Extended Reconstruction Approaches for Saturation Measurements Using Reserved Quantization Indices
Li, Peng; Arildsen, Thomas; Larsen, Torben

Published in:
Proceedings of the 12th IEEE International Symposium on Signal Processing and Information Technology

DOI (link to publication from Publisher):
10.1109/ISSPIT.2012.6621286

Publication date:
2012

Document Version
Peer reviewed version

Link to publication from Aalborg University

Citation for published version (APA):
Abstract—This paper proposes a reserved quantization indices method for saturated measurements in compressed sensing. The existing approaches tailored for saturation effect do not provide a way to identify saturated measurements, which is mandatory in practical implementations. We introduce a method using reserved quantization indices to mark saturated measurements, which is applicable to current quantizer models. Two extended approaches based on the proposed method have been investigated compared to the existing approaches. The investigation shows that saturated measurements can be identified by reserved quantization indices without adding extra hardware resources while maintaining a comparable reconstruction quality to the existing approaches.

Index Terms—Compressed sensing, Quantization index, Saturated measurement, Signal reconstruction

I. INTRODUCTION

Compressed sensing (CS) is a recent signal acquisition and reconstruction technique enabling sampling and compression simultaneously [1], [2]. The compressive sampling paradigm may enable perfect reconstruction with a sub-Nyquist sampling frequency [2]–[5], provided that original signals are known to be sparse (or compressible) in some basis. The compressed sensing technique may relax the requirements of high-speed analog-to-digital converter (ADC) for signals with high frequencies. In practical systems, quantization error (or quantization noise) introduced by ADCs influences the quality of the signal reconstruction, which cannot be neglected in compressed sensing [6]. Therefore, several studies have been done to investigate the effects of quantization in compressed sensing systems [6]–[14].

Some researchers propose modified signal reconstruction algorithms based on the information of quantization, which could improve the quality of signal reconstruction [6]–[11]. Others optimized the design of quantizers according to features of their reconstruction algorithms, which could also benefit the signal reconstruction [12], [13]. Further, since saturation is hard to completely avoid in quantization, unbounded quantization error of saturated samples could significantly impact the reconstruction performance in compressed sensing. Two existing approaches, the rejection approach and the consistent approach, tailored for saturated measurements in compressed sensing were proposed in [14]. The saturated measurements are rejected or be enforced consistency in the two existing approaches, respectively. However, no method was provided to identify the positions of the saturated measurements, which is mandatory in practical implementations, see Fig. 1a.

In this paper, we propose a method reserving one or two of the set of possible quantization indices to mark saturated measurements, see Fig. 1b. It requires no extra hardware for the quantizers, and hence can be applied directly in the existing quantization models. We compare the reconstruction quality of the extended approaches based on the proposed method to the existing approaches. The simulation results show the feasibility of the proposed extended approaches.

II. METHODOLOGY

In this paper, a general CS-based structure is modeled with quantized compressed measurements. All input signals are presented as time-discrete vectors according to general CS theory [3].

A. Signal acquisition

To implement an existing CS technique in signal processing systems, it is necessary to find sparsity in the original signal [2], which means that the original signal should be sparse or compressible:

$$ z = \Psi x, $$

where $z \in \mathbb{R}^{N \times 1}$ is the original signal vector, $\Psi \in \mathbb{C}^{N \times N}$ is the dictionary and $x \in \mathbb{C}^{N \times 1}$ is the sparse vector representation of $z$ in $\Psi$. In this model, $x$ is a vector containing only a few non-zero elements. $x$ is sparse if all other elements are zeros, thus $1 \leq ||x||_0 < N$, where $||x||_0$ is the number of non-zero elements in $x$, or $x$ is compressible if all other
elements are small enough to be accurately approximated by zeros. According to general CS theory [2], a measurement matrix \( \Phi \) is used to sample the original signal, which is described by:

\[
y = \Phi z = \Phi \Psi x,
\]

(2)

where \( y \in \mathbb{C}^{M \times 1} \) is the compressed measurement vector, \( \Phi \in \mathbb{C}^{M \times N} \) is the measurement matrix and \( M \) is the number of compressed measurements. If the product of the measurement matrix \( \Phi \) and the dictionary matrix \( \Psi \) obey the Restricted Isometry Property (RIP) [2], the original signal can be recovered from \( M \) measurements, where \( \|x\|_0 < M \ll N \). In this paper, \( \Phi \) is a random Gaussian matrix, which means all elements in \( \Phi \) are randomly chosen from a collection with Gaussian distribution, to obey the requirement of RIP [2].

