52nd IEEE Conference on Decision and Control
December 10-13, 2013. Florence, Italy

Sufficient Conditions for Optimality of Analog to Digital Converters

Mitra Osqui† Alexandre Megretski‡ Mardavij Roozbehani♯

Abstract— **In this paper we prove optimality of a class of Analog to Digital Converters (ADCs), which can be viewed as generalized Delta-Sigma Modulators (DSMs), with respect to a performance measure that can be characterized as the worst-case average intensity of the signal representation error. An analytic expression for the ADC performance is given. Furthermore, our result proves separation of quantization and control for this class of ADCs subject to some technical conditions.**

I. Introduction and Motivation

Analog to Digital Converters (ADCs) act as the interface between the analog world and digital processors. They are present in almost all digital control and communication systems and modern high-speed data conversion and storage systems. Naturally, the design and analysis of ADCs have, for many years, attracted the attention and interest of researchers from various disciplines across academia and industry. Despite the progress that has been made in this field, the design of optimal ADCs remains an open challenging problem, and the fundamental limitations of their performance are not well understood.

A particular class of ADCs primarily used in high resolution applications is the Delta-Sigma Modulator (DSM). The block diagram of a general DSM is shown in Figure 1, where Q is a quantizer with uniform step size δ and dynamic range R (Figure 2). The classical first- and second-order DSMs given in [1] can be equivalently represented by the block diagram illustrated in Figure 1 with $H(z)$ replaced by

$$H_1(z) = \begin{bmatrix} \dfrac{z}{z-1} \\ \dfrac{-1}{z-1} \end{bmatrix}, \qquad (1)$$

†Mitra Osqui is currently a Visiting Lecturer at the Massachusetts Institute of Technology, Cambridge, MA. E-mail: mitra@mit.edu

‡ Alexandre Megretski is currently a Professor of EECS at LIDS at MIT, Cambridge, MA. E-mail: ameg@mit.edu.

♯ Mardavij Roozbehani is currently a Principal Research Scientist at LIDS at MIT, Cambridge, MA. E-mail: mardavij@mit.edu.

and

$$H_2(z) = \begin{bmatrix} \dfrac{z^2}{(z-1)^2} \\ \dfrac{-2z+1}{(z-1)^2} \end{bmatrix}, \qquad (2)$$

respectively.

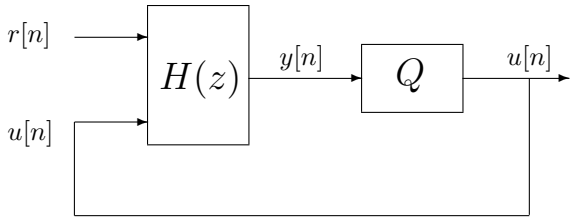

Fig. 1: General Delta Sigma Modulator (DSM)

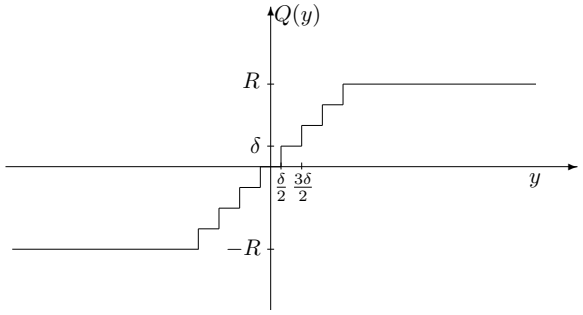

Fig. 2: Quantizer with Uniform Step Size

An extensive body of research on DSMs has appeared in the signal processing literature [1]- [12]. In [13], we cited various different approaches taken for the design and analysis of DSMs. However, to the best of our knowledge, optimality of a general class of ADCs (which also encompasses some classical DSMs) has not been proven. In [14], we presented an exact analytical solution to the optimal ADC for first-order shaping filters, and showed that the classical first-order DSM is identical to our optimal ADC. This result proved the optimality of the

978-1-4673-5716-6/13/$31.00 ©2013 IEEE

classical first-order DSM with respect to the adopted performance measure, and was a step towards understanding the limitations of performance. Moreover, in [14] and [13] we provided a characterization of the solution to the optimal ADC design problem and presented a generic methodology for numerical computation of sub-optimal solutions along with computation of a certified upper bound and lower bound on the performance, respectively.

