Noise-shaping Filter Synthesis for Moving Horizon Optimal Quantisation

Bikash Adhikari[1], John Leth[2] and Arnfinn A. Eielsen[1]

***Abstract*— This paper presents a novel approach for improving the performance of digital-to-analog converters (DACs) using moving-horizon optimal quantisation (MHOQ) with optimal noise-shaping filters and digital calibration. DACs have reduced performance due to non-linearities; principally quantisation and element mismatch. The proposed method integrates an optimal noise-shaping filter for a given reconstruction filter with MHOQ to minimise quantisation error at the filter output. In addition, by including a digital calibration model of the measured element mismatch, the modified MHOQ is able to mitigate the distortion caused by this effect. Simulation results are provided that demonstrate that the approach significantly improves the signal-to-noise-and-distortion ratio (SINAD).**

I. Introduction

In digital signal processing systems, the conversion between physical signals and their digital representations relies on two key components: analogue-to-digital converters (ADCs) and digital-to-analogue converters (DACs). These are enabling technologies with a wide range of applications, including digital audio and video recording [14], adaptive optics [4], interferometry [1], scanning probe microscopy [11], lithography systems [6], and metrology [19].

ADCs convert analogue signals to a digital representation by time-sampling and value-quantisation — and DACs reverse this process. Due to the Nyquist–Shannon sampling theorem [27], anti-aliasing and reconstruction filters are required to reduce artifacts due to time-sampling. Value-quantisation can result in information loss, contributing errors in the reconstructed output signal. The reconstruction filter also helps to mitigate the effect of quantisation error in DACs. Another significant source of error in the reconstructed signal stems from element mismatch, due to imperfections in the physical implementation. This introduces another static non-linear effect, typically represented by a function called the integral non-linearity (INL).

Since DACs are widely used in high-precision applications, improving their performance becomes increasingly important. A promising method for reducing the effect of quantisation error is moving horizon optimal quantisation (MHOQ) [16], which formulates the quantisation problem as a quadratic optimal control problem. MHOQ applies feedback and control input constrained to a finite set using the moving horizon principle to minimise the error. This method significantly reduces the quantisation error, but the assumption of uniform quantisation is quite limited as all practical DACs exhibit element mismatch. The MHOQ method is extended to include a model of the element mismatch in [2]. Since element mismatch is a large source of error, its inclusion into the model provides significant performance gains.

Another method that is closely related to the MHOQ method is noise-shaping quantisation (NSQ) [26]. Using over-sampling and a feedback filter, or noise-shaping filter (NSF), the spectral distribution of the quantisation error can be controlled, constrained by the Bode sensitivity integral and an empirical limit on feedback gain, referred to as Lee's rule [26], [24]. The gain limit is due to the discontinuous nature of the quantiser, and stability can only be determined in a few special cases [15]. The typical use is to shift the error power frequencies higher than the bandwidth of the desired signal; to be attenuated by the reconstruction filter. Techniques like noise-shaping with digital calibration (NSQD) [8] address element mismatch by incorporating a model that consists of measured quantisation levels.

The authors in [16], [3] have shown that the NSQ and MHOQ with prediction horizon $N = 1$ behave identically for uniform quantisation and provide a relation between the NSF and reconstruction filter. This equivalence indicates that MHOQ inherently performs noise-shaping while determining the control signal that minimise quantisation error power after the reconstruction filter. Simulation results using a psychoacoustically optimal noise-shaping filter [29] show significant performance improvement. However, further simulations with a other noise-shaping filters reveal that the performance of MHOQ is highly dependent on the choice of this filter. This highlights the need for a method to consistently find the best performing filter.

In this paper, we propose an MHOQ method that achieves the performance level obtained in [16] for a general class of filters by optimising the noise-shaping filter for the given reconstruction filter. For the given reconstruction filter, we first design an optimal noise-shaping filter that minimises the error power, or variance, using noise-shaping quantisation via the method from [25]. A filter is then derived from the optimal noise-shaping filter to be used in MHOQ. Additionally, the measured element mismatch is incorporated into the quantiser model as in [2], leading to improved DAC output estimates that further enhances the performance.

The design of optimal noise-shaping filter that minimise the variance of the error in the signal of interest has been studied in [7], [25]. A crucial factor in designing a stable

*This work was supported by Research Council of Norway, project FRIPRO 313716.

[1]Bikash Adhikari and Arnfinn A. Eielsen are with Department of Energy and Petroleum Resources, University of Stavanger, Stavanger, Norway firstname.lastname@uis.no

[2]John Leth is with the Department of Electrical Engineering, Aalborg University, Aalborg, Denmark jjl@es.aau.dk

NSQ is limiting the magnitude of the feedback signal [26], [24]. Consequently, the optimisation problem is formulated as minimising the error variance while imposing a constraint on the norm of the feedback signal. The $\mathcal{H}_2$ and $\mathcal{H}_\infty-$ norms are used to capture the error effects, which characterises the variance and maximum absolute value of the error, respectively. These norms are commonly represented in the form of bilinear matrix inequalities (BMIs) which are non-convex and NP-hard to solve. To address this challenge, the BMIs are transformed into linear matrix inequalities (LMIs) through a change of variables [22]. Since, the LMIs are convex and they can be solved using solvers such as SDPT3 [28] or MOSEK [23] via YALMIP [21] or CVX [17].

