Optimality and performance limitations of analog to digital converters

Citation: Osqui, Mitra, Alexandre Megretski, and Mardavij Roozbehani. "Optimality and Performance Limitations of Analog to Digital Converters." Proceedings of the 49th IEEE Conference on Decision and Control (CDC), 2010. 7527–7532. © Copyright 2010 IEEE

As Published: http://dx.doi.org/10.1109/CDC.2010.5718166

Publisher: Institute of Electrical and Electronics Engineers (IEEE)

Persistent URL: http://hdl.handle.net/1721.1/72696

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

49th IEEE Conference on Decision and Control
December 15-17, 2010
Hilton Atlanta Hotel, Atlanta, GA, USA

Optimality and Performance Limitations of Analog to Digital Converters

Mitra Osqui† Alexandre Megretski‡ Mardavij Roozbehani♯

Abstract— **The paper deals with the task of optimal design of Analog to Digital Converters (ADCs). A general ADC is modeled as a causal, discrete-time dynamical system with outputs taking values in a finite set. Its performance is defined as the worst-case average intensity of the filtered input matching error. The design task can be viewed as that of optimal quantized decision making with the objective of optimizing the performance measure. An algorithm based on principles of optimal control is presented for designing general m-dimensional ADCs. The design process involves numerical computation of the candidate value function of the underlying dynamic program, which is computed iteratively, in parallel with the quantization law. A procedure is presented for certifying the numerical solution and providing an upper bound for performance of the designed ADC. Furthermore, an exact analytical solution to the optimal one-dimensional ADC is presented. It is shown that the designed one-dimensional optimal ADC is identical to the classical Delta-Sigma Modulator (DSM) with uniform quantization spacing.**

I. Introduction and Motivation

Analog to Digital Converters (ADCs) are integral components in many digital electronic systems that interact with an analog environment. These include but are not limited to audio/video and music applications, communication devices, and data conversion/storage systems. Design and analysis of ADCs has, justifiably so, received significant attention from various disciplines across academia and industry.

A particular class of ADCs primarily used in high resolution applications is the Delta-Sigma Modulator (DSM). Figure 1, illustrates the classical one-dimensional DSM [1], where Q is a quantizer with uniform step size.

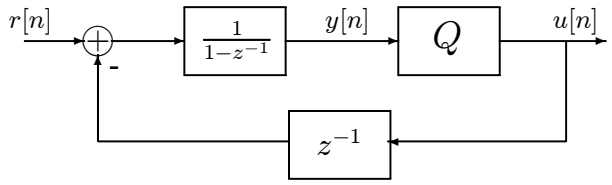

Fig. 1. Classical One-dimensional Sigma-Delta Modulator

An extensive body of research on DSMs has appeared in the signal processing literature. One well known approach is based on linearized additive noise models and filter design for noise shaping [1]-[6]. The underlying assumption for validity of the linearized additive noise model is availability of a relatively high number of bits. Alternative approaches based on a formalism of the signal transformation performed by the quantizer have been exploited for deterministic analysis in [7]-[9]. Some other works that do not use linearized additive noise models are in [10]-[12]. In the control field, relevant work can be found in [13]-[14] and some recent work on optimal quantized control in [15]-[17]. Work on optimal dynamic quantization can be found in [18]. However, to the best of our knowledge, optimality of the classical one-dimensional DSM depicted in Figure 1 has not been verified. One of the contributions of our work is to provide an exact analytical solution to the optimal one-dimensional ADC. Furthermore, we show that our optimal ADC is identical to the classical one-dimensional DSM (Figure 1), hence, proving its optimality with respect to the performance measure adopted in this paper.

In this paper, we are concerned with optimal design of ADCs. The performance of an ADC is evaluated with respect to a cost function which is a measure of the intensity of the error signal (the difference between the input signal and its quantized version) for the worst case input. The design objective is to find a quantization law that minimizes this cost function. We pose the design problem within a nonlinear optimal control framework and show that the optimal quantization law can be characterized via a *control Lyapunov function* satisfying an analog of the Bellman inequality. The value function in the Bellman inequality, and the corresponding control law can be jointly computed via value iteration. For computation of these functions, the state space and the input space are discretized and computations are restricted to finite subsets of the gridded space.

The main contributions of this paper are as follows. First, we provide a characterization of the solution to the optimal ADC design problem for the general m-dimensional case. Second, we present a generic methodology for numerical computation of sub-optimal solutions along with computation of an upper bound on the performance, via post-design verification. Third, we give an exact analytical solution to the optimal one-dimensional ADC.

