity of high-rate scalar quantizers. Thus in practice, oversampling is often preferred, despite the rate-inefficiency. Techniques such as Sigma-Delta quantization, which we discuss in Sec. 2.3, have been developed to improve some of the trade-offs and are often used in conjunction with oversampling.

### 2.2.3 Performance Bounds on Sparse Signals

Scalar quantization in compressive sensing exhibits similar bounds as scalar quantization of oversampled signals. Signals that are  $k$ -sparse in  $\mathbb{R}^n$  belong to a union of  $k$ -dimensional subspaces. When measured using  $m$  linear measurements, they occupy a union of  $k$ -dimensional subspaces of  $\mathbb{R}^m$ ,  $\binom{n}{k}$  of them. Using the same counting argument as above, it is evident that the number of quantization cells intersected, out of the  $L^m$  possible ones, is at most

$$\binom{n}{k} I_{k,m,L} \gtrsim \left( \frac{Lmn}{k^2} \right)^k \quad (15)$$

The resulting error bound is

$$\varepsilon \gtrsim \frac{2^{-B}k}{m} \quad (16)$$

$$\gtrsim \frac{2^{-\frac{R}{m}}k}{m}, \quad (17)$$

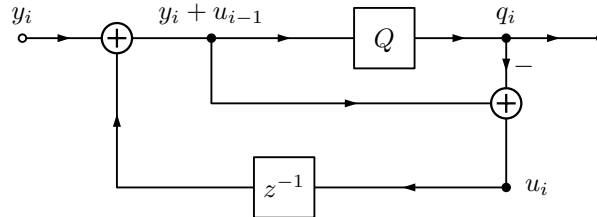
which decays slower than (6) as the rate increases keeping the number of measurements  $m$  constant. Furthermore, as the rate increases with the number measurements  $m$ , keeping  $B$ , the number of bits per measurement constant, the behavior is similar to quantization of oversampled frames: the error can only decay linearly with  $m$ .

These bounds are not surprising, considering the similarities of oversampling and compressive sensing of sparse signals. It should, therefore, be expected that more sophisticated techniques, such as Sigma-Delta ( $\Sigma\Delta$ ) quantization should improve performance, as they do in oversampled frames. However, their application is not as straightforward. The next section provides an overview of  $\Sigma\Delta$  quantization and Sec. 4 discusses in detail how it can be applied to compressive sensing.

## 2.3 Sigma-Delta Quantization

An alternative approach to the scalar quantization techniques detailed in the previous section is feedback quantization. The underlying idea is that the fundamental limits for the reconstruction accuracy discussed above can be overcome if each

quantization step takes into account errors made in previous steps. The most common feedback quantization scheme is  $\Sigma\Delta$  quantization, originally introduced for bandlimited signals in [47] (cf. [46]). A simple  $\Sigma\Delta$  scheme, illustrated in Figure 2.3, shows this feedback structure.



**Fig. 3** A block diagram of a simple 1st order  $\Sigma\Delta$  scheme: The input  $y_i$  is added to the state variable  $u_{i-1}$  (initialized as  $u_0 = 0$ ) and the sum is scalar quantized. Subsequently, the state variable is updated as the difference between the scalar quantizer's input and its output. More complex designs, featuring higher order  $\Sigma\Delta$  quantization with more feedback loops are possible. We discuss such designs in more detail in Section 4.

A motivation in  $\Sigma\Delta$  quantization is that, in some applications, reducing circuit complexity is desirable, even at the expense of a higher sampling rate. Indeed,  $\Sigma\Delta$  designs drastically reduce the required bit depth per sample while allowing for accurate signal reconstruction using simple circuits. In fact, since its introduction,  $\Sigma\Delta$  quantization has seen widespread use (see, *e.g.*, [77] and the references therein) in applications ranging from audio coding to wireless communication.

Nevertheless, a mathematical analysis of  $\Sigma\Delta$  quantization in its full generality has been challenging. A preliminary analysis of simple  $\Sigma\Delta$  schemes for restricted input classes (including constant input and sinusoidal input) was presented in [40] and follow-up works. However, most of these results were limited to linear, or at best low-order polynomial error decay in the oversampling rate. This type of error decay is sub-optimal (albeit better than scalar quantization), and rather far from the optimal exponential error decay. Specifically, a major difficulty that prevented a more comprehensive treatment was understanding the instabilities caused by the positive feedback inherent to the  $\Sigma\Delta$  circuit designs. For example, depending on the design of the  $\Sigma\Delta$  scheme, the state variables could grow without bound. A crucial idea to prevent such phenomena for arbitrary *band-limited* inputs was developed in [32]; their analysis led, for the first-time, to super-polynomial bounds for the error decay. To date, the best known error bounds decay exponentially in the oversampling rate [41, 35]. While this is near-optimal (optimal up to constants in the exponent), it has been shown that with a fixed bit budget per sample, the achievable rate-distortion relationship is strictly worse than for scalar quantization of Nyquist rate samples [63]. That said, increasing the bit budget per sample entails more expensive and complex circuitry, which grows increasingly costly with every added bit (in fact, the best

current quantizers provide a resolution of about 20 bits per sample). Thus, for quantizing bandlimited functions, if one wishes to improve the performance or reduce the cost, one must revert to oversampling-based methods such as  $\Sigma\Delta$  quantization.

The accuracy gain of  $\Sigma\Delta$  quantization is most prominent when a significant oversampling rate and, therefore, a high redundancy of samples is inherent or desired. Such redundant representations can also be encountered in a finite-dimensional discrete context. Namely, this corresponds to a finite frame expansion in the sense of (1.32). This observation served as a motivation to devise  $\Sigma\Delta$  schemes for finite-frame expansions, and the first such construction was provided in [7]. In contrast to oversampled representations of bandlimited signals, which directly correspond to a temporal ordering, finite frames generally do not have an inherent order, nor are the frame vectors necessarily close enough to each other to allow for partial error compensation. Due to this difficulty, the first works on  $\Sigma\Delta$  quantization for finite frame expansions focus on frames with special smoothness properties. Namely, they assume that the frame  $\Phi = \{\phi_j\}_{j=1}^N$  has a well controlled *frame variation*

$$v_\Phi := \sum_{j=1}^{N-1} \|\phi_{j+1} - \phi_j\|_2.$$

The constructions in [7] coupled with (linear) reconstruction via the canonical dual frame (that is, the Moore-Penrose pseudo-inverse of the matrix that generates the redundant representation) was shown to yield an error decay on the order of  $v_\Phi N^{-1}$ , *i.e.*, linear error decay whenever the frame variation is bounded by a constant. By using more sophisticated  $\Sigma\Delta$  schemes these results were later improved to higher order polynomial error decay [6, 12, 13] in the number of measurements, thereby beating the bound (12) associated with scalar quantization. Again, these constructions require certain smoothness conditions on the frame and employ the canonical dual frame for recovery. In a slightly different approach, the design of the feedback and the ordering of the frame vectors has been considered as part of the quantizer design [20, 14].

A new take on the frame quantization problem was initiated in [65, 10] where the authors realized that reconstruction accuracy can be substantially improved by employing an appropriate alternative dual frame (*i.e.*, a different left-inverse) for recovery. At the core of this approach is still a smoothness argument, but this time for the dual frame. Given a frame, an appropriate dual frame, the so-called Sobolev dual, can be obtained by solving a least-squares problem over the space of all duals [10]. Again, this yields polynomial error decay, albeit now in more general settings. Moreover, by optimizing over such constructions, root-exponential error decay can be achieved [60].

While the definition of the Sobolev dual does not require any smoothness of the frame, the concrete examples discussed in the aforementioned works still exclusively focused on smooth frames. Similar results on recovery guarantees for frames without smoothness properties were first obtained for frames consisting of independent standard Gaussian vectors [42] and subsequently generalized to vectors with independent subgaussian entries [61].

The underlying constructions also form the basis for the  $\Sigma\Delta$  quantization schemes for compressed sensing measurements. Details on such schemes are given in Sec. 4. The insight behind the schemes is that the number of measurements taken in compressed sensing is typically larger than the support size by at least a logarithmic factor in the dimension, and there is an interest in choosing it even larger than that, as this induces additional stability and robustness. Thus, once the support of the signal has been identified and only the associated signal coefficients need to be determined, one is dealing with a redundant representation. The goal is now to employ frame quantization schemes to exploit this redundancy.

For typical compressed sensing matrices, any  $k$  columns indeed form a frame; this follows for example from the restricted isometry property. However, as the support of the signal is not known when quantizing the measurements, it is crucial that  $\Sigma\Delta$  quantization is universal. That is, it must not require knowledge regarding which of a given collection of frames (namely, those forming the rows of an  $m \times k$  submatrix of  $A$ ) has been used for encoding. The reconstruction from the resulting digital encodings then typically proceeds in two steps. First the support is identified using standard compressed sensing recovery techniques, just treating the quantization error as noise. In a second step, only the restriction of the measurement matrix to the identified support columns is considered. For the frame consisting of the rows of this matrix, one then applies frame quantization reconstruction techniques. Recovery guarantees for such an approach have been proven for Gaussian measurements [42] and measurements with independent subgaussian entries [61]. It is of great importance that the dual frame used for recovery is chosen properly (*e.g.*, the Sobolev dual), as it follows from the RIP that the frames never have a small frame variation. Here again the recovery error bounds decay polynomially in the number of measurements and beat the analogous bounds for scalar quantization.

Preliminary steps towards a unified approach to support and signal recovery have been considered in [29]. The reconstruction techniques studied in this work, however, intrinsically rely on certain non-convex optimization problems, for which no efficient solution methods are known. Thus the quest remains open for an integrated approach to reconstruction from  $\Sigma\Delta$ -quantized compressed sensing measurements that combines numerical tractability and guaranteed recovery.

### 3 Scalar Quantization and Compressive Sensing

The interplay of scalar quantization and compressed sensing has been widely explored in the literature. In addition to the lower bounds discussed in 2.2.3, there is significant interest in providing practical quantization schemes and reconstruction algorithms with strong performance guarantees.

This part explores these results. Our development considers the following quantized compressed sensing (QCS) model:

$$q = Q(y) = Q(Ax), \quad (18)$$

where  $x \in \mathbb{R}^n$  and  $A \in \mathbb{R}^{m \times n}$ . The sensing matrix can be, for instance, a random Gaussian sensing matrix  $A$  such that  $a_{ij} \sim_{\text{iid}} \mathcal{N}(0, 1)$ . Note that the scaling of the entries of the sensing matrix should be independent of  $m$ . This allows us to fix the design of the scalar quantizer  $Q$  since the dynamic range of the components of  $Ax$  is then independent of the number of measurements. This has no consequence on some of the common requirements the sensing matrix must satisfy, such as the Restricted Isometry Property (see Chap. 1), as soon as an appropriate rescaling of  $A$  is applied. For instance, if  $A$  has RIP of order  $2k$  and if  $A \rightarrow \lambda A$  for some  $\lambda > 0$ , then  $A/\lambda$  has RIP of the same order and the error bound (1.20) in the stability Theorem 1.6 remains unchanged [52].

The first two parts, Sec. 3.1 and Sec. 3.2, focus on the *high resolution assumption* (HRA) that simplifies the QCS model. Under HRA, the quantization bin widths— $\Delta$  or the distance between two consecutive thresholds—are small with respect to the dynamic range of the unquantized input. This allows us to model the quantization distortion  $Q(Ax) - Ax$  as uniform white noise [39]. Determining bounds on its power and moments can better constrain signal reconstruction methods, such as the basis pursuit denoise (BPDN) program [28, 26], which is commonly used for reconstructing signals whose CS measurements are corrupted by Gaussian noise. However, the price to pay is an oversampling in CS measurements.

Sec. 3.3 considers scalar quantizers with *saturation*. Saturation induces information loss in the measurements exceeding the saturation level. However, democracy—a key property of compressive sensing measurements that makes every measurement equally informative—provides robustness against such corruption.

In Sec. 3.4, very low-resolution quantization is studied through 1-bit compressed sensing. In this case, the HRA cannot be assumed anymore—the quantization bins are the two semi-infinite halves of the real line—and the analysis of the QCS model relies on high dimensional geometric arguments.

Finally, Sec. 3.5 studies how noise, either on the signal or on the measurements, can impact the QCS model (18), the reconstruction error and the quantizer trade-offs. In particular, at constant bit budget  $R = mB$ , the total noise power determines the optimal trade-off between quantizer precision and number of measurements.

### 3.1 Uniform Scalar Quantization

First we consider the QCS model (18) using a uniform quantizer with resolution  $\Delta$  and a set of levels  $\mathcal{Q}$ ,

$$q = Q(y) = Q(Ax) \in \mathcal{Q}^m,$$

measuring a signal  $x \in \mathbb{R}^n$  using a sensing matrix  $A \in \mathbb{R}^{m \times n}$ . For simplicity, we assume henceforth that  $x$  is sparse in the canonical basis, *i.e.*,  $\Psi = I$ .

We consider a quantizer  $Q$  that has uniform quantization regions, *i.e.*,  $t_{j+1} - t_j = \Delta$  for all  $j$ , and, setting  $t_j = j\Delta$ , quantization levels  $l_j = \frac{t_j + t_{j+1}}{2} = (j + \frac{1}{2})\Delta$  in  $\mathcal{Q}$ .

By definition, the signal  $x$  satisfies the following *quantization consistency* constraint (QC<sub>u</sub>)

$$\|q - Ax\|_\infty \leq \Delta/2. \quad (\text{QC}_u)$$

From this fact, we can also deduce that

$$\|Ax - q\|_2 \leq \sqrt{m} \|Ax - q\|_\infty \leq \sqrt{m}\Delta/2.$$

This shows that the QCS model can be assimilated to a noisy CS model

$$q = Q(Ax) = Ax + \xi, \quad (19)$$

with a “noise”  $\xi = Q(Ax) - Ax$  of bounded  $\ell_2$ -norm, *i.e.*,  $\|\xi\|_2 \leq \sqrt{m}\Delta/2$ .

The quantization noise power can be further reduced using the high resolution assumption. Under this assumption, the coefficients of  $y$  may lie anywhere in the quantization region determined by the coefficients of  $q$  and it is natural to model the quantization distortion  $\xi$  as a uniform white noise, *i.e.*,

$$\xi_i \sim_{\text{iid}} \mathcal{U}([- \Delta/2, \Delta/2]).$$

Under this model, a simple use of the Chernoff-Hoeffding bound [45] provides, with high probability

$$\|\xi\|_2^2 \leq \varepsilon_2^2 := \frac{\Delta^2}{12}m + \zeta \frac{\Delta^2}{6\sqrt{5}}m^{1/2},$$

for a small constant  $\zeta > 0$ .

