Message-Passing De-Quantization with Applications to Compressed Sensing

Ulugbek S. Kamilov, *Student Member, IEEE*, Vivek K Goyal, *Senior Member, IEEE*, and Sundeep Rangan, *Member, IEEE*

***Abstract*—Estimation of a vector from quantized linear measurements is a common problem for which simple linear techniques are suboptimal—sometimes greatly so. This paper develops message-passing de-quantization (MPDQ) algorithms for minimum mean-squared error estimation of a random vector from quantized linear measurements, notably allowing the linear expansion to be overcomplete or undercomplete and the scalar quantization to be regular or non-regular. The algorithm is based on generalized approximate message passing (GAMP), a recently-developed Gaussian approximation of loopy belief propagation for estimation with linear transforms and nonlinear componentwise-separable output channels. For MPDQ, scalar quantization of measurements is incorporated into the output channel formalism, leading to the first tractable and effective method for high-dimensional estimation problems involving non-regular scalar quantization. The algorithm is computationally simple and can incorporate arbitrary separable priors on the input vector including sparsity-inducing priors that arise in the context of compressed sensing. Moreover, under the assumption of a Gaussian measurement matrix with i.i.d. entries, the asymptotic error performance of MPDQ can be accurately predicted and tracked through a simple set of scalar state evolution equations. We additionally use state evolution to design MSE-optimal scalar quantizers for MPDQ signal reconstruction and empirically demonstrate the superior error performance of the resulting quantizers. In particular, our results show that non-regular quantization can greatly improve rate–distortion performance in some problems with oversampling or with undersampling combined with a sparsity-inducing prior.**

***Index Terms*—analog-to-digital conversion, approximate message passing, belief propagation, compressed sensing, frames, non-regular quantizers, overcomplete representations, Slepian–Wolf coding, quantization, Wyner–Ziv coding**

The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Konstantinos Diamantaras. This material is based upon work supported by the National Science Foundation under Grant Nos. 0729069 and 1116589 and by the DARPA InPho program through the US Army Research Office award W911-NF-10-1-0404. The material in this paper was presented in part at the IEEE International Symposium on Information Theory, St. Petersburg, Russia, July–August 2011, and the 4th IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing, San Juan, Puerto Rico, December 2011.

U. S. Kamilov is with Biomedical Imaging Group, École Polytechnique Fédérale de Lausanne, CH-1015 Lausanne Switzerland (email: ulugbek.kamilov@epfl.ch). This work was completed while he was with the Research Laboratory of Electronics, Massachusetts Institute of Technology.

V. K. Goyal is with the Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: v.goyal@ieee.org).

S. Rangan (email: srangan@poly.edu) is with the Department of Electrical and Computer Engineering, Polytechnic Institute of New York University, Brooklyn, NY 11201 USA (srangan@poly.edu).

I. Introduction

ESTIMATION of a signal from quantized samples is a fundamental problem in signal processing. It arises both from the discretization in digital acquisition devices and the quantization performed for lossy compression.

This paper considers estimation of an i.i.d. vector $\mathbf{x}$ from quantized transformed samples of the form $\mathrm{Q}(\mathbf{z})$ where $\mathbf{z} = \mathbf{A}\mathbf{x}$ is a linear transform of $\mathbf{x}$ and Q is a scalar (componentwise separable) quantization operator. Due to the transform $\mathbf{A}$, the components of $\mathbf{z}$ may be correlated. Even though the traditional transform coding paradigm demonstrates the advantages of expressing the signal with independent components prior to coding [1], quantization of vectors with correlated components nevertheless arises in a range of circumstances. For example, to model oversampled analog-to-digital conversion (ADC), we may write a vector of time-domain samples as $\mathbf{z} = \mathbf{A}\mathbf{x}$, where the entries of the vector $\mathbf{x}$ are statistically independent Fourier components and $\mathbf{A}$ is an oversampled inverse discrete Fourier transform. The oversampled ADC quantizes the correlated time-domain samples $\mathbf{z}$, as opposed to the Fourier coefficients $\mathbf{x}$. Distributed sensing also necessitates quantization of components that are not independent since decorrelating transforms may not be possible prior to the quantization. More recently, compressed sensing has become a motivation to consider quantization of randomly linearly mixed information, and several sophisticated reconstruction approaches have been proposed [2]–[4].

Estimation of a vector $\mathbf{x}$ from quantized samples of the form $\mathrm{Q}(\mathbf{A}\mathbf{x})$ is challenging because the quantization function Q is nonlinear and the transform $\mathbf{A}$ couples, or "mixes," the components of $\mathbf{x}$, thus necessitating joint estimation. Although reconstruction from quantized samples is typically linear, more sophisticated, nonlinear techniques can offer significant improvements in the case of quantized transformed data. A key example ADC, where the improvement from replacing conventional linear estimation with nonlinear estimation increases with the oversampling factor [5]–[13].

We propose a simple reconstruction algorithm called *message-passing de-quantization* (MPDQ) that improves upon the state of the art. The algorithm is based on a recently-developed Gaussian-approximated belief propagation (BP) algorithm called *generalized approximate message passing* (GAMP) [14] or *relaxed belief propagation* [15], [16], which extends earlier methods [17]–[19] to nonlinear output channels. The application of GAMP to de-quantization was first introduced in a conference version of this paper [20] for

regular quantization in a compressive acquisition setting; the present paper provides extensive explanations and simulations for both overcomplete and undercomplete settings, with both regular and non-regular quantization. MPDQ is the first computationally-tractable method for settings with non-regular quantization.

A. Contributions

Gaussian approximations of loopy BP have previously been shown to be effective in several other applications [16]–[19], [21], [22]; for our application to estimation from quantized samples, the extension to general output channels [14], [16] is essential. Using this extension to nonlinear output channels, we show that MPDQ estimation offers several key benefits:

- *General quantizers:* The MPDQ algorithm permits essentially arbitrary quantization functions Q including non-uniform and even non-regular quantizers (i.e. quantizers with cells composed of unions of disjoint intervals) used, for example, in Wyner–Ziv coding [23] and multiple description coding [24]. In Section VI, we will demonstrate that a non-regular modulo quantizer can provide performance improvements for correlated data. We believe that the MPDQ algorithm provides the first tractable estimation method that can exploit such quantizers.
- *General priors:* MPDQ estimation can incorporate a large class of priors on the components of $\mathbf{x}$, provided that the components are independent. For example, in Section VI, we will demonstrate the algorithm on recovery of vectors with sparse priors arising in quantized compressed sensing [2]–[4].
- *Exact characterization with random transforms:* In the case of certain large random transforms $\mathbf{A}$, the componentwise performance of MPDQ can be precisely predicted by a so-called *state evolution* (SE) analysis presented in Section V-A. From the SE analysis, one can precisely evaluate any componentwise performance metric, including mean-squared error (MSE). In contrast, works such as [5]–[13] mentioned above have only obtained bounds or scaling laws.
- *Performance:* Our simulations indicate significantly-improved performance over traditional methods for estimating from quantized samples in a range of scenarios.
- *Computational simplicity:* The MPDQ algorithm is computationally extremely fast. Our simulation and SE analysis indicate good performance with a small number of iterations (10 to 20 in our experience), with the dominant computational cost per iteration simply being multiplication by $\mathbf{A}$ and $\mathbf{A}^T$.
- *Applications to optimal quantizer design:* When quantizer outputs are used as inputs to a nonlinear estimation algorithm, minimizing the MSE between quantizer inputs and outputs is generally not equivalent to minimizing the MSE of the final reconstruction [25]. To optimize the quantizer for the MPDQ algorithm, we use the fact that the MSE under large random mixing matrices $\mathbf{A}$ can be predicted accurately from a set of simple SE equations [14], [15]. We use the SE formalism to optimize the quantizer to minimize the asymptotic distortion after the reconstruction by MPDQ. Note that our use of random $\mathbf{A}$ is for rigor of the SE formalism; the effectiveness of MPDQ does not depend on this.

B. Outline

The remainder of the paper is organized as follows. Section II provides basic background material on quantization and compressed sensing. Section III introduces the problem of estimating a random vector from quantized linear transform coefficients. It concentrates on geometric insights for both the oversampled and undersampled settings. Section IV presents the MPDQ algorithm by formulating the reconstruction problem in Bayesian terms. Note that this Bayesian formulation does not require sparsity of the signal nor specify undersampling or oversampling. Section V describes the use of SE to optimize the quantizers for MPDQ reconstruction. Experimental results are presented in Section VI. Section VII concludes the paper.

C. Notation

Vectors and matrices will be written in boldface type ($\mathbf{A}$, $\mathbf{x}$, $\mathbf{y}$, ...) to distinguish from scalars written in normal weight (m, n, ...). Random and non-random quantities (or random variables and their realizations) are not distinguished typographically since the use of capital letters for random variables would conflict with the convention of using capital letters for matrices (or in the case of quantization, an operator on a vector rather than a scalar). The probability density function (p.d.f.) of random vector $\mathbf{x}$ is denoted $p_{\mathbf{x}}$, and the conditional p.d.f. of $\mathbf{y}$ given $\mathbf{x}$ is denoted $p_{\mathbf{y}|\mathbf{x}}$. When these densities are separable and identical across components, we use p_x for the scalar p.d.f. and $p_{y|x}$ for the scalar conditional p.d.f. Writing $x \sim \mathcal{N}(a, b)$ indicates that x is a Gaussian random variable with mean a and variance b. The resulting p.d.f. is written as $p_x(t) = \phi(t\,;\, a,\, b)$.

II. Background

This section establishes concepts and notations central to the paper. For a comprehensive tutorial history of quantization, we recommend [26]; and for an introduction to compressed sensing, [27].

A. Scalar Quantization

A K-level scalar quantizer $q : \mathbb{R} \to \mathbb{R}$ is defined by its *output levels* or *reproduction points* $\mathcal{C} = \{c_i\}_{i=1}^K$ and *(partition) cells* $\{q^{-1}(c_i)\}_{i=1}^K$. It can be decomposed into a composition of two mappings $q = \beta \circ \alpha$ where $\alpha : \mathbb{R} \to \{1, 2, \ldots, K\}$ is the *(lossy) encoder* and $\beta : \{1, 2, \ldots, K\} \to \mathcal{C}$ is the *decoder*. The boundaries of the cells are called *decision thresholds*. One may allow $K = \infty$ to denote that $\mathcal{C}$ is countably infinite.

