# Finite Range Scalar Quantization for Compressive Sensing

Jason N. Laska [(1)], Petros Boufounos [(2)], and Richard G. Baraniuk[(1)]

(1) Rice University, 6100 Main St., Houston, TX 77005
(2) Mitsubishi Electric Research Laboratories, 201 Broadway Cambridge, MA 02139
laska@rice.edu, petrosb@merl.com, richb@rice.edu

## Abstract:

Analog-to-digital conversion comprises of two fundamental discretization steps: sampling and quantization. Recent results in compressive sensing (CS) have overhauled the conventional wisdom related to the sampling step, by demonstrating that sparse or compressible signals can be sampled at rates much closer to their sparsity rate, rather than their bandwidth. This work further overhauls the conventional wisdom related to the quantization step by demonstrating that quantizer overflow can be treated differently in CS and by exploiting the tradeoff between quantization error and overflow.

We demonstrate that contrary to classical approaches that avoid quantizer overflow, a better finite-range scalar quantization strategy for CS is to amplify the signal such that the finite range quantizer overflows at a pre-determined rate, and subsequently reject the overflowed measurements from the reconstruction. Our results further suggest a simple and effective automatic gain control strategy which uses feedback from the saturation rate to control the signal gain.

## 1. Introduction

Analog-to-digital converters (ADCs) are an essential part of most modern sensing and communications systems. They are the interface between the analog physical world and the digital processing world that extracts the information we are interested in. Ever-increasing demands for information has pushed the requirements on ADCs to their current physical limits. Fortunately, recent theoretical developments in the area of compressive sensing (CS) enable us to significantly extend the capabilities of current ADCs to keep pace with demand.

CS is a framework that allows signals that have sparse representation, i.e., few non-zero elements, or few non-zero coefficients in some basis, to be sampled at a rate close to the sparsity rate, rather than the Nyquist rate. CS employs linear measurement systems and a non-linear reconstruction algorithms to acquire and recover sparse signals.

Most of the CS literature to-date focuses on one particular aspect of ADCs, namely sampling. In this paper we reexamine the other significant aspect, quantization. Specifically, we show that the core tenets of CS enable us to reduce the error due to quantization by allowing the quantizer to saturate more often than usual and removing the saturated measurements from the reconstruction process.

The organization of this paper is as follows. Section 2. presents a brief background on analog-to-digital conversion, compressive sampling, and finite-range quantization. Section 3. presents a brief analysis of finite-range quantization for CS. We show that CS measurements and the quantization error are i.i.d. Gaussian, and analyze the proposed reconstruction strategy. Section 4., presents numerical results that validate our analysis. We conclude with a brief discussion in Sec. 5.

## 2. Background

### 2.1 Analog-to-digital conversion

Analog-to-digital conversion consists of two discretization steps: *sampling*, which converts an analog signal to a set of discrete measurements, and *quantization*, which converts each real-valued measurement to a discrete one chosen from a pre-determined set. Although both steps are necessary to represent a signal in the discrete digital world, classical results due to Shannon and Nyquist demonstrate that the sampling step is information preserving if a sufficient number of samples, i.e., measurements, are obtained. On the other hand quantization always degrades the signal. The system design to goal is to take enough measurements such that the signal does not alias, and to acquire enough bits to limit the quantization distortion.

### 2.2 Finite-range quantization

Scalar quantization is the process of converting the continuous value of the measurements to one of several discrete values through a non-invertible function $R(\cdot)$. In this paper we focus on uniform quantizers with quantization interval $\Delta$. Thus, the quantization points are $q_k = q_0 + k\Delta$, and every scalar $a$ is quantized to the nearest quantization point $R(a) = \operatorname{argmin}_{q_k} |a - q_k|$. For an infinite-range quantizer this implies that the quantization error is bounded by $|a - R(q)| \leq \Delta/2$.