The first approach in modeling and understanding QCS exploited this bound and the development of noise-robust CS approaches to impose a *distortion consistency constraint* (DC<sub>u</sub>) [25]

$$\|q - Ax'\|_2 \leq \varepsilon_2, \quad (\text{DC}_u)$$

on any candidate signal  $x'$  estimating  $x$ . This was indeed a natural constraint to consider since most noise-robust compressed sensing reconstruction methods can incorporate a bounded  $\ell_2$ -norm distortion on the measurements. For instance, the BPDN program can find a solution  $\hat{x}$  of

$$\hat{x} = \arg \min_z \|z\|_1 \quad \text{s.t.} \quad \|q - Az\|_2 \leq \varepsilon_2. \quad (\text{BPDN})$$

Then, if the sensing matrix  $A' = A/\sqrt{m}$  satisfies the RIP with constant  $\delta \leq 1/\sqrt{2}$  on  $2k$  sparse signals, it is known [22] that

$$\|x - \hat{x}\|_2 \lesssim \frac{1}{\sqrt{m}} \varepsilon_2 + \frac{1}{\sqrt{k}} \sigma_k(x)_1 \asymp \Delta + \frac{1}{\sqrt{k}} \sigma_k(x)_1,$$

where  $\sigma_k(x)_1$  is the best  $k$ -term approximation defined in (1.2).

This approach has two drawbacks. First, there is no guarantee that the solution  $\hat{x}$  satisfies the QC<sub>u</sub> constraint above, *i.e.*,  $\|q - A\hat{x}\|_\infty \not\leq \Delta/2$ . This shows that some sensing information has been lost in the reconstruction. Moreover, as described in Sec. 2.2.2, the consistency of the solution helps in reaching the lower bound [38, 84,



50]

$$(\mathbb{E}\|x - \hat{x}\|^2)^{1/2} \gtrsim \frac{k}{m} \Delta$$

in the oversampled setting. Second, from a maximum *a posteriori* standpoint, since every constrained optimisation corresponds to an unconstrained Lagrangian formulation, imposing a small  $\ell_2$ -norm on the residual  $q - A\hat{x}$  can be viewed as enforcing a Gaussian distribution on  $\xi$ , which is not the uniform one expected from the HRA.

To circumvent these two limitations, [52] studied the Basis Pursuit DeQuantizer (BPDQ<sub>p</sub>) program

$$\hat{x}_p = \arg \min_z \|z\|_1 \quad \text{s.t.} \quad \|q - Az\|_p \leq \varepsilon_p, \quad (\text{BPDQ}_p)$$

where  $\varepsilon_p$  must be carefully selected in order for  $x$  to be a feasible point of this new  $\ell_p$ -constraint. If  $\varepsilon_p \rightarrow \Delta$  as  $p \rightarrow \infty$ , the BPDQ<sub>p</sub> solution  $\hat{x}_p$  tends to be consistent with the quantized measurements. But what is the price to pay, *e.g.*, in terms of number of measurements, for being allowed to increase  $p$  beyond 2?

To answer this, we need a variant of the restricted isometry property.

**Definition 1.** Given two normed spaces  $\mathcal{X} = (\mathbb{R}^m, \|\cdot\|_{\mathcal{X}})$  and  $\mathcal{Y} = (\mathbb{R}^n, \|\cdot\|_{\mathcal{Y}})$  (with  $m < n$ ), a matrix  $A \in \mathbb{R}^{m \times n}$  has the Restricted Isometry Property from  $\mathcal{X}$  to  $\mathcal{Y}$  at order  $k \in \mathbb{N}$ , radius  $0 \leq \delta < 1$  and for a normalization  $\mu > 0$ , if for all  $x \in \Sigma_k := \{u \in \mathbb{R}^N : \|u\|_0 \leq k\}$ ,

$$(1 - \delta)^{1/\kappa} \|x\|_{\mathcal{Y}} \leq \frac{1}{\mu} \|Ax\|_{\mathcal{X}} \leq (1 + \delta)^{1/\kappa} \|x\|_{\mathcal{Y}}, \quad (20)$$

the exponent  $\kappa$  depending on the spaces  $\mathcal{X}$  and  $\mathcal{Y}$ . To lighten notation, we write that  $A$  is  $\text{RIP}_{\mathcal{X}, \mathcal{Y}}(k, \delta, \mu)$ .

In this general definition, the common RIP is equivalent to  $\text{RIP}_{\ell_2, \ell_2}^m(k, \delta, 1)$  with  $\kappa = 2$  (see Chap. 1, Eq. (1.10)). Moreover, the  $\text{RIP}_{p, k, \delta'}$  defined in [8] is equivalent to the  $\text{RIP}_{\ell_p, \ell_p}^m(k, \delta, \mu)$  with  $\kappa = 1$ ,  $\delta' = 2\delta/(1 - \delta)$  and  $\mu = 1/(1 - \delta)$ . Finally, the Restricted  $p$ -Isometry Property proposed in [27] is also equivalent to the  $\text{RIP}_{\ell_p, \ell_2}^m(k, \delta, 1)$  with  $\kappa = p$ .

To characterize the stability of BPDQ we consider the space  $\mathcal{X} = \ell_p^m := (\mathbb{R}^m, \|\cdot\|_p)$  and  $\mathcal{Y} = \ell_2^n := (\mathbb{R}^n, \|\cdot\|_2)$  with  $\kappa = 1$ , and we write  $\text{RIP}_p$  as a shorthand for  $\text{RIP}_{\ell_p, \ell_2}^m$ . At first sight, it could seem unnatural to define an embedding of  $\mathcal{X} = \ell_p^m$  in  $\mathcal{Y} = \ell_2^n$  for  $p \neq 2$ , those spaces being not isometrically isomorphic to each other for  $m = n$ . However, the  $\text{RIP}_p$  rather sustains the possibility of an isometry between  $\mathcal{X} \cap A\Sigma_k$  and  $\mathcal{Y} \cap \Sigma_k$ . We will see in Prop. 1 that the existence of such a relation comes with an exponential growth of  $m$  as  $p$  increases, a phenomenon that can be related to Dvoretzky's theorem when specialized to those Banach spaces [69].

From this new characterization, one can prove the following result.

**Theorem 1 ([52, 53]).** *Let  $k \geq 0$ ,  $2 \leq p < \infty$  and  $A \in \mathbb{R}^{m \times n}$  be a  $\text{RIP}_p(s, \delta_s, \mu_p)$  matrix for  $s \in \{k, 2k, 3k\}$  and some normalization constant  $\mu_p > 0$ . If*

$$\delta_{2k} + \sqrt{(1 + \delta_k)(\delta_{2k} + \delta_{3k})(p - 1)} < 1/3, \quad (21)$$

then, for any signal  $x \in \mathbb{R}^n$  observed according to the noisy sensing model  $y = Ax + n$  with  $\|n\|_p \leq \varepsilon_p$ , the unique solution  $\hat{x}_p$  obeys

$$\|x^* - x\| \leq 4 \frac{1}{\sqrt{k}} \sigma_k(x)_1 + 8 \varepsilon_p / \mu_p, \quad (22)$$

where, again,  $\sigma_k(x)_1$  denotes the best  $k$ -term approximation.

This theorem follows by generalizing the fundamental result of Candès in [26] to the particular geometry of Banach spaces  $\ell_p^m$ . It shows that, if  $A$  is  $\text{RIP}_p$  with particular requirement on the  $\text{RIP}_p$  constant, the  $\text{BPDQ}_p$  program is stable under both measurement noise corruption and departure from the strict sparsity model, as measured by  $e_0$ . In particular, under the same conditions, given a measurement noise  $\xi$  and some upper bounds  $\varepsilon_p$  on its  $\ell_p$ -norm, (22) provides the freedom to find the value of  $p$  that minimizes  $\varepsilon_p / \mu_p$ .

This is exactly how QCS signal recovery works. Following Theorem 1 and its stability result (22), we jointly determine a  $\text{RIP}_p$  sensing matrix with known value  $\mu_p$  and a tight error bound  $\varepsilon_p$  on the  $\ell_p$  norm of the residual  $q - Ax$  under HRA. The existence of a  $\text{RIP}_p$  matrix is guaranteed by the following result [52, 53].

**Proposition 1 (RIP<sub>p</sub> Matrix Existence).** *Let a random Gaussian sensing matrix  $A \in \mathbb{R}^{m \times n}$  be such that  $a_{ij} \sim_{\text{iid}} \mathcal{N}(0, 1)$ ,  $p \geq 1$  and  $0 \leq \eta < 1$ . Then,  $A$  is  $\text{RIP}_p(k, \delta_k, \mu_p)$  with probability higher than  $1 - \eta$  when we have jointly  $m \geq 2^{p+1}$  and*

$$m \geq m_0^{\max(p/2, 1)} \quad \text{with } m_0 = O(\delta_k^{-2} (k \log(\frac{n}{k}) + k \log(\delta_k^{-1}) + \log \frac{2}{\eta})). \quad (23)$$

Moreover,  $\mu_p = \Theta(m^{1/p} \sqrt{p+1})$ .

There is thus an exponential price to pay for a matrix  $A$  to be  $\text{RIP}_p$  as  $p$  increases: roughly speaking, for  $p \geq 2$ , we need  $m \geq m_0^{p/2} = O(k^{p/2} \log^{p/2}(n/k))$  measurements for satisfying this property with non-zero probability.

To estimate a tight value of  $\varepsilon_p$  in the case of quantization noise—since, under HRA  $\xi_j \sim_{\text{iid}} \mathcal{U}([-\Delta/2, \Delta/2])$ —we can show that

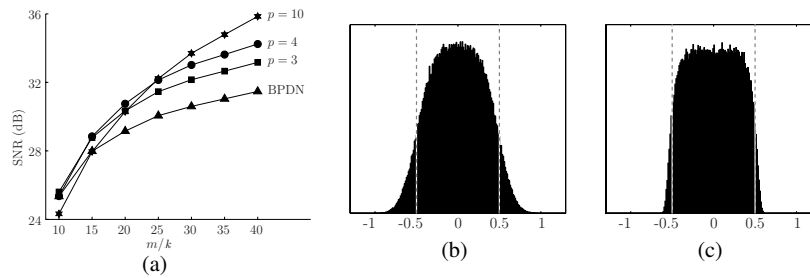
$$\|\xi\|_p \leq \varepsilon_p := \frac{\Delta}{2^{(p+1)^{1/p}}} (m + \zeta(p+1) \sqrt{m})^{\frac{1}{p}}, \quad (24)$$

with probability higher than  $1 - e^{-2\zeta^2}$ . Actually, for  $\zeta = 2$ ,  $x$  is a feasible solution of the  $\text{BPDQ}_p$  fidelity constraint with a probability exceeding  $1 - e^{-8} > 1 - 3.4 \times 10^{-4}$ .

Finally, combining the estimation  $\varepsilon_p$  with the bound on  $\mu_p$ , we find, under the conditions of Prop. 1,

$$\frac{\varepsilon_p}{\mu_p} \lesssim \frac{\Delta}{\sqrt{p+1}}. \quad (25)$$

This shows that, in the high *oversampled sensing scenario* driven by (23), and provided the  $\text{RIP}_p$  constants  $\{\delta_k, \delta_{2k}, \delta_{3k}\}$  satisfy (21), the part of the reconstruction error due to quantization noise behaves as  $O(\Delta/\sqrt{p+1})$ . This is also the error we get if  $x$  is exactly  $k$ -sparse since then  $e_0$  vanishes in (22).



**Fig. 4** (a) Quality of BPDQ $_p$  for different  $m/k$  and  $p$ . (b) and (c): Histograms of  $\Delta^{-1}(A\hat{x} - q)_i$  for  $p = 2$  and for  $p = 10$ , respectively.

If we solve for  $p$ , we can see that the error decays as  $O(\Delta/\sqrt{\log m})$  as  $m$  increases. There is possibly some room for improvements since, as explained in Sec. 2.2.2, the lower bound on reconstruction of sparse signal is  $\Omega(\Delta/m)$ . Beyond scalar quantization schemes, Sec. 4 will also show that much better theoretical error reduction can be expected using  $\Sigma\Delta$  quantization.

Interestingly, we can, however, observe a numerical gain in using BPDQ $_p$  for increasing values of  $p$  when the signal  $x$  is observed by the model (19) and when  $m$  increases beyond the minimal value  $m_0$  needed for stabilizing BPDN (*i.e.*, BPDQ $_2$ ).

This gain is depicted in Fig. 4. The plots on the left correspond to the reconstruction quality, *i.e.*, the value  $\text{SNR} = 20 \log(\|x\|/\|x - \hat{x}_p\|)$  expressed in dB, reached by BPDQ $_p$  for different values of  $p$  and  $m/k$ . The original signal  $x$  has dimension  $n = 1024$  and is  $k$ -sparse in the canonical basis, with support of size  $k = 16$  uniformly random and normally distributed non-zero coefficients. Each point of each curve represents average quality over 500 trials. For each sparse signal  $x$ ,  $m$  quantized measurements were recorded using (19) with a random Gaussian sensing matrix  $A$  and  $\Delta = \|Ax\|_\infty/40$ . The reconstruction was done by solving BPDQ $_p$  with the Douglas-Rachford algorithms [52], an efficient convex optimization method solving constrained programs, such as BPDQ $^1$ , using simpler *proximal operators* [30]. Fig. 4(a) shows that higher oversampling ratio  $m/k$  allows the use of higher  $p$  with significant gain in the reconstruction quality. However, if  $m/k$  is low, *i.e.*, close to  $m/k = 10$ , the best quality is still reached by BPDN. The quantization consistency of the reconstruction, *i.e.*, the original motivation for introducing the BPDQ $_p$  program, can also be tested. This is shown on Fig. 4(b) and Fig. 4(c) where the histograms of the components of  $\Delta^{-1}(A\hat{x}_p - q)$  are represented for  $p = 2$  and  $p = 10$  at  $m/k = 40$ . This histogram for  $p = 10$  is indeed closer to a uniform distribution over  $[-1/2, 1/2]$ , while the one at  $p = 2$  is mainly Gaussian.

<sup>1</sup> The code of BPDQ is freely available at <http://wiki.epfl.ch/bpdq>.

### 3.2 Non-Uniform Scalar Quantization

If the distribution of the measurements is known, quantization distortion can be decreased by adopting a non-uniform scalar quantizer. For instance, when  $A$  is a random Gaussian matrix viewing the signal as fixed and the matrix as randomly drawn, the distribution of the components of  $y = Ax$  is also Gaussian with a variance proportional to the signal energy  $\|x\|_2^2$  (and similarly, for other matrix constructions, such as ones drawn with random sub-Gaussian entries). Assuming the acquired signal energy can be fixed, *e.g.*, using some automatic gain control, the known distribution of the measurements can be exploited in the design of the quantizer, thanks for example to the Lloyd-Max algorithm mentioned in Sec. 2.2 [71]. In particular, the quantization thresholds and levels are then optimally adjusted to this distribution.