A quantizer is called *regular* when each cell is a convex set, i.e., a single interval. Each cell of a regular scalar quantizer thus has a boundary of one point (if the cell is unbounded) or two points (if the cell is bounded). When the input to a

quantizer is a continuous random variable, it suffices to specify the cells of a K-point regular scalar quantizer by its decision thresholds $\{b_i\}_{i=0}^{K}$, with $b_0 = -\infty$ and $b_K = \infty$; the encoder satisfies

$$\alpha(x) = i \qquad \text{for } x \in (b_{i-1},\, b_i),$$

and the output for boundary points can be safely ignored.

The lossy encoder of a non-regular quantizer can be decomposed into the lossy encoder of a regular quantizer followed by a many-to-one integer-to-integer mapping. Suppose K-level non-regular scalar quantizer q' has decision thresholds $\{b'_i\}_{i=0}^{K'}$, and let α be the lossy encoder of a regular quantizer with these decision thresholds. Since q' is not regular, $K' > K$. Let $\alpha' : \mathbb{R} \to \{1, 2, \ldots, K\}$ denote the lossy encoder of q'. Then $\alpha' = \lambda \circ \alpha$, where

$$\lambda : \{1, 2, \ldots, K'\} \to \{1, 2, \ldots, K\}$$

is called a *binning function*, *labeling function*, or *index assignment*. The binning function is not invertible.

The *distortion* of a quantizer q applied to scalar random variable x is typically measured by the MSE

$$D = \mathbb{E}[(x - q(x))^2].$$

A quantizer is called optimal at fixed rate $R = \log_2 K$ when it minimizes distortion D among all K-level quantizers. To optimize scalar quantizers under MSE distortion, it suffices to consider only regular quantizers; a non-regular quantizer will never perform strictly better.

While regular quantizers are optimal for the standard lossy compression problem, non-regular quantizers are sometimes useful when some information aside from $q(x)$ is available when estimating x. Two key examples are Wyner–Ziv coding [23] and multiple description coding [24]. One method for Wyner–Ziv coding is to apply Slepian–Wolf coding across a block of samples after regular scalar quantization [28]; the Slepian–Wolf coding is binning, but across a block rather than for a single scalar. In multiple description scalar quantization [29], two binning functions are used that together are invertible but individually are not. In these uses of non-regular quantizers, side information aids in recovering x with resolution commensurate with K' while the rate is only commensurate with K, with $K' > K$.

A quantizer $\mathrm{Q} : \mathbb{R}^m \to \mathbb{R}^m$ is called a scalar quantizer when it is the Cartesian product of m scalar quantizers $q_i : \mathbb{R} \to \mathbb{R}$. In this paper, Q always represents a scalar quantizer with identical component quantizers q.

B. Compressed Sensing

An important value of the proposed MPDQ framework is that it can exploit non-Gaussian priors on the input vector $\mathbf{x}$. To illustrate this feature, we will apply the MPDQ algorithm to quantization problems in *compressed sensing* (CS) [30]–[32], which considers the reconstruction of sparse vectors $\mathbf{x}$ through randomized linear transforms.

A vector is *sparse* if it has a relatively small number of nonzero components. A central principle of CS is that sparse vectors $\mathbf{x}$ can be reconstructed from certain *underdetermined* linear transforms $\mathbf{z} = \mathbf{A}\mathbf{x}$, where $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $m < n$. Such linear transforms can thus be used as a form of "compression" on $\mathbf{x}$, reducing the vector's dimension from n to a smaller value m. Since many signals are naturally sparse in some domain, there are now a large number of works advocating CS methods in analog front ends prior to quantization to reduce the overall acquisition bit rate. However, properly understanding the rate–distortion performance of such approaches requires that we analyze and design CS reconstruction methods precisely accounting for the effects of quantization on the transform-domain measurements [33].

Analysis and design of CS reconstruction algorithms is challenging, even in the absence of quantization. Most approaches are based on either greedy heuristics (matching pursuit [34] and its variants with orthogonalization [35]–[37] and iterative refinement [38], [39]) and convex relaxations (basis pursuit [40], LASSO [41], Dantzig selector [42], and others). These methods are all nonlinear and their performance can be difficult to precisely characterize, particularly with quantization. Some initial performance bounds for CS reconstruction with quantization can be found in [33], [43]. In [44], high-resolution functional scalar quantization theory was used to design quantizers for LASSO estimation. The papers [2]–[4] consider alternate reconstruction algorithms that use the the partition cells of the quantizers that compose Q. Analyses of these methods produce performance bounds that are not generally tight. Moreover, the results are generally limited to specific sparsity priors as well as regular quantizers.

We will show here that the MPDQ framework enables CS reconstruction for a large class of sparse priors and essentially arbitrary quantization functions. Moreover, the method is computationally simple and, for certain large random transforms, admits an exact performance analysis.

III. Quantized Linear Expansions

This paper focuses on the general quantized measurement abstraction of

$$\mathbf{y} = \mathrm{Q}(\mathbf{A}\mathbf{x}), \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^n$ is a signal of interest, $\mathbf{A} \in \mathbb{R}^{m\times n}$ is a linear *mixing matrix*, and $\mathrm{Q} : \mathbb{R}^m \to \mathbb{R}^m$ is a scalar quantizer. We will be primarily interested in (per-component) MSE $n^{-1}\mathbb{E}[\|\mathbf{x}-\widehat{\mathbf{x}}\|^2]$ for various estimators $\widehat{\mathbf{x}}$ that depend on $\mathbf{y}$, $\mathbf{A}$, and Q. The cases of $m \geq n$ and $m < n$ are both of interest. We sometimes use $\mathbf{z} = \mathbf{A}\mathbf{x}$ to simplify expressions.

A. Overcomplete Expansions

Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ have rank n. Then $\{\mathbf{a}_i\}_{i=1}^{m}$ is a *frame* in $\mathbb{R}^n$, where $\mathbf{a}_i^T$ is row i of $\mathbf{A}$. Rank n can occur only with $m \geq n$, so $\mathbf{A}\mathbf{x}$ is called an *overcomplete expansion* of $\mathbf{x}$. In some cases of interest, the frame may be *uniform*, meaning $\|\mathbf{a}_i\| = 1$ for each i.

Commonly-used *linear reconstruction* forms estimate

$$\widehat{\mathbf{x}} = \mathbf{A}^{\dagger}\mathbf{y} = \mathbf{A}^{\dagger}\mathrm{Q}(\mathbf{A}\mathbf{x}), \tag{2}$$

where $\mathbf{A}^{\dagger} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$ is the pseudoinverse of $\mathbf{A}$. Linear reconstruction generally has MSE inversely proportional to m. For example, suppose the frame is uniform, $\mathbf{A}^T\mathbf{A} = m\mathbf{I}_n/n$,

and $\mathbf{x}$ is an unknown deterministic quantity. By modeling scalar quantization as the addition of zero-mean white noise, one can compute the MSE to be $n\sigma_d^2/m$ [45].

Even when an additive white noise model is accurate [46], the linear reconstruction (2) may be far from optimal. A nonlinear estimate may exploit the boundedness of the *single-sample consistent sets*

$$\mathcal{S}_i(y_i) = \{\mathbf{x} \in \mathbb{R}^n \,|\, q_i(z_i) = y_i\}, \qquad i = 1, 2, \ldots, m.$$

Assuming for now that scalar quantizer q_i is regular and its cells are bounded, the boundary of $\mathcal{S}_i(y_i)$ is two parallel hyperplanes. The full set of hyperplanes obtained for one index i by varying y_i over the output levels of q_i is called a hyperplane wave partition [47], as illustrated for a uniform quantizer in Figure 1(a). The set enclosed by two neighboring hyperplanes in a hyperplane wave partition is called a *slab*; one slab is shaded in Figure 1(a). Intersecting $\mathcal{S}_i(y_i)$ for n distinct indexes specifies an n-dimensional parallelotope as illustrated in Figure 1(b). Using more than n of these single-sample consistent sets restricts $\mathbf{x}$ to a finer partition, as illustrated in Figure 1(c) for $m = 3$.

The intersection

$$\mathcal{S}(\mathbf{y}) = \bigcap_{i=1}^{m} \mathcal{S}_i(y_i)$$

is called the *consistent set*. Since each $\mathcal{S}_i(y_i)$ is convex, one may reach $\mathcal{S}(\mathbf{y})$ asymptotically through a sequence of projections onto $\mathcal{S}_i(y_i)$ using each infinitely often [5], [6].

In a variety of settings, nonlinear estimates achieve MSE inversely proportional to m^2, which is the best possible dependence on m [47]. The first result of this sort was in [5]. When $\mathbf{A}$ is an oversampled discrete Fourier transform matrix and Q is a uniform quantizer, $\mathbf{z} = \mathbf{A}\mathbf{x}$ represents uniformly quantized samples above Nyquist rate of a periodic bandlimited signal. For this case, it was proven in [5] that any $\widehat{\mathbf{x}} \in \mathcal{S}(\mathbf{y})$ has $O(m^{-2})$ MSE, under a mild assumption on $\|\mathbf{x}\|$. This was extended empirically to arbitrary uniform frames in [7], where it was also shown that consistent estimates can be computed through a linear program. The techniques of alternating projections and linear programming suffer from high computational complexity; yet, since they generally find a corner of the consistent set (rather than the centroid), the MSE performance is suboptimal.

Full consistency is not necessary for optimal MSE dependence on m. It was shown in [8] that $O(m^{-2})$ MSE is guaranteed for a simple algorithm that uses each $\mathcal{S}_i(y_i)$ only once, recursively, under mild conditions on randomized selection of $\{\mathbf{a}_i\}_{i=1}^m$. These results were strengthened and extended to deterministic frames in [13].

