HIERARCHICAL DISTRIBUTED SCALAR QUANTIZATION

Petros T. Boufounos

Mitsubishi Electric Research Laboratories, Cambridge, MA 023139
petrosb@merl.com

ABSTRACT

Scalar quantization is the most practical and straightforward approach to signal quantization. However, it has been shown that scalar quantization of oversampled or Compressively Sensed signals can be inefficient in terms of the rate-distortion trade-off, especially as the oversampling rate or the sparsity of the signal increases. Recent theoretical work has provided some insights on improving this trade-off, using non-monotonic quantization functions. This paper builds upon this work to provide a practical hierarchical quantization scheme that enables efficient reconstruction through a hierarchy of convex optimization problems. Our approach generalizes the bit hierarchy—most to least significant bit—of classical multi-bit scalar quantization. We demonstrate experimental results both for dense and sparse signals that demonstrate significant gains and confirm our theoretical analysis.

Keywords— scalar quantization, randomization, oversampling, robustness

1. INTRODUCTION

Digital signal acquisition comprises of two discretization steps: sampling (or measurement) and quantization. Sampling, acquires linear measurements of the signal, such as the signal's instantaneous value or the signal's inner product with a measurement vector. Quantization, maps the continuous-valued measurements of the signal to a set of discrete values, referred to as quantization points. Overall, this discretization process is lossy, i.e., does not preserve all the information in the signal.

While sampling can be designed to preserve all information in the signal, quantization always results to distortion. Several sampling results demonstrate that as long as sufficient samples are obtained, given a class of signals, exact recovery is possible. For example, the Nyquist theorem provides the sampling rate necessary for bandlimited signals, while Compressive Sensing theory provides sampling-rate conditions for exact recovery of sparse signals [1]. However, once the samples are quantized, exact signal recovery is impossible. Thus, the main goal of quantizer design is to reduce the distortion on the signal as much as possible given the available bit-rate.

The most popular quantization method is scalar quantization in which each measurement is quantized independently of the others. This approach is simple and has good performance. It is especially appealing for distributed sensor applications in which each sensor quantizes its own measurement before communicating it to other sensors or to a central node. Unfortunately, present approaches to scalar quantization are suboptimal if the signal is oversampled [2–5]. Specifically, the trade-off between the number of bits used to represent an oversampled signal and the representation error worsens as oversampling increases. In terms of the rate vs. distortion, it is more efficient to allocate more bits per coefficient in a critically sampled representation as opposed to fewer bits per coefficient in an oversampled representation.

Recent theoretical work provides the basis to overcome this trade-off using a non-monotonic quantizer [6]. This work demonstrates that non-monotonic quantizers achieve exponential error decay in the oversampling rate using consistent reconstruction. However, reconstruction from such a quantization method is not straightforward. The resulting optimization problem is non-convex and seems to have combinatorial complexity.

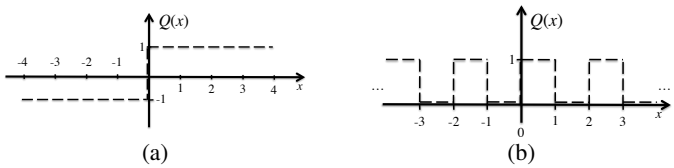

Fig. 1. Examples of Quantization Functions. (a) Typical binary (1-bit) quantization function. (b) Non-monotonic binary quantization function, used in this work.

This paper exploits the theoretical results in [6] to provide a practical hierarchical scheme for distributed oversampled scalar quantization. At every level of the hierarchy we use the non-monotonic quantizer in [6] in a way that ensures that reconstruction at that level becomes a convex problem. Reconstruction at that level reduces the error and the ambiguity in the reconstructed signal, which, in turn, allows the problem in the next level to become convex.

Our approach can be thought of as a generalization of the hierarchy of bits in multi-bit scalar quantization. The hierarchy levels are equivalent to the bit-levels in a multi-bit scalar quantizer, starting from the most significant bit (MSB), all the way to the least significant bit (LSB). Similar to multi-bit scalar quantization, the LSB provides very little information without the context of the more significant bits of the coefficient.