This section shows that the formalism developed in Sec. 3.1 can indeed be adapted to non-uniform scalar quantizer. To understand this adaptation, we exploit a common tool in quantization theory [39]: any non-uniform quantizer can be factored as the composition of a “compression” of the real line over  $[0, 1]$  followed by a uniform quantization of the result that is finally re-expanded on  $\mathbb{R}$ . Mathematically,

$$Q = \mathcal{G}^{-1} \circ Q_\Delta \circ \mathcal{G}, \quad (26)$$

where  $\mathcal{G} : \mathbb{R} \rightarrow [0, 1]$  is the *compressor* and  $\mathcal{G}^{-1} : [0, 1] \rightarrow \mathbb{R}$  is the *expander*, giving the name *compander* as a portemanteau.

In particular, under HRA, the compressor  $\mathcal{G}$  of a distortion optimal quantizer, *i.e.*, one that minimizes  $\mathbb{E}|X - Q(X)|^2$  for a source modeled as a random variable  $X$  with pdf  $\varphi$ , must satisfy

$$\frac{d}{d\lambda} \mathcal{G}(\lambda) = \left( \int \varphi^{1/3}(t) dt \right)^{-1} \varphi^{1/3}(\lambda),$$

and if  $Q$  is an optimal  $B$ -bit quantizer (*e.g.*, obtained by Lloyd-Max method) then  $\Delta = 2^{-B}$  in (26). In this case, the Panter and Dite formula estimates the quantizer distortion as [79]

$$\mathbb{E}|X - Q(X)|^2 \simeq_B \frac{2^{-2B}}{12} \|\varphi\|_{1/3} =: \sigma_{\text{PD}}^2,$$

with  $L_s$ -norm  $\|\varphi\|_s = (\int |\varphi^s(t)| dt)^{1/s}$  and where “ $\simeq_B$ ” means that the relation tends to an equality when  $B$  is large. The rest of this section assumes that the expected distribution is Gaussian, *i.e.*, if  $\varphi \sim \mathcal{N}(0, \sigma_0^2)$  and  $\|\varphi\|_{1/3} = \frac{1}{2}\sqrt{3}\pi\sigma_0^2$ , as it comes by seeing the signal fixed (with known energy) and the Gaussian matrix random in CS.

Compander theory generalizes quantization consistency in the “compressed” domain, *i.e.*,

$$|\mathcal{G}(\lambda) - \mathcal{G}(Q(\lambda))| \leq \Delta/2 = 2^{-B-1}.$$

Therefore, for the right compressor  $\mathcal{G}$ , in the noiseless QCS model (18), the signal  $x$  provides consistency constraints to be imposed on any reconstruction candidate  $x'$ :

$$\|\mathcal{G}(Ax') - \mathcal{G}(q)\|_\infty \leq \Delta/2 = 2^{-B-1}. \quad (\text{QC})$$

This generalizes the uniform quantization consistency (QC<sub>u</sub>) introduced in Sec. 3.1.

The compander formalism is leveraged in [53], to generalize the approach described in Sec. 3.1 to non-uniform quantization. In particular, a new set of parametric constraints are introduced, the  $p$ -Distortion Consistency (or D <sub>$p$</sub> C) for  $p \geq 2$ . These have for limit cases the QC above and the *distortion consistency* constraint (DC) arising from Panter and Dite formula, namely, the constraint imposing any reconstruction candidate  $x'$  to satisfy [31]

$$\|Ax' - q\|_2^2 \leq \varepsilon_{\text{PD}}^2 := m\sigma_{\text{PD}}^2, \quad (\text{DC})$$

with DC asymptotically satisfied by  $x$  when both  $B$  and  $m$  are large.

The D <sub>$p$</sub> C constraint corresponds to imposing that a candidate signal  $x'$  satisfies

$$\|Ax' - Q_p[q]\|_{p,w} = \|Ax' - Q_p[Ax]\|_{p,w} \leq \varepsilon_{p,w}, \quad (\text{D}_p\text{C})$$

where  $\|v\|_{p,w} = \|\text{diag}(w)v\|_p$  is the weighted  $\ell_p$ -norm of  $v \in \mathbb{R}^m$  with weights  $w \in \mathbb{R}_+^m$ , denoting by  $\text{diag}(w)$  the diagonal matrix having  $w$  on its diagonal. The mapping  $Q_p : \mathbb{R}^m \rightarrow \mathbb{R}^m$  is a post-quantization modification of  $q$  characterized componentwise hereafter and such that  $Q_p[q] = Q_p[Ax]$ .

Under HRA, a careful design of  $Q_p$ ,  $w$  and the bounds  $\varepsilon_{p,w}$  ensures that D<sub>2</sub>C amounts to imposing DC on  $x'$  and, that as  $p \rightarrow +\infty$ , D <sub>$p$</sub> C tends to QC [53]. Briefly, if  $q_i$  falls in the quantization bin  $\mathcal{C}_j$ ,  $Q_p(q_i)$  is defined as the minimizer of

$$\min_{\lambda \in \mathcal{C}_j} \int_{\mathcal{C}_j} |t - \lambda|^p \varphi(t) dt.$$

Actually,  $Q_2(q_i) = q_i$  by equivalence with (10), and  $\lim_{p \rightarrow \infty} Q_p(q_i) = \frac{1}{2}(t_j + t_{j+1})$ .

The weights are defined by the quantizer compressor  $\mathcal{G}$  with  $w_i(p) = \frac{d}{dx} \mathcal{G}(Q_p[q_i])^{\frac{p-2}{p}}$ . Moreover, under HRA and asymptotically in  $m$ , an optimal bound  $\varepsilon_p$  reads  $\varepsilon_{p,w}^p = m \frac{2^{-Bp}}{(p+1)2^p} \|\varphi\|_{1/3}$ . For  $p = 2$ ,  $\varepsilon_{2,w} = \varepsilon_{\text{PD}}$  matches the distortion power estimated by the Panter and Dite formula, while for  $p \rightarrow +\infty$ ,  $\varepsilon_{p,w} \rightarrow \frac{1}{2}2^{-B}$ , *i.e.*, half the size of the uniform quantization bins in the domain compressed by  $\mathcal{G}$ .

Similarly to Sec. 3.1, using (D <sub>$p$</sub> C) as a fidelity constraint in the signal reconstruction leads to the definition of a Generalized Basis Pursuit DeNoise program:

$$\hat{x}_{p,w} = \arg \min_{z \in \mathbb{R}^n} \|z\|_1 \text{ s.t. } \|Q_p(q) - Az\|_{p,w} \leq \varepsilon_{p,w}. \quad (\text{GBPDN}(\ell_{p,w}))$$

Ideally, we would like to directly set  $p = \infty$  in order to enforce consistency of  $\hat{x}_{p,w}$  with  $q$ . However, as studied in [53], it is not certain that this limit case minimizes the reconstruction error  $\|x - \hat{x}_{p,w}\|$  as a function of  $p$ , given a certain number of measurements  $m$ .

Actually, the stability of GBPDN can be established from the one of BPDQ (Sec. 3.1) if we impose  $A$  to satisfy the more general RIP $_{p,w,\ell_2}^m$ , as formally defined

in (20). Indeed, for any weighting vector  $w$ , we have always  $\|Q_p(q) - Az\|_{p,w} = \|q' - A'z\|_p$  with  $q' = \text{diag}(w)Q_p(q)$  and  $A' = \text{diag}(w)A$ . Therefore, we know from Theorem 1 that if  $A'$  is  $\text{RIP}_p$ , or equivalently if  $A$  is  $\text{RIP}_{\ell_{p,w}, \ell_2^m}$ , with the additional condition (21) on its RIP constants at different sparsity levels, then the solution of  $\text{GBPDN}(\ell_{p,w})$  will be stable in the sense of (22), *i.e.*,

$$\|\hat{x}_{p,w} - x\| \lesssim \frac{\varepsilon_{p,w}}{\mu_{p,w}} + \frac{\sigma_k(x)_1}{\sqrt{k}}.$$

Compared to the unit weights case (as involved by the  $\text{RIP}_p$ ), a random Gaussian matrix  $A$  with  $a_{ij} \sim_{\text{iid}} \mathcal{N}(0, 1)$  satisfies the  $\text{RIP}_{\ell_{p,w}, \ell_2^m}^m(k, \delta_k, \mu_{p,w})$  with high probability provided that  $m$  grows like  $O((\theta_p \delta_k^{-2} (k \log(n/k))^{p/2})$ . The ratio  $\theta_p := \|w\|_\infty / (m^{-1/p} \|w\|_p)$  depends on the *conditioning* of  $w$ . It is equal to 1 for constant weights (recovering (23)), while it increases with the dynamic range of  $w$ . For the weight  $w(p)$  defined previously and with a Gaussian optimal quantizer,  $\theta_p^{p/2} \simeq_{m,B} \sqrt{p+1}$  asymptotically in  $m$  and  $B$ .

As for the uniform case, a strong (polynomial) oversampling in  $m$  is thus required for satisfying the  $\text{RIP}_{\ell_{p,w}, \ell_2^m}^m$  at  $p > 2$  compared to the minimal number of measurements needed at  $p = 2$ . However, an asymptotic analysis of  $\varepsilon_{p,w} / \mu_{p,w}$  shows that the GBPDN reconstruction error due to quantization for a Gaussian sensing matrix behaves like [53]

$$\|\hat{x}_{p,w} - x\| \lesssim \frac{2^{-B}}{\sqrt{p+1}} + \frac{\sigma_k(x)_1}{\sqrt{k}},$$

This error decay is thus similar to the one found in (25) for uniform QCS with now a direct interpretation in terms of the quantizer bit-depth  $B$ .

Efficient convex optimization methods, like those relying on proximal algorithms [30], can also be used to numerically solve GBPDN. In [53], numerical simulations show that the reconstruction qualities reached in the reconstruction of sparse signals from their non-uniformly quantized measurements behave similarly, with respect to  $p$  and  $m$ , to those observed in Sec. 3.1 for the uniformly quantized CS setting.

We should also remark that beyond QCS, the stability of GBPDN (when  $A$  is  $\text{RIP}_{p,w}$ ) can also be used for reconstructing signals acquired under a (heteroscedastic) noisy sensing model  $y = Ax + \xi$  where  $\xi \in \mathbb{R}^m$  is an additive generalized Gaussian noise with bounded  $\ell_{p,w}$ -norm for some specific weight  $w \in \mathbb{R}_+^m$  [92, 53].

### 3.3 Finite-Range Scalar Quantizer Design

So far we have only considered a scalar quantizer model without saturation. Practical scalar quantizers have a finite range, which implies a saturation level  $\pm S$  and, using  $B$  bits per coefficient, a quantization interval equal to

$$\Delta = S2^{-B+1}. \quad (27)$$

In order to determine the optimal saturation rate, the system designed needs to balance the loss of information due to saturation, as  $S$  decreases, with the increased quantization error due to an increasing quantization interval in (27), as  $S$  increases. In classical systems, this balance requires setting the quantization level relatively close to the signal amplitude to avoid saturation. On the other hand, in compressive sensing systems, the incoherence of the measurements with the sparsity basis of the signal makes them more robust to loss of information and enables higher saturation levels with smaller quantization intervals.

A key property of compressive measurements, which provides the robustness to loss of information, is *democracy*. Intuitively, each measurement contributes an equal amount of information to the reconstruction. If the signal is slightly oversampled, relative to the rate required for CS reconstruction, then any subset with enough measurements should be sufficient to recover the signal. The notion of democracy was first introduced in [23, 43] in the context of information carried in each bit of the representation; the definition below strengthens the concept and formulates it in the context of compressive sensing [33, 66].

**Definition 2.** Let  $A \in \mathbb{R}^{m \times n}$ , and let  $\tilde{m} \leq m$  be given. We say that  $A$  is  $(\tilde{m}, k, \delta_k)$ -democratic if, for all row index sets  $\Gamma$  such that  $|\Gamma| \geq \tilde{m}$ , any matrix  $\tilde{A} = ((A^T)_\Gamma)^T$ , *i.e.*, comprised of a  $\Gamma$ -subset of the rows of  $A$ , satisfies the RIP of order  $k$  with constant  $\delta_k$ .

This definition takes an adversarial view of democracy: a matrix  $A$  is democratic if an adversary can pick any  $d = m - \tilde{m}$  rows to remove from  $A$ , and the remaining matrix still satisfies the RIP. This is a much stronger guarantee than just randomly selecting a subset of the rows to be removed. Such a guarantee is important in the case of saturation robustness because the saturated measurements are the largest ones in magnitude, *i.e.*, potentially the ones most aligned with the measured signal and, presumably, the ones that capture a significant amount of information. Still, despite this strict requirement, randomly generated matrices can be democratic if they have a sufficient number of rows.

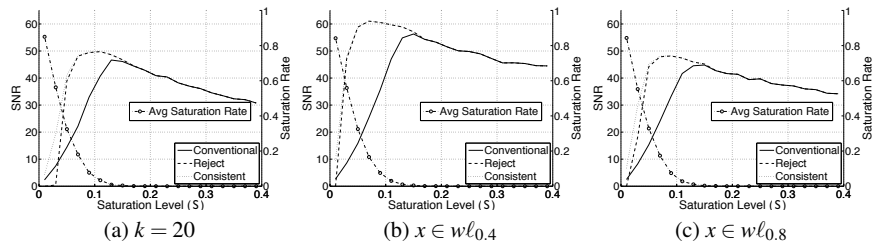
**Theorem 2 ([33]).** Let  $A \in \mathbb{R}^{m \times n}$  with elements  $a_{ij}$  drawn according to  $\mathcal{N}(0, \frac{1}{m})$  and let  $\tilde{m} \leq m$ ,  $k < \tilde{m}$ , and  $\delta \in (0, 1)$  be given. Define  $d = m - \tilde{m}$ . If

$$m = C_1(k + d) \log \left( \frac{n + m}{k + d} \right), \quad (28)$$

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

The practical implication of democratic measurements is that information loss due to saturated measurements can be tolerated.

Saturated measurements are straightforward to detect, since they quantize to the highest or the lowest level of the quantizer. The simplest approach is to treat saturated measurements as corrupted, and reject them from the reconstruction, together with the corresponding rows of  $A$ . As long as the number of saturated measurements



**Fig. 5** Saturation performance using  $\ell_1$  minimization for (a) exactly  $k = 20$ -sparse signals and compressible signals in weak  $\ell_p$  for (b)  $p = 0.4$  and (c)  $p = 0.8$ , with  $n = 1024$ ,  $m = 384$ , and  $B = 4$ . The reconstruction SNR as a function of the saturation level is measured on the left y-axis assuming (solid line) conventional reconstruction, *i.e.*, ignoring saturation, (dotted line) enforcing saturation consistency, and (dashed line) rejecting saturated measurements. The dashed-circled line, measured on the right y-axis, plots the average saturation rate given the saturation level.

is not that large, the RIP still holds and reconstruction is possible using any sparse reconstruction algorithm.

