A BIT-CONSTRAINED SAR ADC FOR COMPRESSIVE ACQUISITION OF FREQUENCY SPARSE SIGNALS

Andrew E. Waters[1,2]*, Charles K. Sestok IV*[2]*, and Richard G. Baraniuk*[1]

[1]Rice University
[2] Texas Instruments

ABSTRACT

We introduce a novel analog-to-digital converter (ADC) based on the traditional successive approximation register. This architecture employs compressive sensing (CS) techniques to acquire and reconstruct frequency sparse signals. One important difference between our approach and traditional CS systems is that our architecture constrains the number of bits used during acquisition rather than the number of measurements. Our system is able to flexibly partition a fixed budget in order to trade the number of measurements it acquires with the quantization depth given to each measurement. We show that this degree of flexibility is particularly advantageous for ameliorating the CS noise folding phenomenon, allowing our ADC significant gains over measurement-constrained compressive sensing systems.

Index Terms— analog-digital conversion, compressed sensing, nonuniform sampling

1. INTRODUCTION

Modern sensing applications are pushing traditional receivers based on the Nyquist criteria to the edge of their sensing capabilities and beyond. This data deluge has forced researchers to consider alternative sampling schemes that can conserve sensing resources by acquiring the desired information in a signal more efficiently. Recent research in the field of compressive sensing (CS) accomplishes this goal for the class of sparse signals. Rather than acquire N samples of a signal at the Nyquist rate, CS attempts to acquire sufficient information from the signal using $M < N$ linear measurements. This allows systems based on compressive sensing to save precious resources over their Nyquist-based counterparts.

The resource savings of CS are tempered in real world applications by the noise folding phenomenon, which observes that any noise power present in the desired signal of interest will be amplified by 3 dB for every octave of compression. Clearly, noise folding can be mitigated by acquiring additional measurements, but this would seem to defeat the purpose of CS entirely. However, in many systems the acquisition cost is not determined by the number of measurements acquired. In an ADC, for example, cost is generally framed in terms of energy consumption and storage complexity, which depends only on the number of bits used during acquisition. A flexible ADC would allow one to trade the number of measurements M acquired with the quantization depth b given to each measurement subject to a total bit constraint B. As shown in the work of Laska, et al [4] such an operating environment provides an opportunity to ameliorate the noise folding burden by acquiring a large number of low-precision measurements. The benefits of this operation can be seen at a high level by considering two sensing extremes: First, for input signals at high SNR, the impact due to noise folding is small, and so the error in the representation is dominated by quantization error. In this case, it is preferable to acquire a small number of measurements at very high bit depth. On the other hand, when the input SNR is low, noise folding becomes a greater concern, and the converter will perform better by shifting its resources towards acquiring more measurements (thus mitigating noise folding) at the expense of having lower precision measurements.

In this work we propose a CS acquisition system based on the SAR ADC that is tailored toward alleviating noise folding. We dub our approach the Bit-Constrained Compressive SAR (BCC-SAR) due to its ability to flexibly trade measurements for quantization depth subject to the constraint $B = Mb$ and with each operating point having equal cost.

Our work shares many similarities with the system described in [5], which is also based on the SAR ADC. This converter acquires a fixed number of measurements with nonuniform quantization depth by randomly sampling in time and forcing closely spaced time samples to sacrifice quantization resolution when interrupted by the acquisition of a new sample. This system has been shown to work well for high SNR signals acquired with a very small number of compressive measurements, with its primary advantage being the simplicity of its implementation and acquisition scheme. However, the converter of [5] possesses some limitations. First, this design does not necessarily lead to the appropriate use of system resources, which are fundamentally in terms of the total

Email: {aew2@rice.edu, sestok@ti.com, richb@rice.edu}; This work was partially supported by the Defense Advanced Research Projects Agency under grant N66001-08-1-2065, AFOSR under grant FA9550-09-1-0432, and the Texas Instruments Leadership University Program.

number of bits rather than the number of measurements. Furthermore, the sampling scheme employed can lead to a wide variance in the bit depth of the measurements which generally leads to suboptimal system performance. Lastly, the authors of [5] consider neither the impact of thermal noise nor that of noise folding in their experiments, which is the driving force behind our architecture.

