Gradient-Descent Algorithm Performance With Reduced Set of Quantized Measurements

Isidora Stanković[1,2], Miloš Brajović[1], Miloš Daković[1], Cornel Ioana[2]
[1] Faculty of Electrical Engineering, University of Montenegro, Podgorica, Montenegro
[2] GIPSA Lab, University of Grenoble Alpes, Grenoble, France
Email: {isidoras, milosb, milos}@ucg.ac.me, cornel.ioana@grenoble-inp.fr

***Abstract*—The quantization (digitalization) of measurements greatly affects the reconstruction performance, especially in algorithms based on the reconstruction in the measurement domain. However, it provides a significant advantage in the hardware implementation sense. In this paper, we analyze the performance of the gradient-based algorithm in the signal reconstruction based on a reduced set of digital measurements. This algorithm is considered as a powerful tool for the reconstruction of various types of signals. The paper investigates the accuracy of the algorithm using B-bit quantized measurements. The reconstruction performance is analyzed through numerical examples.**

***Keywords*—compressive sensing, quantization, reconstruction, sparse signal processing, gradient algorithm**

I. Introduction

The reconstruction of sparse signals has been an attractive research area of signal processing in the last decade. By definition, sparse signals are signals with small number of non-zero components in a transform domain, compared to the total number of components. It can be expected that such signals can be reconstructed with less randomly positioned samples compared to the number of samples required by the traditional way of sampling a signal. The mathematical foundation of such reconstruction procedures is developed within the compressive sensing (CS) framework [1]–[5].

Since the introduction of these concepts, many techniques have been proposed for taking measurements and reconstructing sparse signals [6]–[8]. These approaches have been applied in various everyday applications where signal processing is used. In the cases when signal samples are heavily corrupted, the CS algorithms also demonstrated good performance in the denoising of such signals. Improved results over classical filtering are obtained by excluding highly corrupted samples from the calculation and marking them as unavailable (missing), which then reduces to the concept of signal reconstruction according to the CS theory [4].

Ideally, the measurements used for the reconstruction should be taken accurately, assuming a very large number of bits in their digital form. However, this could be extremely demanding and expensive for hardware implementation. That is the reason why, in practice, the measurements are quantized to a certain level, using a limited number of bits [9]–[14]. Quantized measurements provide robustness, memory efficiency and simplicity in the corresponding sensor design. The most extreme case of quantization is the one-bit quantization, which is very suitable for hardware systems. The disadvantage of this system is that the signal can be recovered up to a constant scalar factor only. This system also requires a larger number of measurements for a successful reconstruction [9]. In this paper, we will focus on the general B-bit quantization of available measurements and the signal reconstruction based on the quantized data.

We consider a gradient-descent based algorithm, from the group of convex relaxation algorithms [7], [8]. This algorithm is characterized with a very interesting idea to perform the reconstruction in the spatial/measurements domain. In this algorithm, the missing samples are considered as variables, and the available samples remain unchanged [7]. The performance of this algorithm in the reconstruction of the sparse signal from a reduced set of quantized samples will be validated using the mean square error (MSE).

The paper is organized as follows. In Section II, the background of CS and quantization are reviewed. In Section III, the considered algorithm is described and in Section IV the performance analysis is presented. The conclusions are given in Section V.

II. Compressive Sensing and Quantization

Consider a discrete-time signal $x(n)$ defined as

$$x(n) = \sum_{k=1}^{N} X(k)\psi_k(n),\ 1 \le n \le N$$

and its transformation domain coefficients $X(k)$, given by

$$X(k) = \sum_{n=1}^{N} x(n)\varphi_n(k),\ 1 \le k \le N.$$

In the vector/matrix notation the signals are

$$\mathbf{x} = \mathbf{\Psi X} \quad \text{and} \quad \mathbf{X} = \mathbf{\Phi x},$$

with $\mathbf{x}$ and $\mathbf{X}$ denoting the signal values and transform coefficient vectors, whereas $\mathbf{\Psi}$ and $\mathbf{\Phi}$ are used for inverse and direct transformation matrices, respectively. We assume that the signal is sparse in the Discrete Fourier Transform (DFT) domain. The elements of the transformation domain matrices are given by

$$\phi_k(n) = e^{-j2\pi(k-1)(n-1)/N}, \quad \text{and } \psi_n(k) \quad = \phi_k^*(n)/N. \quad (1)$$