Figure 3 illustrates the setup used for measuring the performance of the ADC both in this work and in [13]-[14]. The performance of an ADC is evaluated with respect to a cost function which is a measure of the intensity of the error signal e (the difference between the input signal r and its quantized version u) for the worst-case input sequence. The error signal is passed through a shaping filter which dictates the frequency region in which the error is to be minimized. Furthermore, we show that the dynamical system within the optimal ADC is a copy of the shaping filter used to define the performance criteria.

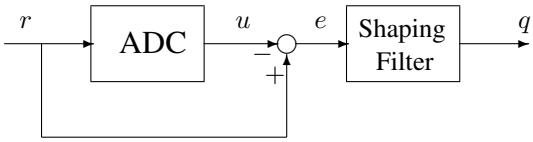

Fig. 3: Setup Used for Measuring the Performance of an ADC

In this paper we provide the optimal solution for arbitrary shaping filters subject to certain technical conditions, which can be used to prove optimality of some higher order DSMs. The contributions of this paper are as follows:

- A general class of ADCs that are optimal with respect to a naturally selected worst-case performance criterion is given along with the exact analytic expression of its performance.
- Subject to certain technical conditions, it is shown that the performance of an ADC designed using a full-state feedback optimal control problem setup is equal to the performance of an ADC designed using the most general problem formulation.
- The optimal control law for this class of ADCs shows separation of quantization and control, that is, quantizing the control law that is optimal in the absence of quantization yields the optimal control law in the presence of quantization.
- Sufficient conditions for separation of quantization and control are given in terms of a lower bound on the quantizer dynamic range and an upper bound on the step size.
- Optimality of certain classical DSMs is shown with respect to specific performance criteria.

In section II the problem formulation is given. Sections III and IV present the approach we take to solve the problem and the main theoretical result, respectively. An illustrative example is given in section V, which utilizes the theoretical result in section IV to provide the conditions for which the classical second-order DSM is optimal.

II. Problem Formulation

The problem setup in this section is taken from [14]. Herein, we use the following notation: given a set P, $\ell_+(P)$ is the set of all one-sided sequences x with values in P, i.e. functions $x : \mathbb{Z}_+ \mapsto P$.

A. Analog to Digital Converters

In this paper, a general ADC is viewed as a causal, discrete-time, non-linear system Ψ, accepting arbitrary inputs in the $[-1, 1]$ range and producing outputs in a fixed finite subset $U \subset \mathbb{R}$, as shown in Figure 4. It is assumed that $\max\{U\} > 1$ and $\min\{U\} < -1$.

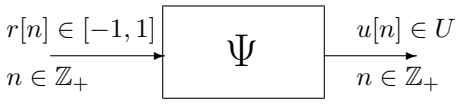

Fig. 4: Analog to Digital Converter as a Dynamical System

Equivalently, an ADC is defined by a sequence of functions $\Upsilon_n : [-1, 1]^{n+1} \mapsto U$ according to

$$\Psi : \; u[n] = \Upsilon_n \left(r[n], r[n-1], \cdots, r[0] \right), \;\; n \in \mathbb{Z}_+. \tag{3}$$

The class of ADCs defined above is denoted by $\mathcal{Y}_U$.