The remainder of this paper is organized as follows: in Section 2 we provide the relevant background on analog-to-digital conversion and compressive sensing. In Section 3 we discuss the relationship of noise folding to optimal quantization depth and argue that quantization depth should be tailored to the input signal to noise ratio to enable optimal performance. We present our ADC architecture in Section 4 and show how it can modify its acquisition strategy in different operating scenarios. We present several numerical simulations in Section 5 to showcase the utility of our approach and provide concluding remarks in Section 6.

2. BACKGROUND

2.1. Analog-to-digital conversion with the SAR ADC

Practical signal acquisition requires quantizing signal samples to a finite number of bits. One popular architecture for accomplishing this is the SAR ADC [3], a block diagram of which is displayed in Figure 1. The S/H circuit latches an input voltage V_{in} and searches for a digital codeword that minimizes the approximation error. A Q-bit SAR converter estimates each sample with up to Q bits of precision. It begins by first estimating the most significant bit (MSB) by activating the MSB and feeding this result to a digital-to-analog converter (DAC). The DAC converts this message into a corresponding analog signal which it then feeds to the input of a binary comparator which determines whether or not the DAC signal is larger or smaller than V_{in}. If the DAC signal is greater, then the MSB is not needed and set to zero; otherwise the MSB remains active throughout the remainder of the decoding process. This process continues from MSB to LSB, with an additional bit of accuracy being added at each clock cycle.

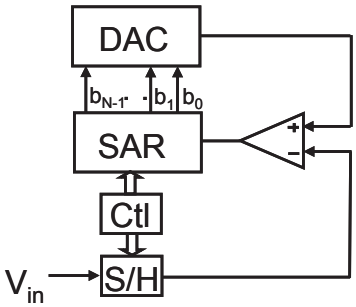


Fig. 1. Block diagram of the SAR ADC.

Due to the SAR requiring Q clock cycles to acquire Q bits of precision, the SAR architecture can naturally be adapted to tradeoff sampling rate and quantization accuracy by choosing to terminate its refinement early in favor of acquiring the next sample sooner.

2.2. Compressive Sensing

Compressive sensing [1] concerns itself with the acquisition of sparse signals at rates close to information rate of the signal. Let $\mathbf{x} \in \mathbb{R}^N$ be a signal that is K-sparse in a transform basis $\mathbf{\Psi}$. Thus, we can write $\mathbf{x} = \mathbf{\Psi s}$ with $\|\mathbf{s}\|_0 = K$. A central result of compressive sensing states that $\mathbf{x}$ can be represented without loss of information from a set of $M < N$ linear measurements computed via a compressive measurement operator $\mathbf{\Phi}$:

$$\mathbf{y} = \mathbf{\Phi x} = \mathbf{\Phi \Psi s} = \mathcal{A}\mathbf{s}$$

provided that $\mathcal{A} = \mathbf{\Phi \Psi}$ satisfies the restricted isometry property (RIP) with RIP constant $\delta < 1$. This condition guarantees that, with high probability, signal norms of all K-sparse signals are preserved by the measurement operator. In this work, we concern ourselves with the Fourier sampling operator $\mathcal{A} = \frac{N}{M}\mathbf{I}_\Omega \Psi$, where $\mathbf{I}_\Omega$ is an $M \times N$ restriction of the identity matrix to the set of rows indexed by Ω and $\mathbf{\Psi}$ is the IDFT basis. When applied to frequency sparse signal, this operator corresponds to random sampling in the time domain. The best known theoretical results for Fourier Sampling are given in [6] which state that $M = |\Omega| = \mathcal{O}(K \log^4 N)$ is sufficient for the RIP to be satisfied.

Signal reconstruction from the observed measurements can be accomplished in a number of ways. A particularly useful method when quantization and/or thermal noise is present in the measurements is to solve a convex program of the form:

$$\min \ \|\mathbf{s}\|_1 \quad \text{subject to} \quad \|\mathbf{y} - \mathcal{A}\mathbf{s}\|_2 \leq \epsilon \tag{1}$$

with the constant ϵ chosen appropriately to bound the distortion in the observed measurements.