However, saturated measurements do contain the information that the measurement is large. In the context of consistent reconstruction, they can be used as constraints in the reconstruction process. If a measurement  $i$  is positively saturated, then we know that  $(Ax)_i \geq S - \Delta$ . Similarly, if it is negatively saturated,  $(Ax)_i \leq -S + \Delta$ . These constraints can be imposed on any reconstruction algorithm to improve performance [66].

Fig. 5 demonstrates the effect of each approach. As demonstrated in the plots, rejecting saturated measurements or treating them as consistency constraints significantly outperforms just ignoring saturation. Furthermore, if saturation is properly taken into account, a distortion optimal finite-range scalar quantizer should be designed with significant saturation rate, often more than 20%. While the figures suggest that saturation rejection and saturation consistency have very similar performance, careful examination demonstrates, as expected, that consistency provides more robustness in a larger range of saturation rates and conditions. A more careful study and detailed discussion can be found in [66]. Furthermore, further gains in the bit-rate can be achieved by coding for the location of the saturated measurements and transmitting those separately [58].

### 3.4 1-Bit Compressive Sensing

The simplest scalar quantizer design to implement in hardware is a 1-bit quantizer, which only computes the sign of its input. Its simplicity makes it quite appealing for compressive sensing systems.

The sensing model of 1-bit CS, first introduced in [19], is very similar to the standard scalar quantization model



$$q = \text{sign}(Ax), \quad (29)$$

where  $\text{sign}(x_i)$  is a scalar function applied element-wise to its input and equals 1 if  $x_i \geq 0$  and  $-1$  otherwise.

One of the challenges of this model is that it is invariant under changes of the signal amplitude since  $\text{sign}(cx) = \text{sign}(x)$  for any positive  $c$ . For that reason, enforcing consistency is not straightforward. A signal can be scaled arbitrarily and still be consistent with the measurements. Thus, a magnitude constraint is typically necessary. Of course, the signal can only be recovered within a positive scalar factor.

Similarly to multi-bit scalar quantization models, the literature in this area focuses on deriving lower bounds for the achievable performance, reconstruction guarantees, as well as practical algorithms to invert this problem.

### 3.4.1 Theoretical Performance Bounds

A lower bound on the achievable performance, can be derived using a similar analysis as in Sec. 2.2.3. The main difference is that the quantization cells are now orthants in the  $m$ -dimensional space, shown in Fig. 6(a), corresponding to each measured sign pattern. Each subspace of the  $\binom{n}{k}$  possible ones intersects very few of those orthants, as shown in Fig. 6(b), *i.e.*, uses very few quantization points. In total, at most  $I \leq 2^k \binom{n}{k} \binom{m}{k}$  quantization cells are intersected by the union of all subspaces [54].

Since the signal amplitude cannot be recovered, the lower bound is derived on  $k$ -dimensional spheres and coverings using spherical caps instead of balls. The derivation ensures that the spherical caps have radius sufficiently large to cover the  $\binom{n}{k}$  spheres. Despite the similarity to the argument in Sec. 2.2.3, this case requires a little bit more care in the derivation; details can be found in [54]. Still, the result is very similar in nature. Defining  $\Sigma_k^* := \{x \in \Sigma_k, \|x\|_2 = 1\}$ , we have:

**Theorem 3 ([54]).** *Given  $x \in \Sigma_k^*$ , any estimation  $\hat{x} \in \Sigma_k^*$  of  $x$  obtained from  $q = \text{sign}(Ax)$  has a reconstruction error of at least*

$$\|\hat{x} - x\| \gtrsim \frac{k}{m + k^{3/2}},$$

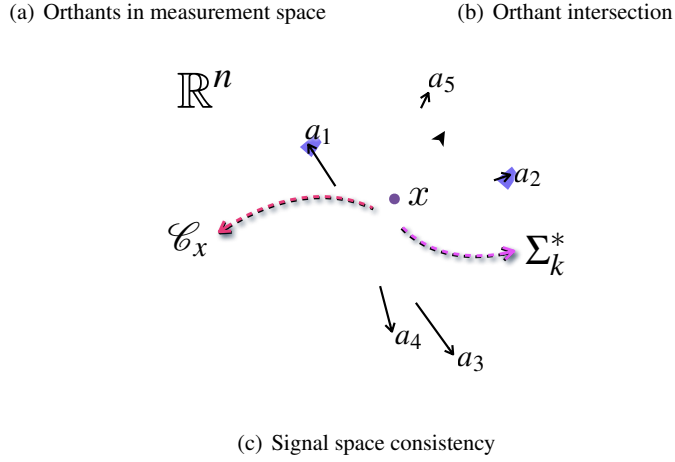
which is on the order of  $\frac{k}{m}$  as  $m$  increases.

If the sensing matrix  $A$  is Gaussian, *i.e.*, if  $a_{ij} \sim_{\text{iid}} \mathcal{N}(0, 1)$ , any  $k$ -sparse signal that has consistent measurements will not be very far from the signal producing the measurements, assuming a sufficient number of them. This guarantee approaches the lower bound of Theorem 3 within a logarithmic factor.

**Theorem 4 ([54]).** *Fix  $0 \leq \eta \leq 1$  and  $\epsilon_o > 0$ . If the number of measurements is*

$$m \geq \frac{2}{\epsilon_o} \left( 2k \log(n) + 4k \log\left(\frac{17}{\epsilon_o}\right) + \log\left(\frac{1}{\eta}\right) \right), \quad (30)$$

then for all  $x, x' \in \Sigma_k^*$  we have that



**Fig. 6** Behavior of 1-bit measurements. (a) The measurement space,  $\mathbb{R}^m$  is separated to high-dimensional orthant, according to the sign of each orthant. (b) Signals in a  $k$ -dimensional space ( $k < m$ ) will only map to a  $k$  dimensional subspace of  $\mathbb{R}^m$  and intersect only a few orthants of the measurement space. (c) The same behavior in the signal space. Each measurement vector defines its orthogonal hyperplane. The measurement sign identifies which side of the hyperplane the signal lies on; all signals in the shaded region have consistent measurements. Newer measurements provide less and less information; the chance of intersecting the consistency region decreases.

$$\|x - x'\|_2 > \epsilon_0 \Rightarrow \text{sign}(Ax) \neq \text{sign}(Ax'), \quad (31)$$

with probability higher than  $1 - \eta$ . Equivalently, if  $m$  and  $k$  are given, solving for  $\epsilon_0$  above leads to

$$\|x - x'\|_2 \lesssim \frac{k}{m} \log \frac{m}{k}, \quad (32)$$

with the same probability.

Fig. 6(c) provides further intuition on these bounds by illustrating how 1-bit measurements operate in the signal space. Specifically, each measurement corresponds to a hyperplane in the signal space, orthogonal to the measurement vector.

The measurement sign determines on which side of the hyperplane the signal lies. Furthermore, the signal is sparse, *i.e.*, lies in  $\Sigma_k^*$ . A consistent sparse reconstruction algorithm can produce any sparse signal in the indicated shaded region.

A new measurement provides new information about the signal only if the corresponding hyperplane intersects the region of consistent signals and, therefore, shrinks it. However, as more measurements are obtained and the consistency region shrinks, newer measurements have lower probability of intersecting that region and providing new information, leading to the  $1/m$  decay of the error.

Consistency can be quantified using the normalized hamming distance between measurements

$$d_H(q, q') = \frac{1}{m} \sum_i q_i \oplus q'_i,$$

where  $\oplus$  denotes the exclusive-OR operator. It is, thus, possible to show that if  $x$  and  $x'$  above differ by no more than  $s$  bits in their 1-bit measurements, *i.e.*, if  $d_H(\text{sign}(Ax), \text{sign}(Ax')) \leq s/m$ , then, with  $m \gtrsim \frac{1}{\varepsilon_0} k \log \max(m, n)$  and with high probability [51],

$$\|x - x'\|_2 \leq \frac{k+s}{k} \varepsilon_0.$$

A bound similar to (32) exists for sign measurements of non-sparse signals in the context of quantization using frame permutations [76]. In particular, reconstruction from sign measurements of signals exhibits (almost surely) an asymptotic error decay rate arbitrarily close to  $O(1/m)$ . However, in contrast to Theorem 4 this result holds only for a fixed signal and not uniformly for all signals of interest.

Note that these results focus on matrices generated using the normal distribution. It has been shown that matrices generated from certain distributions do not perform well in this setting, even though they can be used in standard compressive sensing [82]. For instance, consider a random Bernoulli matrix  $A$  such that  $a_{ij} = 1$  or  $-1$  with equal probability. In this case, the two distinct sparse vectors  $(1, 0, \dots, 0)^T$  and  $(1, \lambda, 0, \dots, 0)^T$  with  $0 \leq \lambda < 1$  are  $\lambda$  apart and they generate the same quantization vector  $q = \text{sign}(A_1)$ , where  $A_1$  is the first column of  $A$ . It is not possible, therefore, to distinguish those two vectors from their 1-bit observations by increasing  $m$  and guarantee that the reconstruction error will decay as measurements increase. This counterexample, however, is exceptional in the sense that such failures can only happen if the signal can have a very large entry. Under mild flatness assumptions on the  $\ell_\infty$ -norm of the signal, arbitrary subgaussian measurements can be utilized [1].

These results establish lower and upper bounds on distances between two sparse signals that have (almost) consistent 1-bit measurements. It is also possible to provide an embedding guarantee similar to the RIP [25]. Since the measurement does not preserve the signal magnitude, we should not expect distances of signals to be preserved. However, the measurements do preserve angles between signals. Defining  $d_S(u, v) = \frac{1}{\pi} \arccos(u^T v)$ ,  $u, v \in S^{n-1}$ , we have:

**Theorem 5 (Binary  $\varepsilon$ -Stable Embedding (B $\varepsilon$ SE) [54]).** *Let  $A \in \mathbb{R}^{m \times n}$  be a random Gaussian matrix such that  $a_{ij} \sim_{\text{iid}} \mathcal{N}(0, 1)$ . Fix  $0 \leq \eta \leq 1$  and  $\varepsilon > 0$ . If the number of measurements satisfies*

$$m \geq \frac{2}{\varepsilon^2} (k \log(n) + 2k \log(\frac{35}{\varepsilon}) + \log(\frac{2}{\eta})), \quad (33)$$

then with probability exceeding  $1 - \eta$

$$d_S(x, x') - \varepsilon \leq d_H(\text{sign}(Ax), \text{sign}(Ax')) \leq d_S(x, x') + \varepsilon, \quad (34)$$

for all  $x, x' \in \Sigma_k^*$ .

In other words, up to an additive distortion that decays as  $\varepsilon \lesssim (\frac{k}{m} \log \frac{mn}{k})^{1/2}$ , the Hamming distance between  $\text{sign}(Ax)$  and  $\text{sign}(Ax')$  tends to concentrate around the angular distance between  $x$  and  $x'$ . Notice that, in contrast to the RIP, a vanishing distance between the quantized measurements of two signals does not imply they are equal, *i.e.*, we observe a (restricted) *quasi-isometry* between  $\Sigma_k^*$  and  $\text{sign}(A\Sigma_k^*)$  instead of the common RIP [49]. This comes from the additive nature of the distortion in (34) and is a direct effect of the inherent ambiguity due to quantization.

This embedding result has been extended to signals belonging to convex sets  $\mathcal{H} \subset \mathbb{R}^n$  provided that their Gaussian mean width

$$w(\mathcal{H}) = \mathbb{E} \sup\{u^T g : u \in \mathcal{H} - \mathcal{H}\}, \quad g \sim \mathcal{N}(0, \mathbf{I}_{n \times n}), \quad (35)$$

with  $\mathcal{H} - \mathcal{H} := \{v - v' : v, v' \in \mathcal{H}\}$ , can be computed [80, 81, 82]. In particular, if

$$m \geq C\varepsilon^{-6} w^2(\mathcal{H})$$

for some constant  $C > 0$ , then (34) holds with high probability for any  $x, x' \in \mathcal{H} \cap S^{n-1}$ . In particular, for

$$\mathcal{H} = K_{n,k} := \{u \in \mathbb{R}^n : \|u\|_1 \leq k^{1/2}, \|u\|_2 \leq 1\},$$

since  $w^2(K_{n,k}) = O(k \log n/k)$  [81], an embedding exists between the set of compressible vectors modeled by  $K_{n,k}$  and  $\{-1, +1\}^m$  provided that  $m \geq C\varepsilon^{-6} k \log n/k$ .

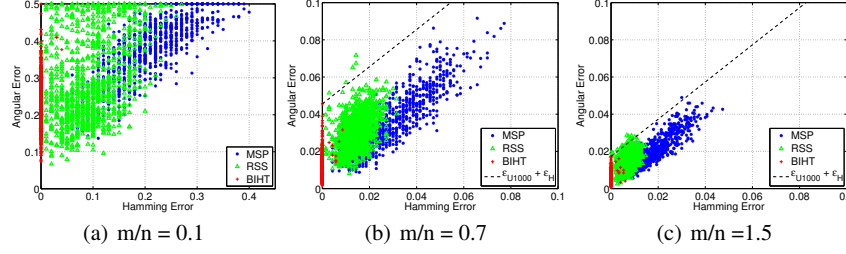
Note that generalizations of these embeddings to non-linear functions other than the sign operator, or to stochastic processes whose expectation is characterizable by such functions, are also possible [81].

### 3.4.2 Reconstruction from 1-Bit Measurements

The original efforts in reconstructing from 1-bit measurements enforced  $\|x\|_2 = 1$  as a reconstruction constraint, formulating the non-convex  $\ell_1$  minimization problem

$$\hat{x} = \arg \min_x \|x\|_1, \text{ s.t. } q = \text{sign}(Ax), \|x\|_2 = 1. \quad (36)$$

Even though the problem is not convex, a number of algorithms have been shown experimentally to converge to the solution [19, 67]. More recently, a number of greedy algorithmic alternatives have also been proposed [15, 3, 54].



**Fig. 7** Angular reconstruction error  $\epsilon_S = d_S(x, \hat{x})$  vs. consistency error  $\epsilon_H = d_H(\text{sign}(A\hat{x}), q)$  for different greedy reconstructions (MSP, RSS and BITH). BIHT returns a consistent solution in most trials. When  $A$  is a B $\epsilon$ SE, (34) predicts that the angular error  $\epsilon_S$  is bounded by the hamming error  $\epsilon_H$  (and conversely) in addition to an offset  $\epsilon$ . This phenomenon is confirmed by an experimental linear trend (in dashed) between the two errors that improves when  $m/n$  increases [54].