### B. Quantization and Saturation

According to existing CS theory [4], [5], an original signal can be reconstructed from a compressed measurement vector \( y \). However, in practice, measurements need to be converted from analog to digital before further processing in a DSP. Since quantization is a necessary part of current ADCs [15], quantization error is an important factor influencing the quality of signal reconstruction. For a general view, uniform quantizers are used in this paper:

\[
y_Q = y + e,
\]

(3)

where \( y \) is the input measurement vector, \( y_Q \) is the quantized measurement vector, and \( e \) is an additive quantization error/noise vector. If we choose a one-dimensional scalar uniform quantizer with a resolution of \( q \) bits/meas., the entire quantization range is divided into \( W \), equal quantization partitions, where \( W = 2^q \). In this work we consider scalar quantization of each of the elements of the vector \( y \).

Usually, the quantization range of a quantizer is chosen based on the type or class of input signals applied. If \( \mu \) is the expectation of input data and a quantization range is bounded by \([\mu - G, \mu + G] \) (\( G > 0 \)), \( G \) is called the saturation level and any input data exceeding this range is saturated. In this paper, input data is quantized to the mid-point values of the quantization partitions in uniform quantizers. Since the measurement matrices are random Gaussian matrices in this paper, the elements of the measurement vector \( y \) are assumed to have a Gaussian distribution too. Therefore, saturation level and saturation rate are correlated [16] as:

\[
G = \sigma \sqrt{2} \cdot \text{erf}^{-1}(1 - r),
\]

(4)

where \( \text{erf}^{-1} \) denotes the inverse error function, \( \sigma \) is the standard deviation of \( y \) and \( r \) (\( 0 \leq r \leq 1 \)) is the saturation rate, i.e., the ratio of the number of saturated measurements to the number of total measurements.

According to (4), quantizers with a small saturation level \( G \) lead to increased saturation rate \( r \) and vice versa. Then, if saturated measurements are allowed to occur in the models of the existing rejection and consistent approaches [14], the average quantization error \( e \) of non-saturated values will become smaller when using the same number of quantization indices \( W \) (due to smaller step-size) at the expense of more saturated values. The quality of the reconstructed signal may be better due to smaller quantization errors [14].

### III. Algorithm

#### A. Reserved quantization indices method

To tailor reconstruction for saturation effects, both the existing rejection and consistent approach need to identify the indices of saturated measurements for further processing [14]. However, the two existing approaches do not provide a method to identify saturated measurements. It is not feasible for the reconstruction algorithms to know this information without any change in the quantizers. Therefore, the reserved quantization indices method is proposed to extend the existing approaches, see Fig. 1. It is achieved by reserving some of the available quantization indices to represent saturated measurement values and then allocating the remaining quantization indices to represent unsaturated measurement values. The number of reserved quantization indices is \( n \in \{1, 2\} \) depending on the specific reconstruction algorithm. Based on the proposed method, we modify the two existing approaches in [14]. The extended approaches provide the feasible ways to tailor for the saturated measurements in CS in practical systems.