One might question the necessity of applying MHOQ after NSQ, given that both cases are identical for a prediction horizon of $N = 1$ and that increasing the horizon does not provide substantial performance improvements [2]. However, it is important to note that model-based methods like MHOQ can capture the dynamics of DACs in ways that noise-shaping quantisation cannot. Other sources of error, such as glitches [12] and slewing effects [13], adversely impact the performance of DACs. This approach serves as a stepping stone for modelling these non-idealities for further performance gains.

The paper is organised as follows: Sec. II outlines the quantisation and error modelling. Secs. III and IV introduce moving horizon optimal quantisation and noise-shaping quantisation, respectively. Sec. V provides the method to synthesise an optimal noise-shaping filter. Sec. VI present the implementation method. Sec. VII present simulation results and Sec. VIII concludes the paper.

A. Notations

The set of the real (symmetric) matrices of dimension $m \times n$ $(n \times n)$ is denoted by $\mathbb{R}^{m\times n}$ $(\mathbb{S}^n)$. The transpose of the matrix X is denoted as $X^\top$. For a symmetric matrix X, positive definiteness is indicated by $X \succ 0$.

II. Quantisation and Error Modelling

A. Uniform quantisation

Let $w = w(t)$ be the input signal, $y = y(t)$ be the quantised output signal, and $\mathbf{Q}$ represent the quantisation operation performed by the quantiser. Moreover, let δ be the quantiser step size. An ideal DAC is modelled by $\mathbf{Q}$, and δ corresponds to the least significant bit (LSB), as DACs typically use binary encoding. For a word-size $\mathcal{B}$ (number of bits), a DAC has $2^\mathcal{B}$ quantisation levels and the step-size is determined by $\delta = \Delta/(2^\mathcal{B} - 1)$, where Δ is the full-scale output range of the DAC. A higher number of levels translates to a smaller quantisation error, but word-size is limited by several practical aspects. The uniform quantiser is defined as having equidistant levels and a mid-tread quantiser $\mathbf{Q}$ can be defined using the truncating operator $\ell = T(w)$ as

$$y = \mathbf{Q}(w) = \delta\ell = \delta T(w) := \delta\left\lfloor \frac{w}{\delta} + \frac{1}{2} \right\rfloor, \tag{1}$$

where $\lfloor\cdot\rfloor$ denotes the floor operator and ℓ is referred to as the level (corresponding to the input w). In the sequel, operation performed by the quantiser $\mathbf{Q}$ will be referred as direct quantisation (DQ). The quantisation of a signal introduces quantisation error, defined as

$$q = \mathbf{Q}(w) - w \tag{2}$$

which is always constrained to be within $\pm\delta/2$. Typically, due to the non-linear behaviour of quantiser due to the truncation operator as shown in equation (1), quantisation is modelled as an additive, zero-mean, and uniformly distributed white noise signal [5], independent of w as follows,

$$y = w + q \tag{3}$$

where w is the input signal and q is the quantisation error.

B. Non-linear quantiser

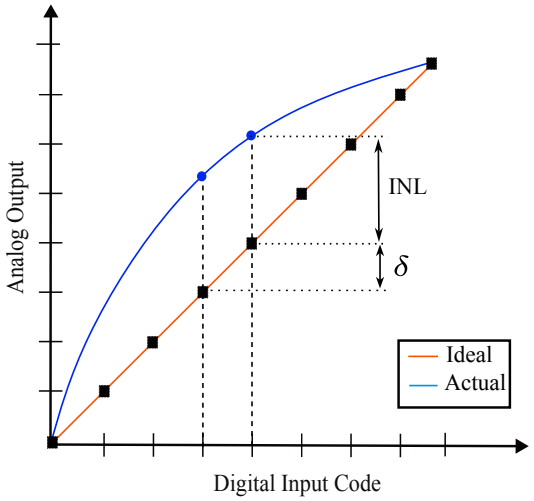

Fig. 1. DAC transfer function: Ideal and non-ideal (actual).

Practical DACs have element mismatch that causes the actual levels to deviate from the ideal, equidistant levels of the uniform quantiser, as illustrated in Fig. 1. This is known as integral non-linearity (INL). Let y be the ideal and $\tilde{y}$ be the actual level of a DAC, then $\mathrm{INL}(\ell)$ represent the deviation at the level $\ell \in \mathbb{N}$ (a static non-linearity), that is,

$$\tilde{y}(\ell) = y(\ell) + \delta\mathrm{INL}(\ell) = \delta\ell + \delta\mathrm{INL}(\ell). \tag{4}$$

The integral non-linearity $\mathrm{INL}(\ell)$, can be represented as

$$\mathrm{INL}(\ell) := \frac{\tilde{y}(\ell) - \delta\ell}{\delta}. \tag{5}$$

The effect of the non-linearity $n(w)$ on the output y due to the input w can be defined as

$$n(w) = \delta\mathrm{INL}(\ell)|_{\ell=T(w)} = \tilde{y}(T(w)) - \delta T(w). \tag{6}$$

Note that the non-linearity $n(w)$ is a discontinuous function due to the truncation operator.

The performance of the DAC is affected by quantisation errors and static non-linearity, among other factors. Therefore, it is essential to mitigate the effects of these

behaviours to enhance DACs performance. The MHOQ is one approach that mitigates the effect of quantisation noise [16] and INL [2] using an optimisation-based method, which will be discussed in the following sections.

III. Moving Horizon Optimal Quantisation

In the moving horizon optimal quantiser, the quantisation problem is cast into the multi-horizon optimisation setting. As the quantiser itself lacks dynamics, the filter is incorporated into the model as shown in Fig. 2, to represent the quantisation problem using optimisation framework.