Quantized overcomplete expansions arise naturally in acquisition subsystems such as ADCs, where m/n represents oversampling factor relative to Nyquist rate. In such systems, high oversampling factor may be motivated by a trade-off between MSE and power consumption or manufacturing cost: within certain bounds, faster sampling is cheaper than a higher number of quantization bits per sample [48]. However, high oversampling does not give a good trade-off between MSE and raw number of bits produced by the acquisition system: combining the proportionality of bit rate R to number of samples m with the best-case $\Theta(m^{-2})$ MSE, we obtain $\Theta(R^{-2})$ MSE; this is poor compared to the exponential decrease of MSE with R obtained with scalar quantization of Nyquist-rate samples.

Ordinarily, the bit-rate inefficiency of the raw output is made irrelevant by recoding, at or near Nyquist rate, soon after acquisition or within the ADC. An alternative explored in this paper is to combat this bit-rate inefficiency through the use of non-regular quantization.

B. Non-Regular Quantization

The bit-rate inefficiency of the raw output with regular quantization is easily understood with reference to Figure 1(c). After y_1 and y_2 are fixed, $\mathbf{x}$ is known to lie in the intersection of the shaded strips. Only four values of y_3 are possible (i.e., the solid hyperplane wave breaks $\mathcal{S}_1(1) \cap \mathcal{S}_2(0)$ into four cells), and bits are wasted if this is not exploited in the representation of y_3.

Recall the discussion of generating a non-regular quantizer by using a binning function λ in Section II-A. Binning does not change the boundaries of the single-sample consistent sets, but it makes these sets unions of slabs that may not even be connected. Thus, while binning reduces the quantization rate, in the absence of side information that specifies which slab contains $\mathbf{x}$ (at least with moderately high probability), it increases distortion significantly. The increase in distortion is due to *ambiguity* among slabs. Taking $m > n$ quantized samples together may provide adequate information to disambiguate among slabs, thus removing the distortion penalty.

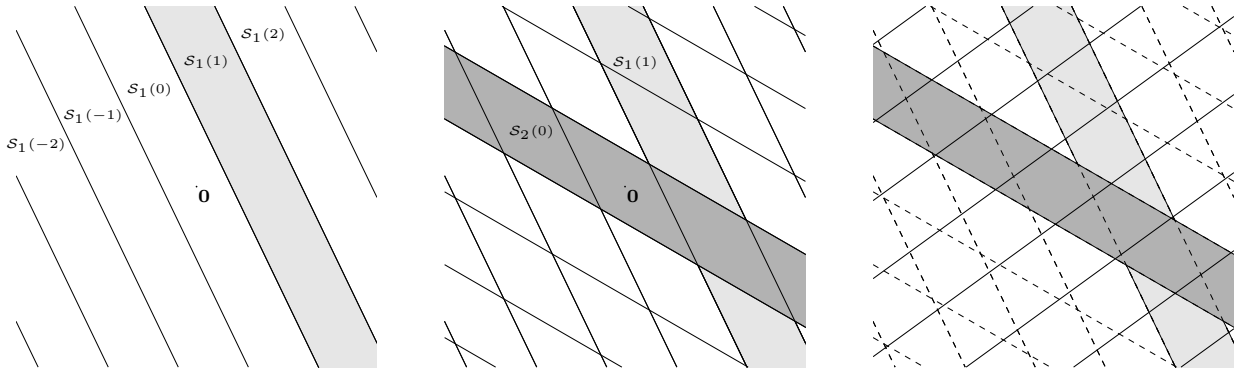
The key concepts in the use of non-regular quantization are illustrated in Figure 2. Suppose one quantized sample y_1 specifies a single-sample consistent set $\mathcal{S}_1(y_1)$ composed of two slabs, such as the shaded region in Figure 2(a). A second quantized sample y_2 will not disambiguate between the two slabs. In the example shown in Figure 2(b), $\mathcal{S}_2(y_2)$ is composed of two slabs, and $\mathcal{S}_1(y_1) \cap \mathcal{S}_2(y_2)$ is the union of four connected sets. A third quantized sample y_3 may now completely disambiguate; the particular example of $\mathcal{S}_3(y_3)$ shown in Figure 2(c) makes $\mathcal{S} = \mathcal{S}_1(y_1) \cap \mathcal{S}_2(y_2) \cap \mathcal{S}_3(y_3)$ a single convex set.

When the quantized samples together completely disambiguate the slabs as in the example, the rate reduction from binning comes with no increase in distortion. The price to pay comes in complexity of estimation.

The use of binned quantization of linear expansions was introduced in [49], where the only reconstruction method proposed is intractable in high dimensions because it is combinatorial over the binning functions. Specifically, using the notation from Section II-A, let the quantizer forming y_i be defined by $(\alpha_i, \beta_i, \lambda_i)$. Then $\lambda_i^{-1}(\beta_i^{-1}(y_i))$ will be a set of possible values of $\alpha_i(z_i)$ specified by y_i. One can try every combination, i.e., element of

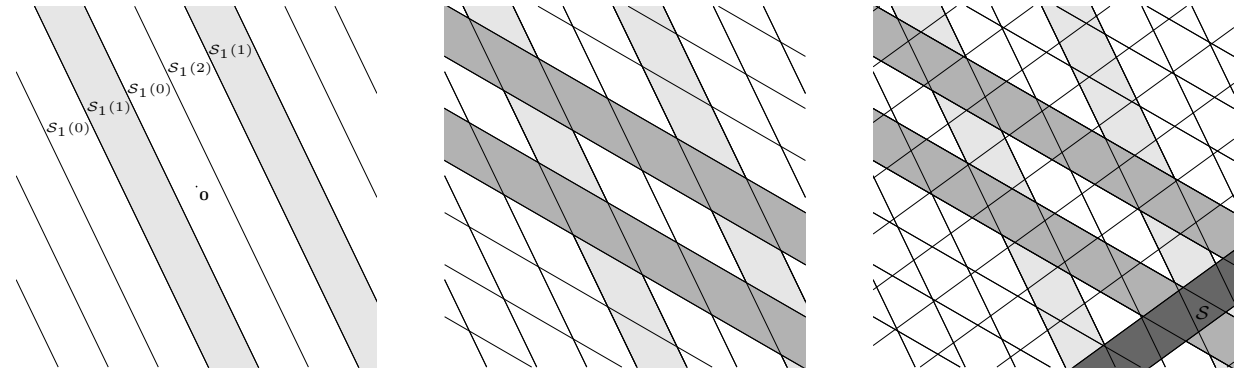
$$\lambda_1^{-1}(\beta_1^{-1}(y_1)) \times \lambda_2^{-1}(\beta_2^{-1}(y_2)) \times \cdots \times \lambda_m^{-1}(\beta_m^{-1}(y_m)), \quad (3)$$

to seek a consistent estimate. If the binning is effective, most combinations yield an empty consistent set; if the slabs are



(a) (b) (c)

Fig. 1: Visualizing the information present in a quantized overcomplete expansion of $\mathbf{x} \in \mathbb{R}^2$ when q_i is a regular quantizer. (a) A single hyperplane wave partition with one single-sample consistent set shaded. (b) Partition boundaries from two hyperplane waves; $\mathbf{x}$ is specified to the intersection of two single-sample consistent sets, which is a bounded convex cell. (c) Partition from part (b) in dashed lines with a third hyperplane wave added in solid lines.



(a) (b) (c)

Fig. 2: Visualizing the information present in a quantized overcomplete expansion of $\mathbf{x} \in \mathbb{R}^2$ when using non-regular (binned) quantizers. (a) A single hyperplane wave partition with one single-sample consistent set shaded. Note that binning makes the shaded set not connected. (b) Partition boundaries from two hyperplane waves; $\mathbf{x}$ is specified to the intersection of two single-sample consistent sets, which is now the union of four convex cells. (c) A third sample now specifies $\mathbf{x}$ to within a consistent set $\mathcal{S}$ that is convex.

disambiguated, exactly one combination yields a nonempty set, which is then the consistent set $\mathcal{S}$. This technique has complexity exponential in m (assuming non-trivial binning). The recent paper [50] provides bounds on reconstruction error for consistent estimation with binned quantization; it does not address algorithms for reconstruction.

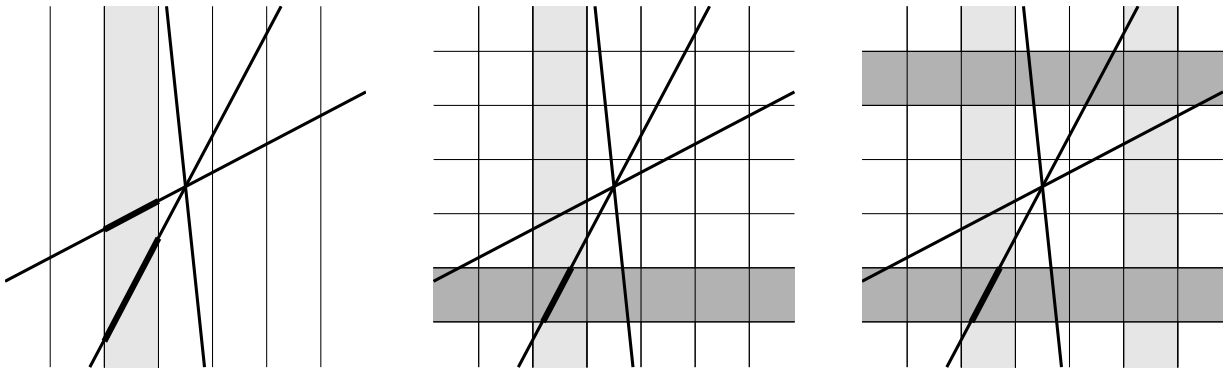
This paper provides a tractable and effective method for reconstruction from a quantized linear expansion with non-regular quantizers. To the best of our knowledge, this is the first such method.