#### B. Rejection approach

1) Existing formulation

The existing rejection approach was proposed by Laska et al. in [14], which is used to accommodate the saturation effect in CS. We define \( S \) as the set of indices of the unsaturated measurements. The vector of unsaturated quantized measurements \( \tilde{y}_Q \) of length \( M_{\tilde{y}} \) (\( M_{\tilde{y}} \leq M \)) and the measurement matrix \( \tilde{\Phi} \) consisting of the rows corresponding to the unsaturated measurements are defined as:

\[
\tilde{y}_Q = \tilde{y}_Q^S, \quad \tilde{\Phi} = \Phi^S.
\]

(5)

The existing rejection approach is then defined as [14]:

\[
\begin{align*}
\min & \quad \| \tilde{x} \|_1 \\
\text{s.t.} & \quad \| \tilde{\Phi} \tilde{\Psi} \tilde{x} - \tilde{y}_Q \|_2 < \epsilon,
\end{align*}
\]

(6)

\[
\epsilon = \sqrt{\tilde{M} + 2\sqrt{2M} \cdot \sigma_e},
\]

(7)

where the threshold \( \epsilon \) is calculated according to [17] that requires estimation of the standard deviation \( \sigma_e \) of the quantization noise \( e \) in (3). In the existing rejection approach, the saturated CS measurements and the corresponding rows in the measurement matrix are discarded. Then, the unsaturated measurements and the corresponding rows in the measurement matrix are used exactly as in the conventional CS reconstruction approach [18]. If the resolution of the quantizer is \( q \) bits/meas., the number of quantization indices for unsaturated measurement is \( W = 2^q \) in this case that how to indicate which values are saturated is not considered in the existing rejection approach.
2) Extended formulation: The existing rejection algorithm needs to know the locations of saturated measurements in the measurement vectors to generate $\tilde{y}_{Q}$ and $\hat{\Phi}$, and this information is necessary for the signal reconstruction. The existing rejection algorithm in [14] does not take into account how to index saturated measurements in practice. Therefore, in the proposed extended rejection approach, one reserved quantization index ($n = 1$) is used to mark all saturated measurements. If the resolution of the quantizer is $q$ bits/meas., and $\tilde{y}_{QR}$ is the vector of the quantized unsaturated measurements, there are $W - 1$ ($W = 2^n$) quantization indices for quantizing each element in $\tilde{y}_{QR}$ in this case. The entire quantization range is divided into $W - 1$ ($W = 2^n$) quantization partitions to quantize the unsaturated measurements. The extended rejection approach is defined as:

$$\tilde{y}_{QR} = y_{QR}^S,$$  \hspace{1cm} (8)

$$\begin{align*}
\min \|\hat{x}\|_1 \\
\text{s.t.} \|\hat{\Phi}\Psi\hat{x} - \tilde{y}_{QR}\|_2 < \epsilon.
\end{align*}$$  \hspace{1cm} (9)

The extended rejection approach is a simple method tailored for identification of saturated measurements which merely requires a change to the quantizer and not the reconstruction method itself.

C. Consistent approach

1) Existing formulation: Since the information in saturated measurements is wasted in the rejection approach, an advanced approach, called the consistent approach, was proposed in [14]. The existing consistent approach makes use of saturated measurements while they are treated differently by enforcing consistency, which means the saturated measurements should be consistent when sampling the original signal and the reconstructed signal. Provided that $S^+$ and $S^-$ represent the sets of indices of the positive saturated measurements and negative saturated measurements, respectively, the measurement matrix $\Phi$ is divided into two sub-matrices: an unsaturated matrix $\Phi_{\Psi} \in \mathbb{C}^{M\times N}$; and a saturated matrix $\Phi_{\Psi} \in \mathbb{C}^{(M-M)\times N}$, such that:

$$\Phi = \begin{bmatrix} \Phi^S+ \\ -\Phi^S- \end{bmatrix}. $$  \hspace{1cm} (10)