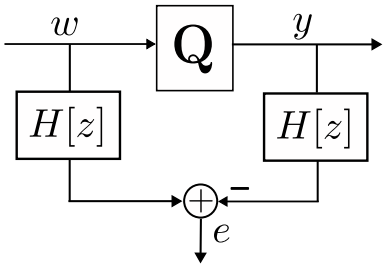

Fig. 2. Incorporation of filter $H[z]$

Consider a $n-$th order stable time-invariant linear filter

$$H[z] = D_h + C_h(zI - A_h)^{-1}B_h \tag{7}$$

where the matrices $A_h \in \mathbb{R}^{n\times n}$, $B_h \in \mathbb{R}^{n\times 1}$, $C_h \in \mathbb{R}^{1\times n}$, and $D_h \in \mathbb{R}$ are related to the impulse response $h(i) = D_h$ for $i = 0$ and $h(i) = C_h A_h^{i-1} B_h$ for $i > 0$ of the filter. Then, the system setup in Fig. 2 leads to a system where the quantised output y is the input to the filter $H[z]$ and the output e is the filtered difference between the input signal w and y. The system can then be represented in state-space form as follows,

$$x(t+1) = A_h x(t) + B_h(y(t) - w(t)) \tag{8}$$

$$e(t) = C_h x(t) + D_h(y(t) - w(t)). \tag{9}$$

Since w is a given reference signal, the quantised output y is here considered as a control input and restricted to belong to the finite set

$$\mathbb{U} = \{y_1, y_2, \ldots, y_{n_\mathbf{U}}\} \tag{10}$$

with $y_i < y_{i+1} \in \mathbb{R}$ and $i = 1, 2, \ldots n_\mathbf{U} - 1$. The optimal control problem is then set up as finding the optimal values of y while minimising the given performance criterion.

Next, let us define the performance criterion at $t = k$ as a quadratic cost,

$$V_N = \sum_{t=k}^{k+N-1} e^2(t) \tag{11}$$

where $e(t)$ is the error defined in (9). The cost function depends on N future values of the control input $y(t)$. For the given state $x(k)$, we seek the optimising sequence of the present and future values of the control input as

$$\mathbf{y}^*(x) := \arg\min_{\mathbf{y}(k)\in\mathbb{U}^N} V_N(x, \mathbf{y}(k)) \tag{12}$$

where

$$\mathbf{y}(k) = \begin{bmatrix} y(k) \\ y(k+1) \\ \cdots \\ y(k+N-1) \end{bmatrix}, \quad \mathbb{U}^N := \mathbb{U} \times \ldots \times \mathbb{U}.$$

Thus the problem of finding the optimal quantisation level is re-cast into a moving horizon optimal quantisation problem where the input to the reconstruction filter has to be chosen from the finite constrained set $\mathbb{U}$. In other words, the objective is to minimise the quadratic cost function (11) such that input belongs to the finite constrained set (10) and satisfies the dynamics (8)-(9).

In order to address the effect of the INL in the signal of interest, we can incorporate INL into the MHOQ, as demonstrated in [2]. INL, which represents the deviation of the actual levels from the ideal ones, is static in nature and can be measured in advance. This data can then be organized into a lookup table that corresponds to the ideal quantisation levels. Then in moving horizon implementation, we optimise over the finite horizon of length N to obtain the optimal values $\mathbf{y}^*(k) = [y^*(k), \ldots, y^*(k+N-1)]^\top$. Then we add the INL corresponding to each level from the lookup table to obtain the actual values that are $\tilde{y}(k), \ldots, \tilde{y}(k+N-1)$. That is,

$$\tilde{y}(k) = y^*(k) + \delta\mathrm{INL}(y^*(k)). \tag{13}$$

As with the moving horizon implementation, we apply the first control $\tilde{y}(k)$ and move the optimisation to the next time horizon (see Fig. 5).

Although the MHOQ method provides significant performance improvements, it has limitations regarding the types of filters that can be used as plant models and get the desired performance improvements. In [16], a psychoacoustically optimal noise-shaping filter [29] is employed to validate the effectiveness of the MHOQ method. However, this approach does not yield similar improvements for all filter choices which is shown in the simulation results.

Next, we propose a method that uses the knowledge of the reconstruction filter $H[z]$ to design the optimal noise-shaping filter [26] based on the result of [25]. But first, let us introduce a noise-shaping quantisation method and its relation with the MHOQ.

IV. Noise-shaping Quantiser

Noise-shaping quantisers can reduce the effective quantisation error by moving quantisation noise to higher frequencies through oversampling and feedback. The reconstruction filter is then used to attenuate the frequency-shaped quantisation noise. It operates by estimating the uniform quantisation error and employing a feedback filter to shape the noise power at the output of the DAC. A block diagram for a noise-shaping quantiser is shown in Fig. 3 where the feedback filter $F[z]$ is called as noise-shaping filter (NSF).