A signal is K-sparse if the number of non-zero coefficients in the transformation domain is much smaller than the total number of coefficients, i.e. $K \ll N$. Following the theory of CS, a K-sparse signal can be reconstructed with less samples than required by the conventional sampling techniques. A random subset of M samples from $x(n)$ at positions $n_i \in \mathbf{M} = \{n_1, n_2, \ldots, n_M\} \subset \mathbf{N} = \{1, 2, \ldots, N\}$ is considered as a set of available signal measurements and denoted by vector $\mathbf{y}$

$$\mathbf{y} = [x(n_1), x(n_2), \ldots, x(n_M)]^T. \quad (2)$$

Each measurement is formed as a linear combination of transform coefficients $X(k)$, that is

$$y(i) = x(n_i) = \sum_{k=1}^{N} X(k)\psi_k(n_i). \quad (3)$$

This system of M measurements can be written as

$$\mathbf{y} = \mathbf{AX}, \quad (4)$$

where $M \times N$ matrix $\mathbf{A}$ is obtained from the transformation matrix $\mathbf{\Psi}$, where the rows at the positions corresponding to the positions of available samples are preserved, while the other $N_Q = N - M$ rows are eliminated from the complete matrix.

For a successful reconstruction, the goal is to minimize a sparsity measure of the transform coefficients in $\mathbf{X}$, corresponding to a candidate problem solution, which satisfies the system of the available measurements equations, $\mathbf{y} = \mathbf{AX}$. The reconstruction can be formulated as the solution of an l_0-norm minimization problem

$$\min \|\mathbf{X}\|_0 \quad \text{subject to} \quad \mathbf{y} = \mathbf{AX}. \quad (5)$$

However, since task (5) is an NP-hard combinatorial problem, the common alternative is to use the l_1-norm to obtain a relaxed formulation of this optimization problem as

$$\min \|\mathbf{X}\|_1 \quad \text{subject to} \quad \mathbf{y} = \mathbf{AX}. \quad (6)$$

It is important to note that the solutions of (5) and (6) are the same if the signal $x(n)$ and its transform $X(k)$ satisfy restricted isometry property with a specified constant [3], [6].

A. Quantization Effect

In the traditional CS framework, the number of bits of the measurements was not normally considered. The aim was to find an exact reconstruction solution of a sparse signal. In recent years, the quantization of the available samples is being also considered, as a very important part of the hardware implementation [12]–[14]. The disadvantage is that the bit-limiting measurements could significantly affect the reconstruction performance of the standard CS approaches. It is assumed that the measurements are stored into B-bit registers, that is

$$\mathbf{y}_B = \text{digital}_B\{\mathbf{AX}\}. \quad (7)$$

Then, the goal is to reconstruct the values of missing samples as realistic as possible. In the most extreme case, one-bit quantization, more measurements are needed $M \gg N$. Note that even for $M \gg N$ the storage requirement could be significantly reduced for these measurements since the total number of bits is reduced.

Note that, for a complex-valued signal $x(n)$, both real and imaginary parts of $\mathbf{y}_B$ are B-bit quantized

$$\mathbf{y}_B = \text{digital}_B\{\Re\{\mathbf{AX}\}\} + j\text{digital}_B\{\Im\{\mathbf{AX}\}\}. \quad (8)$$

When a signal is quantized to B bits, the difference in amplitude which produces the quantization is known as the quantization error defined by $e(n)$. Then, the signal can be assumed as a noisy signal with a uniform noise $e(n)$

$$x_q(n) = x(n) + e(n). \quad (9)$$

The quantization error (noise) is

$$-\Delta/2 < e(n) < \Delta/2, \text{ where } \Delta = 2^{-B}. \quad (10)$$

III. Reconstruction Algorithm

The reconstruction algorithm is an adaptation of the approach presented in [7], modified for digital measurements. The algorithm uses the available measurements as the reference point. While the measurements are not affected, the unavailable samples are iteratively recovered.

