# Report: Synthesis of optimal filter for MHOQ

July 10, 2024

## Contents

# 1 Quantisation

Let $w \in \mathbb{R}$ be the input, $\mathbf{Q}$ be the quantiser and $y \in \mathbb{U}$ the quantiser output, as shown in the additive model of quantisation in the Figure 1. Let us define the quantisation error as

$$q = \mathbf{Q}(w) - w = y - w. \tag{1}$$

The quantisation requires the signal to be mapped to a finite signal where each value of the output $y$ is restriced to

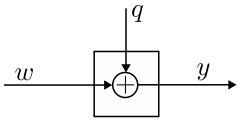


Figure 1: Quantiser additive model

belong to a finite set $\mathbb{U}$. The elements of the set $\mathbb{U}$ represent the quantiser levels and depends on the word-size of the quantiser.

# 2 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. The feedback filter $F(z)$ is designed such that the transfer function $y = (1 - F(z))\epsilon$ is a high-pass filter.

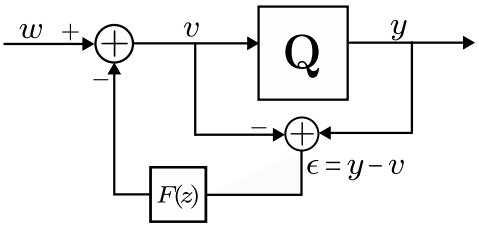


Figure 2: Noise shaping quantiser

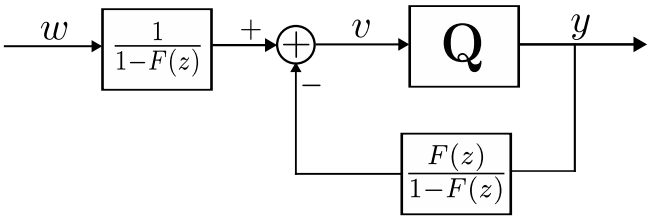


Figure 3: Noise shaping quantiser

In linear analysis, the output is given by

$$Y(z) = \mathbf{STF}.W(z) + \mathbf{NTF}.E(Z) \tag{2}$$

where the signal transfer function $\mathbf{STF} = 1$, noise transfer function $\mathbf{NTF} = (1 - F)$ with $F$ being a noise shaping filter. $F = z^{-1}$ is the special case known as the first-order delta-sigma modulator.

# 3 Moving horizon optimal quantiser (MHOQ)

The design criteria for the MHOQ is the minimization of the perceived errors defined as follows:

$$e(t) = H(z)(w(t) - y(t)) \tag{3}$$

where $H(z)$ is a stable time-invariant linear low-pass filter with the following state-space

$$H(z) = 1 + C(zI - A)^{-1}B \tag{4}$$

The error $e$ then can be written as the output of the following state-space representation of $H$

$$\begin{aligned} x(t+1) &= Ax(t) + B(w(t) - y(t)) \\ e(t) &= Cx(t) + w(t) - y(t) \end{aligned} \tag{5}$$

where $x \in \mathbb{R}^n$ is the state vector. The error $e$ corresponds to the difference between the filtered quantised signal and the filtered input signal.

For moving horison implementation, the optimisation problem is defined as the problem of finding $y \in \mathbb{U}$ that minimises the cost function while satisfying the state equations as follows:

$$y^*(t) = \arg\min_{y(t)} V_N = \sum_{t=k}^{k+N-1} e^2(t) \tag{6}$$

subject to

$$x(t+1) = Ax(t) + B(w(t) - y(t)) \tag{7}$$
$$e(t) = Cx(t) + w(t) - y(t) \tag{8}$$
$$y(t) \in \mathbb{U}. \tag{9}$$

### 3.1 Alternative binary formulation

The optimization problem (6)-(9) can be reformulated as an optimization problem with the binary variables. Let $\mathcal{B}$ be the number of bits. $b_i = \{0, 1\}$ and $Q_i$ , $i = \{0, 1, \ldots, 2^{\mathcal{B}} - 1\}$, be the binary variables and quantisation levels, respectively.