Most of these algorithms attempt to enforce consistency by introducing a one-sided penalty for sign violations

$$J(Az, q) = \|(q \circ Az)_-\|_q, \quad (37)$$

where  $\circ$  is the element-wise product between vectors,  $(y_i)_- = y_i$  if  $y_i$  is negative and 0 otherwise, also applied element-wise, and the  $\ell_q$  norm is typically the  $\ell_1$  or the  $\ell_2$  norm. Typically, a descent step is performed using the gradient of (37), followed by a support identification and sparsity enforcement step. Often, care is taken in selecting the descent step, especially considering the signal is on the unit  $\ell_2$  sphere [67]. Assuming a certain noise level, a maximum likelihood formulation can also be used to remove the norm constraint [3].

For example, the Binary IHT (BIHT), a variation of the popular Iterative Hard Thresholding (IHT) [11], uses the one-sided  $\ell_1$  norm in (37) and follows its subgradient  $\frac{1}{2}A^T(q - \text{sign}(Az))$ . The algorithm is defined by the iteration

$$z^{n+1} = \mathcal{H}_k(z^n + \frac{1}{2}A^T(y - \text{sign}(Az^n))), \quad z^0 = 0, \quad (38)$$

where  $\mathcal{H}_k(\cdot)$  is a hard threshold, keeping the largest  $k$  coefficients of its input and setting the remaining ones to zero.

The BIHT does not have convergence or reconstruction guarantees to a consistent output. Still, as shown in Fig. 7, it works surprisingly well compared to other greedy approaches. Moreover, variations exist to make it more robust to potential binary errors in the knowledge of  $q$  [54] or to extend it to multi-bit scalar quantization [51].

The first iteration of BIHT is a simple truncated back-projection,  $\hat{x}_0 = \mathcal{H}_k(A^T q)$  whose distance to  $x$  is known to decay asymptotically as  $\sqrt{k/m}$  for a Gaussian matrix  $A$  [51, 3]. Furthermore,  $\hat{x}_0$  matches the solution of the (feasible) problem

$$\arg \max_z q^T Az \text{ s.t. } z \in \Sigma_k^*,$$

where maximizing  $q^T Az$  also promotes the 1-bit consistency of  $z$  with  $q$ .

This optimization can be generalized to any convex sets  $\mathcal{K} \subset \mathbb{R}^n$  where  $x$  can lie, such as the set  $\mathcal{K} = K_{n,k}$  of compressible signals with Gaussian width  $w(K_{n,k}) \asymp k \log n/k$  [81]. If  $m \geq C\epsilon^{-2}w(\mathcal{K})^2$  for some  $C > 0$ , and a fixed  $x$  is sensed using (29) with a Gaussian sensing matrix  $A$ , then the solution to

$$\hat{x} = \arg \max_z q^T Az \text{ s.t. } z \in \mathcal{K},$$

satisfies  $\|\hat{x} - x\|^2 = O(\epsilon)$  with high probability. Interestingly, under certain conditions, this holds also for sensing models other than (29), where the sign operator is replaced, for instance, by the logistic function [81].

What makes it difficult to provide reconstruction error estimates for algorithms motivated by the problem (36) is the non-convex constraint  $\|x\|_2 = 1$ , whose convex relaxation allows for the zero solution and is hence meaningless. To overcome this obstacle, it has been proposed in [80, 81] to impose a norm constraint to prevent trivial solutions on the measurements rather than the signal. This results in a different problem, which allows for a meaningful convex relaxation. Namely, since  $q = \text{sign}(Ax)$ , it follows that at the solution  $q^T(Ax) = \|Ax\|_1$ . Thus, by constraining this norm, the following convex problem can be formulated:

$$\hat{x} = \arg \min_x \|x\|_1, \text{ s.t. } q = \text{sign}(Ax), q^T Ax = 1 \quad (39)$$

As shown in [80], the problem in (39) does allow for reconstruction guarantees: If  $m \sim \epsilon^{-5}k \log(n/k)$ , the solution  $\hat{x}$  recovered from quantized Gaussian measurements of a sparse signal  $x$  is such that  $d_S(x, \hat{x}) \leq \epsilon$  with high probability. This holds uniformly for all signals  $x \in \mathbb{R}^n$ . Under flatness assumptions on the signal, recovery guarantees can also be proved for arbitrary subgaussian measurements [1].

### 3.5 Noise, Quantization and Tradeoffs

The sections above were focused on noiseless QCS models. These models only consider the statistical or the geometrical properties of quantization of CS measurements under high or low resolution modes. However, any signal acquisition system is subject to noise corruption before quantization, either on the measurement process or on the signal itself. Such noise can be incorporated in a more general model

$$q = Q(A(x + \xi_x) + \xi_s), \quad (40)$$

where  $\xi_x \in \mathbb{R}^n$  and  $\xi_s \in \mathbb{R}^m$  corrupt the signal and the sensing, respectively, before quantization. Examining the impact of such noise in signal recovery leads to new interesting questions.

In [93] two efficient reconstruction methods are developed for sparse or compressible signals sensed according (40) under sensing noise only, *i.e.*,  $\xi_x = 0$ . The

two approaches are mainly numerical: one relies on a maximum likelihood formulation, built on the quantization model and on a known Gaussian noise distribution, the other follows a least square principle. The two resulting methods are both regularized by an  $\ell_1$ -norm accounting for sparse signal prior. A provably convergent procedure inherited from a fixed point continuation method is used for reconstructing the signal in the two possible frameworks. With their approach, the combined effects of noise and coarse quantization can be jointly handled. Reasonable reconstruction results are achieved even using 1 or 2 bits per measurement.

The case  $\xi_x \neq 0$ ,  $\xi_s = 0$  boils down to an interaction of the well-understood phenomenon of *noise folding* in CS [34] and quantization [68]. Noise-folding in unquantized CS says that under a weak assumption of orthogonality between the rows of  $A$ , the variance of the component  $A\xi_x$  undergoes a multiplication by  $n/m$  compared to the variance  $\sigma_{\xi}^2$  of  $\xi_x$ . This impacts directly the reconstruction error of signals. The corresponding MSE is then  $n/m$  times higher than the noise power, or equivalently, the SNR loses 3 dB each time  $m$  is divided by 2 [34].

An extension of this result to noisy QCS has been provided in [68], assuming the sensing matrix  $A$  is RIP of order  $k$  and constant  $\delta$ . In this case, if  $\xi_x$  is standard normally distributed and if the quantizer has resolution  $B$ , then, under a random signal model where the signal support  $T$  is chosen uniformly at random in  $\{1, \dots, n\}$  and the amplitudes of the non-zero coefficients are standard normally distributed,

$$(1 - \delta)\mathbb{E}\|x - \hat{x}\|^2 = 2^{-2B+1} \frac{k}{m} \mathbb{E}\|x\|^2 + 2(2^{-2B} + 1) \frac{n}{m} \mathbb{E}\|\xi_x|_T\|^2 + km\kappa, \quad (41)$$

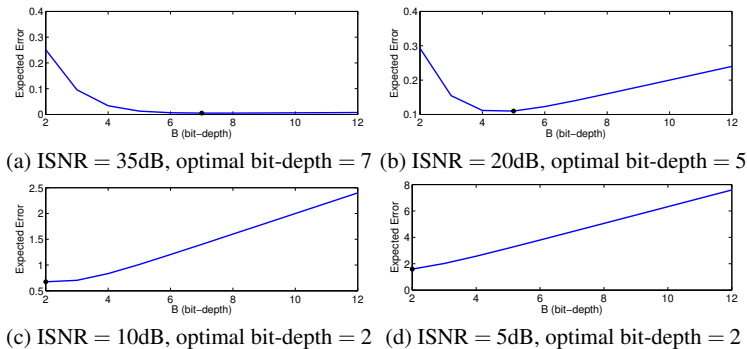
where  $\hat{x} = (A_T^\dagger q)_T$  is the oracle-assisted reconstruction of  $x$  knowing the support  $T$  of  $x$  for each of its realization, and

$$\kappa = \max_{i \neq j} |\mathbb{E}Q(a_i^T(x + \xi_x))Q(a_j^T(x + \xi_x))|,$$

measures the worst correlation between distinct quantized measurements.

In (41), the first term accounts for the quantization error of the signal itself, while the second term represents both the error due to folded signal noise as well as the quantization of that noise. Finally, the third term reflects a distortion due to correlation between quantized measurement. It is expected to be negligible in CS scenarios, specially when  $B$  increases or if a *dithering* is added to  $Q$  [39].

Numerical study of (41) shows that, at constant rate  $R = mB$ , a tradeoff can be expected between a *measurement compression* (MC) regime, where  $m$  is small (but still high enough to guarantee  $A$  to be RIP) and  $B$  is high, and a *quantization compression* (QC) regime, where  $m$  is high compared to the standard CS setting but  $B$  is small. Interestingly, the optimal bit-depth  $B$ , minimizing the expected reconstruction error, depends on the input SNR:  $\text{ISNR} = 20 \log_{10} \|x\| / \|\xi_x\|$ . This is illustrated in Fig. 8 where the evolution of (41) (discarding the effect of the third term) is plotted for four different noise scenarios. The optimal bit depth decays smoothly with the ISNR, suggesting that the QC regime is preferable at low ISNR while MC is clearly better at high ISNR. The general behavior of Fig. 8 is also confirmed on Monte Carlo error estimation of the oracle-assisted reconstruction defined above [68].



**Fig. 8** Upper bound on the oracle-assisted reconstruction error as a function of bit-depth  $B$  and ISNR at constant rate  $R = mB$  [68]. The black dots denote the minimum point on each curve.

## 4 Sigma-Delta Quantization for Compressive Sensing

As mentioned in the introduction,  $\Sigma\Delta$  quantization for compressed sensing fundamentally builds on corresponding schemes for finite frames. Thus before presenting an analysis specific to compressed sensing, we first discuss the finite frame case.

### 4.1 $\Sigma\Delta$ Quantization for Frames

Let  $\Phi \in \mathbb{R}^{n \times N}$  with columns  $\{\phi_j\}_{j=1}^N$  be a frame in the sense of (1.32) and consider the frame expansion

$$c = \Phi^T x$$

of a signal  $x \in \mathbb{R}^n$ . The goal is now to quantize  $c$  as a whole such that the quantized representation  $q$  allows for approximate recovery of  $x$ .  $\Sigma\Delta$  quantization schemes obtain such a  $q$  using a recursive procedure, which we will now explain in detail.

At the core of the schemes is a uniform scalar quantizer  $Q$ , which maps a real number to the closest point in a codebook of the form

$$\mathcal{Q} = \{(\pm j - 1/2)\Delta, j \in \{1, \dots, L\}\}. \quad (42)$$

A  $\Sigma\Delta$  scheme applies such a quantizer sequentially to the entries of  $c$ , taking in each quantization step the errors made in  $r$  previous steps into account. The complexity parameter  $r$  is referred to as the order of the  $\Sigma\Delta$  scheme; it quantifies the trade-off between required storage and achievable accuracy.

A first order  $\Sigma\Delta$  quantization scheme, the simplest such algorithm, hence retains the error only for one step. In the following formalization associated with the so-called *greedy* first order  $\Sigma\Delta$  scheme, the error parameter appears as the state variable  $u_i$ ; it measures the total accumulated error up to step  $i$ . The quantized frame



coefficient vector  $q \in \mathcal{Q}^N$  is computed by running the iteration

$$\begin{aligned} q_i &= Q(u_{i-1} + c_i) \\ u_i &= u_{i-1} + c_i - q_i. \end{aligned} \quad (43)$$

As initialization, one typically uses  $u_0 = 0$ . In matrix-vector notation, the above recurrence relation reads

$$Du = c - q. \quad (44)$$

Here  $D \in \mathbb{R}^{N \times N}$  is the finite difference matrix with entries given in terms of the Kronecker delta by  $D_{ij} = \delta_{i,j} - \delta_{i+1,j}$ , that is,

$$D = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ -1 & 1 & 0 & & 0 \\ 0 & -1 & 1 & & 0 \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & -1 & 1 \end{pmatrix}. \quad (45)$$

The scheme is explicitly designed such that each  $q_j$  partly cancels the error made up to  $q_{j-1}$ . When the signal is approximated as  $\tilde{\Phi}q$  using a dual frame  $\tilde{\Phi} \in \mathbb{R}^{n \times N}$  with columns  $\{\tilde{\phi}_j\}_{j=1}^N$ , this entails that one seeks to compensate an error in the direction of a dual frame vector  $\tilde{\phi}_{j-1}$  using a distortion in the direction of the next dual frame vector  $\tilde{\phi}_j$ . This serves as a motivation to choose a smoothly varying dual frame, *i.e.*, with subsequent dual frame vectors close to each other.

Bounding the reconstruction error using (44) in terms of the operator norm  $\|A\|_{2 \rightarrow 2} := \sup_{\|x\|_2 \leq 1} \|Ax\|_2$ , one obtains

$$\|x - \tilde{\Phi}q\|_2 = \|\tilde{\Phi}(c - q)\|_2 = \|\tilde{\Phi}Du\|_2 \leq \|\tilde{\Phi}D\|_{2 \rightarrow 2} \|u\|_2.$$

The smoothness intuition is reflected in the fact that the columns of  $\tilde{\Phi}D$  are given by  $\tilde{\phi}_j - \tilde{\phi}_{j-1}$ . Thus more precisely, finding a smooth dual frame  $\tilde{\Phi}$  amounts to minimizing  $\|\tilde{\Phi}D\|_{2 \rightarrow 2}$ .

If one is willing to store more than one previous value of the state variable, that is, to consider a higher order  $\Sigma\Delta$  scheme, it is possible to profit from higher order smoothness of the dual frame. Such a generalization of (43) is the greedy  $r$ -th order  $\Sigma\Delta$  scheme, which is associated with the recurrence relation

$$D^r u = c - q. \quad (46)$$

Here, the iteration to compute the quantized coefficients is explicitly given by

$$\begin{aligned}
q_i &= \mathcal{Q}\left(\sum_{j=1}^r (-1)^{j-1} \binom{r}{j} u_{i-j} + c_i\right) \\
u_i &= \sum_{j=1}^r (-1)^{j-1} \binom{r}{j} u_{i-j} + c_i - q_i.
\end{aligned} \tag{47}$$

As before, one initializes  $u_i = 0$ ,  $i \leq 0$ . The reconstruction error is now bounded by

$$\|x - \tilde{\Phi}q\|_2 = \|\tilde{\Phi}(c - q)\|_2 = \|\tilde{\Phi}D^r u\|_2 \leq \|\tilde{\Phi}D^r\|_{2 \rightarrow 2} \|u\|_2. \tag{48}$$

Examining (48), it is advantageous to choose a dual frame that minimizes  $\|\tilde{\Phi}D^r\|_{2 \rightarrow 2}$ , and a  $\Sigma\Delta$  scheme that yields a state-variable sequence with well bounded  $\|u\|_2$ . This motivates the following definitions.