B. Asymptotic Weighted Average Intensity (AWAI) of a Signal

Let $\phi : \mathbb{R} \mapsto \mathbb{R}_+$ be an even, non-negative, and monotonically nondecreasing function on the positive real line; and $G(z)$ be the transfer function of a strictly causal LTI dynamical system L_G with input e and output q:

$$L_G : \begin{cases} x[n+1] = Ax[n] + Be[n], \quad x[0] = 0, \\ q[n] = Cx[n] \end{cases} \tag{4}$$

where A, B, C are given matrices of appropriate dimensions. The Asymptotic Weighted Average Intensity $\eta_{G,\phi}(e)$ of signal e with respect to $G(z)$ and ϕ is given by:

$$\eta_{G,\phi}(e) = \limsup_{N \mapsto \infty} \frac{1}{N} \sum_{n=0}^{N-1} \phi(q[n]). \tag{5}$$

It is assumed without loss of generality that $CB \neq 0$ for $G(z) \not\equiv 0$. Indeed, since $\eta_{G,\phi}$ does not change if $G(z)$ is replaced by $z^k G(z)$, i.e., if $q[n]$ is replaced with $q[n+k]$ in (4), the case when $CB = 0$ can be reduced to the case $CB \neq 0$ by extracting sufficient number of delays from L_G.

C. ADC Performance Measure

The setup used to measure the performance of an ADC is illustrated in Figure 5. The performance measure of $\Psi \in \mathcal{Y}_U$, denoted by $\mathcal{J}_{G,\phi}(\Psi)$, is the worst-case AWAI of the error signal for all input sequences $r \in \ell_+([-1,1])$, that is:

$$\mathcal{J}_{G,\phi}(\Psi) = \sup_{r \in \ell_+([-1,1])} \eta_{G,\phi}(r - \Psi(r)). \tag{6}$$

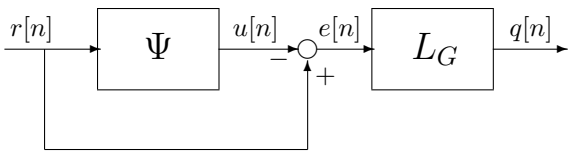

Fig. 5: Setup Used for Measuring the Performance of the ADC

D. ADC Optimization

Given L_G and ϕ, ADC $\Psi_o \in \mathcal{Y}_U$ is considered optimal if $\mathcal{J}_{G,\phi}(\Psi_o) \leq \mathcal{J}_{G,\phi}(\Psi)$ for all $\Psi \in \mathcal{Y}_U$. The corresponding optimal performance measure $\gamma_{G,\phi}(U)$ is defined as

$$\gamma_{G,\phi}(U) = \inf_{\Psi \in \mathcal{Y}_U} \mathcal{J}_{G,\phi}(\Psi). \tag{7}$$

Computation of (7) requires finding a sequence of functions Υ_n, defining ADC Ψ according to (3), that achieves the infimum in (7). This ADC Ψ is called the optimal ADC and its performance is given by $\gamma_{G,\phi}(U)$.

III. Our Approach

In this paper, we show that an optimal ADC can be found by restricting the search over a class of ADCs that are described by a full-state feedback architecture.

A function $K : \mathbb{R}^m \times [-1,1] \mapsto U$ is said to be an admissible controller at performance level $\gamma \in (0,\infty)$ if every triplet of sequences (x_Ψ, u, r) satisfying

$$x_\Psi[0] = 0, \tag{8}$$

$$x_\Psi[n+1] = Ax_\Psi[n] + Br[n] - Bu[n], \tag{9}$$

$$u[n] = K(x_\Psi[n], r[n]), \tag{10}$$

$$q_\Psi[n] = Cx_\Psi[n], \tag{11}$$

also satisfies the dissipation inequality

$$\sup_{N, r \in \ell_+([-1,1])} \sum_{n=0}^{N-1} (\phi(q_\Psi[n]) - \gamma) < \infty. \tag{12}$$

Let γ_o be the maximal lower bound of γ, for which an admissible controller at performance level γ exists. Then K is said to be an optimal controller if (12) is satisfied with $\gamma = \gamma_o$. In this work, we present the optimal control law K along with the corresponding performance $\gamma = \gamma_{G,\phi}(U)$ subject to some technical conditions.

