Coping with saturating projection stages in RMPI-based Compressive Sensing

Mauro Mangia*, Fabio Pareschi†, Riccardo Rovatti*, Gianluca Setti†, Giovanni Frattini ‡
*ARCES – University of Bologna, via Toffano 2/2, Bologna, Italy. {mauro.mangia2,riccardo.rovatti}@unibo.it
†ENDIF – University of Ferrara, via Saragat 1, Ferrara, Italy. {fabio.pareschi,gianluca.setti}@unife.it
‡Texas Instruments – Rozzano (Milano), Italy. giovanni.frattini@ti.com

***Abstract*—Though compressive sensing hinges on extracting linear measurements from the signals to acquire, actual implementations introduce nonlinearities whose effect can be far from negligible. We here address the problem of saturation in the circuit blocks needed by a Random Modulation Pre-Integration architecture.**

To allow a fair a comparison with previous analysis, we rely on a model capturing the essentials of saturations in actual implementations while being able to reproduce more abstract settings considered in the literature.

Based on this, we analyze some methods already proposed to cope with simplified saturation mechanisms, briefly discussing their underlying principles. Finally, we introduce a novel approach that takes into account the more realistic model and, at the cost of an almost negligible hardware overhead, is extremely effective in countering saturation effects.

I. Introduction

Compressive Sensing (CS) [1] is a set of techniques exploiting the intrinsic structure of some signals to allow their acquisition by means of a number of measurements significantly lower than what would be needed in conventional Analog to Digital Converters (ADC).

Though it is sometimes considered in conjunction with other properties (see e.g. [2, 3]), the cardinal assumption allowing CS is that every instance of the signal can be expressed as the linear combination of an extremely limited number of known waveforms, i.e., the signal is *sparse*.

Sparsity tells us that, once expressed in a basis containing those waveforms, every instance of the signal can be identified by a vector of coefficients α with a very small number of non-zero entries.

If m measurements are taken by means of linear operations on the signal (other proposals do exist but are, by now, a strict minority [4]) and collected in the vector y, we may write $y = A\alpha$ from some matrix A.

CS hinges on that fact that, even if m is much less than the number of waveforms needed for the expression of all signal instances, such an equality can be solved for α exploiting its sparsity and some conditions on A that are surprisingly easy to meet.

Since in physical implementations a signal can be only accessed through its time-domain values, measurements must ultimately be linear combinations of the signal samples, i.e. *projections* of the vector of signal samples onto sequences designed to extract the information needed for signal reconstruction. Hence, the design of projection blocks (multiply and accumulate) is a cornerstone of the implementation of most CS systems.

This is particularly true for the straightforward translation of the above general concepts into an acquisition architecture, i.e., the Random Modulation and Pre-Integration (RMPI) scheme [5]. The incoming waveform is multiplied by a possibly high-frequency Pulse Amplitude Modulated (PAM) signal and then integrated over a certain interval of time T to arrive at a measurement to be sampled and converted into a digital word. Different sequences of PAM symbols characterize different measurements.

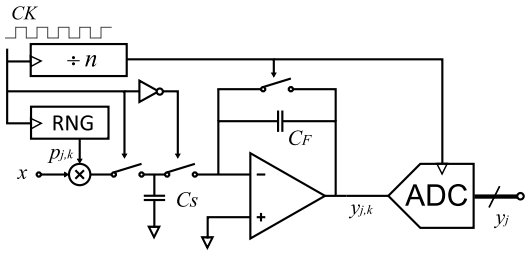

Fig. 1. A switched-capacitor implementation of the block yielding the j-th measurement in an RMPI architecture. The signal is modulated by a PAM with symbols $p_{j,k}$ (produced by a Random Number Generator) and accumulated at the speed of the clock CK. Once every n ticks of CK the accumulation is sampled and converted and then reset to start a new measurement. The working ranges of the OpAmp and of the ADC are limited implying two separate causes of possible saturation.

In the following, we analyze how saturations, that are unavoidable in real implementations, affect the "quality" of RMPI measurements, i.e., the accuracy with which the original signal can be reconstructed based on the information we collect from a projection mechanism in which sums may saturate the range available for circuit operation.