Consider a signal $x(n)$, K-sparse in the DFT domain. Unlike the most CS algorithms, where we use the measurements for the reconstruction procedure, a signal $x_r^{(0)}(n)$ of length N, at the initial iteration $p = 0$, is formed as

$$x_r^{(0)}(n) = \begin{cases} \text{digital}_B\{x(n)\}, & n \in \mathbf{M} \\ 0, & n \in \mathbf{N}_Q \end{cases}, \quad (11)$$

where $\mathbf{M}$ is the set of available sample positions while the set of positions of missing samples is denoted by $\mathbf{N}_Q = \mathbf{N} \setminus \mathbf{M}$. Values of this signal at the available sample positions are equal to original signal $x(n)$, while the values at the positions of the missing samples are set to zero.

Step 1: For each missing sample $n_i \in \mathbf{N}_Q$, two signals

$$\begin{aligned} x_+(n) &= x_r^{(p)}(n) + D\delta(n-n_i) \\ x_-(n) &= x_r^{(p)}(n) - D\delta(n-n_i) \end{aligned} . \tag{12}$$

are formed, by adding a constant value $\pm D$. The reconstruction accuracy depends on D. The value of the largest available measurement is used to initialize parameter D as $D = \max\{|x_r^{(0)}(n)|\}$.

Step 2: Estimate the differential of signal transform measures

$$g(n_i) = \frac{\|X_+(k)\|_1 - \|X_-(k)\|_1}{N} \tag{13}$$

where $X_+(k)$ and $X_-(k)$ present $x_+(n)$ and $x_-(n)$ in the DFT domain, respectively.

Step 3: Form a gradient vector $\mathbf{G}$ of length N. At the available samples positions this vector is zero-valued, indicating that these samples should not be changed. At the missing samples positions, $n_i \in \mathbf{N}_Q$, values of $\mathbf{G}$ are $g(n_i)$, calculated by (13).

Step 4: Update the signal $x_r(n)$ using the gradient vector, and digitize the obtained samples to B-bits, i.e.

$$x_r{}^{(p+1)}(n) = \text{digital}_B\{x_r{}^{(p)}(n) - G(n)\}. \tag{14}$$

These four steps are iteratively repeated until the desired reconstruction precision is achieved. The precision is improved by reducing the value of the parameter D after several iterations. The criterion for such reduction can be conveniently defined as $x_r{}^{(p+1)}(n) = x_r{}^{(p)}(n)$, or alternatively, the oscillatory behavior around the solution can be detected by measuring the angle between the gradients obtained in two successive iterations. The value of this angle close to 180° is an indicator of an oscillatory behavior [7]. The algorithm is stopped when the value D is below the quantization step.

IV. Results

Example 1: Observe a signal defined by

$$x(n) = 3\sin(\tfrac{2\pi}{N}k_1 n) + 2\cos(\tfrac{2\pi}{N}k_2 n) + 0.5\sin(\tfrac{2\pi}{N}k_3 n), \tag{15}$$

of length $N = 64$. The frequency positions are randomly chosen from the range $1, 2, \ldots, N/2$. The sparsity of the signal is $K = 6$. As an example, assume that 25% of samples are unavailable, meaning that $N_Q = 16$. The number of bits is $B = 8$. The original signal, the signal with unavailable samples set to zero, and the reconstructed signal are shown in Fig. 1. The same signal, reconstructed with $B = 4$ bits and $B = 2$ bits are shown in Fig. 2 and Fig. 3, respectively. It is interesting to note that, even when only $B = 2$ bits are used, the algorithm finds the right region of the reconstructed sample.

Example 2: In this example, the performance of the reconstruction is verified by varying the number of bits and missing samples. We assume a signal of the form (15). Typical cases for