In practice quantizers have finite range, dictated by hardware constraints such as the voltage limits of the devices and the finite bit-rate of the quantized representation. Without loss of generality we assume a midrise $B$-bit quantizer that represents a symmetric range of values $|a| < T$, where $T > 0$ is the quantization threshold. The corresponding quantization points are at $q_k =$

$\Delta/2 + k\Delta, k = -2^{B-1}, \ldots, 2^{B-1} - 1$. This assumption implies a quantization interval $\Delta = 2^{-B+1}T$. Any measurement with magnitude greater than $T$ saturates the quantizer and "clips" to magnitude $T$, i.e., it quantizes to the quantization point $T - \Delta/2$.

Most classical quantization error analysis assumes that the measurements are scaled such that the quantizer never clips. This is a sensible quantization strategy for classical approaches using linear reconstruction. In that context, saturation events cause significant signal distortion and are undesirable. For that reason, extreme attention is often devoted to pre-ADC automatic gain control (AGC) systems to ensure that the quantizer saturates only rarely. Under this assumption the analysis of a finite or an infinite range quantizer is equivalent in terms of the quantization error. Thus, an infinite-range quantizer is often assumed for its mathematical simplicity.

### 2.3 Compressive sampling (CS)

The theory of *compressive sampling* (CS) overhauls the conventional wisdom on the sampling process. Specifically, [2] and the references therein show that the number of measurements that are sufficient to exactly reconstruct a sampled signal are significantly fewer than the Shannon-Nyquist rate as long as the signal is sparse, i.e., can be represented with very few non-zero components in some basis.

The key components of CS are *randomized measurements* and *non-linear reconstruction*. Specifically, a Nyquist-rate sampled discrete-time signal $\mathbf{x}$ can be sampled at a lower rate by using a random matrix $\Phi$, of dimension $M \times N$:

$$\mathbf{y} = \Phi\mathbf{x}, \tag{1}$$

and reconstructed exactly, if the signal is $K$-*sparse*, i.e., only has $K$ non-zero components in some basis and the matrix $\Phi$ satisfies the *Restricted Isometry Property* (RIP) [2]:

$$\sqrt{1-\delta_{2K}}\|\mathbf{x}\|_2 \leq \|\Phi\mathbf{x}\|_2 \leq \sqrt{1+\delta_{2K}}\|\mathbf{x}\|_2 \tag{2}$$

for all $2K$-sparse signals $\mathbf{x}$, where $\delta_{2K}$ is the RIP constant of $\Phi$. RIP guarantees that the norm of the measurements does not deviate significantly from the norm of the $K$-sparse signal $\mathbf{x}$.

To reconstruct $\widehat{\mathbf{x}}$ from $\mathbf{y}+\mathbf{n}$, where $\mathbf{n}$ is noise with $\|\mathbf{n}\|_2 = \eta$, we perform the optimization

$$\widehat{\alpha} = \min_{\alpha} \|\alpha\|_1 \text{ s.t. } \|\Phi\Psi\alpha - \mathbf{y}\|_2 < \eta, \quad \widehat{\mathbf{x}} = \Psi\widehat{\alpha} \tag{3}$$

where $\Psi$ is a basis and $\|\alpha\|_1 = \sum_i |\alpha_i|$ is the $\ell_1$ norm of the coefficient vector. Reconstructing using (3) guarantees that the norm of the reconstruction error is bounded by $c\eta$, where $c$ is a system-dependent constant [2].

In this paper we use the two key components of CS, namely randomized measurements and non-linear reconstruction, to overhaul the conventional wisdom on scalar quantization. In the next sections we demonstrate that the CS measurement process makes the quantization error a white noise process. We use that result demonstrate that in the context of non-linear reconstruction it is advantageous to scale the signal such that the quantizer saturates at a positive rate and reject the saturated measurements from the reconstruction.

## 3. Finite-range quantization for CS

The non-linear reconstruction methods used in CS and the *democratic* nature of the measurements, suggests that with only a small performance penalty, we can choose to ignore measurements. Specifically, in this work we choose to deliberately saturate the quantizer and ignore the measurements that saturated. In the analysis that follows we demonstrate the advantages of this approach compared to scaling the measurements such that they do not saturate or incorporating the saturated measurements in the reconstruction.