A projection-block that can be used in an RMPI architecture is modeled in Section II taking into account saturations. Recently proposed methods to cope with saturation in less implementation-oriented settings [6] are then analyzed in Section III to show that real-world saturation prevents abstract properties like *democracy* [6] from exercising its beneficial effects. In Section IV we describe a novel technique that considers the full saturation mechanism and tries to squeeze as much information as possible from each projection.

II. System model

The projection-with-saturation model that we introduce is quite close to the actual behavior of an RMPI implementation, though it is general enough to comprise more abstract models of saturation like [6] and thus allow a straightforward comparison between different approaches.

To exemplify it, consider a switched-capacitor scheme for the implementation of a projection block as reported in Figure 1 whose aim is to highlight the cascade between an OpAmp and an ADC and not to suggest any optimized design. The front switches operate at the frequency of the modulating PAM which is n times faster than the ADC which produces a measurement every T seconds.

In this scheme, two saturation effects has to be considered. The first one is due to the ADC, and it has already been

978-1-4673-0219-7/12/$31.00 ©2012 IEEE

considered in [6]. The second one is due ti the limited output swing of the OpAmp. By indicating with $p_{j,k}$ the value at the k-th switching instant of the PAM signal used for the j-th measurement, we get that the output of the OpAmp (whose delay is assumed to be negligible) at time kT/n for $k = 1, \dots, n$ is

$$y_{j,k} = [\![y_{j,k-1} + K_{\rm INT} p_{j,k} x_k]\!]_{V_{\rm INT}^{\rm sat}} \tag{1}$$

where, $K_{\rm INT} = C_S/C_F$ (in the following we will assume $K_{\rm INT} = 1$), $x_k = \left(k\frac{T}{n}\right)$ for $k = 1, \dots, n$, the capacitor reset at the beginning of each measurement implies $y_{j,0} = 0$ and, for simplicity's sake, we model the symmetric saturation of the OpAmp at $\pm V_{\rm INT}^{\rm sat}$ by means of the function

$$[\![v]\!]_V = \begin{cases} V & \text{if } v > V \\ v & \text{if } |v| \le V \\ -V & \text{if } v < -V \end{cases}$$

that approximates the real saturating trend in which the differential gain continuously decreases from 1 to 0.

With the same notation, the ADC then produces the j-th measurement as

$$y_j = [\![y_{j,n}]\!]_{V_{\rm ADC}^{\rm sat}} \tag{2}$$

where we have neglected the effect of quantization that we assume fine enough not to perturb the substance of our considerations. Actually, the two saturations are enough to make us deviate from the ideal setting in which measurement are a linear function of the signal to acquire.

Note that, though we derived our model from a specific scheme, any implementation of a projection block will imply an analogous two-saturation mechanism as a consequence of the unavoidable cascade of a real-world summing stage and a real-world conversion stage.

To approximate a globally linear behavior, the design of the two levels $V_{\rm INT}^{\rm sat}$ and $V_{\rm ADC}^{\rm sat}$ is commonly done in relation to the expected range of the quantities involved that we assume to be centered in 0. Hence, by defining the ideal unsaturated projection as $\hat{y}_j = \sum_{k=1}^n p_{j,k} x_k$ and $\hat{y}_{\rm rms} = \sqrt{\mathbf{E}\left[\hat{y}_j^2\right]}$ we will set $V_{\rm INT}^{\rm sat} = \gamma_{\rm INT} \hat{y}_{\rm rms}$ and $V_{\rm ADC}^{\rm sat} = \gamma_{\rm ADC} \hat{y}_{\rm rms}$ for two suitably designed numbers $\gamma_{\rm INT} \ge \gamma_{\rm ADC} > 0$. Values of $\gamma_{\rm INT} \ge \gamma_{\rm ADC} \ge 2$ are commonly employed based on the fact that $\hat{y}_j/\sqrt{n}$ is usually well approximated by a Gaussian random variable. From the point of view of maximally exploiting the actual range of physical quantities it is preferable to have $\gamma_{\rm INT}$ as close as possible to $\gamma_{\rm ADC}$ while, clearly, the ideal condition $\gamma_{\rm INT} = \gamma_{\rm ADC} = \infty$ cannot be achieved, and its approximations imply an inadmissible waste of resources in terms of power consumption, hardware complexity, conversion time, etc.

In principle, the information extracted by the m projection blocks is contained in the numbers $y_{j,k}$ in (1). Yet, since the ADC is operated only after n integration steps the only values that can be passed to the reconstruction algorithm are the y_j in (2).