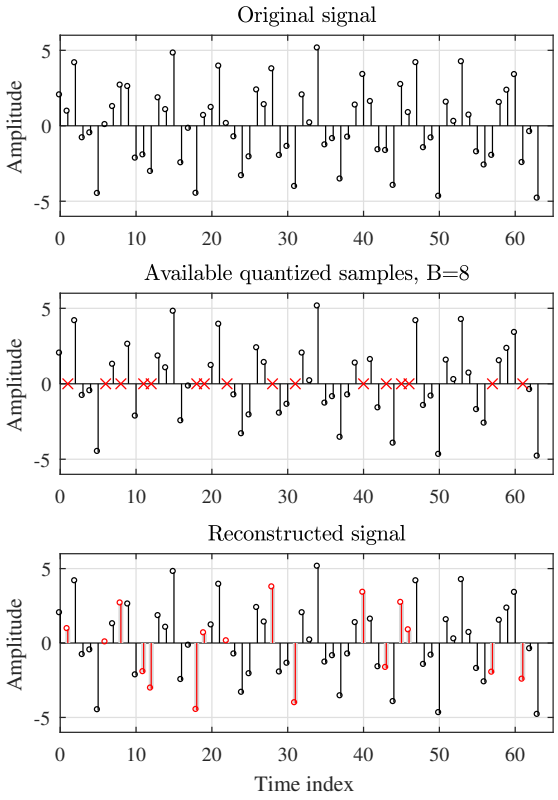

Fig. 1. Gradient-based reconstruction with $B = 8$-bits samples: original signal (top); signal with missing samples set to zero (middle); reconstructed signal (bottom)

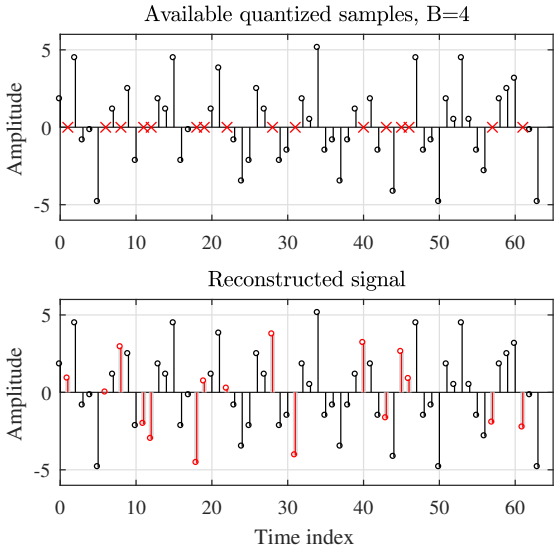

Fig. 2. Gradient-based reconstruction with samples quantized to $B = 4$ bits: signal with missing samples set to zero (top); reconstructed signal (bottom)

TABLE I
MSE RESULTS IN 100 REALIZATIONS WITH VARIOUS NUMBER OF BITS B AND MISSING SAMPLES N_Q

N_Q	$B=4$ $e_q=7.93\times10^{-3}$	$B=8$ $e_q=3.00\times10^{-5}$	$B=12$ $e_q=1.17\times10^{-7}$	$B=16$ $e_q=4.38\times10^{-10}$	$B=20$ $e_q=1.78\times10^{-12}$	$B=24$ $e_q=7.15\times10^{-15}$
8	8.35×10^{-3}	3.15×10^{-5}	1.23×10^{-7}	4.62×10^{-10}	1.89×10^{-12}	7.47×10^{-15}
16	9.29×10^{-3}	3.46×10^{-5}	1.34×10^{-7}	5.04×10^{-10}	2.08×10^{-12}	8.25×10^{-15}
24	1.14×10^{-2}	4.26×10^{-5}	1.63×10^{-7}	6.01×10^{-10}	2.51×10^{-12}	9.90×10^{-15}
32	1.74×10^{-2}	6.96×10^{-5}	2.54×10^{-7}	1.04×10^{-9}	3.96×10^{-12}	1.63×10^{-14}
40	1.08×10^{-1}	5.25×10^{-4}	1.78×10^{-6}	6.45×10^{-9}	3.09×10^{-11}	1.11×10^{-13}

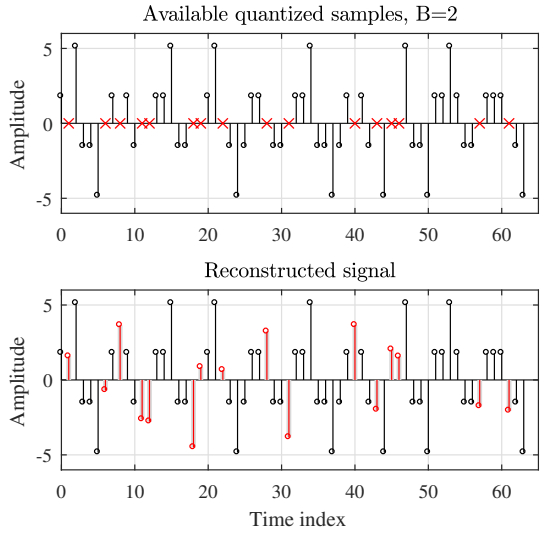

Fig. 3. Gradient-based reconstruction with $B=2$-bits samples: signal with missing samples set to zero (top); reconstructed signal (bottom)