C. Undercomplete Expansions

Maintaining the quantized measurement model (1), let us turn to the case of $m < n$. Since the rank of $\mathbf{A}$ is less than n, $\mathbf{A}$ is a many-to-one mapping. Thus, even without quantization, one cannot recover $\mathbf{x}$ from $\mathbf{Ax}$. Rather, $\mathbf{Ax}$ specifies a proper subspace of $\mathbb{R}^n$ containing $\mathbf{x}$; when $\mathbf{A}$ is in general position, the subspace is of dimension $n - m$. Quantization increases the ambiguity in the value of $\mathbf{x}$, yielding consist sets similar to those depicted in Figures 1(a) and 2(a). However, as described in Section II-B, knowledge that $\mathbf{x}$ is sparse or approximately sparse could be exploited to enable accurate estimation of $\mathbf{x}$ from $\mathrm{Q}(\mathbf{Ax})$.

For ease of explanation, consider only the case where $\mathbf{x}$ is known to be k-sparse with $k < m$. Let $\mathcal{J} \subset \{1, 2, \ldots, n\}$ be the support (sparsity pattern) of $\mathbf{x}$, with $|\mathcal{J}| = k$. The product $\mathbf{Ax}$ is equal to $\mathbf{A}_{\mathcal{J}}\mathbf{x}_{\mathcal{J}}$, where $\mathbf{x}_{\mathcal{J}}$ denotes the restriction of the domain of $\mathbf{x}$ to $\mathcal{J}$ and $\mathbf{A}_{\mathcal{J}}$ is the $m \times k$ submatrix of $\mathbf{A}$ containing the $\mathcal{J}$-indexed columns. Assuming $\mathbf{A}_{\mathcal{J}}$ has rank k (i.e., full rank), $\mathrm{Q}(\mathbf{Ax}) = \mathrm{Q}(\mathbf{A}_{\mathcal{J}}\mathbf{x}_{\mathcal{J}})$ is a quantized *overcomplete* expansion of $\mathbf{x}_{\mathcal{J}}$. All discussion of estimation of $\mathbf{x}_{\mathcal{J}}$ from the previous subsections thus applies, assuming $\mathcal{J}$ is known.

The key remaining issue is that $\mathrm{Q}(\mathbf{Ax})$ may or may not provide enough information to infer $\mathcal{J}$. In an overcomplete representation, most vectors of quantizer outputs cannot occur; this redundancy was used to enable binning in Figure 2, and it can be used to show that certain subsets $\mathcal{J}$ are inconsistent



(a) (b) (c)

Fig. 3: Visualizing the information present in a quantized undercomplete expansion $\mathrm{Q}(\mathbf{Ax})$ of a 1-sparse signal $\mathbf{x} \in \mathbb{R}^3$ when $\mathbf{Ax} \in \mathbb{R}^2$. The depicted 2-dimensional plane represents the vector of measurements $\mathbf{z} = \mathbf{Ax}$. Since $\mathbf{x}$ is 1-sparse, the measurement lies in a union of 1-dimensional subspaces (the angled solid lines); since $\mathbf{x}$ is 3 dimensional, there are three such subspaces. (a) Scalar quantization of z_1 divides the plane of possible values for $\mathbf{z}$ into vertical strips. One particular value of $y_1 = q_1(z_1)$ does not specify which entry of $\mathbf{x}$ is nonzero since the shaded strip intersects all the angled solid lines. For each possible support, the value of the nonzero entry is specified to an interval. (b) Scalar quantization of both components of $\mathbf{z}$ specifies $\mathbf{z}$ to a rectangular cell. In most cases, including the one highlighted, the quantized values specify which entry of $\mathbf{x}$ is nonzero because only one angled solid line intersects the cell. The value of the nonzero entry is specified to an interval. (c) In many cases, including the one highlighted, the quantizers can be non-regular (binned) and yet still uniquely specify which entry of $\mathbf{x}$ is nonzero.

with the sparse signal model. In principle, one may enumerate the sets $\mathcal{J}$ of size k and apply a consistent reconstruction method for each $\mathcal{J}$. If only one candidate $\mathcal{J}$ yields a nonempty consistent set, then $\mathcal{J}$ is determined. This is intractable except for small problem sizes because there are $\binom{n}{k}$ candidates for $\mathcal{J}$.

The key concepts are illustrated in Figure 3. To have an interpretable diagram with $k < m < n$, we let $(k, m, n) = (1, 2, 3)$ and draw the space of unquantized measurements $\mathbf{z} \in \mathbb{R}^2$. (This contrasts with Figures 1 and 2 where the space of $\mathbf{x} \in \mathbb{R}^2$ is drawn.) The vector $\mathbf{x}$ has one of $\binom{n}{k} = \binom{3}{1} = 3$ possible supports $\mathcal{J}$. Thus, $\mathbf{z}$ lies in one of 3 subspaces of dimension 1, which are depicted by the angled solid lines. Scalar quantization of $\mathbf{z}$ corresponds to separable partitioning of $\mathbb{R}^2$ with cell boundaries aligned with coordinate axes, as shown with lighter solid lines.

Only one quantized measurement y_1 is not adequate to specify $\mathcal{J}$, as shown in Figure 3(a) by the fact that a single shaded cell intersects all the subspaces.[1] Two quantized measurements together will usually specify $\mathcal{J}$, as shown in Figure 3(b) by the fact that only one subspace intersects the specified square cell; for fixed scalar quantizers, ambiguity becomes less likely as k decreases, n increases, m increases, or $\|x\|$ increases. Figure 3(c) shows a case where non-regular (binned) quantization still allows unambiguous determination of $\mathcal{J}$.

The naïve reconstruction method implied by Figure 3(c) is to search combinatorially over both $\mathcal{J}$ and the combinations in (3); this is extremely complex. While the use of binning for quantized undercomplete expansions of sparse signals has appeared in the literature, first in [49] and later in [50], to the best of our knowledge this paper is the first to provide a tractable and effective reconstruction method.

[1] Intersections with two subspaces are shown within the range of the diagram.

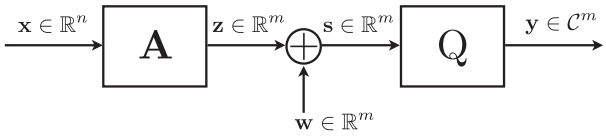


Fig. 4: Quantized linear measurement model considered in this work. Vector $\mathbf{x} \in \mathbb{R}^n$ with an i.i.d. prior is estimated from scalar quantized measurements $\mathbf{y} \in \mathcal{C}^m$. The quantizer input is the sum of $\mathbf{z} = \mathbf{Ax} \in \mathbb{R}^m$ and an i.i.d. Gaussian noise vector $\mathbf{w} \in \mathbb{R}^m$.

IV. Estimation from Quantized Samples

In this section, we provide the Bayesian formulation of the reconstruction problem from quantized measurements and introduce the MPDQ algorithm as a low complexity alternative to the belief propagation.

A. Bayesian Formulation

We now specify more explicitly the class of problems for which we derive new estimation algorithms. Generalizing (1), let

$$\mathbf{y} = \mathrm{Q}(\mathbf{z} + \mathbf{w}) \qquad \text{where} \qquad \mathbf{z} = \mathbf{Ax}, \tag{4}$$

as depicted in Figure 4. The input vector $\mathbf{x} \in \mathbb{R}^n$ is random with i.i.d. entries with prior p.d.f. p_x. The linear mixing matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ is random with i.i.d. entries $a_{ij} \sim \mathcal{N}(0, 1/m)$. The (pre-quantization) additive noise $\mathbf{w} \in \mathbb{R}^m$ is random with i.i.d. entries $w_i \sim \mathcal{N}(0, \sigma^2)$. The quantizer Q is a scalar

quantizer with identical component quantizers q and has K output levels. Note that, the mapping from $\mathbf{z}$ to $\mathbf{y}$ is a separable probabilistic mapping with identical marginals. Specifically, quantized measurement y_i indicates $s_i \in q^{-1}(y_i)$, so each component *output channel* can be characterized as

$$p_{y|z}(y \mid z) = \int_{q^{-1}(y)} \phi\left(t;\, z,\, \sigma^2\right)\, \mathrm{d}t,$$

where ϕ is the Gaussian p.d.f. We then construct the following conditional probability distribution over random vector $\mathbf{x}$ given the measurements $\mathbf{y}$

$$\begin{aligned} p_{\mathbf{x}|\mathbf{y}}\left(\mathbf{x} \,|\, \mathbf{y}\right) &\propto p_{\mathbf{y}|\mathbf{z}}\left(\mathbf{y} \,|\, \mathbf{z}\right) p_{\mathbf{x}}\left(\mathbf{x}\right) \\ &\propto \prod_{i=1}^{m} p_{y|z}\left(y_i \mid z_i\right) \prod_{j=1}^{n} p_x(x_j), \end{aligned} \tag{5}$$

where $\propto$ denotes identity after normalization to unity and $z_i = (\mathbf{A}\mathbf{x})_i$. The posterior distribution (5) of the signal provides a complete statistical characterization of the problem. In particular, we wish to obtain a tractable approximation to the MMSE estimator of $\mathbf{x}$ specified by

$$\widehat{\mathbf{x}}_{\mathrm{MMSE}} = \mathbb{E}\left[\mathbf{x} \,|\, \mathbf{y}\right]. \tag{6}$$

B. Loopy Belief Propagation

Loopy BP [51], [52] is a popular computational method to approximate the MMSE estimator $\widehat{\mathbf{x}}_{\mathrm{MMSE}}$ iteratively. The method is based on computation of marginal probability distributions of $p_{\mathbf{x}|\mathbf{y}}$. To apply loopy BP to the quantization reconstruction problem, construct a bipartite factor graph $G = (V, F, E)$ where V denotes the set of n *variable* or *input* nodes associated with transform inputs x_j, $j = 1, \ldots, n$, and F is the set of m *factor* or *output* nodes associated with the transform outputs z_i, $i = 1, \ldots, m$. The set of (undirected) edges E consist of the pairs (i, j) such that $A_{ij} \neq 0$. Loopy BP passes the following messages along the edges E of the graph:

$$\mu^t_{i \leftarrow j}(x_j) \quad \propto \quad p_x(x_j) \prod_{\ell \neq i} \mu^t_{\ell \rightarrow j}(x_j), \tag{7a}$$

$$\mu^t_{i \rightarrow j}(x_j) \quad \propto \quad \int p_{y|z}(y_i \mid z_i) \prod_{k \neq j} \mu^{t-1}_{i \leftarrow k}(x_j)\, \mathrm{d}\mathbf{x}_{\setminus j}, \tag{7b}$$

where integration is over all the elements of $\mathbf{x}$ except x_j. We refer to messages $\{\mu_{i \leftarrow j}\}_{(i,j) \in E}$ as variable updates and to messages $\{\mu_{i \rightarrow j}\}_{(i,j) \in E}$ as factor updates. BP is initialized by setting $\mu^0_{i \leftarrow j}(x_j) = p_x(x_j)$. The approximate marginal distribution is computed as

$$\hat{p}^t_{x_j|\mathbf{y}}(x_j \mid \mathbf{y}) \propto p_x(x_j) \prod_{i=1}^{m} \mu^t_{i \rightarrow j}(x_j). \tag{8}$$