Actually, any strategy that aims at coping with saturation must at least recognize when saturation is reached by implicit or explicit comparison with the saturation thresholds (in both cases an operation entailing an almost negligible hardware overhead).

For what concerns the saturation of the ADC, the reconstruction algorithm can be informed of the occurrence of either $y_{j,n} > V_{\rm ADC}^{\rm sat}$ or $y_{j,n} < -V_{\rm ADC}^{\rm sat}$.

For what concerns the saturation in the summing stage, when computing the j-th measurement, the reconstruction algorithm can rely on the knowledge of the two *corruption* time instants κ_j^+ and κ_j^-, defined as

$$\kappa_j^{\pm} = \begin{cases} k & \text{if } \begin{array}{l} y_{j,k} = \pm V_{\rm INT}^{\rm sat} \\ |y_{j,l}| < V_{\rm INT}^{\rm sat} \quad l = 1, \dots, k-1 \end{array} \\ \infty & \text{otherwise} \end{cases}$$

Note that if either κ_j^+ or κ_j^- is finite, it indicates that at that point in time, and not before, the summing stage has reached saturation so that, if the sum continues, the result will be corrupted. Hence, the three events "$y_{j,n}$ *is a valid projection*", "κ_j^+ *is finite*", and "κ_j^- *is finite*" are mutually exclusive. Because of this, even considering corruption times, the amount of data passed to the reconstruction algorithm is still made of a single piece of information for each projection.

Moreover, since feeding the PAM signal into the projection block is a clocked operation, the same synchronous counter triggering the ADC can be read to provide the values of κ_j^+ and κ_j^- when they are finite.

Overall, at time n, the outcome of the operations of the block computing the j-th projection can be encoded in a certain number of inequalities in terms of the values of the PAM signals $p_{j,k}$ and of the signal $x_k = x(kT/n)$.

First, the corruption times give us

$$-V_{\rm INT}^{\rm sat} < \sum_{k=1}^{l} p_{j,k} x_k < V_{\rm INT}^{\rm sat} \tag{3}$$

that holds for $l = 1, \dots, \min\{\kappa_j^+, \kappa_j^-, n\} - 1$. Then, depending on saturations, we may also say that

- when $\kappa_j^{\pm} = \infty$ and $|y_{j,n}| \le V_{\rm ADC}^{\rm sat}$, then

$$\sum_{k=1}^{n} p_{j,k} x_k = y_j \tag{4}$$

- when $\kappa_j^{\pm} = \infty$ and $y_{j,n} > +V_{\rm ADC}^{\rm sat}$, then

$$V_{\rm ADC}^{\rm sat} < \sum_{k=1}^{n} p_{j,k} x_k < V_{\rm INT}^{\rm sat} \tag{5}$$

- when $\kappa_j^{\pm} = \infty$ and $y_{j,n} < -V_{\rm ADC}^{\rm sat}$, then

$$-V_{\rm INT}^{\rm sat} < \sum_{k=1}^{n} p_{j,k} x_k < -V_{\rm ADC}^{\rm sat} \tag{6}$$

- when $\kappa_j^+ < \infty$ and $\kappa_j^- = \infty$, then

$$\sum_{k=1}^{\kappa_j^+} p_{j,k} x_k \simeq +V_{\rm INT}^{\rm sat} \tag{7}$$

- when $\kappa_j^+ = \infty$ and $\kappa_j^- < \infty$, then

$$\sum_{k=1}^{\kappa_j^-} p_{j,k} x_k \simeq -V_{\rm INT}^{\rm sat} \tag{8}$$

Note that (4) is the equality holding in non-saturating systems (i.e., the only one employed in classical CS theory) that here holds only when saturations do not occur.

The possibility of ADC saturation is taken into account by (5) and (6), while (7) and (8) (that hold only approximatively since they are written for a time instant at which the signal is already slightly corrupted) collect information from projections that saturate the range of intermediate sums.

In addition to these mutually exclusive inequalities, (3) encodes the fact that everything runs fine until corruption.

In any case, (3)-(8) are all linear equalities or inequalities and this allows us to plug the information coming from saturation-prone projections into a straightforward generalization of a celebrated reconstruction algorithm, i.e., of one of the procedures used to reproduce the original signal from the measurements [1].