The non-linear behaviour of the quantiser makes the analysis of NSQ difficult to tackle and the common approach involves approximation by the linearised architecture, i.e.,

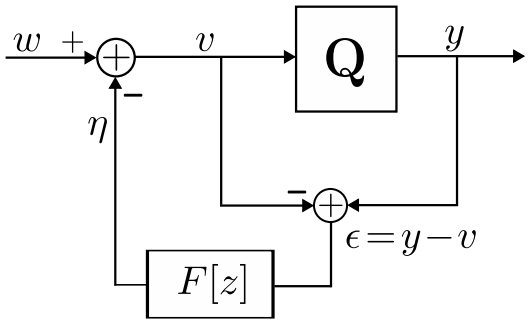

Fig. 3. Noise-shaping quantiser

the additive model (3). Using the additive model allows us to decompose the output of the NSQ in Fig. 3 as follows,

$$y = w + (1 - F[z])\epsilon. \tag{14}$$

Ideally, we want the output of the NSQ to be unaltered and identical to the input, however the output in (14) shows it is effected by the quantisation noise. The term $(1 - F[z])$ is know as the noise-transfer function (NTF) and we define it as follows,

$$R[z] := 1 - F[z]. \tag{15}$$

The noise-shaping filter $F[z]$ is typically choosen such that the noise transfer function $R[z]$ is a high-pass filter, which shifts noise to higher frequencies outside the signal bandwidth. This enables the noise to be effectively removed by a low-pass filter.

The MHOQ also perform a noise-shaping in obtaining the optimal quantisation by optimally switching between the quantisation levels. Moreover, the MHOQ with prediction horizon $N = 1$ and the noise-shaping quantisation behave identically in the uniform quantisation case. It is shown in the moving horizon implementation of the quantisation [16], that the filter $H[z]$ and the noise-shaping filter $F[z]$ are related as follows,

$$F[z] = \frac{H[z] - 1}{H[z]}. \tag{16}$$

Next, we present a method to synthesize an optimal noise-shaping filter $F[z]$ for the filter $H[z]$ based on the method from [25]. Then, using the relationship (16), we obtain the modified filter and use it in the MHOQ implementation.

V. Synthesis of Optimal Noise-shaping Filter

In this section, we present a method to synthesise an optimal noise-shaping filter for a given reconstruction filter $H[z]$ which mitigates the effect of quantisation noise at the output. The objective is to design an optimal noise-shaping function that minimises the variance of the filtered quantisation error e in the output. However, the design criteria impose constraints on the NTF i.e., $R[z]$ to ensure that the NSQ operation achieves optimal performance. Therefore, we reconfigure the NSQ using the $R[z]$ along with the reconstruction filter $H[z]$, as illustrated in the Fig. 4.

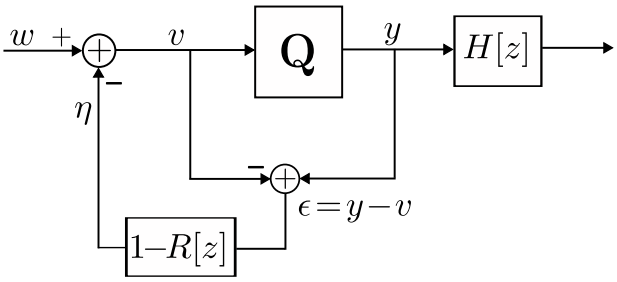

Fig. 4. Noise-shaping quantiser with reconstruction filter $H[z]$.

Then from equations (14) and (15), the NSQ output is $y = w + R[z]\epsilon$ and from Fig. 4 the filtered output using the reconstruction filter $H[z]$ is

$$H[z]y = H[z]w + H[z]R[z]\epsilon. \tag{17}$$

Let us denote the error between the filtered input and the quantized output signal by e, defined as follows:

$$e := H[z](y - w). \tag{18}$$

Then, from equations (17) and (18), we have,

$$e = H[z]R[z]\epsilon. \tag{19}$$

The equation (19) shows that the error in the output signal due to quantisation noise can be reduced by designing the filter $R[z]$ with the knowledge of filter $H[z]$. Recall that the e in (18) is identical to the error defined in (9) in Sec. III.

If the quantisation error ϵ is modelled as a uniform random variable with zero mean and variance σ_ϵ^2 and the error is i.i.d., then the variance of the error at the output is

$$\sigma_e^2 = \|H[z]R[z]\|_2^2 \sigma_\epsilon^2. \tag{20}$$

Moreover, the practical realisation of the NSQ requires at least one clock-period delay which implies that $R[z]$ is proper and rational. Furthermore, in order to prevent the overloading of the quantiser and render it unstable, the practical designs tend to rely on an empirical rule based on [9], [26] that limits the gain of the NTF as follows,

$$\|R[z]\|_\infty < \gamma_\eta. \tag{21}$$

The Lee criterion requires the value of the constant γ_η to be less than 2 and is typically set at 1.5 for practical purpose [9]. Thus, the objective is to design an optimal noise-transfer function $R[z]$ that minimise the variance of the error at output (20) while satisfying the constraint (21).