3. SIGNAL MODEL AND NOISE FOLDING

3.1. Signal Model

We consider the following signal observation model:

$$\mathbf{y}_B = \mathcal{Q}_b\big(\mathcal{A}(\mathbf{s} + \mathbf{n})\big), \tag{2}$$

where $\mathcal{A}$ is the Fourier Sampling operator described in Section 2.2 and $\mathcal{Q}_b$ is a b-bit scalar quantizer. The frequency domain signal $\mathbf{s}$ is K-sparse with coefficients drawn i.i.d. from $\mathcal{N}(0, \sigma_s^2)$. The noise term $\mathbf{n}$ is statistically white with covariance matrix $\Sigma_{\mathbf{n}} = \sigma_{\mathbf{n}}\mathbf{I}$ and is uncorrelated with $\mathbf{s}$. We define the input signal to noise ratio (ISNR) as:

$$\text{ISNR} := 10\log_{10}\left(\frac{\|\mathbf{s}\|_2^2}{\|\mathbf{n}\|_2^2}\right). \quad (3)$$

3.2. Bit Depth vs Noise Folding

The noise term $\mathbf{n}$ in (2) is added directly to the desired signal $\mathbf{s}$ and processed by $\mathcal{A}$. As shown in [2], the covariance of $\mathcal{A}\mathbf{n}$ is given by $\frac{N}{M}\sigma_{\mathbf{n}}^2\mathbf{I}$. The consequence of this is that the measurement operator $\mathcal{A}$ doubles the variance of the noise for every octave of compression that it provides. This additional noise variance naturally degrades reconstruction performance as well. We can bound the effect of this noise by defining $\mathbf{z} = \mathcal{A}\mathbf{s} - \mathbf{y}$ and applying Theorem 4.1 of [2] from which we can write:

$$\frac{K\sigma_{\mathbf{z}}}{1+\delta} \leq \mathbb{E}\left(\|\mathbf{s}-\widehat{\mathbf{s}}\|_2^2\right) \leq \frac{K\sigma_{\mathbf{z}}}{1-\delta}, \quad (4)$$

where $\widehat{\mathbf{s}}$ is the oracle reconstructed signal. A consequence of (4) is that the expected reconstruction performance increases with a corresponding decrease in the error per measurement. For the observation model of (2) we have two sources of distortion: the folded thermal noise and error induced by the scalar quantizer. Here we provide analysis for these distortion terms which will motivate our proposed architecture. First, the folded thermal noise provides an expected squared distortion of $\frac{N}{M}\sigma_n^2$ to our signal. The exact distortion for the scalar quantizer is, unfortunately, not amenable to direct analysis. However, assuming that the quantizer is range limited to $R = \|\mathcal{A}\mathbf{s}\|_\infty$ the mean square distortion is $\approx \mathcal{O}(2^{-2b})$, subject to the constraint $B = Mb$.

These relations provide us our fundamental tradeoff. As we increase the resolution per sample, we decrease the quantization error per sample. This however, also decreases the number of measurements that we can acquire which increases the distortion due to noise folding. Optimally, we would like to operate at a point on the constraint $B = Mb$ where the distortion due to noise folding is approximately equal to the distortion due to quantization. This point of operation depends on the ISNR. At high ISNR, the distortion due to quantization noise is higher than the distortion due to noise folding. In this case, we would be more concerned with acquiring highly precise measurements at the expense of obtaining a large number of measurements. At low SNR, the opposite is true; noise folding presents more distortion and so we acquire a large number of low precision measurements.

4. SAMPLING SCHEME

The BCC-SAR provides flexibility in choosing an operating point (M, b) on the $B = Mb$ curve, which can be used to minimize the error per measurement. Changing the operating point of the BCC-SAR requires only small adjustments to the SAR ADC of Figure 1. The scope of this modification is limited to controlling the time instances and quantization depth of the sampling function. Let $\Omega = \{\omega_1, \omega_2, \ldots, \omega_M\}$ denote the set of M time instance at which we choose to sample. Due to the overall bit budget B and the serial nature in which the SAR converter acquires additional bits of precision, Ω must satisfy two constraints. First, we must ensure that for any pair of time instances $\omega_i, \omega_j \in \Omega$ that $|\omega_i - \omega_j| > b$ for all $i \neq j$. Second, we must ensure that $|\Omega| \cdot b \leq B$.