To describe it, say that any instance of the signal $x(t)$ can be expressed as a linear combination $\sum_j \alpha_j s_j(t)$ for some coefficients collected in a vector α and some waveforms whose samples are arranged as the columns of the matrix S. The n samples $x_k = x(kT/n)$ for $k = 1, \ldots, n$ are collected in the vector $x = S\alpha$.

In an ideal linear setting, the vector y containing the measurements can be obtained as $y = Px = PS\alpha$ if P is the $m \times n$ matrix whose rows list the values $p_{j,k}$ of the PAM waveforms used for projections. It can be proved [7, 8] that, under suitable conditions on the matrix $A = PS$, one is able to retrieve the most sparse vector α in agreement with the projections by minimizing the $\mathbb{L}_1$ norm of α subject to the constraints $y = A\alpha$.

Though no analogous formal guarantee exists, it is sensible to extend such a method to the case in which the information that can be extracted from the projections comes also in the form of inequalities. In particular, since $x = S\alpha$, one can rewrite (3)-(8) in terms of the variables α and try to solve

$$\begin{array}{ll} \min & \sum_j |\alpha_j| \\ \text{s.t.} & \text{some choices from (3)-(8)} \end{array} \tag{9}$$

where the actual choices distinguish various approaches.

The analysis of the performance of some of these approaches, two known and one introduced here, will be done by simulation in a normalized setting, analogous to others commonly employed to test CS systems.

In particular we set $n = 256$ and build S with samples of sinusoids and cosinusoids with frequencies from $1/T$ to $127/T$ plus a bias term, all scaled to achieve unit energy. The vector α has 6 non-zero entries drawn at random according to a uniform distribution in $[0, 1]$, and then scaled so that the sample vector x has unit energy. The sample vector is perturbed by additive white Gaussian noise to yield a system input vector with a finite intrinsic Signal-to-Noise Ratio (SNR) of 25 dB. The PAM waveforms are obtained by modulating rectangular pulses of duration $T/256$ by a stream of independent random variables taking values in $\{-1, 1\}$ each with equal probability. The number of measurements is $m = 64$.

To evaluate the system performance, the result of the reconstruction algorithm is matched against the original noiseless vector $S\alpha$ and the energy of the error is compared to the energy of the signal to get a Reconstruction SNR (RSNR).

Performance are established by computing the Probability of Correct Reconstruction (PCR), i.e., the probability that the RSNR is within 10 dB from the RSNR of an ideal system with $\gamma_{\text{INT}} = \gamma_{\text{ADC}} = \infty$, and with the Average RSNR (ARSNR).

III. Dropping or Bounding Allegedly Democratic Projections

In [6] systems with $\gamma_{\text{INT}} = \infty$ are addressed in which only the ADC saturation may affect otherwise fully linear

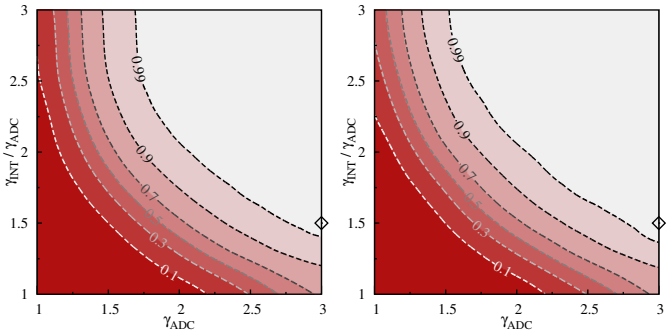

Fig. 2. PCR as a function of γ_{ADC} and $\gamma_{\text{INT}}/\gamma_{\text{ADC}}$ (good implementations should be close to $\gamma_{\text{INT}}/\gamma_{\text{ADC}} = 1$) for SPD (a), and SPB (b).

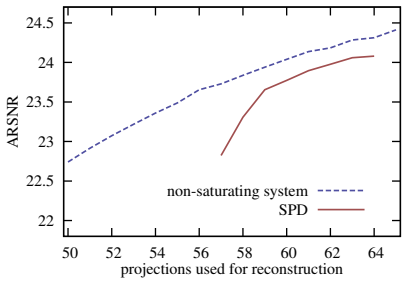

Fig. 3. Comparison between the ARSNR of two systems as a function of the number of projections entering the reconstruction algorithm: an ideal system (upper curve) and a system with $\gamma_{\text{INT}} = \infty$ and $\gamma_{\text{ADC}} = 2.2$ in which $m = 64$ projections are computed SPD is adopted and only the non saturating ones are used for reconstruction (lower curve).

measurements $y = A\alpha$.