A. Optimisation Problem

The optimisation problem is set as minimising the upper bound of the error $\|e\|_2$ subject to the constraint on the noise transfer function $\|R[z]\|_\infty$ as follows,

$$\min_{R[z] \in \mathbb{RH}_\infty} \gamma_e \tag{22}$$

subject to

$$\|H[z]R[z]\|_2 < \gamma_e \tag{23}$$

$$\|R[z]\|_\infty < \gamma_\eta \tag{24}$$

where $\mathbb{RH}_\infty$ is the set of proper stable rational transfer functions. This setup allows us to represent the constraints in the optimisation problem (22)-(24) using bilinear matrix inequalities (BMIs). These BMIs can then be converted to linear matrix inequalities (LMIs) using a change of variables. First, recall the state space representation of $H[z]$ from Sec. III and let the state-space representation $R[z]$ be denoted by $(A_r, B_r, C_r, 1)$, respectively. Let us define $H_R := H[z]R[z]$ and in the sequel we only use H_R to represent the composite system. Then, the state-space realization of H_R is

$$H_R : \begin{cases} x(k+1) & = Ax(k) + B\epsilon(k) \\ e(k) & = Cx(k) + D\epsilon(k) \end{cases} \tag{25}$$

where

$$A = \begin{bmatrix} A_h & B_hC_r \\ \mathbf{0} & A_r \end{bmatrix} \in \mathbb{R}^{2n\times 2n},\ B = \begin{bmatrix} B_h \\ B_r \end{bmatrix} \in \mathbb{R}^{2n\times 1}, \tag{26}$$

$$C = \begin{bmatrix} C_h & D_hC_r \end{bmatrix} \in \mathbb{R}^{1\times 2n},\ D = D_h \in \mathbb{R}. \tag{27}$$

where $x \in \mathbb{R}^n$.

From (20) and (25), we see that the variance of the error e is given in terms of the $\mathcal{H}_2$-norm of the system H_R. If A is Schur (stable), then there exists a positive semi-definite matrix $P \in \mathbb{S}^{2n}$ such at

$$P = A^\top PA + BB^\top \tag{28}$$

and the $\mathcal{H}_2$-norm can be calculated as follows,

$$\|H_R\|_2^2 = CPC^\top + DD^\top. \tag{29}$$

Moreover, using (28) and (29), the $\mathcal{H}_2$ performance of the systems H_R can be characterised using the following Lemma.

Lemma 1 ($\mathcal{H}_2 - \mathbf{Performance}$): $\|H_R\|_2 < \gamma_e$ if and only if there exists a matrix $P \in \mathbb{S}^{2n}$ such that

$$\begin{bmatrix} P & PA & PB \\ A^\top P & P & \mathbf{0} \\ B^\top P & \mathbf{0} & \mathbf{1} \end{bmatrix} \succ 0, \tag{30}$$

$$\begin{bmatrix} \mu_e & C & D \\ C^\top & P & \mathbf{0} \\ D^\top & \mathbf{0} & \mathbf{1} \end{bmatrix} \succ 0, \quad \mu_e = \gamma_e^2. \tag{31}$$

The $\mathcal{H}_2$ condition requires the matrix A to be Schur since the fundamental Lyapunov inequality appears in the matrix inequality in (30).

Lemma 2 ($\mathcal{H}_\infty - \mathbf{Performance}$): $\|H_R\|_\infty < \gamma_e$ if and only if there exists a matrix $P \in \mathbb{S}^{2n}$ such that

$$\begin{bmatrix} P & PA & PB & \mathbf{0} \\ A^\top P & P & \mathbf{0} & C^\top \\ B^\top P & \mathbf{0} & \mu_e & D^\top \\ \mathbf{0} & C & D & \mu_e \end{bmatrix} \succ 0, \qquad \mu_e = \gamma_e^2. \tag{32}$$

If we set $C_h = 0$ and $D_h = 1$, we get, $H[z] = 1$. Thus the constraint (24) is satisfied if and only if there exists a positive definite matrix P such that [25]

$$\begin{bmatrix} P & PA & PB & \mathbf{0} \\ A^\top P & P & \mathbf{0} & \tilde{C}^\top \\ B^\top P & \mathbf{0} & \mu_\eta & \mathbf{1} \\ \mathbf{0} & \tilde{C} & \mathbf{1} & \mu_\eta \end{bmatrix} \succ 0 \tag{33}$$

where $\tilde{C} = \begin{bmatrix} 0 & C_r \end{bmatrix}$ and $\mu_\eta = \gamma_\eta^2$.

The matrix inequalities (30) and (33) are bilinear matrix inequalities (BMI) due to the product between the system matrix P with A and B, respectively. BMIs are not convex and NP-hard to solve, but they can be converted to convex LMIs and can be solved numerically. The non-convex BMIs can be converted to convex LMIs using a change of variables [22] as follows:

1) Change of Variables:: Let the order of $H[z]$ be n and the set of $n\times n$ positive define matrices be denoted by $\mathrm{PD}(n)$. Denote by $\mathcal{P}$ the set of variables $\mathcal{P} = \{P_f, P_g, W_f, W_g, L\}$ where $P_f \in \mathrm{PD}(n)$, $P_g \in \mathrm{PD}(n)$, $W_f \in \mathbb{R}^{1\times n}$, $W_g \in \mathbb{R}^{n\times 1}$, and $L \in \mathbb{R}^{n\times n}$. Then define the following matrix values function on $\mathcal{P}$:

$$\begin{aligned} M_A &:= \begin{bmatrix} A_hP_f + B_hW_f & A_h \\ L & P_gA_h \end{bmatrix}, \qquad M_B := \begin{bmatrix} B_h \\ W_g \end{bmatrix} \\ M_C &:= \begin{bmatrix} C_hP_f + D_hW_f & C_h \end{bmatrix}, \qquad M_P := \begin{bmatrix} P_f & I_n \\ I_n & P_g \end{bmatrix}. \end{aligned} \tag{34}$$