**Definition 3.** Let  $\Phi \in \mathbb{R}^{n \times N}$  be a frame and  $r$  be a positive integer. Then the  $r$ -th order *Sobolev dual* of  $\Phi$  is given by

$$\tilde{\Phi}^{(r)} := \arg \min \|\tilde{\Phi}D^r\|_{2 \rightarrow 2} = (D^{-r}\Phi)^\dagger D^{-r}, \tag{49}$$

where the minimum is taken over all dual frames of  $\Phi$ .

**Definition 4.** A  $\Sigma\Delta$  scheme with a codebook  $\mathcal{Q}$  is *stable* if there exist constants  $C_1$  and  $C_2$  such that whenever  $\|c\|_\infty \leq C_1$  we have  $\|u\|_\infty \leq C_2$ .

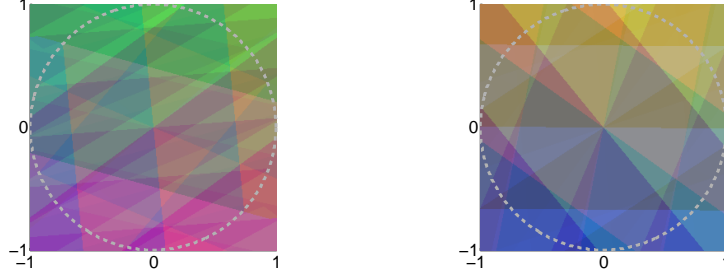
In general, designing and proving the stability of  $\Sigma\Delta$  quantization schemes of arbitrary order can be quite difficult if the number of elements in the associated codebook is held fixed. This challenge is especially difficult in the case of 1-bit quantizers and overcoming it is the core of the contributions of [32, 41, 35], where stable  $\Sigma\Delta$  quantization schemes of arbitrary order are designed. On the other hand, if the number of elements in the codebook (42) is allowed to increase with order, then even the simple greedy  $\Sigma\Delta$  schemes (47) are stable, as the following proposition shows (see, e.g., [13]).

**Proposition 2.** *The greedy  $r$ -th order  $\Sigma\Delta$  scheme (47) associated with the  $2L$ -level scalar quantizer (42) is stable, with  $\|u\|_\infty \leq \Delta/2$ , whenever  $\|c\|_\infty \leq \Delta(L - 2^{r-1} + 2^{-1})$ .*

*Proof.* The proof is by induction. We begin by rewriting (47) in terms of auxiliary state variables  $u_i^{(j)}$ ,  $j = 1, \dots, r$  and  $u_i^{(0)} = c_i - q_i$  as

$$\begin{aligned}
q_i &= \mathcal{Q}\left(\sum_{j=1}^r u_{i-1}^{(j)} + c_i\right) \\
u_i^{(j)} &= u_{i-1}^{(j)} + u_i^{(j-1)}, \quad j = 1, \dots, r
\end{aligned} \tag{50}$$

with  $u_0^{(j)} = 0$  for  $j = 1, \dots, r$ . Note that with this notation  $u_i^{(r)} = u_i$ . Now suppose that  $|u_{i-1}^{(j)}| \leq 2^{r-j}\Delta/2$  for all  $j \in \{1, \dots, r\}$ , then  $|\sum_{j=1}^r u_{i-1}^{(j)}| \leq (2^r - 1)\Delta/2$ . Since by the  $\Sigma\Delta$  iterations we have  $u_i^{(j)} = \sum_{k=1}^j u_{i-1}^{(k)} + c_i - q_i$  we deduce that



**Fig. 9** The first order  $\Sigma\Delta$  (left) and scalar quantization (right) cells associated with 2-bit quantization of  $\Phi^T x$  where  $x$  is in the unit ball of  $\mathbb{R}^2$  and  $\Phi$  is a  $2 \times 15$  Gaussian random matrix.

$$|u_i^{(r)}| = \left| \sum_{k=1}^r u_{i-1}^{(k)} + c_i - Q\left(\sum_{k=1}^r u_{i-1}^{(k)} + c_i\right) \right| \leq \Delta/2$$

provided  $\|c\|_\infty \leq \Delta(L - 2^{r-1} + 1/2)$ . Moreover, by (50),  $|u_i^{(j)}| \leq 2^{r-j}\Delta/2$ .

Working with stable  $r$ -th order  $\Sigma\Delta$  schemes and frames with smoothness properties, and employing the Sobolev dual for reconstruction, it was shown in [10] that the reconstruction error satisfies  $\|x - \tilde{\Phi}q\|_2 \leq C_{r,\Phi} N^{-r}$ , where the constant  $C_{r,\Phi}$  depends only on the quantization scheme and the frame. Such results for  $\Sigma\Delta$ -quantization show that its error decay rate breaks the theoretical  $\sim 1/N$  lower bound of scalar quantization described in the introduction. Fig. 9 helps to illustrate why such a result is possible. It shows the quantization cells associated with two bit quantization of  $\Phi^T x$ , where  $x$  in the unit ball of  $\mathbb{R}^2$ , using both first order  $\Sigma\Delta$  quantization and scalar quantization. For vectors belonging to a given cell, the worst case error achieved by an optimal decoder is proportional to the diameter of the cell. The figure shows that the cells resulting from  $\Sigma\Delta$  quantization are smaller than those resulting from scalar quantization, indicating the potential for a smaller reconstruction error. For a detailed overview of  $\Sigma\Delta$  quantization of frame expansions see, *e.g.*, [83].

The existing recovery algorithms for  $\Sigma\Delta$  quantized compressed sensing measurements rely on a two stage algorithm. In the first stage, the signal support is recovered and in the second stage, the signal coefficients are estimated using the Sobolev dual of the frame associated with the recovered support.

## 4.2 Finding the Signal Support

Let  $q \in \mathcal{Q}^m$  be the  $r$ -th order  $\Sigma\Delta$  quantization of the compressed sensing measurements  $y = Ax \in \mathbb{R}^m$  associated with the sparse vector  $x \in \Sigma_k$  and the measurement matrix  $A \in \mathbb{R}^{m \times n}$ . In order to preserve the codebook definition (42), we assume as in Sec. 3 that the scaling of the entries of  $A$  is independent of  $m$ .

The goal of the first stage of the reconstruction algorithm is to recover  $T := \text{supp}(x)$ . To that end, following [42] we will use a (standard) compressed sensing decoder  $\mathcal{D} : \mathbb{R}^m \rightarrow \mathbb{R}^n$  that has uniform robustness guarantees for matrices with an appropriate RIP constant. For such a decoder and an arbitrary scalar  $\kappa$

$$x \in \Sigma_k \text{ and } \gamma \in \mathbb{R}^m : \|\gamma\|_2 \leq \kappa\sqrt{m} \implies \|\mathcal{D}(Ax + \gamma) - x\|_2 \leq C\kappa. \quad (51)$$

For example, if  $\mathcal{D}(Ax + \gamma)$  is the output of an  $\ell_1$ -minimization algorithm such as Basis Pursuit DeNoising (BPDN), it satisfies (51) with constant  $C := C(\delta, k)$  when the matrix  $A$  (more precisely  $A/\sqrt{m}$ ) satisfies an appropriate restricted isometry property [26]. As the next proposition shows, robust decoders allow recovering the support of a sparse vector when its smallest non-zero entry is above the error level.

**Proposition 3.** *Let  $\mathcal{D}$  be a compressed sensing decoder satisfying (51) and let  $x \in \Sigma_k$  with  $T := \text{supp}(x)$ . Define  $\hat{x} := \mathcal{D}(Ax + \gamma)$ . If  $\min_{i \in T} |x_i| > 2C\kappa$  then the largest  $k$  coefficients of  $\hat{x}$  are supported on  $T$ .*

*Proof.* First, note that for all  $i \in T$ , (51) yields  $|\hat{x}_i - x_i| \leq C\kappa$ . Since  $\min_{i \in T} |x_i| > 2C\kappa$ , the reverse triangle inequality gives  $|\hat{x}_i| > C\kappa$  for all  $i$  in  $T$ . On the other hand, (51) also ensures that  $|\hat{x}_i| \leq C\kappa$  for all  $i \in T^c$ .

A sharper version of this argument appears in [42] but Proposition 3 is sufficient for our purposes. In particular, consider an  $r$ th order greedy  $\Sigma\Delta$  quantization associated with a codebook  $\mathcal{Q}$  having  $2L$  elements. Applying such a scheme to  $Ax$  yields a quantized vector  $q$  satisfying  $\|q - Ax\|_2 \leq \frac{\Delta}{2} 2^r \sqrt{m}$  provided

$$L > \|Ax\|_\infty / \Delta + 2^{r-1} - 1/2. \quad (52)$$

Thus assuming that  $A/\sqrt{m}$  has appropriate RIP constants, Proposition 3 shows that using a decoder satisfying (51), the support  $T$  of  $x \in \Sigma_k \subset \mathbb{R}^n$  can be accurately recovered provided  $|z_i| > 2^r C \Delta$  for all  $i \in T$ . What remains is to choose the number of levels  $L$  in the codebook to satisfy (52); this in turn requires an estimate of  $\|Ax\|_\infty$ .

To that end, we now consider subgaussian measurement matrices, *i.e.*, matrices whose entries are subgaussian random variables as defined below.

**Definition 5.** Let  $\xi$  be a Gaussian random variable drawn according to  $\mathcal{N}(0, \sigma^2)$ . If a random variable  $\eta$  satisfies  $P(|\eta| > t) \leq eP(|\xi| > t)$  for all  $t$ , then we say  $\eta$  is subgaussian with parameter  $\sigma > 0$ .

Examples of subgaussian random variables include Gaussian, Bernoulli, and bounded random variables, as well as their linear combinations. For matrices populated with such subgaussian entries, the following proposition from [61] gives a

bound on  $\|Ax\|_\infty$  when the non-zero entries of  $x$  are restricted to a fixed support  $T$  so that  $Ax = \Phi^T x_T$  for a frame  $\Phi$  associated with the support.

**Proposition 4.** *Let  $\widehat{\Phi}$  be a  $k \times m$  subgaussian matrix with mean zero, unit variance, and parameter  $\sigma$ , where  $k < m$ . Let  $\Phi = \frac{1}{\sqrt{m}} \widehat{\Phi}$  and fix  $\alpha \in (0, 1)$ . Then, with probability at least  $1 - e^{-\frac{1}{4}m^{1-\alpha}k^\alpha}$ , we have for all  $m > C^{\frac{1}{1-\alpha}}k$  and  $x \in \mathbb{R}^k$*

$$\|\Phi^T x\|_\infty \leq e^{1/2} \left(\frac{m}{k}\right)^{-\frac{\alpha}{2}} \|x\|_2. \quad (53)$$

Here  $C$  is a constant that may depend on  $\sigma$ , but is independent of  $k$  and  $\alpha$ .

Taking a union bound over all the  $\binom{n}{k}$  submatrices of  $A$  of size  $m \times k$  yields an identical uniform bound on  $\|Ax\|_\infty$ , which holds for sparse vectors  $x$  with high probability, provided  $m > Ck(\log n)^{\frac{1}{1-\alpha}}$ .

Thus an  $r$ -th order greedy  $\Sigma\Delta$  scheme with sufficiently many quantization levels allows the recovery of a sparse signal's support from its compressed sensing measurements. Equipped with this knowledge, we can estimate the signal coefficients using the Sobolev dual of the frame associated with the recovered support.

### 4.3 Recovering the Signal Coefficients

We continue to consider Gaussian or subgaussian measurement matrices, now assuming that the support  $T$  of the signal  $x$  has been identified. Our goal is to approximate the coefficients  $x_i, i \in T$ . With high probability, the matrix  $A/\sqrt{m}$  has the restricted isometry property of order  $2k$  and level  $\delta_{2k} \leq 1/\sqrt{2}$  provided one takes at least on the order of  $k \log(n/k)$  measurements. Then the matrix  $A_T/\sqrt{m}$  restricted to the columns indexed by  $T$  is close to an isometry and its rows hence form a frame. Consequently, the measurement vector is the associated frame expansion of  $x_T$ , and  $q$  is the corresponding  $\Sigma\Delta$  frame quantization.

As shown in Sec. 4.1, it is advantageous to reconstruct  $x$  from the  $r$ -th order  $\Sigma\Delta$  quantization  $q$  of the measurement vector  $Ax$  using the Sobolev dual  $\widetilde{A}_T^{(r)}$  of  $A_T$ , see (48) and (49). A possible bound for the reconstruction error is then proportional to  $\|\widetilde{A}_T^{(r)} D^r\|_{2 \rightarrow 2}$ . Thus to show a uniform recovery guarantee, one needs a bound for this quantity which is uniform over all potential support sets  $T$ . In the initial work [42], dealing with Gaussian compressed sensing matrices, the approach to proving such a bound consisted of explicitly controlling the lowest singular value of  $D^{-r}A_T$ . Their approach utilized the unitary invariance of the Gaussian measure to identify the distribution of the singular values of the random matrix  $D^{-r}A_T$  with those of  $S_{D^{-r}}\Psi$ , where  $S_{D^{-r}}$  is a diagonal matrix whose entries are the singular values of  $D^{-r}$ , and  $\Psi$  is a Gaussian matrix. This, coupled with bounds on the singular values of  $D^{-r}$ , allowed [42] to derive bounds that held with probability high enough to survive a union bound over all  $\binom{n}{k}$  Gaussian submatrices of  $A$ . In [61], this approach was

extended to subgaussian matrices. Herein, to prove such a bound on  $\|\tilde{A}_T^{(r)} D^r\|_{2 \rightarrow 2}$ , we follow the simpler, RIP-based approach presented in [37].

To that end, let  $E = U_E S_E V_E^T$  be the singular value decomposition (SVD) of any matrix  $E$  (for some orthogonal matrices  $U_E$  and  $V_E$ ) where the matrix  $S_E$  is diagonal with (ordered) diagonal entries  $\sigma_j(E)$ . We denote also  $\sigma_{\min}(E) := \sigma_1(E)$  the smallest singular value of  $E$ . Then the following proposition (see, e.g., [42]) holds.

**Proposition 5.** *There are positive constants  $C_1(r)$  and  $C_2(r)$ , independent of  $m$ , such that*

$$C_1(r) \left(\frac{m}{j}\right)^r \leq \sigma_j(D^{-r}) \leq C_2(r) \left(\frac{m}{j}\right)^r, \quad j = 1, \dots, m. \quad (54)$$

Denote by  $P_\ell$  the  $\ell \times m$  matrix that maps a vector to its first  $\ell$  components. Moreover, denote by  $\tilde{\Sigma}_k(A, \mathcal{D}) \subset \Sigma_k$  the set of  $k$ -sparse signals  $x$  whose support can be recovered from  $q$  with the decoder  $\mathcal{D}$  as in Proposition 3. The following theorem describes the reconstruction performance.