The next section, provides an overview of scalar quantization. It serves as a quick reference and establishes the notation. In Sec. 3 we present our hierarchical quantization approach and our reconstruction algorithm, together with some discussion and connections with classical scalar quantization. Finally, Sec. 4 presents experimental results that verify our approach.

2. BACKGROUND

2.1. Scalar Quantization

A scalar quantizer operates directly on individual scalar signal measurements without taking into account any information on the value or the quantization level of nearby measurements. Specifically, the generation of the m^{th} quantization bit from the quantized signal $\mathbf{x} \in \mathbb{R}^K$ is performed using

$$y_m = \langle \mathbf{x}, \phi_m \rangle \tag{1}$$

$$q_m = Q\left(\frac{y_m}{\Delta_m} + w_m\right) = Q\left(p_m\right), \tag{2}$$

where ϕ_m is the measurement vector used to produce a scalar measurement y_m, which is subsequently scaled by a precision parameter Δ_m, dithered by the additive dither w_m and quantized by the quantization function $Q(\cdot)$. The intermediate variable p_m denotes the scaled, dithered measurement before the quantization. The measurements are index by $m = 1, \ldots, M$, where M is the total number of quantized coefficients acquired. The precision parameter is usually not explicit in the literature but is incorporated as a design parameter of the quantization function $Q(\cdot)$. We make it explicit here in anticipation of our development.

A more compact, vectorized form of (1) and (2) will often be more convenient in our discussion

$$\mathbf{y} = \Phi\mathbf{x} \tag{3}$$

$$\mathbf{q} = \mathbf{Q}\left(\Delta^{-1}\mathbf{y} + \mathbf{w}\right) = \mathbf{Q}\left(\mathbf{p}\right), \tag{4}$$

where $\mathbf{y}$, $\mathbf{w}$ $\mathbf{p}$ and $\mathbf{q}$ are vectors containing the measurements, the dither coefficients, the intermediate and the quantized values, respectively, Δ is a diagonal matrix with the precision parameters Δ_m in its diagonal, $\mathbf{Q}(\cdot)$ denotes the scalar quantization function, applied element-by-element on its input, and Φ is an $M \times K$ measurement matrix that contains the measurement vectors ϕ_m in its rows.

For a binary 1-bit quantizer, the focus of this paper, the quantization function $Q(\cdot)$ is typically the one shown in Fig. 1(a). The scaling performed by the precision parameter Δ_m controls the trade-off between quantization accuracy and the number of quantization bits. The non-monotonic quantizer discussed in [6] is shown in Fig. 1(b). Here Δ_m controls the precision width of each quantization interval. Smaller Δ_m increases the precision but requires more measurements to guarantee reconstruction and makes the reconstruction significantly more complex.

2.2. Reconstruction from Quantized Measurements

A reconstruction algorithm, denoted $R(\cdot)$, uses the quantized representation generated by the signal to produce a signal estimate $\widehat{\mathbf{x}} = R(\mathbf{q})$. The performance of the quantizer and the reconstruction algorithm is measured in terms of the reconstruction distortion, typically measured using the ℓ_2 distance: $d = \|\mathbf{x} - \widehat{\mathbf{x}}\|_2$. The goal of the quantizer and the reconstruction algorithm is to minimize the average or the worst case distortion given a probabilistic or a deterministic model of the acquired signals.

While the simplest reconstruction approach is to substitute the quantized value in standard reconstruction approaches for unquantized measurements, it is in general suboptimal [3–5]. A better approach is to use consistent reconstruction, a method that enforces that the reconstructed signal quantizes to the same value as the acquired one, i.e., satisfies

$$\mathbf{q} = \mathbf{Q}\left(\Delta^{-1}\Phi\widehat{\mathbf{x}} + \mathbf{w}\right) \tag{5}$$

Consistent reconstruction was originally proposed for oversampled frames in [4], where it was shown to outperform linear reconstruction. Subsequently consistent reconstruction, or approximations of it, have been shown in various scenarios to improve Compressive Sensing reconstruction from quantized measurements [6–14].