$$y^*(t) = \arg\min_{y(t)} V_N = \sum_{t=k}^{k+N-1} e^2(t) \tag{10}$$

subject to

$$x(t+1) = Ax(t) + B(w(t) - y(t)) \tag{11}$$
$$e(t) = Cx(t) + (w(t) - y(t)) \tag{12}$$
$$y(t) = \sum_{i=0}^{2^{\mathcal{B}}-1} Q_i b_i, \quad \sum_{i=0}^{2^{\mathcal{B}}-1} b_i = 1, \quad b_i = \{0, 1\}. \tag{13}$$

## 4 Noise Transfer Function(NTF)

The frequency response of the noise transfer functions due to butterworth filters at different cutoff frequencies are shown in the figure Fig. 4. In the figure, we can see that the net area under the curve remain the same. In Fig. 5 the frequency reponse of the low pass filter is plotted along with that of the noise transefer function. This observation shows that the better performance can be achieved by increasing the cutoff frequency during MHOQ while keeping the cutoff frequency of the reconstruction as same. The simulation results in the following table confirm this observation.

Table 1: ENOB at different cutoff frequencies. Reconstruction filter: Butterworth LPF with $n = 2$, $Fc = 100$kHz and $Fs = 1$Mhz.

| Fc | 100 kHz | 200 kHz | 300 kHz | 400 kHz | 500 kHz |
|---|---|---|---|---|---|
| ENOB | 3.981 | 5.307 | 7.817 | 10.481 | 10.936 |

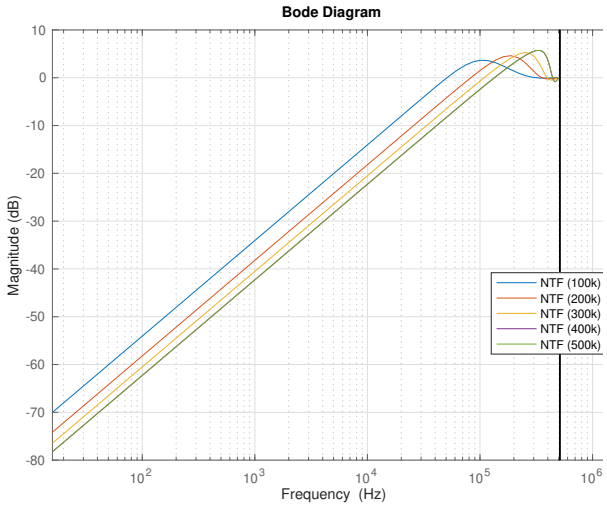


Figure 4: Frequency response of NTF for different cutoff frequency

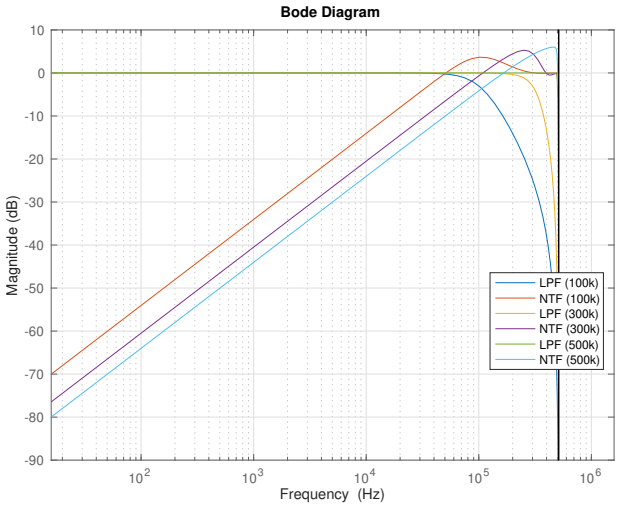


Figure 5: Frequency response of LPF and NTF for different cutoff frequency

# 5 Spectrum of quantisation noise

The sampling frequency effects the spectrum of the quantisation noise and consequently the ENOB as shown in the following figures.