We can generate such a sample set as follows: Let $S = \{1, 2, \ldots, N\}$ denote the original index set and let $P = \pi(S)$ denote a pseudorandom permutation of S. We construct Ω iteratively by moving in order through P. At the i^{th} iteration, we add $P_i = S_{\pi_i}$ to Ω. To enforce the constraint that all elements of Ω must lie at a distance no less than b samples from its nearest neighbors, the algorithm removes the elements $\{S_{\pi_i} - b + 1 \ldots S_{\pi_i} - 1, S_{\pi_i} + 1, \ldots, S_{\pi_i} + b - 1\}$ from P. This ensures that the sampling set Ω will be able to sample all entries to the desired precision.

In practice, it may not be possible to operate at an arbitrary point (M, b) such that $B = Mb$, especially as $B \to N$. This is due to the possibility that the construction of Ω will exhaust the set of feasible sample points before exhausting the total bit budget B. This is a function of the random nature in which samples are chosen; it is possible that our sampling pattern Ω will not be sufficiently dense over the index set $\{1, 2, \ldots, N\}$, resulting in a partition that includes a large number of unusable time intervals. Thus, for large values of B it may only be possible to operate at a point (M, b) with $B \geq Mb$. In this event, one could opt to acquire lower precision measurements with the remaining bits, or to increase the quantization depth at each measurement.

5. EXPERIMENTS

We now present a series of numerical simulations that validate the BCC-SAR architecture. We first demonstrate the advantage of the BCC-SAR to the measurement-constrained compressive SAR proposed in [5] (hereafter referred to as the MCC-SAR). We consider in all trials a randomly drawn frequency sparse signal with $K = 2$ and $N = 1024$ with no additive noise (infinite ISNR). We observe samples according to the model (2) with the quantizer range R set to an oracle value of $\|\mathcal{A}\mathbf{s}\|_\infty$ and with each sample allowed to take a maximum of 16 bits/sample. We vary the number of measurements taken by the MCC-SAR over the range $M \in [50, 100, 150, 200, 250, 300]$. After observing measurements with the MCC-SAR, we reconstruct the signal using the program specified in [5]:

$$\widehat{\mathbf{s}} = \arg\min \ \|\mathbf{s}\|_1 \quad \text{subject to} \quad \|\mathbf{W}\left(\mathbf{y} - \mathcal{A}\mathbf{s}\right)\|_2 \leq \sqrt{M} \quad (5)$$

and calculate the reconstruction error $e = \|\frac{\mathbf{s}-\widehat{\mathbf{s}}}{\mathbf{s}}\|$. We then observe the same signal with the BCC-SAR using a bit budget B equal to the observed number of bits used by the MCC-SAR at a quantization depth $b = 16$. In general, the BCC-

SAR will acquire fewer measurements than the MCC-SAR. We again reconstruct according to (5) with $\mathbf{W} = \mathbf{I}$ and again measure the reconstruction error. For each value of M, we repeat our experiment 100 times and compute an average reconstruction error $\bar{e}$. We then compute the average reconstructed signal to noise ratio (RSNR) via $\text{RSNR} = 10\log_{10}\left(\frac{1}{\bar{e}}\right)$.

We display our results in Figure 2, where it is clear that the BCC-SAR consistently achieves a higher average output RSNR than the MCC-SAR. This is due to its obtaining samples of consistent precision while still maintaining a sufficient number of measurements to enable reconstruction. By contrast, low resolution samples that often occur in the MCC-SAR tend to dominate the reconstruction error.