The organization of this paper is as follows. Section II provides a rigorous problem formulation. The main contributions are presented in Sections III and IV. Section III describes the design and verification methodology for the general m-dimensional case, followed by the optimal one-dimensional ADC in Section IV. A two-dimensional example is presented in Section V. Section VI concludes the paper.

This work was supported in part by the NSF grant 0835947 and the Los Alamos National Laboratory's Information Science and Tech. Inst.

† Mitra Osqui is currently a Ph.D. candidate at the department of EECS, Laboratory for Information and Decision Systems (LIDS) 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 research scientist at LIDS at MIT, Cambridge, MA. E-mail: mardavij@mit.edu.

II. Problem Formulation

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 2.

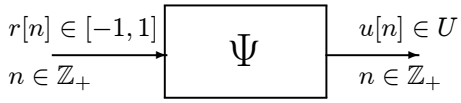

Fig. 2. Analog to Digital Converter as a Dynamical System

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

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

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

B. Asymptotic Weighted Average (AWA) of a Signal

The Asymptotic Weighted Average (AWA) of a signal w is denoted by $\rho_{G,\phi}\left(w\right)$, which depends on the transfer function $G\left(z\right)$ of a strictly causal LTI dynamical system L_G and a non-negative function $\phi : \mathbb{R} \to \mathbb{R}_+$:

$$\rho_{G,\phi}\left(w\right) = \limsup_{N\to\infty} \frac{1}{N+1} \sum_{n=0}^{N} \phi\left(q\left[n\right]\right), \quad (2)$$

where, as shown in Figure 3, the sequence q is the output of the dynamical system L_G defined as:

$$\begin{aligned} x\left[0\right] &= 0, & (3)\\ x\left[n+1\right] &= Ax\left[n\right] + Bw\left[n\right], \qquad \forall n \in \mathbb{Z}_+ & (4)\\ q\left[n\right] &= Cx\left[n\right]. & (5) \end{aligned}$$

where A, B, C are given matrices of appropriate dimensions.

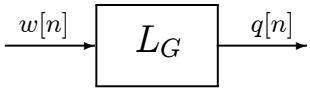

Fig. 3. Strictly Proper LTI Filter $G(z)$

C. ADC Performance Measure

The setup that we use to measure the performance of an ADC is illustrated in Figure 4. The performance measure of $\Psi \in \mathcal{Y}_U$, denoted by $\mathcal{J}_{G,\phi}\left(\Psi\right)$, is the worst-case AWA of the error signal for all input sequences $r\left[n\right] \in [-1,1]$, that is:

$$\mathcal{J}_{G,\phi}\left(\Psi\right) = \sup_{r\in\Omega} \rho_{G,\phi}\left(\Psi\left(r\right) - r\right) \quad (6)$$

where $\Omega = \{r : \; \mathbb{Z}_+ \to [-1,1]\}$ is the set of all possible input signals.

The motivation for using the AWA as a measure for the quality of analog to digital conversion is that when $\phi\left(\cdot\right) = \left|\cdot\right|^2$ and L_G is a strictly stable dynamical system with transfer function $G\left(z\right)$, the AWA can be interpreted as the average power of the filtered input for signals which are sums of sinusoids:

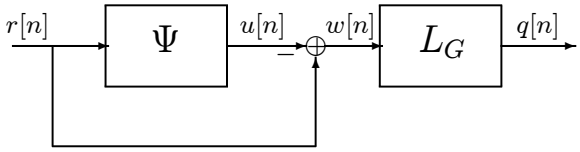

Fig. 4. Setup Used for Measuring the Performance of the ADC

$$w\left[n\right] = \sum_{k=0}^{\infty} w_k e^{j\omega_k n} \Rightarrow \rho_{G,\phi}\left(w\right) = \sum_{k=0}^{\infty} \left|w_k\right|^2 \left|G\left(e^{j\omega_k}\right)\right|^2. \quad (7)$$

Therefore, the AWA allows for penalizing the input of the filter over the frequency range of interest via the pass-band of $G\left(z\right)$. An alternative measure can be obtained with $\phi\left(\cdot\right) = \left|\cdot\right|$, which is attractive due to its simplifying properties. In this case, the AWA represents the average amplitude of the filtered input signal.