As a first option, the simple Saturated Projection Dropping approach (SPD) is analyzed, i.e., the use of (9) with equalities like (4) only for the non-saturated projections.

Since it is sensible to assume that randomized measurement matrices A exhibit a substantial degree of "democracy" (i.e. the ability of behaving almost equally well as an approximate isometry for sparse vectors when some rows are deleted [6]), it is expected that the non-saturated projections deliver just as much information about the signal as any other projection, thus implying a small and very smooth degradation of performance as saturation takes place.

Instead of dropping, Saturated Projection Bounding (SPB) is also proposed, a method that plugs (5) or (6) into (9) in addition to (4) whenever appropriate.

Figure 2 reports the contour plot of the relationship between PCR and the two parameters γ_{ADC} and $\gamma_{\text{INT}}/\gamma_{\text{ADC}}$ (darker colors correspond to lower PCRs) for both SPD (a) and SPB (b).

Note that, even for very high values of $\gamma_{\text{INT}}/\gamma_{\text{ADC}}$ for which the probability of an undetected corruption at the summing stage vanishes, performance degradation is always substantial and is only mildly countered by SPB: in both SPD and SPB, a system aiming at 99% of PCR, while keeping the two saturation thresholds as close as possible, should reserve a dynamic range $> 3\hat{y}_{\text{rms}}$ for the ADC and $> 1.5 \times 3\hat{y}_{\text{rms}} = 4.5\hat{y}_{\text{rms}}$ for the summing stage (the highlighted point in Figure 2-(a) and (b)).

Democracy does not seem to unfold its beneficial effect as shown in Figure 3 in which we analyze what happens

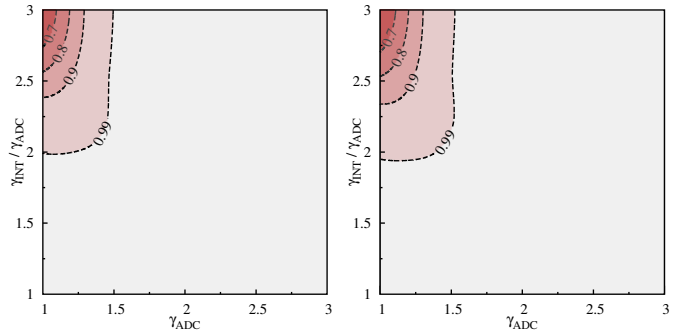

Fig. 4. PCR as a function of γ_{ADC} and $\gamma_{\mathrm{INT}}/\gamma_{\mathrm{ADC}}$ (good implementations should be close to $\gamma_{\mathrm{INT}}/\gamma_{\mathrm{ADC}} = 1$) for SHC (a), and SFC (b)

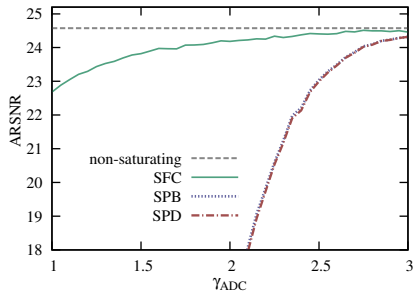

Fig. 5. Comparison between the ARSNR for SPD, SPB, SFC when $\gamma_{\mathrm{INT}}/\gamma_{\mathrm{ADC}} = 1.2$. The SPB and SPD curves are almost superimposing and can hardly be distinguished.

when $m = 64$ projections are computed and the saturated ones discarded to plot (lower curve) the ARSNR as a function of the number of the non-discarded projections that enter the reconstruction algorithm. In the same Figure, the upper curve shows the ARSNR for an ideal system working on an equal number of projections.

Clearly, democracy cannot be advocated as a guideline to cope with saturation even in the $\gamma_{\mathrm{INT}} = \infty$ case. This is because dropping saturated projections means discarding the measurements that carry most energy and thus more information on the signal to acquire. In fact, performance degrades even when no saturation takes place since a finite γ_{ADC} limits the maximum energy of the projection used for reconstruction.