Finally, the component $\hat{x}^t_j$ of the estimate $\widehat{\mathbf{x}}^t$ is computed as

$$\hat{x}^t_j = \int_{\mathbb{R}} x\, \hat{p}^t_{x_j|\mathbf{y}}(x \mid \mathbf{y}) \mathrm{d}x. \tag{9}$$

When the graph G induced by the matrix $\mathbf{A}$ is cycle free, the BP outputs will converge to the true marginals of the posterior density $p_{\mathbf{x}|\mathbf{y}}$. However, for general $\mathbf{A}$, loopy BP is only approximate – the reader is referred to the references above for a general discussion on the performance of loopy BP. We will discuss the performance of the specific variant of loopy BP used in MPDQ in detail in Section V-A. What is important here is the computational complexity of loopy BP: Direct implementation of loopy BP is impractical for the de-quantization problem unless $\mathbf{A}$ is very sparse. For dense $\mathbf{A}$, the algorithm must compute the marginal of a high-dimensional distribution at each measurement node; i.e., the integration in (7b) is over many variables. Furthermore, integration must be approximated through some discrete quadrature rule.

C. Message-Passing De-Quantization

To overcome the computational complexity of loopy BP, the proposed MPDQ algorithm uses a Gaussian approximation. Gaussian approximations of loopy BP have been used in successfully in CDMA multiuser detection [15], [17], [18] and, more recently, in compressed sensing [14], [16], [22]. We apply the specific generalized approximate message passing (GAMP) method in [14], which allows for nonlinear output channels. The approximations are based on a Central Limit Theorem and other second-order approximations at the measurement nodes. Details can be found in [14]. Here, we simply restate the algorithm as applied to the specific de-quantization problem.

Given the measurements $\mathbf{y} \in \mathcal{C}^m$, the measurement matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, the noise variance σ^2, the mapping q of the scalar quantizer, and the prior $p_{\mathbf{x}}$, the MPDQ estimation proceeds as follows:

1) *Initialization:* Set $t = 1$ and evaluate

$$\widehat{\mathbf{x}}^0 = \mathbb{E}\left[\mathbf{x}\right], \tag{10a}$$

$$\mathbf{v}^0_x = \mathrm{var}\left[\mathbf{x}\right], \tag{10b}$$

$$\hat{\mathbf{s}}^0 = 0, \tag{10c}$$

where the expected value and the variance are with respect to the prior $p_{\mathbf{x}}$.

2) *Factor update:* First, compute the linear step

$$\hat{\mathbf{p}}^t = \mathbf{A}\widehat{\mathbf{x}}^{t-1} - \mathbf{v}^t_p \bullet \hat{\mathbf{s}}^{t-1}, \tag{11a}$$

$$\mathbf{v}^t_p = \left(\mathbf{A} \bullet \mathbf{A}\right) \mathbf{v}^{t-1}_x, \tag{11b}$$

where $\bullet$ denotes the Hadamard product (i.e. component-wise multiplication). Then, evaluate the nonlinear step

$$\hat{\mathbf{s}}^t = \mathrm{E}_1\left(\mathbf{y}, \hat{\mathbf{p}}^t, \mathbf{v}^t_p + \sigma^2\mathbf{e}; q\right), \tag{12a}$$

$$\mathbf{v}^t_s = \mathrm{V}_1\left(\mathbf{y}, \hat{\mathbf{p}}^t, \mathbf{v}^t_p + \sigma^2\mathbf{e}; q\right), \tag{12b}$$

where $\mathbf{e}$ is an all-ones vector. The scalar functions E_1 and V_1 are applied component-wise and given by

$$\mathrm{E}_1(y, \hat{p}, v_p; q) = \frac{1}{v_p}\left(\mathbb{E}\left[z | z \in q^{-1}(y)\right] - \hat{p}\right), \tag{13a}$$

$$\mathrm{V}_1(y, \hat{p}, v_p; q) = \frac{1}{v_p}\left(1 - \frac{\mathrm{var}\left[z | z \in q^{-1}(y)\right]}{v_p}\right). \tag{13b}$$

The expected value and the variance are evaluated with respect to $z \sim \mathcal{N}(\hat{p}, v_p)$.

3) *Variable update:* First, compute the linear step

$$\hat{\mathbf{r}}^t = \widehat{\mathbf{x}}^{t-1} + \mathbf{v}_r^t \bullet \left(\mathbf{A}^{\mathrm{T}} \hat{\mathbf{s}}^t\right), \tag{14a}$$

$$\mathbf{v}_r^t = \left((\mathbf{A} \bullet \mathbf{A})^{\mathrm{T}} \mathbf{v}_s^t\right)^{-1}. \tag{14b}$$

Then, evaluate the nonlinear step

$$\widehat{\mathbf{x}}^t = \mathrm{E}_2\left(\hat{\mathbf{r}}^t, \mathbf{v}_r^t; p_x\right), \tag{15a}$$

$$\mathbf{v}_x^t = \mathrm{V}_2\left(\hat{\mathbf{r}}^t, \mathbf{v}_r^t; p_x\right), \tag{15b}$$

where the scalar functions E_2 and V_2 are applied component-wise and given by

$$\mathrm{E}_2(\hat{r}, v_r; p_x) = \mathbb{E}\left[x \mid \hat{r}\right], \tag{16a}$$

$$\mathrm{V}_2(\hat{r}, v_r; p_x) = \mathrm{var}\left[x \mid \hat{r}\right]. \tag{16b}$$

The expected value and the variance are evaluated with respect to $p_{x|\hat{r}}(\cdot \mid \hat{r}) \propto \phi(\cdot; \hat{r}, v_r) p_x(\cdot)$. This is essentially a scalar AWGN denoising problem with noise $w \sim \mathcal{N}(0, v_r)$.

4) Set $t \leftarrow t+1$ and proceed to step 2).

For each iteration $t = 1, 2, 3, \ldots$, the proposed update rules produce estimates $\widehat{\mathbf{x}}^t$ of the true signal $\mathbf{x}$. Thus the algorithm reduces the intractable high-dimensional integration to a sequence of matrix-vector products and scalar nonlinearities. Note that scalar inequalities (13a) and (13b) are easy to evaluate since they admit closed-form expressions in terms of $\mathrm{erf}\,(z) = \frac{2}{\sqrt{\pi}} \int_0^z \mathrm{e}^{-t^2}\,\mathrm{d}t$. Depending on the prior distribution p_x, the scalar inequalities (16a) and (16b) either admit closed-form expressions or can be implemented as a look-up table.

V. Quantizer Optimization

A remarkable fact about MPDQ is that, under large random transforms, the MPDQ performance can be precisely predicted by a scalar state evolution (SE) analysis presented in this section. One can then apply the SE analysis to the design of MSE optimal quantizers under MPDQ reconstruction. Superior reconstruction performance of MPDQ with such quantizers is numerically confirmed in the Section VI.

A. State Evolution for MPDQ

The equations (10)–(16) are easy to implement, however they provide us no insight into the performance of the algorithm. The goal of the SE equation is to describe the asymptotic behavior of MPDQ under large random measurement matrices $\mathbf{A}$. For $t \geq 1$, it is defined as a recursion

$$\tau^t = \mathrm{F}_{\mathrm{SE}}\left(\tau^{t-1}; \beta, p_x, q, \sigma^2\right), \tag{17}$$

where the scalar function F_{SE} implicitly depends on $\beta = n/m$, the prior distribution p_x, the mapping q of the scalar quantizer, AWGN variance σ^2, and is given by

$$\mathrm{F}_{\mathrm{SE}}(\tau; \beta, p_x, q, \sigma^2) = \overline{\mathrm{V}}_2\left(\frac{1}{\overline{\mathrm{V}}_1\left(\tau; \beta, q, \sigma^2\right)}; p_x\right), \tag{18a}$$

$$\overline{\mathrm{V}}_1(\nu; \beta, q, \sigma^2) = \mathbb{E}\left[\mathrm{V}_1\left(y, \hat{p}, \beta\nu + \sigma^2; q\right)\right], \tag{18b}$$

$$\overline{\mathrm{V}}_2(\nu; p_x) = \mathbb{E}\left[\mathrm{V}_2\left(\hat{r}, \nu; p_x\right)\right], \tag{18c}$$

where V_1 and V_2 are defined in (13b) and (16b), respectively. The recursion is initialized by setting $\tau^0 = \mathrm{var}[x]$, with $x \sim p_x$. The expectation in (18b) is taken over $p_{y|z}$ and $(z, \hat{p}) \sim \mathcal{N}(0, \mathbf{C}(\nu))$, where covariance matrix is given by

$$\mathbf{C}\left(\nu\right) = \begin{pmatrix} \beta\tau^0 & \beta\tau^0 - \nu \\ \beta\tau^0 - \nu & \beta\tau^0 - \nu \end{pmatrix}. \tag{19}$$

Similarly, the expectation in (18c) is taken over the scalar random variable $\hat{r} = x + w$, with $x \sim p_x$ and $w \sim \mathcal{N}(0, \nu)$.