D. ADC Optimization

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

$$\gamma_{G,\phi}\left(U\right) = \inf_{\Psi\in\mathcal{Y}_U} \mathcal{J}_{G,\phi}\left(\Psi\right). \quad (8)$$

In the remainder of this paper we present a framework for:

(i) Designing an optimal one-dimensional ADC and finding the optimal value of (8).

(ii) Designing a sub-optimal m-dimensional ADC and finding an upper bound for (8).

III. Our Approach

We search for the optimal ADC within the class of time-invariant state-space models and associate the optimal ADC design problem with a full-information feedback control problem. The setup is depicted in Figure 5, where the function $K : \mathbb{R}^m \times [-1,1] \to U$ is said to be an admissible controller if every triplet of sequences (x_Ψ, u, r) satisfying

$$\begin{aligned} x_\Psi\left[0\right] &= 0, & (9)\\ x_\Psi\left[n+1\right] &= Ax_\Psi\left[n\right] + Br\left[n\right] - Bu\left[n\right], & (10)\\ u\left[n\right] &= K\left(x_\Psi\left[n\right], r\left[n\right]\right), & (11)\\ q_\Psi\left[n\right] &= Cx_\Psi\left[n\right], & (12) \end{aligned}$$

also satisfies

$$\sup_{N, r\in\Omega} \sum_{n=0}^{N} \left(\phi\left(q_\Psi\left[n\right]\right) - \gamma\right) < \infty \quad (13)$$

for some $\gamma \in [0,\infty)$. Note that if (13) holds, then $\mathcal{J}_{G,\phi}\left(\Psi\right) \leq \gamma$. Let γ_o be the maximal lower bound of γ, for which an admissible controller exists. Then K is said to be an optimal controller if (13) is satisfied with $\gamma = \gamma_o$.

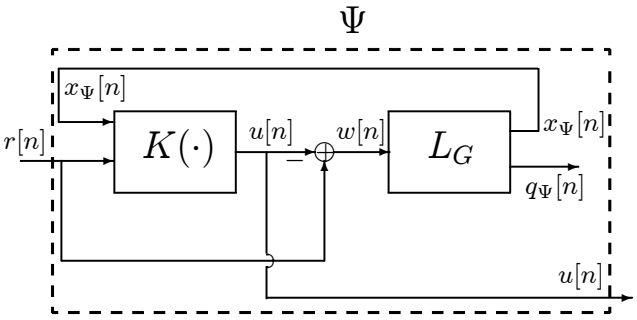

Fig. 5. Full State-Feedback Control Setup

A. The Bellman Inequality

The solution to a well-posed state-feedback optimal control problem can be characterized as the solution to the associated Bellman equation [19]-[22]. Herein, standard techniques will be used to show that a controller K such that (13) is achieved exists, if and only if the solution to an analog of the Bellman equation exists. The formulation will be made more precise as follows. Define function $\sigma : \mathbb{R}^m \to \mathbb{R}$ as

$$\sigma_\gamma(x) = \phi(q_\Psi) - \gamma \tag{14}$$

The control sequence $u[n]$ satisfying (11) results in an output sequence $q_\Psi[n]$ satisfying (13), if and only if there exists a function $V : \mathbb{R}^m \to \mathbb{R}_+$, such that the inequality

$$V(x) \geq \sigma_\gamma(x) + \sup_{r\in[-1,1]} \min_{u\in U} V(Ax + Br - Bu) \tag{15}$$

holds for all $x \in \mathbb{R}^m$ (Theorem 3). We refer to inequality (15) as the Bellman inequality.

B. Numerical Solutions to the Bellman Inequality

In this section we outline our approach for numerical computation of the value function V and the control function K. The numerical search for V satisfying (15) involves a finite-dimensional parameterization of V defined over a bounded, control-invariant subset of the state space. A control invariant set of system (10) is a subset $\mathcal{I} \subset \mathbb{R}^m$ such that:

$$\forall x \in \mathcal{I},\ \exists u \in U :\ Ax + Br - Bu \in \mathcal{I},\ \forall r \in [-1,1]. \tag{16}$$

At the first step, a bounded set $\mathcal{I}$ satisfying (16) is constructed. Then, uniform grids are created for both the state space and the input space. In this paper, these are uniformly-spaced, discrete subsets of the Euclidean space, and are defined precisely as follows. The set

$$\mathbb{G} = \{i\Delta \mid i \in \mathbb{Z}\}$$

is a grid on $\mathbb{R}$, where $D = 1/\Delta$ is a positive integer. The corresponding grid on $\mathcal{I} \times [-1,1]$ is $\Gamma \times \Gamma_r$, where

$$\begin{aligned} \Gamma &= \mathbb{G}^m \cap \mathcal{I}, \\ \Gamma_r &= \mathbb{G} \cap [-1,1]. \end{aligned}$$