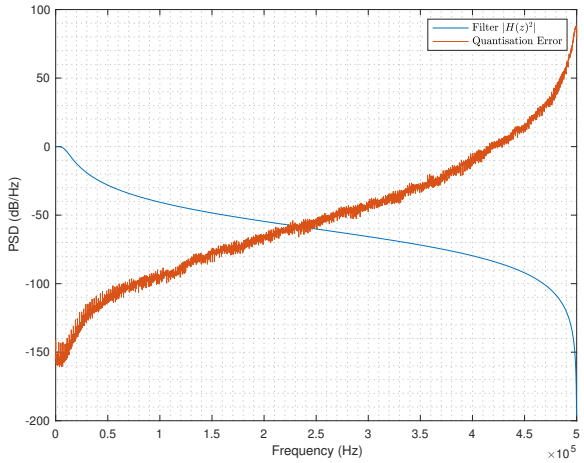


Figure 6: Butterworth Frequency response and frequency spectrum of quantisation noise:
**Fc** = 10 kHz, **Fs** = 1 MHz, **ENOB** = 16.58, **INL**

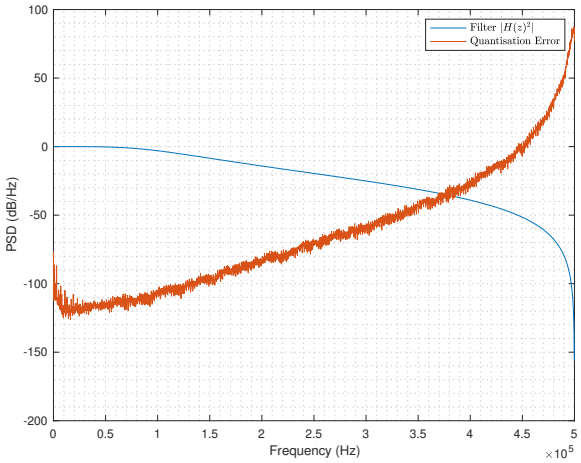


Figure 7: Butterworth Frequency response and frequency spectrum of quantisation noise:
**Fc** = 100 kHz, **Fs** = 1 MHz, **ENOB** =7.43, **INL**

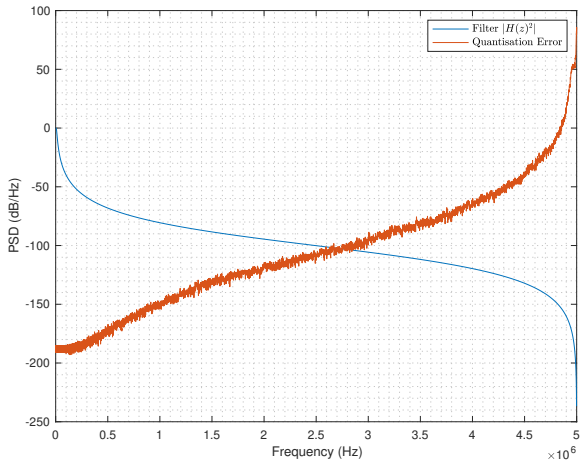


Figure 8: Butterworth Frequency response and frequency spectrum of quantisation noise:
**Fc** = 10 kHz, **Fs** = 10 MHz, **ENOB** = 25.12, **INL**

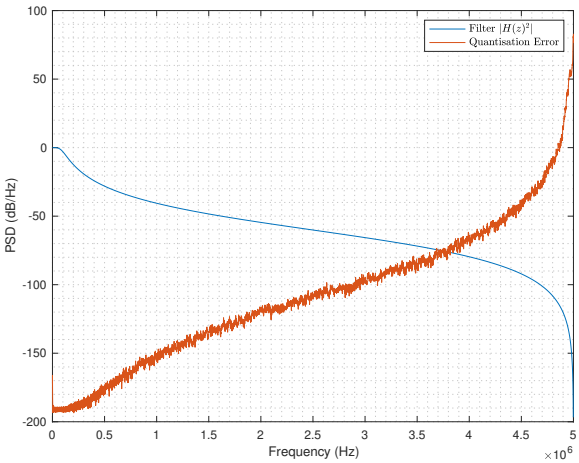


Figure 9: Butterworth Frequency response and frequency spectrum of quantisation noise:
**Fc** = 100 kHz, **Fs** = 10 MHz, **ENOB** = 17.03, **INL**