2.3. Reconstruction Rate and Distortion Performance

The performance of scalar quantizers is typically measured by their rate vs. distortion trade-off, i.e., how increasing the number of bits used by the quantizer affects the distortion on the measurement signal due to quantization. In this paper we focus on the worst-case distortion, i.e.,

$$d = \max_{\mathbf{x}} \left\|\mathbf{x} - R\left(\mathbf{Q}\left(\Delta^{-1}\Phi\mathbf{x} + \mathbf{w}\right)\right)\right\|_2, \tag{6}$$

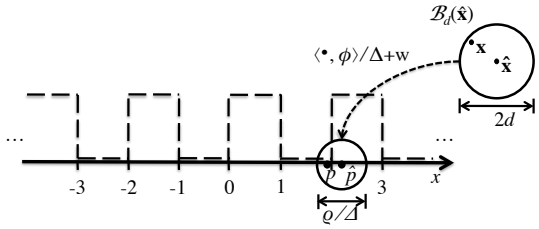

Fig. 2. Quantization convexity. To ensure the problem is convex, the projection of the ball of ambiguity around the estimated signal through the measurement process should incorporate at most one transition of the quantization function.

where $\widehat{\mathbf{x}} = R\left(\mathbf{Q}\left(\Delta^{-1}\Phi\mathbf{x} + \mathbf{w}\right)\right)$ is the signal reconstructed from the quantization of $\mathbf{x}$.

Under this sampling model, there are two ways to increase the bit-rate and reduce the quantization distortion: increase the number of bits used per quantized coefficient or increase the number of measurements at a fixed number of bits per coefficient. Using the former, exponential reduction in the reconstruction error is possible as a function of the bit-rate

$$d = O(c^r), c \leq 1, \tag{7}$$

where $r = MB$ is the total rate used to represent the signal at M measurements and B bits per measurement. Using the latter with the scalar quantizer in Fig 1(a), the distortion cannot reduce at a rate faster than linear with respect to the oversampling rate. At a fixed number of bits per measurement, this is proportional to the bit-rate

$$d = \Omega(1/M), \tag{8}$$

much slower than the rate in (7). This rate is achieved by consistent reconstruction but not by linear reconstruction [3–5].

Exponential error decay in the oversampling rate can be achieved using the non-monotonic quantizer in Fig 1(b). Specifically, [6] demonstrates that with very high probability the worst-case reconstruction error satisfies

$$d \leq C\left(\frac{3}{4}\right)^{M/2K}, \tag{9}$$

where C depends on the desired confidence and the size of the set of signals of interest. The decay is achieved by assuming all the measurements use the same quantization precision parameter Δ, which is set as a function of the desired accuracy. Unfortunately, as Δ becomes smaller compared to the size of the set of signals of interest the reconstruction problem becomes significantly harder.

3. HIERARCHICAL SCALAR QUANTIZATION

3.1. Quantization Hierarchy

Our approach in this paper is to use the scalar quantizer in Fig. 1(b) hierarchically, such that the reconstruction of each level requires the solution of a convex problem and provides guarantees for reconstruction error at that level.

At each level, l, we measure and quantize the signal according to (3) and (4) using the quantizer in Fig. 1(b).

$$\mathbf{q}_l = \mathbf{Q}\left(\Delta_l^{-1}\Phi_l\mathbf{x} + \mathbf{w}_l\right) = \mathbf{Q}\left(\mathbf{p}_l\right) \tag{10}$$

Similar to [6], Φ_l contains i.i.d., normally distributed entries with zero mean and variance $1/\sqrt{K}$, and the dither $\mathbf{w}_l$ contains uniformly distributed entries in $[0, 1]$. However, at each level we ensure the problem

Fig. 3. Decomposition of the multi-bit quantization hierarchy. The most significant bit (MSB, left) provides coarse information. The least significant bit (LSB, center) provides refinement information, but in a non-convex, ambiguous, set. The LSB becomes a convex set when restricted by the convex set defined by the MSB, and the information can be easily decoded (right).

is convex by selecting the precision parameter Δ_l according to the reconstruction guarantees of the previous level.