The next step is to create a finite-dimensional parametrization of V. In this paper, the search is performed over the class of piecewise constant (PWC) functions assuming a constant value over a *tile*. A *tile* in $\mathbb{G}^n$, $n \in \mathbb{N}$ is defined as the smallest hypercube formed by 2^n points on the grid, and thus, has $2n$ faces (the faces are hypercubes of dimension $n-1$). By convention, we assume that the n faces that contain the lexicographically smallest vertex are closed, and the rest are open. The union of all such tiles covers $\mathbb{R}^n$ and their intersection is empty. Let T_i denote the i^{th} tile over the grid $\mathbb{G}^m$, and $\mathcal{T}$ the set of all tiles that fully lie within $\mathcal{I}$, and N_T the number of all such tiles:

$$\mathcal{T} = \{T_i \mid i \in \{1, 2, \cdots, N_T\}\}.$$

The PWC parametrization of V is as follows

$$V(x) = V_i,\ \forall x \in T_i,\ i \in \{1, 2, \cdots, N_T\} \tag{17}$$

where, $V_i \in \mathbb{R}_+$. Let R_j be the j^{th} tile in the input space:

$$R_j = [r_j, r_{j+1}),\ \forall j \in \{1, \cdots, 2D\} \tag{18}$$

where, $r_1 = -1$, and $r_j - r_{j-1} = \Delta, \forall j \in \{2, \cdots, 2D\}$. We then use (17) to find a function V satisfying the following discretized version of (15)

$$V(x) \geq \sigma_\gamma(x) + \sup_{r\in\Gamma_r} \min_{u\in U} V(Ax + Br - Bu),\ \forall x \in \Gamma. \tag{19}$$

The corresponding control function $K_\Gamma : \Gamma \times \Gamma_r \to U$ is given by

$$K_\Gamma(x, r) = \arg\min_{u\in U} V(Ax + Br - Bu),\ \forall x \in \Gamma,\ \forall r \in \Gamma_r. \tag{20}$$

Using the control function $K_\Gamma : \Gamma \times \Gamma_r \to U$ and (18), a piecewise constant function $K : \mathcal{T} \times [-1,1] \to U$ is constructed as follows

$$K(x, r) = K_\Gamma(l_i, r_j),\ \forall x \in T_i,\ \forall r \in R_j \tag{21}$$

where, l_i and r_j are the lexicographically smallest vertices of T_i and R_j respectively.

In Subsection III-C we show how to search for functions V and K_Γ satisfying (19) and (20), and in Subsection III-D we show how to verify that the candidate functions V and K (the piecewise constant extension of K_Γ defined by (21)) satisfy (15).

C. Searching for Numerical Solutions

The *discrete* Bellman inequality (19) is solved via value iteration. The algorithm is initialized at $\Lambda_0(x) = 0, \forall x \in \mathcal{T}$, and at stage $k+1$ it computes a piecewise constant function $\Lambda_{k+1} : \mathcal{T} \to \mathbb{R}_+$ satisfying

$$\Lambda_{k+1}(x) = \max\left\{0,\ \sigma_\gamma(x) + \sup_{r\in\Gamma_r} \min_{u\in U} \Lambda_k(Ax + Br - Bu)\right\}. \tag{22}$$

When the iteration converges, it converges pointwise to a limit $\Lambda : \mathcal{T} \to \mathbb{R}_+$, where the limit satisfies, for all $x \in \Gamma$, the equality

$$\Lambda(x) = \max\left\{0,\ \sigma_\gamma(x) + \sup_{r\in\Gamma_r} \min_{u\in U} \Lambda(Ax + Br - Bu)\right\} \tag{23}$$

The smallest γ for which (22) converges is found through line search. We take $V(x) = \Lambda(x)$, for all $x \in \mathcal{T}$. The optimal control law at all the grid points is given by (20), and its piecewise constant extension to $\mathcal{T}$ by (21).