The analysis is based on three distinct results:

1. CS measurements approximately follow an i.i.d. Gaussian distribution, making the quantization error a well characterized white noise process.
2. Clipping without quantization followed by dropping the saturated measurements preserves the signal norm and the RIP.
3. Once quantization is introduced, the signal-to-quantization noise ratio can be minimized by selecting a positive saturation rate and rejecting the saturated measurements.

The subsequent sections state and sketch the proofs for these results and their consequences. Due to space limitations, we defer complete proofs and extended analysis to future publications.

### 3.1 Distribution of CS measurements

We assume the measurement matrix $\Phi$ in (1) is randomly generated using a zero-mean sub-Gaussian distribution with variance $1/M$. Under this assumption, all the measurements $y_i = \sum_j (\Phi)_{i,j} x_j$ are i.i.d. zero-mean random variables with variance $\|\mathbf{x}\|_2^2/M$. Using the Lyapunov variant of the Central Limit Theorem, it is also straightforward to show that as the dimension $N$ of the signal $\mathbf{x}$ increases the $y_i$ become normally distributed. The statement becomes non-asymptotic if the elements of $\Phi$ are themselves distributed as a Gaussian. Our initial experiments show that commonly used CS matrix families reach asymptotic behavior even for small $N$.

The implications of this statement are threefold:

1. The expected number of measurements exceeding in magnitude a threshold $T\|\mathbf{x}\|_2/\sqrt{M}$ is $2Q(T)$, where $Q(x) = \frac{1}{\sqrt{2\pi}}\int_x^{+\infty} e^{-t^2/2}dt$ is the tail integral of the standard Gaussian distribution.
2. The ratio of $T\|\mathbf{x}\|_2/\sqrt{M}$ determines the saturation rate. Thus, scaling the signal such that a specific saturation rate is achieved provides a very effective gain control strategy.
3. The quantization error is a white process, although it is correlated to the measurements.

We should note that in the sequel only the ratio $T\sqrt{M}/\|\mathbf{x}\|_2$ is relevant. This ratio is the threshold we select by varying the parameter $T$. The $\sqrt{M}$ factor reflects that in practical systems the variance of the elements of the measurement matrix is not a function of the number of measurements. The normalization by $\|\mathbf{x}\|_2$ reflects that in practice automatic gain control or prior signal knowledge is used to determine the proper gain in the input.

### 3.2 Analysis of finite-range CS measurements

In this section we introduce clipping at threshold $T\|x\|_2/\sqrt{M}$, without quantization. We reject the clipped measurements and demonstrate that if the remaining measurements, denoted using $\widetilde{\mathbf{y}}$, are sufficient in number, the measurement process still satisfies the RIP and preserves the norm of $K$-sparse signals. We use the notation $\widetilde{(\cdot)}$ to denote the relevant quantities after the saturated measurements are dropped: $\widetilde{M}$ is the number of remaining measurements and $\widetilde{\Phi}$ the mutilated measurement matrix corresponding to the remaining measurements.

Assuming the result of Sec. 3.1, the expected number of saturated measurements is $2MQ(T)$. The remaining $\widetilde{M}$ measurements follow a truncated Gaussian distribution:

$$\widetilde{y_i} \propto \begin{cases} \mathcal{N}\left(y_i; 0, \frac{\|x\|_2^2}{M}\right), & |y_i| < \frac{T\|\mathbf{x}\|_2^2}{\sqrt{M}} \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

Thus, the expected norm of $\widetilde{\mathbf{y}}$ is equal to:

$$E\{\|\widetilde{\mathbf{y}}\|_2^2\} = M(1-2Q(T))\sigma_T^2, \quad (5)$$

where $\sigma_T^2$ is the variance of (4). Thus, the scaled system

$$G\widetilde{\mathbf{y}} = G\widetilde{\Phi}\mathbf{x} \quad (6)$$

$$G = \left(\frac{\|\mathbf{x}\|_2^2}{M(1-2Q(T))\sigma_T^2}\right)^{1/2} \quad (7)$$

$$= \left(\frac{\sqrt{2\pi}}{\sqrt{2\pi}(1-2Q(T)) - 2Te^{-T^2/2}}\right)^{1/2} \quad (8)$$

preserves the expected value of the norm of the signal. It is also straightforward to demonstrate that the density of the norm of the signal concentrates around its expected value with very high probability, in manner similar to [1, 3].