Assume that each level l provides a worst-case error guarantee of d_l. Then the signal of interest $\mathbf{x}$ is within a ball of radius d_l from the signal $\widehat{\mathbf{x}}_l$ reconstructed at that level: $\mathbf{x} \in \mathcal{B}_{d_l}(\widehat{\mathbf{x}}_l)$,where $\mathcal{B}_d(\mathbf{x}) \equiv \{\mathbf{x}' \,|\|\mathbf{x}-\mathbf{x}'\|_2 \leq d\}$ denotes the ball of radius d around $\mathbf{x}$. We denote the diameter of this ball, as projected through the measurement vectors ϕ at level $l+1$, using ρ_{l+1}:

$$\rho_{l+1} \equiv \max_{\mathbf{x}_1,\mathbf{x}_2 \in \mathcal{B}_{d_l}(\widehat{\mathbf{x}}_l)} |\langle \phi, \mathbf{x}_1 - \mathbf{x}_2 \rangle|. \tag{11}$$

It is straightforward to show that the reconstruction problem is convex at level l if the precision parameter at that level, Δ_l, is set such that $\rho_l < \Delta_l$. This is because at most one transition of the quantization function will occur within this projection, as shown in Fig. 2. Our goal in this paper is to obtain a sufficient number of measurements at each level l such that we can guarantee convexity when setting

$$\Delta_{l+1} = \alpha \Delta_l, \tag{12}$$

for some scaling factor $\alpha < 1$.

Since the measurement vectors are i.i.d., normally distributed, random vectors with variance $1/\sqrt{K}$, their norm is concentrated around 1. In particular,

$$P(\|\phi\|_2 \geq 2) = \gamma\left(\frac{K}{2}, K\right), \tag{13}$$

where $\gamma(\cdot,\cdot)$ is the regularized upper incomplete gamma function. This probability decreases extremely fast and becomes negligible for any reasonable value of K. Therefore, it is safe to assume that the projection will at most double the radius of $\mathcal{B}_{d_l}(\widehat{\mathbf{x}}_l)$, i.e., that

$$\rho_{l+1} \leq 4 d_l. \tag{14}$$

Of course, for large values of K, we can tighten this bound significantly. To ensure convexity, we set Δ_l such that $d_l \leq \alpha \Delta_l / 4$; we choose equality, i.e.,

$$\Delta_l = 4 d_l / \alpha. \tag{15}$$

To guarantee the convexity is maintained though all levels of the hierarchy, we desire to scale d_l at every level linearly, similar to Δ_l,

$$d_{l+1} \leq \alpha d_l. \tag{16}$$

We determine the number of measurements necessary to achieve this scaling using the results in [6]. Specifically, we can show that after M_l measurements, the probability that (16) holds, is greater than

$$P_{(16)} \geq 1 - (3c)^K \left(\frac{1}{2} + \frac{1}{2} e^{-\frac{\pi \alpha^2}{4\sqrt{2}K}} + \frac{\alpha}{c} + \gamma\left(\frac{K}{2}, K\right)\right)^{M_l}, \tag{17}$$

for any constant c. For the right choice of c this probability approaches 1 exponentially fast in the number of measurements at each level, M_l.

This hierarchical process can be though of as the distributed generalization to multi-bit scalar quantization. Consider the 2-bit representation for the binary coefficient 10, as dissected in Fig. 3. The coarse-level MSB splits the region of interest in two convex regions (positive and negative), as shown on the left of the figure. The fine-level LSB splits the region of interest in two non-convex regions, as shown on the center of the figure. Using the information that the MSB is equal to 1, the non-convex set corresponding to the LSB equal to 0 can be restricted to a convex set, and the coefficient can be easily decoded. In terms of the hierarchical quantization scheme we describe above, this would be equivalent to each level in the hierarchy using the same measurement vectors and dither, and setting $\alpha = 1/2$.

3.2. Reconstruction Algorithm

To reconstruct the signal we incorporate each level hierarchically and use a consistent reconstruction algorithm, such as the one in [4], after the incorporation of each level. The reconstruction algorithm uses the quantization interval consistent with the quantization point as a constraint. Thus, it enforces the constraint

$$q_{\min} \leq \langle \phi, \widehat{\mathbf{x}} \rangle / \Delta + w \leq q_{\max} \tag{18}$$

on the reconstructed signal $\widehat{\mathbf{x}}$, for all the incorporated measurement vectors ϕ, where $[q_{\min}, q_{\max}]$ is the consistent reconstruction interval.