D. Certifying Piecewise Constant Solutions

So far in this paper, we have shown how to obtain a piecewise constant value function along with the corresponding piecewise constant control law as candidate solutions to the Bellman inequality (15). In this section, we show how to formally certify validity of these candidate solutions and provide a proof that the obtained value function and the candidate control law satisfy (15). Define

$$\eta(r,x) = V(x) - \sigma_\gamma(x) - V(Ax + Br - BK(x,r)), \tag{24}$$

$$\eta(x) = \sup_{r\in[-1,1]} \eta(r,x). \tag{25}$$

If

$$\min_{x\in\mathcal{T}} \eta(x) \geq 0 \tag{26}$$

then V satisfies the Bellman inequality (15) for all $x \in \mathcal{T}$. The verification procedure is as follows:

1) Define
$$\begin{aligned} v_i &= V(x),\ x \in T_i,\ i \in \{1,2,\cdots,N_T\}, \\ \sigma_i &= \max_{x\in T_i} \sigma_\gamma(x). \end{aligned}$$
2) Define
$$\begin{aligned} u_{ij} &= K(x,r),\ x \in T_i,\ r \in R_j. \\ Y_{ij} &= \{Ax + Br - Bu_{ij} \mid x \in T_i,\ r \in R_j\}, \end{aligned}$$
and find all the tiles that intersect with Y_{ij}
$$\Theta_{ij} = \{T_k \mid T_k \cap Y_{ij} \neq \{\emptyset\},\ k \in \{1,2,\cdots,N_T\}\}.$$
3) Compute
$$\widetilde{v}_{ij} = \max_{s\in\Theta_{ij}} v_s$$
4) Verify that
$$v_i - \sigma_i - \widetilde{v}_{ij} \geq 0 \tag{27}$$
5) Repeat steps 2-4 for all input tiles $r \in R_j$, $j \in \{1,\cdots,2D\}$.
6) Repeat steps 1-5 for all tiles $x \in T_i$, $i \in \{1,2,\cdots,N_T\}$.
7) An arbitrarily close approximation for the smallest value of γ for which (27) is satisfied for all i, j, can be found through line search.

IV. Theoretical Statements

In this section, we present an exact analytical solution to the one-dimensional optimal ADC design problem. We show that the classical Delta Sigma Modulator (DSM) is optimal with respect to the performance criterion defined in this paper. Lastly, in subsection IV-B, we present a theorem that establishes the link between the full information feedback control problem and the Bellman inequality (15).

A. One-dimensional Optimal ADC

We analyze the scalar case for $|A| \leq 1$, and assume, without loss of generality, that $B = C = 1$ in (3)–(5). Let $\delta \in (0,\infty)$ and $M \in \mathbb{N}$ be such that $M\delta > 1$. Define the set U and function $K : \mathbb{R} \to U$ as

$$U = \{m\delta \mid m \in \mathbb{Z},\ |m| \leq M\} \tag{28}$$

$$K(\theta) = \min\left\{\mu : |\theta - \mu| = \min_{u\in U} |\theta - u|\right\} \tag{29}$$

We show in theorems 1 and 2 that the ADC $\Psi_o \in \mathcal{Y}_U$ defined as

$$x_\Psi[0] = 0 \tag{30}$$

$$x_\Psi[n+1] = Ax_\Psi[n] + r[n] - u[n] \tag{31}$$

$$u[n] = K(Ax_\Psi[n] + r[n]) \tag{32}$$

is optimal.

Theorem 1: Let $\widehat{\Psi} \in \mathcal{Y}_U$ be the ADC defined by (30)–(32) with $|A| = 1$ and K defined by (28)–(29), and let $\phi(\cdot) = f(|\cdot|)$ for some monotonically nondecreasing function $f : [0,\infty) \to [0,\infty)$. Then $\widehat{\Psi}$ is an optimal ADC in the sense that

$$\mathcal{J}_{G,\phi}(\Psi) \geq \mathcal{J}_{G,\phi}(\widehat{\Psi}) = \phi(\delta/2) \qquad \forall \Psi \in \mathcal{Y}_U. \tag{33}$$

Proof: Due to symmetry of the input and control sets, it is sufficient to present a proof for the case where $A = 1$. First, we show by induction that

$$\left|x_{\widehat{\Psi}}[n]\right| \leq \delta/2, \qquad \forall n \in \mathbb{Z}_+. \tag{34}$$