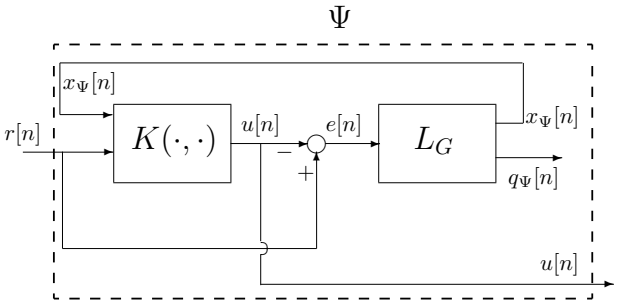

Fig. 6: Full-State Feedback Control Setup

Consider the case when the output of the ADC can take any value in $\mathbb{R}$, i.e., there is no quantization in the ADC; then the optimal control law that minimizes γ in (12) is trivially given by

$$u[n] = (CB)^{-1} CAx_\Psi[n] + r[n]. \tag{13}$$

This choice of control law in conjunction with $x[0] = 0$ gives $q[n] = 0$ for all $n \geq 0$. Thus,

$$\sup_{N, r \in \ell_+([-1,1])} \sum_{n=0}^{N-1} \phi(q_\Psi[n]) = 0.$$

Hence, the smallest γ satisfying (12) is $\gamma = 0$. However, it is not obvious that in the presence of quantization (i.e., the set of values that control u can take belongs to a finite set) the optimal control decision would be to quantize the control given in (13). The main contribution of this paper is to study the conditions that the uniform quantizer Q must satisfy in order for the optimal control decision to be given by

$$u[n] = Q((CB)^{-1}CAx_\Psi[n] + r[n]). \tag{14}$$

We provide sufficient conditions for separation of quantization and control given in terms of a lower bound on the quantizer dynamic range and an upper bound on the step size. That is, R_0 and δ_0 are computed (which are functions of the coefficients of the shaping filter L_G which in turn defines the performance measure) such that for a sufficiently large quantizer range $R > R_0$ and sufficiently small quantizer step size $\delta < \delta_0$, the optimal control law is given by quantizing the control law that would be optimal in the absence of quantization. These conditions will be referred to as the *optimality conditions* where optimality is with respect to the specific worst-case performance criterion defined in terms of the shaping filter L_G.

IV. Class of Optimal ADCs

For $\delta \in (0, 2]$ and $M \in \mathbb{N} \cup \{\infty\}$, define the set U_M and function $K_M : \mathbb{R} \to U_M$ as

$$U_M = \{m\delta \mid m \in \mathbb{Z},\ |m| \leq M\} \tag{15}$$

$$K_M(\theta) = \min\left\{\arg\min_{u \in U_M} |\theta - u|\right\}, \tag{16}$$

where the function K_M represents a nearest neighbor quantization scheme. Consider the ADC $\widehat{\Psi} \in \mathcal{Y}_{U_M}$ defined by

$$L_{\widehat{\Psi}} : \begin{cases} x_{\widehat{\Psi}}[0] = 0, \\ x_{\widehat{\Psi}}[n+1] = Ax_{\widehat{\Psi}}[n] + Br[n] - Bu[n], \\ q_{\widehat{\Psi}}[n] = Cx_{\widehat{\Psi}}[n], \end{cases} \tag{17}$$

with the control law

$$u[n] = K_M\left((CB)^{-1}CAx_{\widehat{\Psi}}[n] + r[n]\right). \tag{18}$$

The control decision $u[n]$ in (18) minimizes $|q_{\widehat{\Psi}}[n+1]|$ at every time instance n. Theorem 1 below states that if $M\delta$ is large enough and δ is small enough, then the ADC defined above is optimal. Thus, an interpretation of this theorem is that a greedy algorithm, i.e., an algorithm that minimizes the cost function at each time instant without taking into account future expectations, is optimal subject to certain conditions. Furthermore, optimality of the control decision in (18) indicates that there is a separation of quantization and control whenever the quantizer step size is sufficiently small and the quantizer range is sufficiently large, i.e., the control law that is optimal in the presence of quantization is given by quantizing the control law that is optimal in the absence of quantization.