One of the main results of [14], which is an extension of the analysis in [19], was to demonstrate the convergence of the error performance of the GAMP algorithm to the SE equations. Specifically, these works consider the case where $\mathbf{A}$ is an i.i.d. Gaussian matrix, $\mathbf{x}$ is i.i.d. with a prior p_x and $m, n \to \infty$ with $n/m \to \beta$. Then, under some further technical conditions, it is shown that for any fixed iteration number t, the empirical joint distribution of the components $(x_j, \widehat{x}_j^t)$ of the unknown vector $\mathbf{x}$ and its estimate $\widehat{\mathbf{x}}^t$ converges to a simple scalar equivalent model parameterized by the outputs of the SE equations. From the scalar equivalent model, one can compute any asymptotic componentwise performance metric. It can be shown, in particular, that the asymptotic MSE is given simply by τ^t. That is,

$$\tau^t = \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} |x_j - \widehat{x}_j^t|^2 = \lim_{n\to\infty} \frac{1}{n} \|\mathbf{x} - \widehat{\mathbf{x}}^t\|^2. \tag{20}$$

Thus, τ^t can be used as a metric for the design and analysis of the quantizer, although other non-squared error distortions could also be considered. Although our simulations will consider dense transforms $\mathbf{A}$, similar SE equations can be derived for certain large sparse matrices [15]–[18]. In this case, when the fixed points of the SE equations are unique, then, it can in fact be shown that the approximate message passing method is, in fact, mean-squared error optimal.

To conclude, despite the fact that the prior on $\mathbf{x}$ may be non-Gaussian and the quantizer function Q is nonlinear, one can precisely characterize the exact asymptotic behavior of MPDQ at least for large random transforms.

B. Optimization

Ordinarily, quantizer designs depend on the distribution of the quantizer input, with an implicit aim of minimizing the MSE between the quantizer input and output. Often, only uniform quantizers are considered, in which case the "design" is to choose the loading factor of the quantizer. When quantized data is used as an input to a nonlinear function, overall system performance may be improved by adjusting the quantizer designs appropriately [25]. In the present setting, conventional quantizer design minimizes $\mathbb{E}[\|\mathbf{z} - \mathrm{Q}(\mathbf{z})\|^2]$, but minimizing $\mathbb{E}[\|\mathbf{x} - \widehat{\mathbf{x}}\|^2]$ is desired instead.

The SE description of MPDQ performance facilitates the desired optimization. By implementing the SE equations for MPDQ, we can make use of the convergence result (20) to recast our optimization problem to

$$\mathrm{Q}^* = \arg\min_{\mathrm{Q}} \left\{ \lim_{t\to\infty} \tau^t \right\}, \tag{21}$$

where minimization is done over K-level scalar quantizers. Based on (20), the optimization is equivalent to finding the

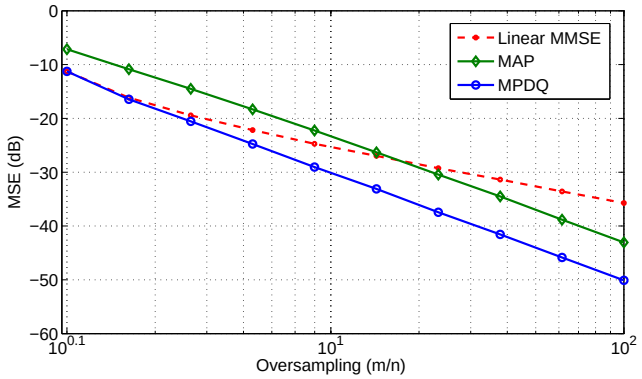


Fig. 5: Performance comparison for oversampled observation of a jointly Gaussian signal vector (no sparsity). MPDQ outperforms linear MMSE and MAP estimators.

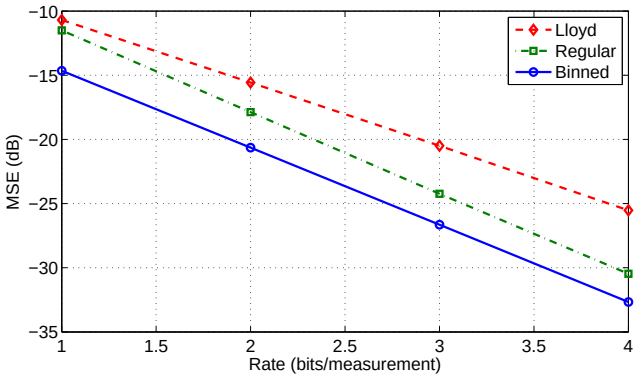


Fig. 6: Performance comparison of MPDQ with optimal uniform quantizers under Gaussian prior for regular and binned quantizers.

quantizer that minimizes the asymptotic MSE. In the optimization (21), we have considered the limit in the iterations, $t \to \infty$. One can also consider the optimization with a finite t, although our simulations exhibit close to the limiting performance with a relatively small number of iterations.

It is important to note that the SE recursion behaves well under quantizer optimization. This is due to the fact that SE is independent of actual output levels and small changes in the quantizer boundaries result in only minor change in the recursion (see (18b)). Although closed-form expressions for the derivatives of τ^t for large t's are difficult to obtain, we can approximate them by using finite difference methods. Finally, the recursion itself is fast to evaluate, which makes the scheme in (21) practically realizable under standard optimization methods.

VI. Experimental Results

A. Overcomplete Expansions

Consider overcomplete expansion of $\mathbf{x}$ as discussed in Section III-A. We generate the signal $\mathbf{x}$ with i.i.d. elements from the standard Gaussian distribution $x_j \sim \mathcal{N}(0,1)$. We form the measurement matrix $\mathbf{A}$ from i.i.d. zero-mean Gaussian random variables. To concentrate on the degradation due to quantization we assume noiseless measurement model (1); i.e., $\sigma^2 = 0$ in (4).

Figure 5 presents squared-error performance of three estimation algorithms while varying the oversampling ratio m/n and holding $n = 100$. To generate the plot we considered estimation from measurements discretized by a 16-level regular uniform quantizer. We set the granular region of the quantizer to $[-3\sigma_z, 3\sigma_z]$, where $\sigma_z^2 = n/m$ is the variance of the measurements. For each value of m/n, 200 random realizations of the problem were generated; the curves show the median-squared error performance over these 200 Monte Carlo trials. We compare error performance of MPDQ against two other common reconstruction methods: linear MMSE and maximum a posteriori probability (MAP). The MAP estimator was implemented using quadratic programming (QP).

The MAP estimation is type of *consistent reconstruction* method proposed in [5]–[13]; since the prior is a decreasing function of $\|\mathbf{x}\|$, the MAP estimate $\widehat{\mathbf{x}}$ is the vector consistent with $\mathrm{Q}(\mathbf{A}\widehat{\mathbf{x}})$ of minimum Euclidean norm. In the earlier works, it is argued that consistent reconstruction methods offer improved performance over linear estimation, particularly at high oversampling factors. We see in Figure 5 that MAP estimation does indeed outperform linear MMSE at high oversampling. However, MPDQ offers significantly better performance than both LMMSE and MAP, with more than 5 dB improvement for many values of m/n. In particular, this reinforces that MAP is suboptimal because it finds a corner of the consistent set, rather than the centroid. Moreover, the MPDQ method is actually computationally simpler than MAP, which requires the solution to a quadratic program.

With Figure 6 we turn to a comparison among quantizers, all with MPDQ reconstruction, $n = 100$, $m = 200$, and $\mathbf{x}$ and $\mathbf{A}$ distributed as above. To demonstrate the improvement in rate–distortion performance that is possible with non-regular quantizers, we consider simple *uniform modulo* quantizers

$$\mathrm{Q}(z) = \left\lfloor \frac{z}{\Delta} \right\rfloor \mod N, \tag{22}$$

where Δ is the size of the quantization cells. These quantizers map the entire real line $\mathbb{R}$ to the set $\{0, 1, \ldots, N-1\}$ in a periodic fashion.

We compare three types of quantizers: those optimized for MSE of the measurements (*not* the overall reconstruction MSE) using Lloyd's algorithm [26], regular uniform quantizers with loading factors optimized for reconstruction MSE using SE analysis, and (non-regular) uniform modulo quantizers with Δ optimized for reconstruction MSE using SE analysis. The last two quantizers were obtained by solving (21) via the standard SQP method found in MATLAB. The uniform modulo quantizer achieves the best rate–distortion performance, while the performance of the quantizer designed with Lloyd's algorithm is comparatively poor. The stark suboptimality of the latter is due to the fact that it optimizes the MSE only between quantizer inputs and outputs, ignoring the nonlinear estimation algorithm following the quantizer.

It is important to point out that, without methods such as MPDQ, estimation with a modulo quantizer such as (22) is not even computationally possible in works such as [5]–[13], since

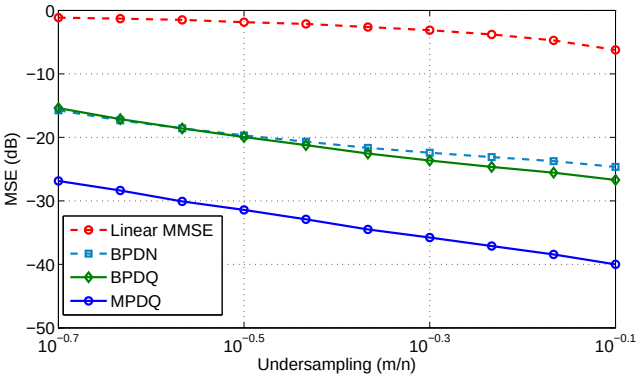


Fig. 7: Performance comparison of MPDQ with LMMSE, BPDN, and BPDQ (with moment $p = 4$) for estimation from compressive measurements.