For $n = 0$, inequality (34) follows from (30). Assuming that (34) holds for $n = t$, we have

$$\left|x_{\widehat{\Psi}}[t] + r[t]\right| \leq (M + 0.5)\,\delta, \tag{35}$$

which follows immediately from $A = 1$, $|r[n]| \leq 1$ and $M\delta > 1$.
Define $\mathcal{F} = \{1-M,\ 2-M,\ \cdots,\ M-2,\ M-1\}$. The control law given in (29) can be equivalently expressed as:

$$K(\theta) = \begin{cases} -M\delta, & \theta \leq (-M+0.5)\,\delta \\ k\delta, & (k-0.5)\,\delta < \theta \leq (k+0.5)\,\delta,\ k \in \mathcal{F} \\ M\delta, & (M-0.5)\,\delta < \theta \end{cases} \tag{36}$$

Using (35) and (36), it can be verified that the control law can be expressed as $u[t] = k\delta$, where, integer $k \in [-M, M]$ is such that

$$(k-0.5)\,\delta < x_{\widehat{\Psi}}[t] + r[t] \leq (k+0.5)\,\delta.$$

Therefore,

$$-\delta/2 < x_{\widehat{\Psi}}[t] + r[t] - u[t] \leq \delta/2,$$

which is equivalent to (34) with $n = t+1$. Since both systems L_G and $L_{\widehat{\Psi}}$ have the same input and $x_{\widehat{\Psi}}[0] = x[0] = 0$, condition (34) implies that

$$\left|x_{\widehat{\Psi}}[n]\right| \leq \delta/2, \qquad \forall n \in \mathbb{Z}_+.$$

Therefore,

$$\sup_{N,r\in\Omega}\sum_{n=0}^{N}\left(\phi\left(x_{\widehat{\Psi}}\left[n\right]\right)-\phi(\delta/2)\right)\leq 0<\infty,$$

which implies that

$$\mathcal{J}_{G,\phi}(\widehat{\Psi})\leq\phi\left(\delta/2\right).$$

In order to complete the proof, we need to show that no ADC can achieve a better performance than $\phi\left(\delta/2\right)$. It is sufficient to show that for all $\Psi\in\mathcal{Y}_U$, there exist an input sequence r, and an integer T, such that

$$\left|x_{\Psi}\left[n+T\right]\right|\geq\delta/2,\qquad\forall n\in\mathbb{Z}_{+}. \tag{37}$$

Let function $\varphi:\mathbb{R}\rightarrow(-\delta/2,\delta/2]$ be defined according to:

$$\varphi\left(x\right)=y:\left((x-y)/\delta\right)\in\mathbb{Z}, \tag{38}$$

and denote $\widetilde{x}_{\Psi}=\varphi\left(x_{\Psi}\right)$. Applying $\varphi\left(\cdot\right)$ to both sides of (31) yields:

$$\begin{aligned}\widetilde{x}_{\Psi}\left[n+1\right] &= \varphi\left(x_{\Psi}\left[n\right]+r\left[n\right]+u\left[n\right]\right) &(39)\\ &= \varphi\left(\widetilde{x}_{\Psi}\left[n\right]+r\left[n\right]\right), &(40)\end{aligned}$$

where (40) follows from the definition of $\varphi\left(\cdot\right)$. The intuition behind defining $\varphi\left(\cdot\right)$ as in (38) is to obtain the dynamical system (40), which is independent of the control input. Since

$$\left|x_{\Psi}\left[n\right]\right|\geq\left|\widetilde{x}_{\Psi}\left[n\right]\right|, \tag{41}$$

inequality (37) will be satisfied if there exist r and T such that

$$\widetilde{x}_{\Psi}\left[n+T\right]=\delta/2,\qquad\forall n\in\mathbb{Z}_{+}. \tag{42}$$

Define

$$r\left[t\right]=\begin{cases}\dfrac{1}{T}\left(\dfrac{1}{2}\delta-\widetilde{x}_{\Psi}\left[n\right]\right), & \forall t\in[n,n+T)\\ 0, & \forall t\geq n+T\end{cases} \tag{43}$$

where $T\in\mathbb{Z}$ is such that

$$\left|r\left[t\right]\right|\leq 1,\ \forall t\in[n,n+T). \tag{44}$$

It then follows from (40) and (43), that (42) is satisfied. ■

Theorem 2: Let $\delta\leq 2$, and $\widehat{\Psi}\in\mathcal{Y}_U$ be the ADC defined by (30)−(32) with $|A|<1$, and K and $\phi\left(\cdot\right)$ defined as in Theorem 1. Then, $\widehat{\Psi}$ is an optimal ADC in the sense that (33) holds.

Proof: Omitted for brevity. ■

Remark 1: Since the input is bounded in magnitude by one, the assumption $\delta\leq 2$ is reasonable.