It is also possible to demonstrate that the resulting $\widetilde{\Phi}$, which is now signal-dependent, preserves the RIP for all $K$-sparse signals, as long as $\widetilde{M} = O(K\log(N/K))$, or equivalently $M = O(K\log(N/K)/(1-2Q(T))$. The proof is beyond the scope of this paper [5]. However, it is important since it guarantees recovery of the signal, and the robustness to noise we need in the next section.

### 3.3 Quantization noise

In this section we quantize the thresholded measurements using quantization interval $\Delta = 2^{-B+1}T\|\mathbf{x}\|_2/\sqrt{M}$:

$$R(\widetilde{\mathbf{y}}) = \widetilde{\mathbf{y}} + \widetilde{\epsilon_Q}, \quad (9)$$

where $\widetilde{\epsilon_Q}$ is the vector of the quantization error. From the results of Sec. 3.1 and the distribution of the measurements after thresholding it follows that $\epsilon_Q$ is a white random vector with elements distributed as a wrapped truncated Gaussian random variable and bounded by $\pm\Delta/2$. For small quantization intervals the distribution is well approximated by a uniform distribution in the same interval, with variance $\Delta^2/12$ [6]. Assuming a unit norm input $\mathbf{x}$ the expected squared norm of the quantization error is:

$$E\{\|\widetilde{\epsilon_Q}\|_2^2\} = M(1-2Q(T))\Delta^2/12 \quad (10)$$

$$= 2^{-2B}(1-2Q(T))T^2/3. \quad (11)$$

It can also be shown that for large $M$ the measure of this norm concentrates around its mean. When properly scaled with the $G$ in (8), the quantization error becomes:

$$E\{\|G\widetilde{\epsilon_Q}\|_2^2\} = \frac{\sqrt{2\pi}2^{-2B}}{3}\frac{T^2}{\sqrt{2\pi} - \frac{Te^{-T^2/2}}{(1-2Q(T))}}, \quad (12)$$

which suggests an optimal threshold $T$ that minimizes the error.

If the RIP is guaranteed, the norm of reconstruction error can be bounded by $c\|G\widetilde{\epsilon_q}\|_2^2$ with very high probability [2]. For most practical applications, the minimizing $T$ in (12) is not sufficient to guarantee RIP, and therefore we select the smallest $T$ that does.

A similar analysis can be performed if we keep all the saturated measurements. In this case the RIP always holds and the measurement error is equal to:

$$E\{\|\epsilon_Q\|_2^2\} = \quad (13)$$

$$= M\left((1-2Q(T))\frac{\Delta^2}{12} + \frac{2Q(T)\|\mathbf{x}\|_2^2}{M}\sigma_{\text{trunc}}^2\right), \quad (14)$$

$$= \|\mathbf{x}\|_2^2\left((1-2Q(T))\frac{2^{-2B}}{3} + 2Q(T)\sigma_{\text{trunc}}^2\right), \quad (15)$$

where $\sigma_{\text{trunc}}^2$ is the variance of the tail distribution for a standard Gaussian random variable, as truncated by the saturation. Detailed analysis of this can be found in [4]. At $T$ decreases, both $\sigma_{\text{trunc}}$ and $Q(T)$ increases, which means the error due to the saturated measurements increases at the error due to the unsaturated measurements decreases. The optimal $T$ in this case minimizes (15).

The two strategies can be compared to select the optimal given the operating conditions. Especially in low-bit conditions, reducing the quantization interval pays off in terms of the error. However, the tail effects cause a significant penalty if we keep the measurements, and the better strategy is to discard them. As we discuss in the next section in our extensive simulations under a large variety of practical conditions discarding the measurements performs better than using them.