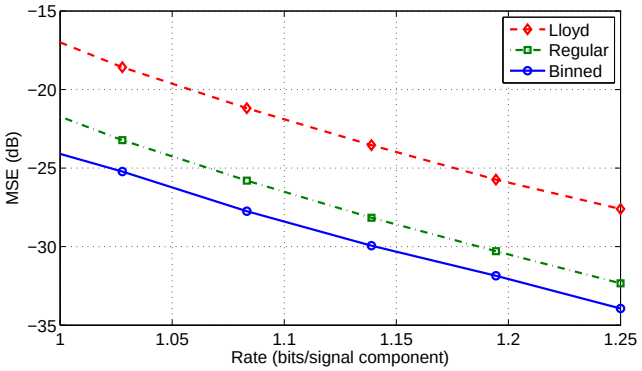


Fig. 8: Performance comparison of MPDQ with optimal uniform quantizers under Gauss-Bernoulli prior for regular and binned quantizers.

the consistent set is non-convex and consists of a disjoint union of convex sets. Beyond the performance improvements, we believe that MPDQ provides the first computationally-tractable and systematic method for such non-convex quantization reconstruction problems.

B. Compressive Sensing with Quantized Measurements

We next consider estimation of an n-dimensional sparse signal $\mathbf{x}$ from $m < n$ random measurements—a problem considered in quantized compressed sensing [2]–[4]. We assume that the signal $\mathbf{x}$ is generated with i.i.d. elements from the Gauss–Bernoulli distribution

$$x_j \sim \begin{cases} \mathcal{N}(0, 1/\rho), & \text{with probability } \rho; \\ 0, & \text{with probability } 1-\rho, \end{cases} \tag{23}$$

where ρ is the sparsity ratio that represents the average fraction of nonzero components of $\mathbf{x}$. In the following experiments we assume $\rho = 1/32$. Similarly to the overcomplete case, we form the measurement matrix $\mathbf{A}$ from i.i.d. Gaussian random variables and we assume no additive noise ($\sigma^2 = 0$ in (4)).

Figure 7 compares MSE performance of MPDQ with three other standard reconstruction methods. In particular, we consider linear MMSE and the Basis Pursuit DeNoise (BPDN) program [53]

$$\widehat{\mathbf{x}} = \underset{\mathbf{x}\in\mathbb{R}^n}{\arg\min} \|\mathbf{x}\|_1 \text{ s.t. } \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_p \leq \epsilon,$$

where $p = 2$ and $\epsilon \in \mathbb{R}_+$ is the parameter representing the noise power. In the same figure, we additionally plot the error performance of the Basis Pursuit DeQuantizer (BPDQ)[2] of moment p, proposed in [3], which solves the problem above for $p \geq 2$. It has been argued in [3] that BPDQ offers better error performance compared to the standard BPDN as the number of samples m increases with respect to the sparsity k of the signal $\mathbf{x}$.

We obtain the curves by varying the ratio m/n and holding $n = 1024$. We perform estimation from measurements obtained from a 16-level regular uniform quantizer with granular region of length $2\|\mathbf{A}\mathbf{x}\|_\infty$ centered at the origin.

[2]The source codes for the BPDQ algorithm can be downloaded from http://wiki.epfl.ch/bpdq

The figure plots the median of the squared error from 1000 Monte Carlo trials for each value of m/n. For basis pursuit methods we optimize the parameter ϵ for the best squared error performance; in practice this oracle-aided performance would not be achieved. The top curve (worst performance) is for linear MMSE estimation; and middle curves are for the basis pursuit estimators BPDN and BPDQ with moment $p = 4$. As expected, BPDQ achieves a notable 2 dB reduction in MSE compared to BPDN for high values of m, however MPDQ significantly outperforms both methods over the whole range of m/n. Note also that MPDQ is significantly faster than both BPDN and BPDQ. For example, in Figure 7 the average reconstruction times—across all realizations and undersampling rates—were 7.45, 19.95, and 4.52 seconds for BPDN, BPDQ, and MPDQ, respectively.

In Figure 8, we compare the performance of MPDQ under three quantizers consider before: those optimized for MSE of the measurements using Lloyd's algorithm, and regular and non-regular quantizers optimized for reconstruction MSE using SE analysis. Note that MPDQ is the first tractable reconstruction method for compressive sensing that handles non-regular quantizers. We assume the same $\mathbf{x}$ and $\mathbf{A}$ distributions as above. We plot MSE of the reconstruction against the rate measured in bits per component of $\mathbf{x}$. For each rate and for each quantizer, we vary the ratio m/n for the best possible performance. We see that, in comparison to regular quantizers, binned quantizers with MPDQ estimation achieve much lower distortions for the same rates. This indicates that binning can be an effective strategy to favorably shift rate–distortion performance of the estimation.

VII. Conclusions

We have presented message-passing de-quantization as an effective and efficient algorithm for estimation from quantized linear measurements. The proposed methodology is general, allowing essentially arbitrary priors and quantization functions. In particular, MPDQ is the first tractable and effective method for high-dimensional estimation problems involving non-regular scalar quantization. In addition, the algorithm is computationally extremely simple and, in the case of large ran-

dom transforms, admits a precise performance characterization using a state evolution analysis.

The problem formulation is Bayesian, with an i.i.d. prior over the components of the signal of interest $\mathbf{x}$; the prior may or may not induce sparsity of $\mathbf{x}$. Also, the number of measurements may be more or less than the dimension of $\mathbf{x}$, and the quantizers applied to the linear measurements may be regular or not. Experiments show significant performance improvement over traditional reconstruction schemes, some of which have higher computational complexity. Moreover, using extensions of GAMP such as hybrid approximate message passing [54], [55], one may also in the future be able to consider quantization of more general classes of signals described by general graphical models. MATLAB code for experiments with GAMP is available online [56].

Despite the improvements demonstrated here, we are not advocating quantized linear expansions as a compression technique—for the oversampled case or the undersampled sparse case; thus, comparisons to rate–distortion bounds would obscure the contribution. For regular quantizers and some fixed oversampling $\beta = m/n > 1$, the MSE decay with increasing rate is $\sim 2^{-2R/\beta}$, worse than the $\sim 2^{-2R}$ distortion–rate bound. For a discussion of achieving exponential decay of MSE with increasing oversampling, while the quantization step size is held constant, see [57]. For the undersampled sparse case, [33] discusses the difficulty of recovering the support from quantized samples and the consequent difficulty of obtaining near-optimal rate–distortion performance. Performance loss rooted in the use of a random transformation $\mathbf{A}$ is discussed in [58].

REFERENCES

[1] V. K. Goyal, "Theoretical foundations of transform coding," *IEEE Signal Process. Mag.*, vol. 18, no. 5, pp. 9–21, Sep. 2001.

[2] A. Zymnis, S. Boyd, and E. Candès, "Compressed sensing with quantized measurements," *IEEE Signal Process. Lett.*, vol. 17, no. 2, pp. 149–152, Feb. 2010.

[3] L. Jacques, D. K. Hammond, and J. M. Fadili, "Dequantizing compressed sensing: When oversampling and non-Gaussian constraints combine," *IEEE Trans. Inform. Theory*, vol. 57, no. 1, pp. 559–571, Jan. 2011.

[4] J. N. Laska, P. T. Boufounos, M. A. Davenport, and R. G. Baraniuk, "Democracy in action: Quantization, saturation, and compressive sensing," *Appl. Comput. Harm. Anal.*, vol. 31, no. 3, pp. 429–443, Nov. 2011.

[5] N. T. Thao and M. Vetterli, "Reduction of the MSE in R-times oversampled A/D conversion from $O(1/R)$ to $O(1/R^2)$," *IEEE Trans. Signal Process.*, vol. 42, no. 1, pp. 200–203, Jan. 1994.

[6] ——, "Deterministic analysis of oversampled A/D conversion and decoding improvement based on consistent estimates," *IEEE Trans. Signal Process.*, vol. 42, no. 3, pp. 519–531, Mar. 1994.

[7] V. K. Goyal, M. Vetterli, and N. T. Thao, "Quantized overcomplete expansions in $\mathbb{R}^N$: Analysis, synthesis, and algorithms," *IEEE Trans. Inform. Theory*, vol. 44, no. 1, pp. 16–31, Jan. 1998.

[8] S. Rangan and V. K. Goyal, "Recursive consistent estimation with bounded noise," *IEEE Trans. Inform. Theory*, vol. 47, no. 1, pp. 457–464, Jan. 2001.

[9] Z. Cvetković, "Resilience properties of redundant expansions under additive noise and quantization," *IEEE Trans. Inform. Theory*, vol. 49, no. 3, pp. 644–656, Mar. 2003.

[10] J. J. Benedetto, A. M. Powell, and Ö. Yilmaz, "Sigma–Delta ($\Sigma\Delta$) quantization and finite frames," *IEEE Trans. Inform. Theory*, vol. 52, no. 5, pp. 1990–2005, May 2006.

[11] B. G. Bodmann and V. I. Paulsen, "Frame paths and error bounds for sigma-delta quantization," *Appl. Comput. Harm. Anal.*, vol. 22, no. 2, pp. 176–197, Mar. 2007.

[12] B. G. Bodmann and S. P. Lipshitz, "Randomly dithered quantization and sigma–delta noise shaping for finite frames," *Appl. Comput. Harm. Anal.*, vol. 25, no. 3, pp. 367–380, Nov. 2008.

[13] A. M. Powell, "Mean squared error bounds for the Rangan–Goyal soft thresholding algorithm," *Appl. Comput. Harm. Anal.*, vol. 29, no. 3, pp. 251–271, Nov. 2010.

[14] S. Rangan, "Generalized approximate message passing for estimation with random linear mixing," arXiv:1010.5141v1 [cs.IT]., Oct. 2010.

[15] D. Guo and C.-C. Wang, "Random sparse linear systems observed via arbitrary channels: A decoupling principle," in *Proc. IEEE Int. Symp. Inform. Theory*, Nice, France, Jun. 2007, pp. 946–950.

[16] S. Rangan, "Estimation with random linear mixing, belief propagation and compressed sensing," in *Proc. Conf. on Inform. Sci. & Sys.*, Princeton, NJ, Mar. 2010, pp. 1–6.

[17] J. Boutros and G. Caire, "Iterative multiuser joint decoding: Unified framework and asymptotic analysis," *IEEE Trans. Inform. Theory*, vol. 48, no. 7, pp. 1772–1793, Jul. 2002.

[18] D. Guo and C.-C. Wang, "Asymptotic mean-square optimality of belief propagation for sparse linear systems," in *Proc. IEEE Inform. Theory Workshop*, Chengdu, China, Oct. 2006, pp. 194–198.