# 6 Synthesis of optimal noise-shaping filter

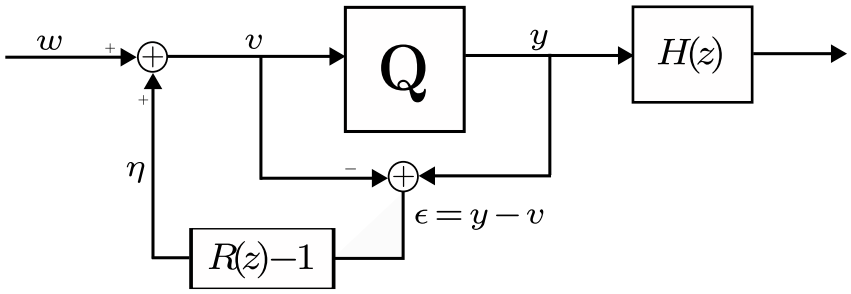


Figure 10: Noise shaping quantiser and a filter $H(z)$.

In noise shaping quantiser with error-feedback structure as shown in the Figure 10, the input to the quantiser is $v = w + \eta = w + (R(z) - 1)\epsilon$ and feedback error $\epsilon$ is

$$\epsilon = y - v = y - w - (R(z) - 1)\epsilon. \tag{14}$$

Then quantisaion noise defined as $q := y - w$ can be expressed as

$$q = y - w = R(z)\epsilon. \tag{15}$$

Then the effect of the quantisation error on the system $H(z)$ can be expressed as

$$e = H(z)R(z)\epsilon. \tag{16}$$

and it shows that we can reduce the error in the plant output by properly designing the noise shaping filter $R(z)$ with the knowledge of the plant $H(z)$.

The objective is to design stable noise-shaping filter such that it minimises the effect of the quantisation noise in the plant output. A constraint on the error feedback signal should be imposed to prevent the quantiser from overloading and achieve a stable noise-shaping quantiser as

$$\eta = (R(z) - 1)\epsilon.$$

- **Design Problem:** For a fixed pair (p,q) [1],

$$\begin{aligned} &\min_{R(z)\in\mathbb{RH}_\infty} \|e\|_p \\ \text{subject to.}\quad &R(\infty) = 1, \\ &\|\eta\|_q < \gamma_\eta \end{aligned} \tag{17}$$

where $\mathbb{RH}_\infty = \mathbb{R} \cap \mathbb{H}_\infty$ is the set of proper stable rational transfer functions.

## 6.1 Optimal noise shaping filter with constraint

The optimization problem is setup as the minimization of the upper bound of the $\|e\|_p$ and $\|\eta_q\|$ as follows:

$$\min_{R(z)\in\mathbb{RH}_\infty} \gamma_e \tag{18}$$

subject to $R(\infty) = 1$ and

$$\|H(z)R(z)\|_{ind.p} < \gamma_e \tag{19}$$

$$\|R(z) - 1\|_{ind.q} < \gamma_\eta \tag{20}$$

where $\|.\|_{ind.r}$ is the induced norm and $\gamma_e$ and $\gamma_\eta$ are the upper bound of $\|H(z)R(z)\|_{ind.p}$ and $\|R(z) - 1\|_{ind.q}$ , respectively.

**State space representation**

### 6.1.1 System

Denoting the state-space representation of the $H(z)$ and $R(z)$ as $(A_h, B_h, C_h, D_h)$ and $(A_r, B_r, C_r, 1)$ respectively, the state space realization of $H(z)R(z)$ is

$$x_{k+1} = Ax_k + B\epsilon_k \tag{21}$$
$$e_k = Cx_k + D\epsilon_k \tag{22}$$

where

$$A = \begin{bmatrix} A_h & B_hC_r \\ \mathbf{0} & A_r \end{bmatrix} \qquad B = \begin{bmatrix} B_h \\ B_r \end{bmatrix} \qquad C = \begin{bmatrix} C_h & D_hC_r \end{bmatrix} \qquad D = D_h. \tag{23}$$