IV. Checking saturation history

What we propose, as in the old common saying, is to exploit "*everything but the Oink!*", i.e., to suffer the almost negligible hardware overhead implicit in corruption checking and be able to plug into (9) whatever equality or inequality from (3)-(8) agrees with the measurement outcomes. Let us indicate this approach as Saturation History Checking (SHC).

Note that considering (3) implies that the number of constraints introduced in (9) is of the order of $m \times n$. Yet, especially for small values of l, (3) materializes slack constraints since corruption is highly improbable when only few samples are summed.

Then, we may choose to avoid (3) and select one among (4)-(8) for each projection thus reducing to Saturation Final Checking (SFC).

Figure 4 shows how, despite the relevant difference in computational effort, SHC and SFC offer substantially the same performance thus making SFC the best approach. When corruption is properly handled, a 99% PCR can be easily reached for $\gamma_{\mathrm{INT}}/\gamma_{\mathrm{ADC}} \simeq 1$ and for very small values of γ_{ADC}, thus allowing an extremely effective implementation. The approach suffers from a small drawback only for γ_{INT} large and γ_{ADC} small, i.e. when many equations like (5) and (6) are used in (9), as only a limited amount of information can be achieved from them.

Actually, the reason for both the poor performance of SPD/SPB and for the superior and equivalent performance of SHC/SFC is quite simple. In fact, note that Figure 4-(a) concerns a system in which all pieces of information are exploited. This keeps the PCR very high but for large $\gamma_{\mathrm{INT}}/\gamma_{\mathrm{ADC}}$ and small γ_{ADC}, i.e., for systems in which the most probable outcome of a projection is a saturation of the ADC without previous corruption. For these systems, most of the constraints in (9) are of the kind (3), (5) and (6), i.e., inequalities. This suggests the somehow trivial conclusion that performance is directly related to the number of equalities introduced into (9) whose number is increased neither by SPD nor by SPB and is the same in SHC and in SFC.

V. Conclusion

Based on a model describing the saturations that must be taken into account when designing projections blocks for RMPI architectures, as well as on extensive simulations we have compared the performance of some methods dealing with saturated measurements.

Two of these methods (SPD and SPB) have been recently proposed while the other two (SHC and SFC) are novel and rely on a negligible hardware overhead to monitor saturation and maximize the amount of information (in terms of equalities) that is available at the signal reconstruction phase.

Figure 5 compares SPD, SPB and SFC in terms of ARSNR for reasonable sizing of the working ranges of the circuit (namely for $\gamma_{\mathrm{INT}}/\gamma_{\mathrm{ADC}} = 1.2$ and γ_{ADC} from 1 to 3) using as a reference the performance of a system with no saturation. The improvement of the newly proposed method is evident from straightforward visual inspection.

References

[1] E. J. Candes and M. B. Wakin, "An Introduction to Compressive Sampling," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 21–30, Mar. 2008.

[2] M. Mangia, R. Rovatti, and G. Setti, "Analog-to-information conversion of sparse and non-white signals: Statistical design of sensing waveforms," in *Proceedings of 2011 IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2011, pp. 2129–2132.

[3] M. Mangia, J. Haboba, R. Rovatti, and G. Setti, "Rakeness-based approach to Compressed Sensing of ECGs," in *Proceedings of 2011 IEEE International Symposium on Biomedical Circuits and Systems (BIOCAS)*, Nov. 2011, pp. 424–427.

[4] J. Haboba, M. Mangia, R. Rovatti, and G. Setti, "An architecture for 1-bit localized compressive sensing with application to eeg," in *Proceedings of 2011 IEEE International Symposium on Biomedical Circuits and Systems (BIOCAS)*, Nov. 2011, pp. 137–140.

[5] J. N. Laska, S. Kirolos, M. F. Duarte, T. S. Ragheb, R. G. Baraniuk, and Y. Massoud, "Theory and Implementation of an Analog-to-Information Converter using Random Demodulation," in *Proceedings of 2007 IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2007, pp. 1959–1962.

[6] J. N. Laska, P. T. Boufounos, M. A. Davenport, and R. G. Baraniuk, "Democracy in Action: Quantization, Saturation, and Compressive Sensing," *Applied and Computational Harmonic Analysis*, vol. 31, pp. 429–443, Nov. 2011.

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

[8] E. J. Candes, J. K. Romberg, K. Justin, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Mathematics*, vol. 59, no. 8, pp. 1207–1223, Aug. 2006.