## 4. Experimental validation

### 4.1 Experimental setup

**Signal model**: We study the performance of our approach using signals sparse in the frequency domain: in each trial $K$ non-zero Fourier coefficients $\alpha_n$ are drawn from an i.i.d. Gaussian distribution, normalized to have unit norm, and randomly assigned to $K$ frequency bins out of the $N$-dimensional space. The sampled signal $\mathbf{x}$ is the DFT of the generated Fourier coefficients. Beyond quantization we do not include additional noise sources. In addition to exactly sparse signals, we have performed extensive simulations with compressible signals and confirmed similar results. However, compressible signals are beyond the scope of this paper.

**Measurement matrix**: For each trial a measurement matrix is generated using a Rademacher distribution: each element is drawn independently to be $+1$ or $-1$ with equal probability. Our extended experimentation, not shown here in the interest of space, shows that our results are robust to large variety of measurement matrix classes.

**Reconstruction metric**: We report the reconstruction *signal-to-noise ratio* (SNR) in decibels (dB):

$$\text{SNR} \triangleq 10\log\left(\frac{\|\mathbf{x}\|_2^2}{\|\mathbf{x}-\widehat{\mathbf{x}}\|_2^2}\right), \quad (16)$$

where $\widehat{\mathbf{x}}$ denotes the reconstructed signal.

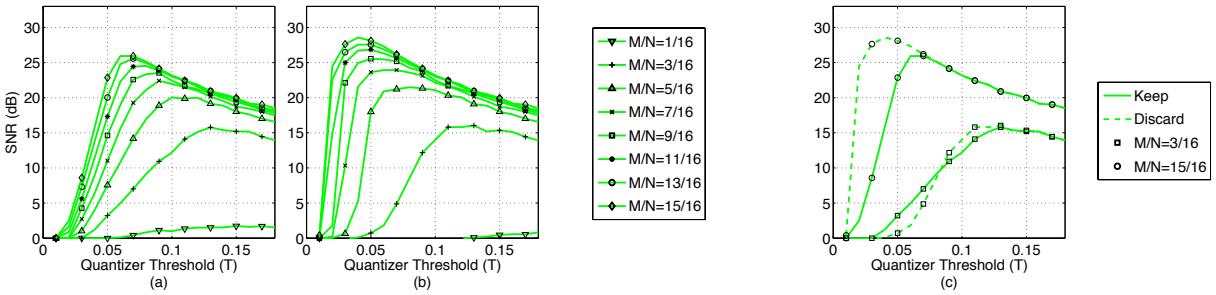


Figure 1: Reconstruction SNR (dB) vs. quantizer saturation threshold ($T$) using a 4-bit quantizer and downsampling rate $\frac{M}{N} = \frac{1}{16} \dots \frac{13}{16}$ when (a) the saturated measurements are used for reconstruction and (b) the saturated measurements are discarded before reconstruction. (c) Side-by-side comparison of (a) and (b) for $\frac{M}{N} = \frac{3}{16}$ and $\frac{15}{16}$: by lowering the threshold $T$ and rejecting saturated measurements, we achieve the highest reconstruction SNR.

### 4.2 Experimental results

We performed extensive simulations with a variety of signal parameters. Due to space limitations, we present here the results for $N = 2048$, $K = 60$, and $B = 4$ which are typical of the system performance. In our experiments we vary $M$ such that $\frac{M}{N} = \frac{1}{16} \dots \frac{15}{16}$ and the threshold $T$ in the range $[0, 0.18]$. For each parameter combination we repeat 100 trials, each trial with a different signal $\mathbf{x}$ and matrix $\Phi$ as described in Sec. 4.1.

For each trial we quantize the measurements using a finite-range quantizer and use them to reconstruct the signal (a) by incorporating the saturated measurements in the reconstruction and (b) by discarding the saturated measurements before reconstruction. Both cases use the linear program (3) with the appropriate value for $\eta$. We denote the reconstructed signal with $\widehat{\mathbf{x}}_{\text{keep}}$ and $\widehat{\mathbf{x}}_{\text{discard}}$, respectively.