1) Optimality of the Classical DSM: Figure 6 shows the feedback interconnection for the optimal ADC defined by (30)−(32) with $A=B=C=1$, and the optimal control function K defined by (28)−(29). The transfer functions from r to y, denoted by $H_{ry}\left(z\right)$ in our design and by H_{ry}^{DSM} in the classical DSM (Figure 1) are given by

$$H_{ry}\left(z\right)=H_{ry}^{\text{DSM}}\left(z\right)=\frac{z}{z-1}.$$

Similarly, the transfer functions from u to y in both designs are given by:

$$H_{uy}\left(z\right)=H_{uy}^{\text{DSM}}\left(z\right)=\frac{-1}{z-1}.$$

Furthermore, the optimal control law defined in (29), is essentially a quantizer with uniform step size. Therefore, the classical one-dimensional DSM with a uniform quantizer is identical to our design, and hence, is optimal with respect to the performance criterion adopted in this paper.

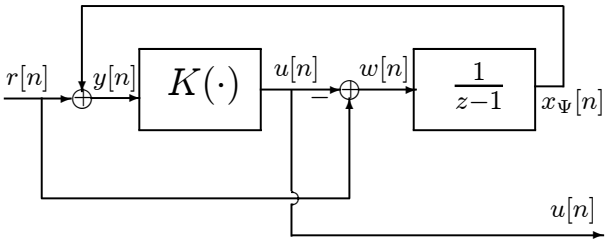

Fig. 6. Optimal one-dimensional ADC

B. Bellman Inequality Theorem

Theorem 3: Let $U\subset\mathbb{R}$ be a finite set, and function $\sigma_{\gamma}:\mathbb{R}^{m}\rightarrow\mathbb{R}$ be such that

$$\inf_{x\in\mathbb{R}^{m}}\sigma_{\gamma}\left(x\right)=-\gamma>-\infty. \tag{45}$$

Then for every $\gamma\in[0,\infty)$, the following statements are equivalent:

(i) There exists a function $\alpha:\mathbb{R}^{m}\rightarrow\mathbb{R}_{+}$ and a sequence of functions $F_{n}:\left[-1,1\right]^{n+1}\times\mathbb{R}^{m}\rightarrow U$, such that the inequality

$$\sum_{n=0}^{N}\sigma_{\gamma}\left(x\left[n\right]\right)\leq\alpha\left(x[0]\right) \tag{46}$$

holds for all N and all functions $x:\mathbb{Z}_{+}\rightarrow\mathbb{R}^{m}$, $r:\mathbb{Z}_{+}\rightarrow[-1,1]$, and $u:\mathbb{Z}_{+}\rightarrow U$ satisfying (10), (12) and

$$u[n]=F_{n}\left(r\left[n\right],\cdots,r\left[0\right],x\left[0\right]\right).$$

(ii) There exists a function $V:\mathbb{R}^{m}\rightarrow\mathbb{R}_{+}$ such that (15) holds for every $x\in\mathbb{R}^{m}$.

(iii) The sequence of functions $\Lambda_{k}:\mathbb{R}^{m}\rightarrow\mathbb{R}_{+}$ defined by $\Lambda_{0}\left(x\right)\equiv 0$,

$$\Lambda_{k+1}\left(x\right)=\max\left\{0,\ \sigma_{\gamma}\left(x\right)+\sup_{r\in[-1,1]}\min_{u\in U}\Lambda_{k}\left(Ax+Br-Bu\right)\right\}$$

converges pointwise to a limit $\Lambda:\mathbb{R}^{m}\rightarrow\mathbb{R}_{+}$.

Moreover, when (i)−(iii) hold, $\Lambda\left(\cdot\right)$ satisfies

$$\Lambda\left(x\right)=\max\left\{0,\ \sigma_{\gamma}\left(x\right)+\sup_{r\in[-1,1]}\min_{u\in U}\Lambda\left(Ax+Br-Bu\right)\right\}$$

and there exists a function $K:\mathbb{R}^{m}\times[-1,1]\rightarrow U$ such that every triplet of sequences (x,u,r) satisfying (10)−(12) also satisfies (46).