Let

$$q_{\widehat{\Psi}}[n+1] = \sum_{i=0}^{k} a_i q_{\widehat{\Psi}}[n-i] + \sum_{j=0}^{k} b_j (r[n-j] - u[n-j]) \tag{19}$$

be the difference equation that is equivalent to (17). Let $\mathcal{F}$ be the causal LTI system with transfer function

$$F(z) = \frac{1}{\sum_{j=0}^{k} b_j z^{-j}}. \tag{20}$$

Let $\{c_l\}_{l=0}^{\infty}$ be the unit sample response of system (19), i.e.

$$F(z) = \sum_{l=0}^{\infty} c_l z^{-l}, \quad \text{for} \quad |z| > R_0 \tag{21}$$

where $R_0 \in \mathbb{R}$ is the maximal absolute value of the largest pole of $F(z)$ in (20).

Theorem 1: Let $\widehat{\Psi} \in \mathcal{Y}_{U_M}$ be the ADC defined by (17)–(18) with $CB \neq 0$ and K_M defined by (15)–(16). Let

$$\beta = \left[|CB|\frac{\delta}{2}\left(\sum_{i=0}^{k}|a_i| + 1\right) + \sum_{j=0}^{k}|b_j|\right]\sum_{l=0}^{\infty}|c_l|, \tag{22}$$

where $\{a_i\}_{i=0}^{k}$ and $\{b_j\}_{j=0}^{k}$ are defined by (19) and $\{c_l\}_{l=0}^{\infty}$ is defined by (20)–(21). Let $M\delta$ be such that $M\delta > 1$ and

$$M\delta > \beta - \delta. \tag{23}$$

Let $f : [0, \infty) \to [0, \infty)$ be a monotonically nondecreasing function and $\phi(q) = f(|q|)$. Then $\widehat{\Psi}$ is an optimal ADC in the sense that

$$\mathcal{J}_{G,\phi}(\Psi) \geq \mathcal{J}_{G,\phi}(\widehat{\Psi}) = \phi(|CB|\delta/2) \quad \forall \Psi \in \mathcal{Y}_{U_M}. \tag{24}$$

Proof: Please see: arxiv.org/pdf/1207.4967.pdf ■

The optimal ADC architecture presented in Figure 6 along with the optimal control law given in (18) can be equivalently represented by Figure 7 and equation (25), where Q is a uniform quantizer with step size δ and saturation level $M\delta$ satisfying (23) and $G(z)$ is the transfer function of the shaping filter L_G.

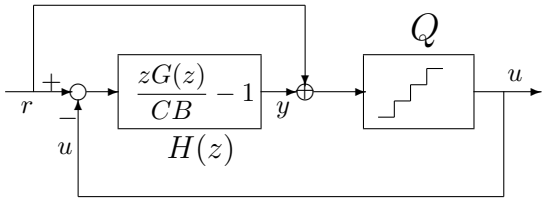

Fig. 7: Optimal ADC Architecture, where $G(z) = C(zI - A)^{-1}B$ is the transfer function of L_G

$$H(z) = (CB)^{-1}zG(z) - 1 = (CB)^{-1}C(zI - A)^{-1}AB \quad (25)$$

Figure 7 has a DSM architecture and thus with proper selection of the shaping filter L_G, many standard DSMs that satisfy the optimality conditions in Theorem 1 can be proven optimal. Applying Theorem 1 for shaping filter $L_G = 1/(z-1)$, gives optimality of the first-order DSM for any uniform quantizer Q with $M\delta > 1$ and $\delta \in (0, 2]$, which is in agreement with the the result in [14]. In the next section, we discuss optimality of some second-order classical DSMs. Of course, with an appropriate selection of L_G, Theorem 1 can be similarly applied to provide sufficient conditions on the quantizer range and step size to guarantee optimality of third- and higher-order classical DSMs.