The results are shown in Fig. 1, which plots the average reconstruction SNR versus the quantizer dynamic range $T$ for a variety of $\frac{M}{N}$. In particular, Figs. 1 (a) and (b) display the SNR of $\widehat{\mathbf{x}}_{\text{keep}}$ and $\widehat{\mathbf{x}}_{\text{discard}}$, respectively. Figure 1 (c) compares the two approaches for the two extreme cases of $\frac{M}{N} = \frac{3}{16}$ and $\frac{M}{N} = \frac{15}{16}$.

The plots demonstrate that lowering the threshold $T$ such that the saturation rate is non-zero achieves a higher reconstruction SNR compared to scaling such that no measurements clip. Furthermore, rejecting saturated measurements performs better than incorporating them in the reconstruction. This is best illustrated in Fig. 1 (c): the optimal point on the dashed line, which corresponds to discarding saturated measurements, exhibits better SNR than the optimal point on the solid line, which corresponds to incorporating saturated measurements. As expected, the curves coincide when the saturation rate is effectively zero.

We also performed this experiment for larger values of $K$ and $B$. As expected with higher $B$, we achieve less performance gain. As $B$ grows, the quantization error goes down and thus reducing the quantization interval by dropping measurements is less effective. As $K$ increases, rejecting measurements remains an optimal strategy. However, when $K$ is large enough such that the non-saturated measurements do not satisfy RIP, our method performs worse than incorporating the saturated measurements.

## 5. Discussion

Our results demonstrate that CS overthrows the conventional wisdom on finite range quantization. Specifically the common practice of scaling the signal such that the ADC does not overflow is not optimal in light of the non-linear reconstruction. Our results demonstrate that allowing the signal to saturate is advantageous because it decreases the quantization interval in the unsaturated measurements. The non-linear reconstruction methods allow us to discard saturated measurements and prevent the saturation error from affecting the reconstruction process.
Our results further suggests a simple automatic gain control (AGC) strategy, in which the deviation of the average clipping rate from the desired one is used as a feedback to modify the gain. Since the desired clipping rate is non-zero, the feedback is symmetric and increases the gain if the clipping rate is too low. In comparison, classical AGC systems rely on the clipping rate only when the gain is too high and should be reduced. Since in such systems a zero clipping rate is the desired behavior, the AGC needs to rely on other signal features to ensure the gain is sufficient to provide a good signal-to-quantization noise ratio.

## 6. Acknowledgments

The work was supported by grants NSF CCF-0431150, CCF-0728867, CNS-0435425, and CNS-0520280, DARPA/ONR N66001-08-1-2065, ONR N00014-07-1-0936, N00014-08-1-1067, N00014-08-1-1112, and N00014-08-1-1066, AFOSR FA9550-07-1-0301, ARO MURI W311NF-07-1-0185, and the Texas Instruments Leadership University Program.

## References:

[1] R. G. Baraniuk, M. A. Davenport, R. DeVore, and M. Wakin. A simple proof of the Restricted Isometry Property for random matrices. In *Constructive Approximation*, volume 28(3), pages 253–263, Dec 2008.

[2] E. Candes. Compressive sampling. In *Int. Congress of Mathematics*, volume 3, pages 1433–1452, 2006.

[3] S. Dasgupta and A. Gupta. An elementary proof of the Johnson-Lindenstrauss lemma. In *U.C. Berkeley Tech. Rep.*, volume TR-99-006, 1999.

[4] G. A. Gray and G. W. Zeoli. Quantization and saturation noise due to analog-to-digital conversion. In *IEEE Trans. on Aerospace and Electronic Systems*, pages 222–223, Jan 1971.

[5] J. N. Laska, P. Boufounos, M. A. Davenport, and R. G. Baraniuk. Democracy in action: finite-range scalar quantization for compressive sensing. In *To be submitted*, 2009.

[6] A. B. Sripad and D. L. Snyder. A necessary and sufficient condition for quantization errors to be uniform and white. In *IEEE Trans. on Acoustics, Speech, and Signal Processing*, volume ASSP-25, pages 442 – 448, 1977.