**Theorem 6 ([37]).** *Let  $A \in \mathbb{R}^{m \times n}$  be a matrix such that for a fixed  $\ell \leq m$ , the  $\ell \times n$  matrix  $\frac{1}{\sqrt{\ell}} P_\ell V_{D^{-r}}^T A$  has restricted isometry constant  $\delta_k \leq \delta$ . Then the following holds uniformly for all  $x \in \tilde{\Sigma}_k(A, \mathcal{D})$ .*

*If  $x$  has support  $T$ ,  $q$  is the  $r$ -th order  $\Sigma\Delta$  quantization of  $Ax$ , and  $\hat{x} := \tilde{A}_T^{(r)} q$ , then*

$$\|x - \hat{x}\|_2 \leq \frac{\Delta}{C(r)\sqrt{(1-\delta)}} \left(\frac{m}{\ell}\right)^{-r+\frac{1}{2}},$$

where  $C(r) > 0$  is a constant depending only on  $r$  and  $\Delta$  is the quantization step size.

*Proof.* As the SVD of  $D^{-r}$  provides  $D^{-r} = U_{D^{-r}} S_{D^{-r}} V_{D^{-r}}^T$ , the smallest singular value of  $D^{-r} A_T$  satisfies

$$\begin{aligned} \sigma_{\min}(D^{-r} A_T) &= \sigma_{\min}(S_{D^{-r}} V_{D^{-r}}^T A_T) \\ &\geq \sigma_{\min}(P_\ell S_{D^{-r}} V_{D^{-r}}^T A_T) \\ &= \sigma_{\min}((P_\ell S_{D^{-r}} P_\ell^T)(P_\ell V_{D^{-r}}^T A_T)) \\ &\geq \sigma_\ell(D^{-r}) \sigma_{\min}(P_\ell V_{D^{-r}}^T A_T), \end{aligned}$$

To bound  $\sigma_{\min}(P_\ell V_{D^{-r}}^T A_T)$  uniformly over all support sets  $T$  of size  $k$  we simply note that if  $\frac{1}{\sqrt{\ell}} P_\ell V_{D^{-r}}^T \Phi$  has restricted isometry constant  $\delta_k \leq \delta$  then  $\sigma_{\min}(P_\ell V_{D^{-r}}^T A_T)$  is uniformly bounded from below by

$$\sqrt{\ell} \sqrt{1 - \delta}. \quad (55)$$

The theorem follows by applying (48), (54), (55) as

$$\frac{1}{\sigma_{\min}(D^{-r} A_T)} \|u\|_2 \leq \frac{\Delta}{C(r)\sqrt{(1-\delta)}} \left(\frac{m}{\ell}\right)^{-r+\frac{1}{2}} \quad (56)$$

□

The above theorem can be applied almost directly to Gaussian compressed sensing matrices. If  $A$  is a Gaussian matrix with independent zero mean and unit variance entries, then by rotation invariance so is the matrix  $P_\ell V_{D-r}^T A$ . Regarding the choice of  $\ell$ , note from Theorem 6 that the smaller  $\ell$  is, the better the bound. On the other hand  $\ell$  has to be large enough for  $\frac{1}{\sqrt{\ell}}(P_\ell V_{D-r}^T \Phi)$  to have the restricted isometry constant  $\delta_k \leq \delta$ . This prompts the choice  $\ell \asymp k \log n$ , as then  $\frac{1}{\sqrt{\ell}}(P_\ell V_{D-r}^T \Phi)$  has the restricted isometry constant  $\delta_k < \delta$  with high probability, as discussed in Chapter 1 (see, *e.g.*, Theorem 1.5). In particular, if

$$m \gtrsim k(\log n)^{\frac{1}{1-\alpha}}, \quad \alpha \in (0, 1)$$

and

$$\ell \asymp k \log n$$

then

$$\frac{m}{\ell} \asymp \frac{m}{k \log n} = \left(\frac{m}{k}\right)^\alpha \cdot \left(\frac{m}{k(\log n)^{\frac{1}{1-\alpha}}}\right)^{1-\alpha} \gtrsim \left(\frac{m}{k}\right)^\alpha$$

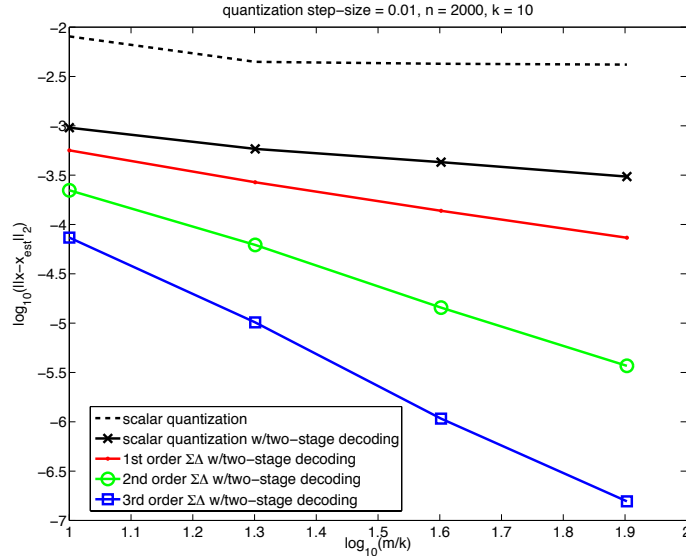
Applying Theorem 6 directly, we obtain

$$\|x - \hat{x}\|_2 \lesssim \Delta \left(\frac{m}{k}\right)^{-\alpha(r-\frac{1}{2})}.$$

This essentially recovers the result in [42] and a similar, albeit more technical argument for subgaussian matrices, using either bounds on tail probabilities for quadratic forms [44, 87] or bounds for suprema of chaos processes [59] recovers the analogous result in [61].

To illustrate the advantage of using  $\Sigma\Delta$  schemes for quantizing compressed sensing measurements we conduct a numerical experiment with  $k$ -sparse signals in  $\mathbb{R}^n$ , as we vary the number of measurements  $m$ . We fix  $k = 10$ ,  $n = 1000$ , and the quantization step-size  $\Delta = 0.01$ . We draw  $m \times n$  Gaussian matrices  $A$  for  $m \in \{100, 200, 400, 800\}$  and quantize the measurements  $Ax$  using scalar quantization and  $r$ th order  $\Sigma\Delta$  schemes with  $r = 1, 2, 3$ . We then use the two-stage reconstruction method described herein to obtain an approximation  $\hat{x}$  of  $x$  using its quantized measurements. Repeating this experiment 30 times, we compute the average of the reconstruction error  $\|x - \hat{x}\|_2$  for each of the quantization methods and plot them against the oversampling ratio  $m/k$  in Fig. 10.

In summary, using Gaussian and subgaussian compressed sensing matrices recovery of sparse signals from their  $\Sigma\Delta$  quantized measurements is possible. More importantly, the reconstruction error decays polynomially in the number of measurements and thus outperforms the (at best) linear error decay that can be achieved with scalar quantization. This improvement comes at the cost of introducing memory elements, and feedback, into the quantization procedure.



**Fig. 10** Average errors over 30 experiments. The figure shows the reconstruction errors resulting from scalar quantization, using  $\ell_1$ -minimization for reconstruction (dashed line). Also corresponding to scalar quantization, the figure shows the errors resulting from reconstructing via the two-stage algorithm described herein (solid black line), using the canonical dual of the frame corresponding to the recovered support in the second stage. It also shows the reconstruction errors resulting from 1st, 2nd, and 3rd order  $\Sigma\Delta$  quantization respectively. These errors decay as  $(m/k)^{-r}$  for  $r = 1, 2, 3$  respectively, slightly outperforming the theoretical predictions presented here.

## 5 Discussion and Conclusion

Quantization is an essential component of any acquisition system, and, therefore, an important part of compressive sensing theory and practice. While significant work has been done in understanding the interaction of quantization and compressive sensing, there are several open problems and questions.

One of the most interesting open problems is the interaction of quantization with noise. While the discussion and references in Sec. 3.5 provides some initial results and theoretical analysis, a comprehensive understanding is still missing. An understanding of the optimal bit allocation and the optimal quantizer design, uniform or non-uniform scalar, or  $\Sigma\Delta$ , given the noise level, as well as the robustness of the reconstruction to noise and quantization is still elusive.

While  $\Sigma\Delta$  can be used to improve the rate efficiency of compressive sensing, compared to scalar quantization, the performance is still not comparable to the state-of-the-art in conventional  $\Sigma\Delta$  methods. For example, conventional  $\Sigma\Delta$  quantization of band-limited functions can achieve error that decays exponentially as the sampling rate increases, not currently possible with existing compressive sensing  $\Sigma\Delta$ . Furthermore, the analysis in Sec. 4 does not hold for 1-bit quantization, often de-



sirable in practical systems due to its simplicity. Such an extension has significant practical importance.

Even with  $\Sigma\Delta$  approaches, the rate efficiency of compressive sensing systems is not ideal. As evident from the fundamental bounds in Sec. 2, compressive sensing is not rate-efficient compared to classical methods such as transform coding. In other words, while compressive sensing is very promising in building sensing systems because it can significantly reduce the number of measurements and the sampling burden, it is not a good data compression approach if the measurements have already been obtained and the achievable bit-rate is important. That said, due to the intimate connection between frame quantization and quantization for compressed sensing, promising results in the finite frames context, e.g., [48] can inform future developments in compressed sensing.

The potential encoding simplicity of a compressive sensing system is very appealing. Acquiring generalized linear measurements and quantizing them can be less complex than typical transform-coding approaches and much more attractive in low-power and computationally-restricted sensing applications. The complexity is shifted to the reconstruction, which, in many applications, can bear significantly more computational complexity. Nevertheless, the rate inefficiency of compressive sensing can be a barrier in such applications.

A number of promising approaches have been proposed to overcome this barrier using modifications of the quantizer that produce non-contiguous quantization regions [78, 16, 17, 57]. Initial theoretical analysis and experimental results are promising. However, our understanding is still limited. One of the drawbacks of such approaches is that the reconstruction is no longer convex and, therefore, not as simple to provide guarantees for.

Alternatively, recent work on adaptive quantization strategies has shown that error decay exponential in the bit-rate can be achieved, even using a 1-bit quantizer, at the cost of adaptivity in the measurements and – in contrast with the methods presented in this chapter – significant computation at the encoder. Specifically, [5] shows that adaptively choosing the threshold of a 1-bit quantizer allows the error to decay exponentially with the number of measurements. The cost is that the thresholds are updated by solving an  $\ell_1$  minimization problem, or running an iterative hard thresholding scheme. It is thus interesting to quantify the tradeoff between computational complexity at the quantizer, and achievable reconstruction accuracy.

Another important aspect is that while the best recovery guarantees in compressed sensing are obtained for Gaussian and subgaussian measurement matrices, which are also mainly considered in this article, applications usually require structured matrices, such as subsampled Fourier matrices, e.g., as a model for subsampled MRI measurements [72], or subsampled convolution, e.g., as a model for coded aperture imaging [73]. In both cases, when the subsampling is randomized, near-optimal recovery guarantees are known for unquantized compressed sensing [86, 59]. Combined with quantization, however, hardly anything is known for such matrices. Such results would be of great importance to move the approaches discussed in this survey closer to the application scenarios.

Quantization is also important when considering randomized embeddings, an area of research intimately related to compressive sensing [4, 62]. Embeddings are transformations that preserve the geometry of the space they operate on; reconstruction of the embedded signal is not necessarily the goal. They have been proven quite useful, for example, in signal-based retrieval applications, such as augmented reality, biometric authentication and visual search [70, 21, 85].

These applications require storage or transmission of the embedded signals, and, therefore, quantizer design is very important in controlling the rate used by the embedding. Indeed, significant analysis has been performed for embeddings followed by conventional scalar quantization, some of it in the context of quantized compressive sensing [54, 81, 82] or in the study of quantized extensions to the Johnson Lindenstrauss Lemma [55, 70, 85, 49]. Furthermore, since reconstruction is not an objective anymore, non-contiguous quantization is more suitable, leading to very interesting quantized embedding designs and significant rate reduction [21]. In this context, quantization can also provide to significant computation savings in the retrieval, leading to Locality Sensitive Hashing (LSH) and similar methods [2].

**Acknowledgements** Petros T. Boufounos is exclusively supported by Mitsubishi Electric Research Laboratories. Laurent Jacques is a Research Associate funded by the Belgian F.R.S.-FNRS. Felix Kraher acknowledges support by the German Science Foundation (DFG) in the context of the Emmy-Noether Junior Research Group KR 4512/1-1 “RaSenQuaSP”. Rayan Saab is an assistant professor of mathematics with the University of California, San Diego.

## References

1. 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.
2. A. Andoni and P. Indyk. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. *Commun. ACM*, 51(1):117–122, 2008.
3. S. Bahmani, P. T. Boufounos, and B. Raj. Robust 1-bit compressive sensing via Gradient Support Pursuit. *arXiv preprint arXiv:1304.6627*, 2013.
4. R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. *Constr. Approx.*, 28(3):253–263, 2008.
5. R. Baraniuk, S. Foucart, D. Needell, Y. Plan, and M. Wooters. Exponential decay of reconstruction error from binary measurements of sparse signals. *arXiv preprint arXiv:1407.8246*, 2014.
6. J. J. Benedetto, A. M. Powell, and Ö. Yılmaz. Second-order Sigma-Delta ( $\Sigma\Delta$ ) quantization of finite frame expansions. *Appl. Comput. Harmon. Anal.*, 20(1):126–148, 2006.
7. J. J. Benedetto, A. M. Powell, and Ö. Yılmaz. Sigma-Delta ( $\Sigma\Delta$ ) quantization and finite frames. *IEEE Trans. Inform. Theory*, 52(5):1990–2005, 2006.
8. R. Berinde, A. C. Gilbert, P. Indyk, H. Karloff, and M. J. Strauss. Combining geometry and combinatorics: a unified approach to sparse signal recovery. In *Proc. 46th Annu. Allerton Conf. Commun. Control Comput.*, pages 798–805. IEEE, 2008.
9. T. Blu, P-L Dragotti, M. Vetterli, P. Marziliano, and L. Coulot. Sparse sampling of signal innovations. *IEEE Signal Process. Mag.*, 25(2):31–40, 2008.
10. J. Blum, M. Lammers, A. M Powell, and Ö. Yılmaz. Sobolev duals in frame theory and Sigma-Delta quantization. *J. Fourier Anal. Appl.*, 16(3):365–381, 2010.