Next, define

$$P^{-1} := \begin{bmatrix} P_f & S_f \\ S_f & S_f \end{bmatrix},\ T := \begin{bmatrix} P_f & I_n \\ S_f & \mathbf{0} \end{bmatrix},\ S_f := P_f - P_g^{-1} (\succ 0) \tag{35}$$

then we have,

$$M_P = T^\top PT. \tag{36}$$

If the matrices (A_r, B_r, C_r) are given by

$$\begin{aligned} A_r &:= \left[B_hW_f - P_g^{-1}(L - P_gA_hP_f)\right] S_f^{-1} \\ B_r &:= \left[B_h - P_g^{-1}W_g\right] \\ C_r &:= W_fS_f^{-1} \end{aligned} \tag{37}$$

then (A, B, C) satisfy,

$$M_A = T^\top PAT, \quad M_B = T^\top PB, \tag{38}$$

$$M_C = CT, \quad M_P = T^\top PT \quad \text{and} \quad M_{\tilde{C}} = \tilde{C}T. \tag{39}$$

Now, we multiply (30) with the transformation matrix $\mathrm{diag}(T, T, I)$ on the right and its transpose on the left, we multiply (31) by $\mathrm{diag}(I, T, I)$ on the right and its transpose on the left and we multiply (33) by $\mathrm{diag}(T, T, I, I)$ on the right and by its transpose on the left. Then, we obtain the following optimization problem equivalent to (22)-(24)

$$\min_{R[z]\in\mathbb{RH}_\infty} \gamma_e \tag{40}$$

subject to

$$\begin{bmatrix} M_P & M_A & M_B \\ M_A^\top & M_P & \mathbf{0} \\ M_B^\top & \mathbf{0} & \mathbf{I} \end{bmatrix} \succ 0, \tag{41}$$

$$\begin{bmatrix} \mu_e & M_C & D \\ M_C^\top & M_P & \mathbf{0} \\ D^\top & \mathbf{0} & \mathbf{I} \end{bmatrix} \succ 0, \tag{42}$$

$$\begin{bmatrix} M_P & M_A & M_B & \mathbf{0} \\ M_A^\top & M_P & \mathbf{0} & M_{\tilde{C}}^\top \\ M_B^\top & \mathbf{0} & \mathbf{I} & \mathbf{I} \\ \mathbf{0} & M_{\tilde{C}} & \mathbf{I} & \mu_\eta \end{bmatrix} \succ 0. \tag{43}$$

where $\mu_e = \gamma_e^2$ and $\mu_\eta = \gamma_\eta^2$. Substituting the values of M_P, M_A, M_B, M_C and $M_{\tilde{C}}$, BMIs are converted to LMIs and the LMIs are solved for the variables $\{P_f, P_g, W_f, W_g, W_h, L\}$. Then from the optimal solution, the noise-shaping filter can be reconstructed using the expression in (37) and using the relation in (15).

VI. Implementation

This section presents the MHOQ implementation using the optimal noise-shaping obtained in Sec. V. For a given reconstruction filter $H[z]$, let us denote the optimal noise-transfer function obtained solving the optimisation problem (40)-(43) as $R^*[z]$ and equivalently the optimal noise-shaping filter as $F^*[z]$ using equation (15). Then, the corresponding filter obtained using the relationship in equation (16) is denoted as $H^*[z]$. Then, the block diagram showing the implementation of the MHOQ is shown in Fig. 5.

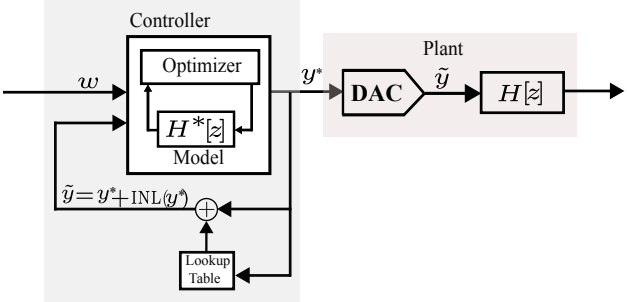

Fig. 5. Implementation of the MHOQ.

VII. Simulation Results

In this section, we present numerical simulations that illustrates the merits of the proposed method above. The MHOQ that minimise the error (11) with the state constraints (8)-(9) is formulated as a mixed integer quadratic programming problem (MIQP). The state variable $x \in \mathbb{R}^n$ is continuous whereas the control variable y is integer valued and constrained to belong to the finite set $\{0, 1, 2, .., 2^{\mathcal{B}} - 1\}$, where $\mathcal{B}$ is word-size (number of bits) of the DAC. The MIQP problem is then solved numerically using the Gurobi optimizer [18]. As discussed in [20], a general figure-of-merit for DACs is the signal-to-noise-and-distortion ratio (SINAD) defined as

$$\mathrm{SINAD} = 20 \log_{10} \left(\frac{\sigma_s}{\sigma_t} \right) \mathrm{dB}$$

where σ_s is the standard deviation of the input signal w and σ_t is the standard deviation of the noise and distortion in the output signal. The standard deviations are usually found via a power spectral density estimate. Note that the theoretical value of SINAD for an over-sampled signal is [20]

$$\mathrm{SINAD} = 6.02\mathcal{B} + 1.76 + 10 \log_{10}(\mathrm{OSR}) \quad \mathrm{dB} \qquad (44)$$

where $\mathrm{OSR} = F_s/2\mathrm{BW}$ is an oversampling ratio defined as the ratio of the sampling frequency F_s and desired bandwidth BW for the input signal w.