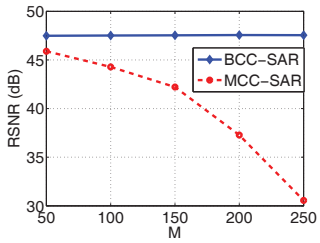
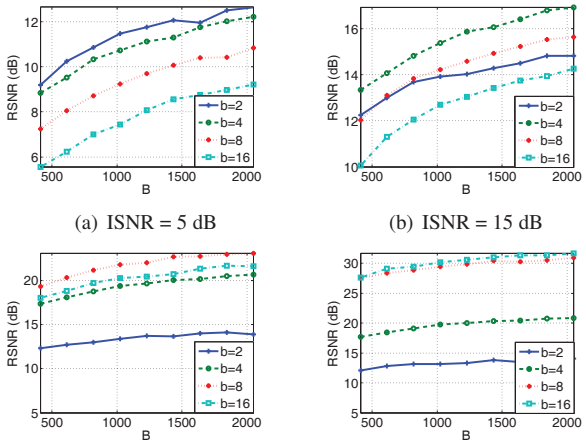


Fig. 2. Comparison of the MCC-SAR converter of [5] for a fixed number of measurements with the BCC-SAR on an equivalent bit budget.

Next, we examine the tradeoff between quantization depth and the number of measurements acquired for various values of ISNR. We consider signals with length $N = 4096$ and sparsity level $K = 6$. We simulate various values of the bit-budget $B \in (\frac{N}{10}, \frac{N}{2})$, values of $b \in \{2, 4, 8, 16\}$, and values of ISNR $\in \{5, 15, 30, 35\}$ dB. At each combination of parameters we reconstruct the signal using oracle-based reconstruction and calculate the resulting RSNR averaged over 100 trials. As expected, we obtain better performance at high ISNR by acquiring a smaller number of high precision measurements. In contrast, at low ISNR we obtain better performance by acquiring a large number of measurements at reduced bit depth. In particular we note gains at low ISNR of nearly 4 dB by using coarser quantization with a larger number of measurements.

6. CONCLUSION

We have presented a bit-constrained SAR ADC that acquires and reconstructs frequency sparse signals via random time sampling. Our fundamental resource is given in terms of bits rather than in terms of measurements as is typical in CS literature. Our architecture is quite flexible in that it can intelligently partition its bit budget to create either many coarsely quantized measurements or a smaller number of high precision measurements. This allows our converter to effectively



(a) ISNR = 5 dB (b) ISNR = 15 dB

(c) ISNR = 30 dB (d) ISNR = 35 dB

Fig. 3. Tradeoff between measurements and quantization for a fixed bit budget B at various values of ISNR

diminish the effects of noise folding for real world CS applications.

The fundamental notion of steering the bit-constrained sampling operator sampling operator to improve reconstruction performance is certainly noteworthy. It further raises the question of whether other situations (possibly outside of the domain of compressive sensing) exist where quantization steering can provide tangible system benefits. Is it be possible to utilize some (possibly time evolving) side information from the underlying signal to steer the sampling process and improve the fidelity of our representation? Lastly, would any achievable gains outweigh the cost of acquiring and utilizing this side information?

7. REFERENCES

[1] E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inform. Theory*, 52(2):489–509, Feb. 2006.

[2] M.A. Davenport, J.N. Laska, J.R. Treichler, and R.G. Baraniuk. The pros and cons of compressive sensing for wideband signal acquisition: Noise folding vs. dynamic range. *Arxiv preprint arXiv:1104.4842*, 2011.

[3] W.A. Kester, editor. *Data conversion handbook*. Newnes, 2005.

[4] J.N. Laska and R.G. Baraniuk. Regime change: Bit-depth versus measurement-rate in compressive sensing. *Arxiv preprint arXiv:1110.3450v1*, 2011.

[5] C. Luo and J. McClellan. Compressive sampling with a successive approximation adc architecture. In *IEEE Int. Conf. Acoust., Speech, and Signal Processing*, pages 2590–2593, Prague, Czech Republic, 2011.

[6] M. Rudelson and R. Vershynin. On sparse reconstruction from fourier and gaussian measurements. *Comm. Pure Appl. Math.*, 61(8):1025–1045, 2008.