Since the quantization function in Fig. 1(b) is non-monotonic, the set of consistent reconstruction is not a continuous interval but a non-convex set. The goal of incorporating a hierarchy level is to select the a convex subset of the consistent subset, according to the reconstruction estimate from the already incorporated levels. In terms of the scalar quantization analogy of Fig. 3, incorporating the LSB uses the information from the already incorporated MSB to select the right interval from the two possible, consistent with the LSB equal to 0, as shown in the middle figure.

To incorporate a hierarchy level l we use the estimate $\widehat{x}_{l-1}$ from the previous iteration, measure it using the measurement matrix Φ_l and quantize it according to the acquisition parameters

$$\widehat{\mathbf{y}}_l = \Phi_l \widehat{\mathbf{x}}, \tag{19}$$

$$\widehat{\mathbf{q}}_l = Q(\Delta^{-1} \widehat{\mathbf{y}}_l + \mathbf{w}_l) = Q(\widehat{\mathbf{p}}_l). \tag{20}$$

We then compare the quantized estimate vector $\widehat{\mathbf{q}}_l$ to the acquired quantized vector $\mathbf{q}_l$. For the quantization values in $\widehat{\mathbf{q}}_l$ that are consistent with the quantized values in $\mathbf{q}_l$ we use the quantization interval in which the corresponding value in $\widehat{\mathbf{p}}_l$ lies. For the quantization values in $\widehat{\mathbf{q}}_l$ that are inconsistent with the quantized values in $\mathbf{q}_l$ we use the quantization interval which is closest to $\widehat{\mathbf{p}}_l$ and consistent with $\mathbf{q}_l$. For example, if the estimate $\widehat{\mathbf{x}}$ projects to $\widehat{p}$ as shown in Fig. 2, it will be inconsistent with the actual measurement p. Since the interval $[1,2]$ is the closest interval to $\widehat{p}$ consistent with $q = 0$, we select this interval for the consistent reconstruction algorithm. The conditions discussed in the previous section guarantee this is the correct choice. The algorithm is initialized with $\widehat{\mathbf{x}} = 0$ before the first (coarsest) level of the hierarchy is incorporated.

Once we determine the appropriate quantization intervals at hierarchy level l we perform consistent reconstruction using all the information (i.e, measurement vectors, scaling parameters, dither, and resolved quantization intervals) from levels 1 to l to produce the next signal estimate $\widehat{\mathbf{x}}_l$, and use it to incorporate the next level in the hierarchy of measurements.

Of course, it is possible to incorporate this approach to a number of optimization-based reconstruction algorithms, such as the basis pursuit commonly used in Compressive Sensing applications. In this case, the optimization algorithm is executed for every level of the hierarchy, subject to the consistent reconstruction constraints, to produce the signal estimate.

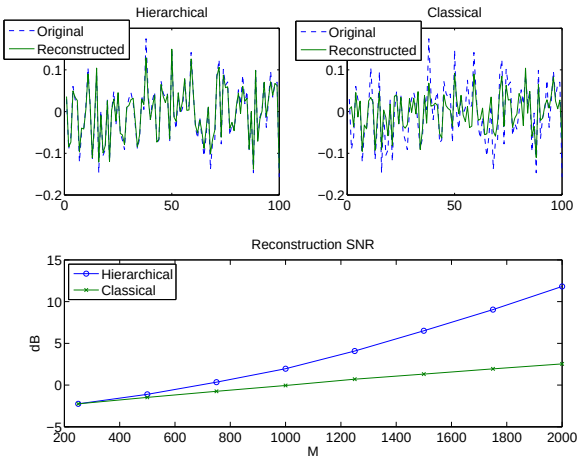

Fig. 4. Reconstruction results for non-sparse signals. Top: sample reconstruction using hierarchical (left) and classical (right) quantization. Bottom: Reconstruction SNR for classical and hierarchical approaches. The marks on the line denote the level points of the hierarchical approach.