11. T. Blumensath and M. Davies. Iterative hard thresholding for compressive sensing. *Appl. Comput. Harmon. Anal.*, 27(3):265–274, 2009.
12. B. G. Bodmann and V. I. Paulsen. Frame paths and error bounds for Sigma-Delta quantization. *Appl. Comput. Harmon. Anal.*, 22(2):176–197, 2007.
13. B. G. Bodmann, V. I. Paulsen, and S. A. Abdalbaki. Smooth frame-path termination for higher order Sigma-Delta quantization. *J. Fourier Anal. Appl.*, 13(3):285–307, 2007.
14. P. T. Boufounos. *Quantization and Erasures in Frame Representations*. D.Sc. Thesis, MIT EECS, Cambridge, MA, January 2006.
15. P. T. Boufounos. Greedy sparse signal reconstruction from sign measurements. In *Proc. Asilomar Conf. on Signals Systems and Comput.*, Asilomar, California, Nov. 2009.
16. P. T. Boufounos. Hierarchical distributed scalar quantization. In *Proc. Int. Conf. Sampling Theory and Applications (SampTA)*, Singapore, May 2-6 2011.
17. P. T. Boufounos. Universal rate-efficient scalar quantization. *IEEE Trans. Inform. Theory*, 58(3):1861–1872, 2012.
18. P. T. Boufounos and R. G. Baraniuk. Quantization of sparse representations. In *Rice University ECE Department Technical Report 0701*. Summary appears in *Proc. Data Compression Conference (DCC)*, Snowbird, UT, March 27-29 2007.
19. P. T. Boufounos and R. G. Baraniuk. 1-bit compressive sensing. In *Proc. Conf. Inform. Science and Systems (CISS)*, Princeton, NJ, March 19-21 2008.
20. P. T. Boufounos and A. V. Oppenheim. Quantization noise shaping on arbitrary frame expansions. *EURASIP J Adv. Signal Proc.*, page 053807, 2006.
21. P. T. Boufounos and S. Rane. Efficient coding of signal distances using universal quantized embeddings. In *Proc. Data Compression Conference (DCC)*, Snowbird, UT, March 20-22 2013.
22. T. T. Cai and A. Zhang. Sparse representation of a polytope and recovery of sparse signals and low-rank matrices. *IEEE Trans. Inform. Theory*, 60(1):122–132, 2014.
23. A. Calderbank and I. Daubechies. The pros and cons of democracy. *IEEE Trans. Inform. Theory*, 48(6), 2002.
24. E. Candès and J. Romberg. Encoding the  $\ell_p$  ball from limited measurements. In *Proc. Data Compression Conference (DCC)*, Snowbird, UT, March 28-30 2006.
25. E. Candès, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.*, 59(8):1207–1223, 2006.
26. E. J. Candès. The restricted isometry property and its implications for compressed sensing. *C. R. Acad. Sci., Ser. I*, 346:589–592, 2008.
27. R. Chartrand and V. Staneva. Restricted isometry properties and nonconvex compressive sensing. *Inverse Problems*, 24(3):1–14, 2008.
28. S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic Decomposition by Basis Pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1998.
29. E. Chou. Non-convex decoding for sigma delta quantized compressed sensing. In *Proc. Int. Conf. Sampling Theory and Applications (SampTA 2013)*, pages 101–104, Bremen, Germany, 2013.
30. P. L. Combettes and J-C Pesquet. Proximal splitting methods in signal processing. In *Fixed-point algorithms for inverse problems in science and engineering*, pages 185–212. Springer, 2011.
31. W. Dai, H. V. Pham, and O. Milenkovic. Distortion-Rate Functions for Quantized Compressive Sensing. Technical Report arXiv:0901.0749, 2009.
32. I. Daubechies and R. DeVore. Approximating a bandlimited function using very coarsely quantized data: A family of stable sigma-delta modulators of arbitrary order. *Ann. Math.*, pages 679–710, 2003.
33. M. A. Davenport, J. N. Laska, P. T. Boufounos, and R. G. Baraniuk. A simple proof that random matrices are democratic. Technical report, Rice University ECE Department Technical Report TREE-0906, Houston, TX, November 2009.
34. M. A. Davenport, J. N. Laska, J. Treichler, and R. G. Baraniuk. The pros and cons of compressive sensing for wideband signal acquisition: Noise folding versus dynamic range. *IEEE Trans. Signal Proc.*, 60(9):4628–4642, 2012.

35. P. Deift, C. S. Güntürk, and F. Kraher. An optimal family of exponentially accurate one-bit sigma-delta quantization schemes. *Comm. Pure Appl. Math.*, 64(7):883–919, 2011.
36. D. E. Edmunds and H. Triebel. *Function spaces, entropy numbers, differential operators*. Cambridge Univ. Press, Cambridge, UK., 1996.
37. J. Feng and F. Kraher. An RIP approach to Sigma-Delta quantization for compressed sensing. *IEEE Signal Proc. Lett.*, 21(11):1351–1355, 2014.
38. V. K Goyal, M. Vetterli, and N. T. Thao. Quantized overcomplete expansions in  $\mathbb{R}^N$ : Analysis, synthesis, and algorithms. *IEEE Trans. Inform. Theory*, 44(1):16–31, 1998.
39. R. M. Gray and D. L. Neuhoff. Quantization. *IEEE Trans. Inform. Theory*, 44(6):2325–2383, 1998.
40. R.M. Gray. Oversampled sigma-delta modulation. *IEEE Trans. Comm.*, 35(5):481–489, 1987.
41. C. S. Güntürk. One-bit sigma-delta quantization with exponential accuracy. *Comm. Pure Appl. Math*, 56(11):1608–1630, 2003.
42. 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. *Found. Comput. Math.*, 13(1):1–36, 2013.
43. S. Güntürk. *Harmonic analysis of two problems in signal compression*. PhD thesis, Program in Applied and Computation Mathematics, Princeton University, Princeton, NJ, Sept. 2000.
44. D. L. Hanson and F. T. Wright. A bound on tail probabilities for quadratic forms in independent random variables. *Ann. Math. Stat.*, 42(3):1079–1083, 1971.
45. W. Hoeffding. Probability inequalities for sums of bounded random variables. *J. Am. Stat. Assoc.*, 58(301):13–30, 1963.
46. H. Inose and Y. Yasuda. A unity bit coding method by negative feedback. *Proc. IEEE*, 51(11):1524–1535, Nov 1963.
47. H. Inose, Y. Yasuda, and J. Murakami. A telemetering system by code modulation –  $\Delta$  –  $\Sigma$  modulation. *IRE Trans. Space El. Tel.*, SET-8(3):204–209, Sept 1962.
48. M. Iwen and R. Saab. Near-optimal encoding for sigma-delta quantization of finite frame expansions. *J. Fourier Anal. Appl.*, 19(6):1255–1273, 2013.
49. L. Jacques. A quantized Johnson Lindenstrauss lemma: The finding of buffon’s needle. *arXiv preprint arXiv:1309.1507*, 2013.
50. L. Jacques. Error decay of (almost) consistent signal estimations from quantized random gaussian projections. *arXiv preprint arXiv:1406.0022*, 2014.
51. L. Jacques, K. Degraux, and C. De Vleeschouwer. Quantized iterative hard thresholding: Bridging 1-bit and high-resolution quantized compressed sensing. In *Proc. Intl. Conf. Sampling Theory and Applications (SampTA 2013)*, arXiv:1305.1786, pages 105–108, Bremen, Germany, 2013.
52. L. Jacques, D. K. Hammond, and M. J. Fadili. Dequantizing compressed sensing: When oversampling and non-gaussian constraints combine. *IEEE Trans. Inform. Theory*, 57(1):559–571, January 2011.
53. L. Jacques, D. K. Hammond, and M. J. Fadili. Stabilizing nonuniformly quantized compressed sensing with scalar companders. *IEEE Trans. Inform. Theory*, 5(12):7969 – 7984, January 2013.
54. 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 Trans. Inform. Theory*, 59(4):2082–2102, 2013.
55. W. B. Johnson and J. Lindenstrauss. Extensions of lipschitz mappings into a hilbert space. *Contemp. Math.*, 26(189-206):1, 1984.
56. U. Kamilov, V.K. Goyal, and S. Rangan. Optimal quantization for compressive sensing under message passing reconstruction. In *Proc. IEEE International Symposium on Information Theory (ISIT)*, pages 459–463, 2011.
57. U.S. Kamilov, V.K. Goyal, and S. Rangan. Message-passing de-quantization with applications to compressed sensing. *IEEE Trans. Signal Proc.*, 60(12):6270–6281, Dec 2012.
58. V. Kostina, M. F Duarte, S. Jafarpour, and R. Calderbank. The value of redundant measurement in compressed sensing. In *Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, pages 3656–3659, 2011.

59. F. Krahmer, S. Mendelson, and H. Rauhut. Suprema of chaos processes and the restricted isometry property. *Comm. Pure Appl. Math.*, 67(11):1877–1904, 2014.
60. F. Krahmer, R. Saab, and R. Ward. Root-exponential accuracy for coarse quantization of finite frame expansions. *IEEE Trans. Inform. Theory*, 58(2):1069–1079, 2012.
61. F. Krahmer, R. Saab, and Ö. Yilmaz. Sigma-delta quantization of sub-gaussian frame expansions and its application to compressed sensing. *Inform. Inference*, 3(1):40–58, 2014.
62. F. Krahmer and R. Ward. New and improved johnson-lindenstrauss embeddings via the restricted isometry property. *SIAM J. Math. Anal.*, 43(3):1269–1281, 2011.
63. F. Krahmer and R. Ward. Lower bounds for the error decay incurred by coarse quantization schemes. *Appl. Comput. Harmonic Anal.*, 32(1):131–138, 2012.
64. T. Kühn. A lower estimate for entropy numbers. *J. Approx. Theory*, 110(1):120–124, 2001.
65. M. Lammers, A. M. Powell, and Özgür Yilmaz. Alternative dual frames for digital-to-analog conversion in sigma-delta quantization. *Adv. Comput. Math.*, 32(1):73–102, 2010.
66. J. Laska, P. Boufounos, M. Davenport, and R. Baraniuk. Democracy in action: Quantization, saturation, and compressive sensing. *Appl. Comput. Harmon. Anal.*, 31(3):429–443, November 2011.
67. J. Laska, Z. Wen, W. Yin, and R. Baraniuk. Trust, but verify: Fast and accurate signal recovery from 1-bit compressive measurements. *IEEE Trans. Signal Proc.*, 59(11):5289–5301, 2010.
68. J. N. Laska and R. G. Baraniuk. Regime change: Bit-depth versus measurement-rate in compressive sensing. *IEEE Trans. Signal Proc.*, 60(7):3496–3505, 2012.
69. M. Ledoux. *The concentration of measure phenomenon*. American Mathematical Soc., 2005.
70. M. Li, S. Rane, and P. T. Boufounos. Quantized embeddings of scale-invariant image features for mobile augmented reality. In *Proc. IEEE Int. Workshop on Multimedia Signal Processing (MMSP)*, Banff, Canada, Sept. 17–19 2012.
71. S. Lloyd. Least squares quantization in PCM. *IEEE Trans. Inform. Theory*, 28(2):129–137, March 1982.
72. M. Lustig, D. Donoho, and J.M. Pauly. Sparse MRI: The application of compressed sensing for rapid MRI imaging. *Magn. Reson. Med.*, 58(6):1182–1195, 2007.
73. R. F. Marcia and R. M. Willett. Compressive coded aperture superresolution image reconstruction. In *Proc. Intl. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, pages 833–836. IEEE, 2008.
74. J. Max. Quantizing for minimum distortion. *IEEE Trans. Inform. Theory*, 6(1):7–12, March 1960.
75. M. Mishali and Y. C. Eldar. Sub-Nyquist sampling. *IEEE Signal Proc. Mag.*, 28(6):98–124, 2011.
76. H. Q. Nguyen, V.K. Goyal, and L.R. Varshney. Frame permutation quantization. *Appl. Comput. Harmon. Anal.*, 2010.
77. S. R. Norsworthy, R. Schreier, G. C. Temes, et al. *Delta-Sigma data converters: theory, design, and simulation*, volume 97. IEEE press New York, 1996.
78. R. J. Pai. Nonadaptive lossy encoding of sparse signals. M.eng. thesis, MIT EECS, Cambridge, MA, August 2006.
79. P. F. Panter and W. Dite. Quantization distortion in pulse-count modulation with nonuniform spacing of levels. *Proc. IRE*, 39(1):44–48, 1951.
80. Y. Plan and R. Vershynin. One-bit compressed sensing by linear programming. *Comm. Pure Appl. Math.*, 66(8):1275–1297, 2013.
81. Y. Plan and R. Vershynin. Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *IEEE Trans. Inform. Theory*, 59(1):482–494, 2013.
82. Y. Plan and R. Vershynin. Dimension reduction by random hyperplane tessellations. *Discrete & Computational Geometry*, 51(2):438–461, 2014.
83. A. M. Powell, R. Saab, and Ö. Yilmaz. Quantization and finite frames. In *Finite Frames*, pages 267–302. Springer, 2013.
84. A. M. Powell and J. T. Whitehouse. Error bounds for consistent reconstruction: Random polytopes and coverage processes. *arXiv preprint arXiv:1405.7094*, 2013.

85. S. Rane, P. T. Boufounos, and A. Vetro. Quantized embeddings: An efficient and universal nearest neighbor method for cloud-based image retrieval. In *Proc. SPIE Applications of Digital Image Processing XXXVI*, 2013. 885609.
86. M. Rudelson and R. Vershynin. On sparse reconstruction from Fourier and Gaussian measurements. *Comm. Pure Appl. Math.*, 61:1025–1045, 2008.
87. M. Rudelson and R. Vershynin. Hanson-wright inequality and sub-gaussian concentration. *Electron. Comm. Probab.*, 18:1–9, 2013.
88. C. Schütt. Entropy numbers of diagonal operators between symmetric Banach spaces. *J. Approx. Theory*, 40(2):121–128, 1984.
89. J. Z. Sun and V. K. Goyal. Optimal quantization of random measurements in compressed sensing. In *Proc. IEEE International Symposium on Information Theory (ISIT)*, pages 6–10, 2009.
90. N. T. Thao and M. Vetterli. Lower bound on the mean-squared error in oversampled quantization of periodic signals using vector quantization analysis. *IEEE Trans. Inform. Theory*, 42(2):469–479, March 1996.
91. N.T. Thao and M. Vetterli. Reduction of the MSE in R-times oversampled A/D conversion  $O(1/R)$  to  $O(1/R^2)$ . *IEEE Trans. Signal Proc.*, 42(1):200–203, Jan 1994.
92. M. K. Varanasi and B. Aazhang. Parametric generalized Gaussian density estimation. *J. Acoust. Soc. Am.*, 86:1404–1415, 1989.
93. A. Zymnis, S. Boyd, and E. Candes. Compressed sensing with quantized measurements. *IEEE Signal Proc. Lett.*, 17(2):149–152, 2010.