V. Optimality of Some Classical Second-Order DSMs

The architecture of the optimal ADC (Figure 7) indicates that a proper selection of the shaping filter L_G, which is used to define the performance measure, will prove the optimality of some classical DSMs with respect to that performance measure. This is illustrated in this section for the classical second-order DSM. Furthermore, it is possible to design a second- and higher-order ADCs with superior performance to the classical DSM by selecting a sharper noise-shaping filter.

Consider the setup in Figure 5 with the transfer function of L_G as $G(z) = z/(z-1)^2$. Let the state-space matrices for L_G be given by

$$A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 \end{bmatrix}, \quad D = 0$$

The difference equation for L_G is given by:

$$q[n+1] = 2q[n] - q[n-1] + r[n] - u[n].$$

Thus, the non-zero coefficients are $a_0 = 2$, $a_1 = -1$, and $b_0 = 1$; and the transfer function defined in (20) is $F(z) = 1$. Therefore, $c_0 = 1$. We have $CB = 1$, and thus from (22) we have

$$\beta = 2\delta + 1.$$

Hence, the condition on the range of the quantizer is

$$M\delta > \delta + 1.$$

The optimal ADC Ψ (shown in Figure 8) is given by

$$L_\Psi : \begin{cases} x_\Psi[0] = 0, \\ x_\Psi[n+1] = Ax_\Psi[n] + Br[n] - Bu[n], \\ q_\Psi[n] = Cx_\Psi[n], \end{cases} \quad (26)$$

with the control law

$$u[n] = K_M\left(CAx_\Psi[n] + r[n]\right), \quad (27)$$

where K_M is defined by (16).

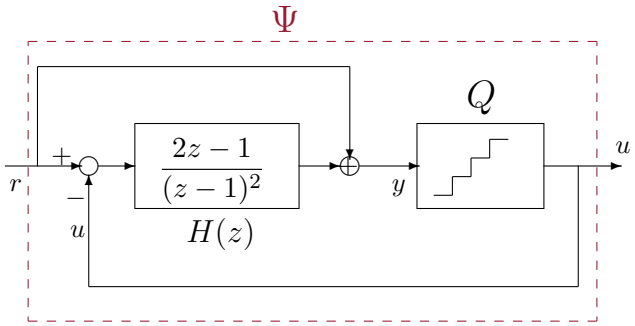

Fig. 8: Optimal ADC when the performance is defined for shaping filter L_G with transfer function $G(z) = \frac{z}{(z-1)^2}$

According to Theorem 1, the performance of this optimal ADC is given by

$$\mathcal{J}_{G,\phi}(\Psi) = \phi(\delta/2). \quad (28)$$

In Figure 8, the transfer functions from r to y and from u to y are denoted by $H_{ry}(z)$ and $H_{uy}(z)$,

respectively. It is clear that these transfer functions are given by:

$$H_{ry}(z) = \frac{z^2}{(z-1)^2},$$
$$H_{uy}(z) = \frac{-2z+1}{(z-1)^2},$$

which are equal to the transfer function of the loop filter for the second-order classical DSM given in (2). Therefore, for the shaping filter $L_G = z/(z-1)^2$, and any uniform quantizer Q with step size $\delta \leq 2$ and the magnitude of the largest value of the quantizer being larger than $1+\delta$, the classical second-order DSM is optimal.