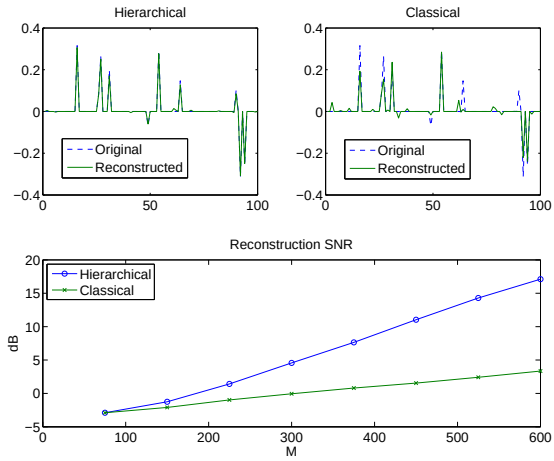

Fig. 5. Reconstruction results for sparse signals. Top: sample reconstruction using hierarchical (left) and classical (right) quantization. Bottom: Reconstruction SNR for classical and hierarchical approaches. The marks on the line denote the level points of the hierarchical approach.

4. EXPERIMENTAL RESULTS

This section presents experimental results that confirm our analysis. Specifically, we consider both dense and sparse signals, and we compare our results to consistent reconstruction from classical 1-bit scalar quantization with the same number of measurements.

Figure 4 shows the experimental results for randomly generated dense signals in a 100-dimensional space. We used 8 hierarchy levels, with 250 measurements (i.e., bits) per level. At each level we compared the performance classical 1-bit scalar quantization and consistent reconstruction using the same number of measurements. The scaling factor α in our simulations was 1/1.4. The trials were averaged over 100 iterations. The top of the figure plots a sample signal reconstruction using hierarchical (left) and classical (right) 1-bit quantization, compared to the original. The bottom plot the average reconstruction SNR as a function of the number of measurements for the two approaches. The markers in the curve, every 250 measurements, indicate each level of the hierarchical scalar quantization. It is evident from the figure that our hierarchical approach significantly outperforms classical quantization.

Figure 5 shows the same results for 10-sparse signals in a 100-dimensional space. We used 8 hierarchy levels, with 75 measurements per level. The scaling factor α was set to 1/1.55. As above, hierarchical reconstruction significantly outperforms the classical approach.

The results were consistent for other signal dimensions and sparsity levels. As expected from the theory, our experimental results confirm the improved performance of hierarchical quantization. Of course, significant further work is necessary on the optimal choice of parameters, such as the number of measurements per level, the number of levels, and the choice of α.

5. REFERENCES

[1] E. Candès, J. Romberg, and T. Tao, "Stable Signal Recovery from Incomplete and Inaccurate Measurements," *Comm. Pure and Applied Math.*, vol. 59, no. 8, pp. 1207–1223, Aug. 2006.

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

[3] 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. Info. Theory*, vol. 42, no. 2, pp. 469–479, Mar. 1996.

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

[5] P. Boufounos and R. Baraniuk, "Quantization of sparse representations," in *Proc. Data Compression Conference (DCC)*, Mar. 2007, pp. 378 –378.

[6] P. T. Boufounos, "Universal rate-efficient scalar quantization," 2010, preprint, http://arxiv.org/abs/1009.3145.

[7] P. Boufounos and R. G. Baraniuk, "One-Bit Compressive Sensing," in *Proc. 42nd annual Conference on Information Sciences and Systems (CISS)*, Princeton, NJ, Mar 19-21 2008.

[8] J. Laska, P. Boufounos, M. Davenport, and R. Baraniuk, "Democracy in action: Quantization, saturation, and compressive sensing," *Preprint*, 2009.

[9] J. Laska, P. Boufounos, and R. Baraniuk, "Finite-range scalar quantization for compressive sensing," in *Proc. Sampling Theory and Applications (SampTA)*, Marseille, France, May 2009.

[10] L. Jacques, D. Hammond, and M. Fadili, "Dequantizing compressed sensing: When oversampling and non-gaussian contraints combine," *Preprint*, 2009.

[11] A. Zymnis, S. Boyd, and E. Candès, "Compressed sensing with quantized measurements," *Preprint*, 2009.

[12] W. Dai, H. Pham, and O. Milenkovic, "Distortion-rate functions for quantized compressive sensing," *Preprint*, 2009.

[13] P. Boufounos, "Greedy sparse signal reconstruction from sign measurements," in *Conference Record of the Forty-Third Asilomar Conference on Signals, Systems and Computers, 2009*, Nov. 2009, pp. 1305 –1309.

[14] ——, "Reconstruction of sparse signals from distorted randomized measurements," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Dallas, TX, March 14-19 2010.