Then, the existing consistent approach is defined as [14]:

$$\begin{align*}
\min \|\hat{x}\|_1 \\
\text{s.t.} \|\Phi\Psi\hat{x} - \tilde{y}_{Q}\|_2 < \epsilon, \\
\text{and } \hat{\Phi}\Psi\hat{x} \geq G \cdot 1,
\end{align*}$$  \hspace{1cm} (11)

where $1 \in \mathbb{C}^{(M-M)\times 1}$ is a vector of ones. The existing consistent approach still uses the unsaturated measurements like conventional $l_1$ norm algorithms [18]. Like the existing rejection approach, if the resolution of the quantizer is $q$ bits/meas., the number of quantization indices for unsaturated measurement is $W = 2^n$ in this case that how to indicate which values are saturated is not considered in the existing consistent approach. The existing consistent approach generally provides a better signal reconstruction quality than the existing rejection approach [14]. However, it should be noted that the total computational complexity also increases in the existing consistent approach [14].

2) Extended formulation: The information of indices of positive saturated measurements and negative saturated measurements is necessary in the above signal reconstruction method. However, the acquisition and recording of this information is not included in the existing consistent approach. Therefore, in the proposed extended consistent approach, two reserved quantization indices ($n = 2$) are used to represent positive and negative saturated measurements, respectively. If the quantizer resolution is $q$ bits/meas. for the entire system, and $\tilde{y}_{QC}$ is the quantized unsaturated measurements vector, the number of quantization indices for $\tilde{y}_{QC}$ is then $W - 2$ ($W = 2^n$). The extended consistent approach is defined as:

$$\tilde{y}_{QC} = y_{QC}^S,$$  \hspace{1cm} (12)

$$\begin{align*}
\min \|\hat{x}\|_1 \\
\text{s.t.} \|\Phi\Psi\hat{x} - \tilde{y}_{QC}\|_2 < \epsilon, \\
\text{and } \hat{\Phi}\Psi\hat{x} \geq G \cdot 1.
\end{align*}$$  \hspace{1cm} (13)

Reserving two quantization indices to mark positive and negative saturated measurements in the proposed approach, respectively, the saturated measurements can be identified for use in the reconstruction algorithm. When the quantizer has a high resolution $q$, lacking two quantization indices for unsaturated values may represents only a minor decrease in the available number of quantization levels and should cause only minor increase of quantization error.

IV. Simulation and Empirical Results

In this section, numerical results are presented to evaluate two proposed extended approaches. We compare the extended approaches in (III-B2, III-C2) to the existing approaches in (III-B1, III-C1) by the reconstruction quality defined in terms of Normalized Mean Square Error (NMSE) $\rho$:

$$\rho = \frac{\|\tilde{z} - z\|_2^2}{\|z\|_2^2} = \frac{\|\Psi\hat{x} - \Psi x\|_2^2}{\|\Psi x\|_2^2},$$  \hspace{1cm} (14)

where $z$ is the original signal, $x$ is its sparse representation, $\tilde{z}$ is the reconstructed signal and $\tilde{x}$ is the reconstructed sparse representation. If $\rho_1$ and $\rho_2$ are two values of NMSE based on the same specification in the existing and extended approaches (rejection or consistent), respectively, the difference of reconstruction quality for the two approaches is evaluated by:

$$\left|\frac{\rho_2 - \rho_1}{\rho_1}\right| \times 100\%.$$  \hspace{1cm} (15)

For a general view, we test above four approaches in the scenarios of different saturation rates $r$, quantizer resolutions $q$ and numbers of measurements $M$.

In all simulations, multi-tone signals are used as original signals, which are sparse in the frequency domain. We stress the fact that the multi-tone signal is just an example of all possible kinds of original signals for the approaches in our simulations, which helps us to focus the analysis on the effects
of the proposed methods. The tones are randomly located in a range of \([0, 500]\) Hz in the frequency domain. The minimum guard space is \(10\) Hz between each tone. According to (1) and (2), we apply the four approaches to multi-tone signals randomly generated according to the following specifications: size of sparse vector \(N = 1000\); number of tones \(K = 10\); number of compressed measurements \(M \in \{80, 160\}\). The expectation of compressed measurements, \(\mu\), is zero in all simulations. The dictionary \(\Psi\) is an inverse discrete Fourier transform (DFT) matrix in our simulations according to (1). The measurement matrix \(\Phi\) has i.i.d. zero-mean Gaussian entries \(\sim \mathcal{N}(0, 1/M)\). We test uniform quantizers with resolution \(q \in \{2, 4, 6, 8\}\) bits/meas. In each simulation, we use 1000 randomly generated multi-tone signals as training signals to estimate the saturation level \(G\) for the uniform quantizer according to different saturation rate \(r \in \{0.5\%, 1\%, 2\%, 4\%, 8\%, 16\%\}\), see (4). In this paper, each simulation repeats 1000 times with randomly generated original signal and measurement matrix in each iteration and all numerical results shown in following figures use the average values. The open source optimizers SPGL1 [19], [20] and CVX [21], [22] are used in our simulations.