Similary, the variance of the error $e$ under the white noise assumption at time $k$ in state-space form is given by $\mathbb{H}_2$-norm as

$$\|H(z)R(z)\|_2^2 = \sum_{k=0}^{\infty} \|CA^kB\|_2^2 + DD^\top. \tag{24}$$

Moreover, if $A$ is Schur matrix, then there exists a positive semi-definite solution $P$ of the discrete Lyapunov equation defined as

$$P = A^\top PA + BB^\top \tag{25}$$

and the squared $\mathbb{H}_2$ norm is given by

$$\|H(z)R(z)\|_2^2 = CPC^\top + DD^\top. \tag{26}$$

### 6.1.2 Constraint 1

Then $\|H(z)R(z)\|_2 < \gamma_e$ if and only if there exist a positive definite matrix $P$ such that

$$(\mathbf{BMI}) \qquad \begin{bmatrix} P & PA & PB \\ A^\top & P & \mathbf{0} \\ B^\top & \mathbf{0} & 1 \end{bmatrix} \succ 0 \tag{27}$$

$$(\mathbf{LMI}) \qquad \begin{bmatrix} \mu_e & C & D \\ C^\top & P & \mathbf{0} \\ D^\top & \mathbf{0} & 1 \end{bmatrix} \succ 0 \tag{28}$$

$$\mu_e = \gamma_e^2. \tag{29}$$

### 6.1.3 Constraint 2

The variance of the noise shaping filter is given by

$$E\{|\eta_k|^2\} = \|R(z) - 1\|_2^2 = \sum_{k=1}^{\infty} \|\tilde{C}A^kB\|_2^2, \tag{30}$$

where $\tilde{C} = \begin{bmatrix} \mathbf{0} & C_r \end{bmatrix}$. Then $\|R(z) - 1\|_2^2 < \gamma_\eta$ if and only if there exist a positive definite matrix $P$ that satisfies

$$(\mathbf{BMI}) \qquad \begin{bmatrix} P & PA & PB \\ A^\top & P & \mathbf{0} \\ B^\top & \mathbf{0} & 1 \end{bmatrix} \succ 0 \tag{31}$$

$$(\mathbf{LMI}) \qquad \begin{bmatrix} \mu_\eta & \tilde{C} \\ \tilde{C}^\top & P \end{bmatrix} \succ 0 \tag{32}$$

$$\mu_\eta = \gamma_\eta^2. \tag{33}$$

### 6.1.4 Optimization problem: State Space formulation

Thus the optimization problem can be written as follows,

$$\min_{R(z) \in \mathbb{RH}_\infty} \gamma_e \tag{34}$$

subject to $R(\infty) = 1$ and

$$\begin{bmatrix} P & PA & PB \\ A^\top & P & \mathbf{0} \\ B^\top & \mathbf{0} & 1 \end{bmatrix} \succ 0 \tag{35}$$

$$\begin{bmatrix} \mu_e & C & D \\ C^\top & P & \mathbf{0} \\ D^\top & \mathbf{0} & 1 \end{bmatrix} \succ 0 \tag{36}$$

$$\begin{bmatrix} \mu_\eta & \tilde{C} \\ \tilde{C}^\top & P \end{bmatrix} \succ 0 \tag{37}$$

$$\mu_e = \gamma_e^2, \mu_\eta = \gamma_\eta^2. \tag{38}$$

## 6.2 LMI Synthesis: Convert BMIs to convex LMIs

BMIs are not convex and NP hard to solve, but they can be converted to convex LMIs and can be solved numerically whereas the LMIs are convex. The non-convex BMIs can be converted to convex LMIs using change of variables [2].