VI. CONCLUSION

In this work, it was shown that there exists separation of quantization and control for a class of ADCs. Sufficient conditions were given in terms of the quantizer range and step size, which when satisfied, resulted in the optimal discrete control law being exactly equal to the quantized version of the optimal control law for the linear system in the absence of quantization. An analytic description for a class of optimal ADCs, which were shown to have DSM-like architecture, along with the exact analytic expression for its performance was provided. As a consequence of this result, certain DSMs were proven optimal.

REFERENCES

[1] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-Time Signal Processing*. Prentice-Hall, 1999.

[2] M. Derpich, E. Silva, D. Quevedo, and G. Goodwin, "On optimal perfect reconstruction feedback quantizers," *IEEE Transactions on Signal Processing*, vol. 56, no. 8, pp. 3871 –3890, Aug 2008.

[3] S. Ardalan and J. Paulos, "An analysis of nonlinear behavior in delta - sigma modulators," *IEEE Transactions on Circuits and Systems*, vol. 34, no. 6, pp. 593 – 603, jun 1987.

[4] A. Marques, V. Peluso, M. S. Steyaert, and . W. M. Sansen, "Optimal parameters for $\Delta\Sigma$ modulator topologies," *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, vol. 45, no. 9, pp. 1232–1241, Sep. 1998.

[5] R. Schreier and G. Temes, *Understanding Delta-Sigma Data Converters*. IEEE Press, John Wiley and Sons, Inc, 2005.

[6] S. Norsworthy, R. Schreier, and G. C. Temes, *Delta-Sigma Data Converters: Theory, Design, and Simulation*. IEEE Press, John Wiley and Sons, Inc, 1997.

[7] N. T. Thao and M. Vetterli, "A deterministic analysis of oversampled A/D conversion and $\Sigma\Delta$ modulation," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, pp. 468–471, Apr. 1993.

[8] ——, "Deterministic Analysis of Oversampled A/D Conversion and Decoding Improvement Based on Consistent Estimates," *IEEE Transactions on Signal Processing*, vol. 42, no. 3, pp. 519–531, Mar. 1994.

[9] N. T. Thao, "The Tiling Phenomenon in $\Sigma\Delta$ Modulation," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 51, no. 7, pp. 1365 – 1378, Jul. 2004.

[10] D. Quevedo and G. Goodwin, "Multistep optimal analog-to-digital conversion," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 52, no. 3, pp. 503 – 515, march 2005.

[11] P. Steiner and W. Yang, "A framework for analysis of high-order sigma-delta modulators," *Circuits and Systems II: Analog and Digital Signal Processing, IEEE Transactions on*, vol. 44, no. 1, pp. 1 –10, jan 1997.

[12] H. Wang, "A geometric view of sigma; delta; modulations," *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, vol. 39, no. 6, pp. 402 –405, jun 1992.

[13] M. Osqui, A. Megretski, and M. Roozbehani, "Lower bounds on the performance of analog to digital converters," in *Decision and Control and European Control Conference (CDC-ECC), 2011 50th IEEE Conference on*, dec. 2011, pp. 1036 –1041.

[14] ——, "Optimality and Performance Limitations of Analog to Digital Converters," *Conference on Decision and Control*, pp. 7527–7532, Dec. 2010.

[15] A. Megretski, "Robustness of finite state automata," in *Multidisciplinary Research in Control: The Mohammed Dahleh Symposium 2002*, ser. Lecture Notes in Control and Information Sciences, L. Giarre and B. Bamieh, Eds. Springer, 2003, vol. 289, pp. 147–160.

[16] M. Osqui, M. Roozbehani, and A. Megretski, "Semidefinite Programming in Analysis and Optimization of Performance of Sigma-Delta Modulators for Low Frequencies," *American Control Conference*, pp. 3582–3587, Jul. 2007.

[17] K. Zhou, J. C. Doyle, and K. Glover, *Robust and Optimal Control*. Prentice Hall, 1996.

[18] T. Basar and P. Bernhard, H^∞ - *Optimal Control and related Minimax Design Problems: A Dynamic Game Approach*. Birkhauser, 1995.