[19] M. Bayati and A. Montanari, "The dynamics of message passing on dense graphs, with applications to compressed sensing," *IEEE Trans. Inform. Theory*, vol. 57, no. 2, pp. 764–785, Feb. 2011.

[20] U. Kamilov, V. K. Goyal, and S. Rangan, "Optimal quantization for compressive sensing under message passing reconstruction," in *Proc. IEEE Int. Symp. Inform. Theory*, St. Petersburg, Russia, Jul.–Aug. 2011, pp. 390–394.

[21] T. Tanaka and M. Okada, "Approximate belief propagation, density evolution, and neurodynamics for CDMA multiuser detection," *IEEE Trans. Inform. Theory*, vol. 51, no. 2, pp. 700–706, Feb. 2005.

[22] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proc. Nat. Acad. Sci.*, vol. 106, no. 45, pp. 18 914–18 919, Nov. 2009.

[23] A. D. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Trans. Inform. Theory*, vol. IT-22, no. 1, pp. 1–10, Jan. 1976.

[24] V. K. Goyal, "Multiple description coding: Compression meets the network," *IEEE Signal Process. Mag.*, vol. 18, no. 5, pp. 74–93, Sep. 2001.

[25] V. Misra, V. K. Goyal, and L. R. Varshney, "Distributed scalar quantization for computing: High-resolution analysis and extensions," *IEEE Trans. Inform. Theory*, vol. 57, no. 8, pp. 5298–5325, Aug. 2011.

[26] R. M. Gray and D. L. Neuhoff, "Quantization," *IEEE Trans. Inform. Theory*, vol. 44, no. 6, pp. 2325–2383, Oct. 1998.

[27] E. J. Candès and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 21–30, Mar. 2008.

[28] Z. Liu, S. Cheng, A. D. Liveris, and Z. Xiong, "Slepian–Wolf coded nested lattice quantization for Wyner–Ziv coding: High-rate performance analysis and code design," *IEEE Trans. Inform. Theory*, vol. 52, no. 10, pp. 4358–4379, Oct. 2006.

[29] V. A. Vaishampayan, "Design of multiple description scalar quantizers," *IEEE Trans. Inform. Theory*, vol. 39, no. 3, pp. 821–834, May 1993.

[30] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inform. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[31] E. J. Candès and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?" *IEEE Trans. Inform. Theory*, vol. 52, no. 12, pp. 5406–5425, Dec. 2006.

[32] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[33] V. K. Goyal, A. K. Fletcher, and S. Rangan, "Compressive sampling and lossy compression," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 48–56, Mar. 2008.

[34] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3397–3415, Dec. 1993.

[35] S. Chen, S. A. Billings, and W. Luo, "Orthogonal least squares methods and their application to non-linear system identification," *Int. J. Control*, vol. 50, no. 5, pp. 1873–1896, Nov. 1989.

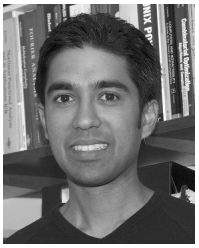
[36] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Conf. Rec. 27th Asilomar Conf. Sig., Sys., & Comput.*, vol. 1, Pacific Grove, CA, Nov. 1993, pp. 40–44.

[37] G. Davis, S. Mallat, and Z. Zhang, "Adaptive time-frequency decomposition," *Optical Eng.*, vol. 33, no. 7, pp. 2183–2191, Jul. 1994.

[38] D. Needell and J. A. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Appl. Comput. Harm. Anal.*, vol. 26, no. 3, pp. 301–321, May 2009.
[39] W. Dai and O. Milenkovic, "Subspace pursuit for compressive sensing signal reconstruction," *IEEE Trans. Inform. Theory*, vol. 55, no. 5, pp. 2230–2249, May 2009.
[40] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comp.*, vol. 20, no. 1, pp. 33–61, 1999.
[41] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Royal Stat. Soc., Ser. B*, vol. 58, no. 1, pp. 267–288, 1996.
[42] E. J. Candès and T. Tao, "The Dantzig selector: Statistical estimation when p is much larger than n," *Ann. Stat.*, vol. 35, no. 6, pp. 2313–2351, Dec. 2007.
[43] E. J. Candès and J. Romberg, "Encoding the ℓ_p ball from limited measurements," in *Proc. IEEE Data Compression Conf.*, Snowbird, UT, Mar. 2006, pp. 33–42.
[44] J. Z. Sun and V. K. Goyal, "Optimal quantization of random measurements in compressed sensing," in *Proc. IEEE Int. Symp. Inform. Theory*, Seoul, Korea, Jun.–Jul. 2009, pp. 6–10.
[45] V. K. Goyal, J. Kovačević, and J. A. Kelner, "Quantized frame expansions with erasures," *Appl. Comput. Harm. Anal.*, vol. 10, no. 3, pp. 203–233, May 2001.
[46] H. Viswanathan and R. Zamir, "On the whiteness of high-resolution quantization errors," *IEEE Trans. Inform. Theory*, vol. 47, no. 5, pp. 2029–2038, Jul. 2001.
[47] 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*, vol. 42, no. 2, pp. 469–479, Mar. 1996.
[48] R. H. Walden, "Analog-to-digital converter survey and analysis," *IEEE J. Sel. Areas Comm.*, vol. 17, no. 4, pp. 539–550, Apr. 1999.
[49] R. J. Pai, "Nonadaptive lossy encoding of sparse signals," Master's thesis, Massachusetts Inst. of Tech., Cambridge, MA, Aug. 2006.
[50] P. T. Boufounos, "Universal rate-efficient scalar quantization," *IEEE Trans. Inform. Theory*, vol. 58, no. 3, pp. 1861–1872, Mar. 2012.
[51] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufmann Publ., 1988.
[52] C. M. Bishop, *Pattern Recognition and Machine Learning*, ser. Information Science and Statistics. New York, NY: Springer, 2006.
[53] E. J. Candès, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Commun. Pure Appl. Math.*, vol. 59, no. 8, pp. 1207–1223, Aug. 2006.
[54] S. Rangan, A. K. Fletcher, V. K. Goyal, and P. Schniter, "Hybrid approximate message passing with applications to structured sparsity," arXiv:1111.2581 [cs.IT], Nov. 2011.
[55] ——, "Hybrid generalized approximate message passing with applications to structured sparsity," in *Proc. IEEE Int. Symp. Inform. Theory*, Cambridge, MA, Jul. 2012, pp. 1241–1245.
[56] S. Rangan *et al.*, "Generalized approximate message passing," SourceForge.net project gampmatlab, available on-line at `http://gampmatlab.sourceforge.net/`.
[57] Z. Cvetković and M. Vetterli, "Error-rate characteristics of oversampled analog-to-digital conversion," *IEEE Trans. Inform. Theory*, vol. 44, no. 5, pp. 1961–1964, Sep. 1998.
[58] A. K. Fletcher, S. Rangan, and V. K. Goyal, "On the rate-distortion performance of compressed sensing," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process.*, vol. III, Honolulu, HI, Apr. 2007, pp. 885–888.

Ulugbek S. Kamilov (S'11) received his M.Sc. degree in Communications Systems from the École Polytechnique Fédérale de Lausanne (EPFL), Switzerland, in 2011.

In 2007–08, he was an exchange student in Electrical and Computer Engineering at Carnegie Mellon University. In 2008, he worked as a research intern at the Telecommunications Research Center in Vienna, Austria. In 2009, he worked as a software engineering intern at Microsoft. In 2010–11, he was a visiting student at the Massachusetts Institute of Technology. In 2011, he joined the Biomedical Imaging Group at EPFL where he is currently working toward his Ph.D. His research interests include message-passing algorithms and the application of signal processing techniques to various biomedical problems.



Vivek K Goyal (S'92–M'98–SM'03) received the B.S. degree in mathematics and the B.S.E. degree in electrical engineering from the University of Iowa, where he received the John Briggs Memorial Award for the top undergraduate across all colleges. He received the M.S. and Ph.D. degrees in electrical engineering from the University of California, Berkeley, where he received the Eliahu Jury Award for outstanding achievement in systems, communications, control, or signal processing.

He was a Member of Technical Staff in the Mathematics of Communications Research Department of Bell Laboratories, Lucent Technologies, 1998–2001; and a Senior Research Engineer for Digital Fountain, Inc., 2001–2003. He has been with the Massachusetts Institute of Technology since 2004. He is coauthor of the forthcoming textbooks *Foundations of Signal Processing* and *Fourier and Wavelet Signal Processing* (both Cambridge University Press). His research interests include computational imaging, sampling, quantization, and source coding theory.

Dr. Goyal is a member of Phi Beta Kappa, Tau Beta Pi, Sigma Xi, Eta Kappa Nu and SIAM. He was awarded the 2002 IEEE Signal Processing Society Magazine Award and an NSF CAREER Award. As a research supervisor, he is co-author of papers that won student best paper awards at IEEE Data Compression Conference in 2006 and 2011 and IEEE Sensor Array and Multichannel Signal Processing Workshop in 2012. He served on the IEEE Signal Processing Society's Image and Multiple Dimensional Signal Processing Technical Committee 2003–2009. He is a Technical Program Committee Co-chair of IEEE ICIP 2016 and a permanent Conference Co-chair of the SPIE Wavelets and Sparsity conference series.

Sundeep Rangan (M'02) received the B.A.Sc. degree from the University of Waterloo, Canada, and the M.S. and Ph.D. degrees from the University of California, Berkeley, all in electrical engineering. He held postdoctoral appointments at the University of Michigan, Ann Arbor, and Bell Labs. In 2000, he co-founded (with four others) Flarion Technologies, a spin-off of Bell Labs, that developed Flash OFDM, one of the first cellular OFDM data systems. In 2006, Flarion was acquired by Qualcomm Technologies, where Dr. Rangan was a Director of Engineering involved in OFDM infrastructure products. He joined the Department of Electrical and Computer Engineering at the Polytechnic Institute of New York University in 2010, where he is currently an Associate Professor. His research interests are in wireless communications, signal processing, information theory and control theory.