the number of bits are $B=4,8,12,16,20,24$. Moreover, the number of missing samples is varied as $N_Q=8,16,24,32,40$. The mean square error (MSE)

$$\text{MSE}=\text{mean}(||\mathbf{x}-\mathbf{x}_r||_2^2) \tag{16}$$

will be used as the reference for successful reconstruction, where $\mathbf{x}_r$ is the reconstructed signal and $\mathbf{x}$ is the original signal (in vector form). The MSE values of the reconstruction are averaged in 100 realizations and shown in Table I. The average square quantization error in the measurements is given as e_q.

The results imply that the number of bits greatly affects the reconstruction process. For example, in the case when $B=4$ bits are used, and depending on the scenario, a decent reconstruction can be assumed. It is interesting to observe that $B=8$ bits provides a fair reconstruction result. Hence, in practical cases, it delivers an efficient compromise between the number of bits and the reconstruction performance.

V. CONCLUSIONS

The performance of the gradient-based algorithm in the CS reconstruction based on a reduced set of quantized measurements was analyzed. Unlike many conventional CS algorithms, which perform the reconstruction in the transformation domain, the gradient-based algorithm greatly depends on the measurement values. The algorithm has shown a great accuracy in the reconstruction based on B-bit quantized samples. The reconstruction precision was validated numerically. Our future work is focused on the adaptation and improvement of the algorithm, taking into account the application context.

REFERENCES

[1] D. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, 2006, pp. 1289–1306.

[2] R. Baraniuk, "Compressive Sensing," *IEEE Signal Processing Magazine*, vol. 24, no. 4, pp. 118-121, July 2007.

[3] E. J. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. on Information Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[4] LJ. Stanković, S. Stanković, M. Amin, "Missing Samples Analysis in Signals for Applications to L-estimation and Compressive Sensing," *IEEE Signal Processing Letters*, vol. 94, Jan 2014, pp. 401-408, 2014.

[5] E. Candès, J. Romberg and T. Tao. "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory*, vol. 52, pp. 489–509, 2006.

[6] LJ. Stanković, *Digital Signal Processing with Selected Topics*. CreateSpace Independent Publishing Platform, An Amazon.com Company, November, 2015.

[7] L. Stanković, M. Daković, S. Vujović, "Adaptive variable step algorithm for missing samples recovery in sparse signals," *IET Signal Processing*, vol.8, no.3, 2014, pp.246–256, doi: 10.1049/iet-spr.2013.0385

[8] M.A.T. Figueiredo, R.D. Nowak and S.J. Wright, "Gradient Projection for Sparse Reconstruction: Application to Compressed Sensing and Other Inverse Problems," *IEEE Journal on Selected Topics in Signal Processing*, 2007

[9] P. Boufounos, R. Baraniuk, "1-bit Compressive Sensing," *42nd Annual Conf. on Information Sciences and Systems*, Princeton, NJ, USA, 2008.

[10] P. T. Boufounos, "Greedy sparse signal reconstruction from sign measurements," *2009 Conf. Record of the 43rd Asilomar Conference on Signals, Systems and Computers*, CA, USA, 2009. sparse vectors," *IEEE Transactions of Information Theory*, vol. 59. no. 4, pp. 2082–2102, 2011.

[11] I. Stanković, M. Brajović, M. Daković, and L. Stanković, "Complex-Valued Binary Compressive Sensing," *26th Telecommunications Forum (TELFOR 2018)*, November 20 - 21, 2018, Belgrade, Serbia

[12] W. Dai, O. Milenkovic, "Information Theoretical and Algorithmic Approaches to Quantized Compressive Sensing," *IEEE Transactions on Communications*, vol. 59, no. 7, July 2011.

[13] P.T. Boufounos, L. Jacques, F. Krahmer, R. Saab, "Quantization and Compressive Sensing," in: Boche H., Calderbank R., Kutyniok G., Vybial J. (eds) Compressed Sensing and its Applications. Applied and Numerical Harmonic Analysis. Birkh auser, Cham, pp. 193-237, 2013.

[14] H.M. Shi, M. Case, X. Gu, S. Tu, D. Needell, "Methods for quantized compressed sensing," *Information Theory and Applications Workshop (ITA) 2016*, California, USA, 2016.