Proof: Proof omitted for brevity. ■

V. Numerical Example

Let matrices A, B,and C of the dynamical system L_G (3)−(5) be given by

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

Let $U = \{-1.5, 0, 1.5\}$ and $\sigma_\gamma = |Cx| - \gamma$. The control invariant set $\mathcal{I}$ is selected according to the following procedure. The eigenvalues of A are $\lambda_0 = 0$ and $\lambda_1 = 1$, with corresponding left eigenvectors $p_0 = \begin{bmatrix} 1 & -1 \end{bmatrix}$ and $p_1 = \begin{bmatrix} 1 & 0 \end{bmatrix}$. Multiplying both sides of (10) by p_0 yields

$$\begin{aligned} p_0 x_\Psi[n+1] &= \lambda_0 p_0 x_\Psi[n] + p_0 B(r[n] - u[n]) \\ &= r[n] - u[n]. \end{aligned}$$

Hence,

$$|p_0 x_\Psi| \le 2.5,\ \forall r \in [-1, 1],\ \forall u \in U. \tag{47}$$

Similarly, in the direction of the eigenvector corresponding to the pole on the unit circle we have:

$$\begin{aligned} p_1 x_\Psi[n+1] &= \lambda_1 p_1 x_\Psi[n] + p_1 B(r[n] - u[n]) \\ &= p_1 x_\Psi[n] + r[n] - u[n]. \end{aligned}$$

In this case, an invariant strip of the form (47) does not exist, however, it can be verified that for all $r \in [-1, 1]$, there exists a control $u \in U$, such that $|p_1 x_\Psi| \le 0.75$. However, this bound is too restrictive on the values control can take and it must be increased to allow for the optimal control action to be found by the method. We found by trial and error that the following invariant set is adequately large:

$$\mathcal{I} = \{x_\Psi \mid |p_1 x_\Psi| \le 3,\ |p_0 x_\Psi| \le 2.5\}.$$

Next, a grid spacing of $\Delta = 1/64$ was selected. Following the procedures outlined in subsections III-B and III-C, the smallest γ for which the iteration in (22) converges to the limit Λ in (23), is $\gamma = 1$. For this example the iteration converges in 10 steps. Finally the design is verified using the method outlined in subsection III-D, which gives the upper bound on the performance: $\gamma_{G,\phi}(U) \le 1.274$.

VI. Conclusion

The problem of design of optimal Analog to Digital Converters (ADC) was considered. The optimal ADC design problem was associated with a full information feedback optimal control problem. An algorithm was presented for numerical computation of solutions to the underlying Bellman inequality, followed by computation—via post-design verification—of an upper bound on the quality of the design. Finally, an exact analytical solution to the optimal one-dimensional ADC was presented and it was shown that the optimal design is identical to the classical Delta-Sigma Modulator (DSM) with uniform quantization.

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, Sept. 1998.

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

[6] S. R. Norsworthy, R. Schreier, and G. C. Themes, Eds., *Delta-Sigma Data Converters: Theory, Design,and Simulation*. IEEE Press, 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, July 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] Y. Wang and S. Boyd, "Performance bounds for linear stochastic control," *Systems and Control Letters*, vol. 58, no. 3, pp. 178 – 182, 2009.

[14] ——, "Performance bounds and suboptimal policies for linear stochastic control via lmis," 2009. [Online]. Available: http://www.stanford.edu/~boyd/papers/gen_ctrl_bnds.html

[15] F. Bullo and D. Liberzon, "Quantized control via locational optimization," *IEEE Transactions on Automatic Control*, vol. 51, no. 1, pp. 2 – 13, jan. 2006.

[16] R. Brockett and D. Liberzon, "Quantized feedback stabilization of linear systems," *IEEE Transactions on Automatic Control*, vol. 45, no. 7, pp. 1279 –1289, July 2000.

[17] N. Elia and S. Mitter, "Stabilization of linear systems with limited information," *IEEE Transactions on Automatic Control*, vol. 46, no. 9, pp. 1384 –1400, sep 2001.

[18] S.-i. Azuma and T. Sugie, "Synthesis of optimal dynamic quantizers for discrete-valued input control," *IEEE Transactions on Automatic Control*, vol. 53, no. 9, pp. 2064 –2075, oct. 2008.

[19] J. W. Helton and M. R. James, *Extending H-infinity Control to Nonlinear Systems: Control of Nonlinear Systems to Achieve Performance Objectives*. Society for Industrial Mathematics (SIAM), 1999.

[20] D. P. Bertsekas, *Dynamic Programming and Optimal Control, Vol. I*. Athena Scientific, 2005.

[21] ——, *Dynamic Programming and Optimal Control, Vol. II*. Athena Scientific, 2007.

[22] R. Bellman, *Dynamic Programming*. Dover Publications, 2003.

[23] 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.

[24] 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, July 2007.