The simulation was carried out using a 8-bit DAC, thus the control variable y is restricted to belong to the set $\{1, 2, \ldots, 255\}$. The reference signal is $w(t) = A \sin(2\pi f t)$, with signal frequency $f = 1$ kHz, the signal amplitude $A = 2^{\mathcal{B}}$ and the sampling frequency $F_s = 1$ MHz. A second-order Butterworth filter $H[z]$ with cut-off frequency $F_c = 100$ kHz is used as a reconstruction filter. Other notations follows as described in Sec. VI. The quantisation methods in the simulations which are abbreviated as follows:

DQ: direct uniform quantisation (Sec. II),
NSQ: noise-shaping quantisation (Sec. IV)
MHOQ: moving horizon optimal quantisation (Sec. III)

The MHOQ is performed for different prediction horizon $N = 1, 2, 3, 4$. Moreover, the simulations were performed for the DACs with and without the INL denoted as **DAC-INL** and **DAC-Ideal**, respectively. The DAC-INL represent the DAC devices that are effected by the element-mismatch where the actual levels are deviated from the ideal ones. The INL used in the simulation is measured from the actual DAC and stored in the lookup table.

It is worth noting that the prediction horizon of 1 is particularly important in power electronics and signal processing applications and widely employed in such application [10]. This is because the associated optimization problem is non-convex, as the input signal is constrained to a finite set, which increases the computational burden with a larger prediction horizon. Therefore, in the simulation results, the unit prediction horizon will be used as a reference for assessing the performance of the proposed method.

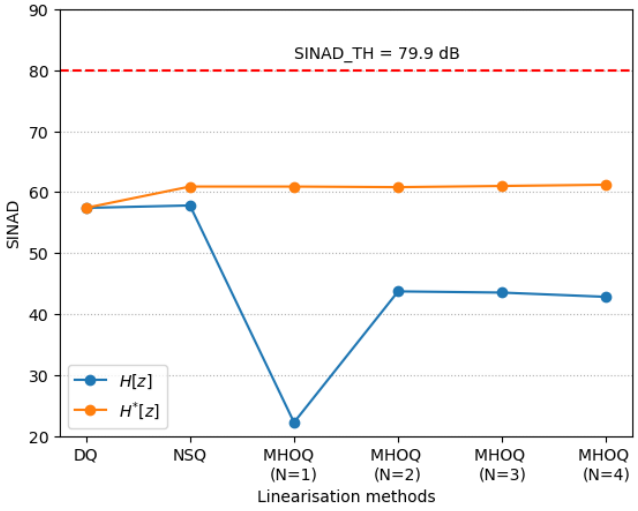

Fig. 6. SINAD for **DAC-Ideal** using different linearisation methods: $H[z]$ vs $H^*[z]$.

The simulations were carried out using the $H[z]$ and $H^*[z]$ as the model for the optimiser in Fig. 5 and the results are shown for the ideal and non-ideal DACs in Fig. 6 and Fig. 7, respectively. The theoretical SINAD value of $\mathrm{SINAD}_{\mathrm{th}} = 79.9$ dB is obtained using the expression in (44) where $\mathcal{B} = 8$, and $\mathrm{OSR} = F_s/(2 * f) = 1e6/(2e3)$.

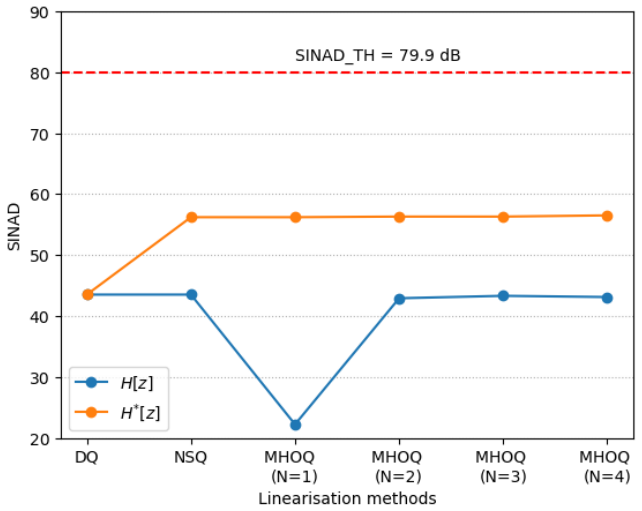

Fig. 7. SINAD for **DAC-INL** using different linearisation methods: $H[z]$ vs $H^*[z]$.

In the SINAD plots in both Fig. 6 and Fig. 7, the performance decrease at $N = 1$ with the use of $H[z]$. It improves for increase in prediction horizon but it is well below the one reached using the NSQ. However, the use of $H^*[z]$ shows the significant performance improvement in both the ideal and non-ideal case. Moreover, the NSQ and MHOQ $(\mathrm{N} = 1)$ behave identically, as mentioned in Sec. IV, with the use of the optimal noise-shaping filter.

The simulation results show that the use of an optimal noise-shaping filter in moving horizon implementation improves the performance significantly. Most notably, the performance at horizon length $N = 1$ improves significantly. Recall that for application in signal process unit prediction horizon is adopted. The simulation using higher prediction horizons are also performed but the results shows that there is no significant improvement in the performance.

VIII. Conclusions

The effectiveness of combining moving horizon optimal quantization with an optimal noise-shaping filter to reduce the quantisation error in digital-to-analogue converters (DACs) was demonstrated. This comes in addition to the performance increase seen by incorporating a model of the element mismatch seen in practical DACs. Simulation results validated the performance improvements across different filter configurations. The performance level appears to be marginally better than more traditional noise-shaping quantisation, but the moving horizon framework allows for models of other non-linear effects seen in DACs, such as glitches and slewing, to be included. Future work will extend the DAC model to account for additional effects, further enhancing the applicability of this approach in high-precision applications.