Fig. 2 shows the reconstruction quality, in terms of NMSE, versus saturation rate using the existing and extended rejection approaches. Four different quantizer resolutions are tested for both approaches. The simulation results show that the extended rejection approach has a comparable reconstruction performance to the existing rejection approach for high quantizer resolutions. In Fig. 2a, the maximum difference of reconstruction quality between the two approaches is only approximately \(6\%\) for the case of \(q = 4, 6, 8\) bits/meas. Relatively big difference of reconstruction quality between the two approaches is observed for the case of \(q = 2\) bits/meas. This is due to that the quantization error in low quantizer resolutions becomes relatively larger when there is one quantization partition less in the extended rejection approach. In Fig. 2b, the curves for the two approaches hold a consistent similarity. The maximum difference is approximately \(12\%\) when \(q = 4\) or 6 bits/meas. When \(q = 8\) bits/meas., the extended rejection approach even provides a better reconstruction quality than the existing rejection approach. This is an added benefit for the extended approach. It should be noticed that quantization error is only one of the factors which influence the signal reconstruction quality [3], [6]. Due to the non-linearity of reconstruction algorithms in compressed sensing [2], it is possible that the proposed extended approaches exceed the existing approaches in some cases.

Fig. 3 shows the reconstruction quality, in terms of NMSE, versus saturation rate using the existing and extended consistent approaches. The number of quantization indices for unsaturated measurements is two fewer in the extended consistent approach. However, like results in Fig. 2, the extended consistent approach with high quantization resolutions provides a comparable reconstruction performance to the existing consistent approach. In Fig. 3a, the maximum difference of reconstruction quality between the two consistent approaches is below \(11\%\) when \(q = 6\) or 8 bits/meas. In Fig. 3b, the maximum difference is below \(6\%\) when \(q = 6\) or 8 bits/meas.

It is noticed that both proposed extended approaches provide more accurate signal reconstruction with larger numbers of compressed measurements \(M\) or higher quantizer resolutions \(q\). Choosing the optimum saturation rate may significantly improve the signal reconstruction quality for both extended approaches. The extended consistent approach is more robust to larger saturation rates than the extended rejection approach. The main decline of reconstruction quality in the proposed extended approaches occurs when using low resolution quantizers. However, since the two existing approaches cannot provide accurate signal reconstruction with low resolution quantizers, e.g., \(q = 2\) bits/meas., it is not a significant drawback for the extended approaches. In practical implementations, low resolution quantizers are not recommended to be applied to the extended rejection or consistent approach.
V. Conclusions

A reserved quantization indices method is proposed for saturation effects in compressed sensing. It is capable of identifying saturated measurements, which has not been included in the existing approaches. This is achieved by allocating a few of the available quantization indices to saturated measurements. Two extended approaches based on the reserved quantization indices method can be directly applied in the existing quantizer models. The investigation has been done by comparing the reconstruction quality of the extended approaches to the existing approaches. The simulation results indicate that the extended approaches are feasible and more practicable to realize with existing ADCs. The reserved quantization indices method holds the potential of practical quantizer implementations in compressed sensing.

VI. ACKNOWLEDGEMENT

The work is financed by The Danish Council for Strategic Research under grant number 09-067056.

REFERENCES