References

[1] B. P. Abbott, R. Abbott, R. Adhikari, et al. Ligo: the laser interferometer gravitational-wave observatory. *Reports on Progress in Physics*, 72(7):076901, 2009.

[2] B. Adhikari, R. van der Rots, J Leth, and A. A. Eielsen. Linearisation of digital-to-analog converters by model predictive control. *IFAC-PapersOnLine*, 58(18):92–98, 2024. 8th IFAC Conference on Nonlinear Model Predictive Control NMPC 2024.

[3] S. Azuma and T. Sugie. Optimal dynamic quantizers for discrete-valued input control. *Automatica*, 44(2):396–406, 2008.

[4] J. M. Beckers. Adaptive Optics for Astronomy: Principles, Performance, and Applications. *Annual Review of Astronomy and Astrophysics*, 31(1):13–62, 1993.

[5] W. R. Bennett. Spectra of quantized signals. *The Bell System Technical Journal*, 27(3):446–472, 1948.

[6] A. Biswas, I. S. Bayer, A. S. Biris, T. Wang, E. Dervishi, and F. Faupel. Advances in top–down and bottom–up surface nanofabrication: Techniques, applications & future prospects. *Advances in Colloid and Interface Science*, 170(1–2):2–27, 2012.

[7] S. Callegari and F. Bizzarri. Output filter aware optimization of the noise shaping properties of delta-sigma modulators via semi-definite programming. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 60(9):2352–2365, 2013.

[8] T. Cataltepe, A. R. Kramer, L. E. Larson, G. C. Temes, and R. H. Walden. Digitally corrected multi-bit $\Sigma\Delta$ data converters. In *IEEE International Symposium on Circuits and Systems*, volume 1, pages 647–650, 1989.

[9] K.C.-H. Chao, S. Nadeem, W.L. Lee, and C.G. Sodini. A higher order topology for interpolative modulators for oversampling a/d converters. *IEEE Transactions on Circuits and Systems*, 37(3):309–318, 1990.

[10] P. Cortes, M. P. Kazmierkowski, R. M. Kennel, D. E. Quevedo, and J. Rodriguez. Predictive control in power electronics and drives. *IEEE Transactions on Industrial Electronics*, 55(12):4312–4324, 2008.

[11] G. E. Fantner, P. Hegarty, J. H. Kindt, G. Schitter, G. A. G. Cidade, and P. K. Hansma. Data acquisition system for high speed atomic force microscopy. *Review of Scientific Instruments*, 76(2):026118, 2005.

[12] A. Faza, J. Leth, and A. A. Eielsen. Criterion for Sufficiently Large Dither Amplitude to Mitigate Non-linear Glitches. In *IEEE Conference on Control Technology and Applications*, pages 970–977. IEEE, 2023.

[13] D. M. Freeman. Slewing distortion in digital-to-analog conversion. *Journal of the Audio Engineering Society*, 25:178–183, april 1977.

[14] D. Goedhart, R. J. van de Plassche, and E. F. Stikvoort. Digital-to-analog conversion in playing a compact disc. *Phillips tech. Rev.*, 40(6):174–179, 1982.

[15] M. Goodson, Bo Zhang, and R. Schreier. Proving stability of delta-sigma modulators using invariant sets. In *IEEE International Symposium on Circuits and Systems*, volume 1, pages 633–636, 1995.

[16] G. C. Goodwin, D. E. Quevedo, and D. McGrath. Moving-horizon optimal quantizer for audio signals. *Journal of the Audio Engineering Society*, 51(3):138–149, 2003.

[17] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 2.1, March 2014.

[18] Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual, 2023.

[19] C. A. Hamilton and Y. H. Tang. Evaluating the uncertainty of Josephson voltage standards. *Metrologia*, 36(1):53–58, 1999.

[20] W. Kester, editor. *Data Conversion Handbook*. Elsevier, 2005.

[21] J. Löfberg. Yalmip : A toolbox for modeling and optimization in matlab. In *In Proceedings of the CACSD Conference*, Taipei, Taiwan, 2004.

[22] I. Masubuchi, A. Ohara, and N. Suda. Lmi-based controller synthesis: a unified formulation and solution. *International Journal of Robust and Nonlinear Control: IFAC-Affiliated Journal*, 8(8):669–686, 1998.

[23] MOSEK, ApS. The MOSEK optimization toolbox for MATLAB manual. Version 9.0, 2019.

[24] Marko Neitola. Lee's Rule Extended. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 64(4):382–386, 2017.

[25] S. Ohno and M. R. Tariq. Optimization of noise shaping filter for quantizer with error feedback. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 64(4):918–930, 2017.

[26] R. Schreier, G. C. Temes, et al. *Understanding delta-sigma data converters*, volume 74. IEEE press Piscataway, NJ, 2005.

[27] C.E. Shannon. Communication in the presence of noise. *Proceedings of the IRE*, 37(1):10–21, 1949.

[28] R. Tutuncu, Kim-Chuan Toh, and Michael Todd. Sdpt3 - a matlab software package for semidefinite-quadratic-linear programming, version 3.0, 09 2001.

[29] Robert A. Wannamaker. Psychoacoustically optimal noise shaping. *Journal of the Audio Engineering Society*, 40:611–620, july 1992.