### Change of Variables:

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

$$\begin{aligned} M_A &:= \begin{bmatrix} A_h P_f + B_h W_f & A_h \\ L & P_g A_h \end{bmatrix} \\ M_B &:= \begin{bmatrix} B_h \\ W_g \end{bmatrix} \\ M_C &:= \begin{bmatrix} C_h P_f + D_h W_f & C_h \end{bmatrix} \\ M_P &:= \begin{bmatrix} P_f & I_n \\ I_n & P_g \end{bmatrix}. \end{aligned} \tag{39}$$

Next, define

$$P^{-1} := \begin{bmatrix} P_f & S_f \\ S_f & S_f \end{bmatrix}, \qquad U := \begin{bmatrix} P_f & I_n \\ S_f & \mathbf{0} \end{bmatrix} \qquad \text{and} \tag{40}$$

$$S_f := P_f - P_g^{-1} (\succ 0) \tag{41}$$

then we have,

$$M_P = U^\top P U. \tag{42}$$

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

$$\begin{aligned} A_r &:= \left[ B_h W_f - P_g^{-1}(L - P_g A_h P_f) \right] S_f^{-1} \\ B_r &:= \left[ B_h - P_g^{-1} W_g \right] \\ C_r &:= W_f S_f^{-1} \end{aligned} \tag{43}$$

then $(A, B, C)$ satisfy,

$$M_A = U^\top P A U \tag{44}$$

$$M_B = U^\top P B \tag{45}$$

$$M_C = C U \tag{46}$$

$$M_p = U^\top P U. \tag{47}$$

Multiplying with the transformation $\phi = \mathrm{diag}(U, U, 1)$ form the RHS and $\phi^\top$ from the LHS the BMI condition 35 takes the following form:

$$\begin{bmatrix} M_P & M_A & M_B \\ M_A^\top & M_P & \mathbf{0} \\ M_B^\top & \mathbf{0} & \mathbf{I} \end{bmatrix} \succ 0. \tag{48}$$

Similarly, using the transformation $\text{diag}(1, U, 1)$, the LMI condition 36 is

$$\begin{bmatrix} \mu_e & M_C & D^\top \\ M_C^\top & M_P & \mathbf{0} \\ D & \mathbf{0} & \mathbf{I} \end{bmatrix} \succ 0 \tag{49}$$

and finally the constraint 37 using the transformation is

$$\begin{bmatrix} \gamma_\eta^2 & M_{\tilde{C}} \\ M_{\tilde{C}}^\top & M_P \end{bmatrix} \succ 0 \tag{50}$$

with $M_{\tilde{C}} := \tilde{C}U$. The LMI conditions 48, 49 and 50 are convex and the minimization of $\gamma_e$ with these constraints is a convex optimization problem.

### 6.3 Optimization Problem:

With the change of the variables the BMI is converted to the LMI and since all the LMIs are convex, the optimization problem is a convex optimization problem. Also, since the objective is linear and the constraints are LMIs and with linearity constraints, the optimization problem is Semi-Definite Program (SDP). Such SDP can be solved numerically using CVX. The optimization problem takes the following form,

$$\min \mu_e = \gamma_e^2 \tag{51}$$

subject to,

$$\begin{bmatrix} M_P & M_A & M_B \\ M_A^\top & M_P & \mathbf{0} \\ M_B^\top & \mathbf{0} & \mathbf{I} \end{bmatrix} \succ 0 \tag{52}$$

$$\begin{bmatrix} \mu_e & M_C & D^\top \\ M_C^\top & M_P & \mathbf{0} \\ D & \mathbf{0} & \mathbf{I} \end{bmatrix} \succ 0. \tag{53}$$

$$\begin{bmatrix} \mu_\eta & M_{\tilde{C}} \\ M_{\tilde{C}}^\top & M_P \end{bmatrix} \succ 0 \tag{54}$$

where $\mu_e$ is the variance of the output error and $\mu_\eta$ the variance of the feedback error.

# 7 Simulation: Optimal noise shaping

**Notations:**

Let us denote the transfer function of a Butterworth low pass filter as $H(z)$. Then from Figure 10, the noise transfer function (NTF) is denoted by $R(z)$ and comparing Figure 2 and Figure 10, the noise shaping transfer function $F(z) = 1 - R(z)$. Moreover, it is shown in the moving horizon implementation of the quantisation [3], the low-pass filter and the noise shaping filter are related as follows,

$$F(z) = \frac{H(z) - 1}{H(z)} \qquad \Leftrightarrow \qquad H(z) = \frac{1}{1 - F(z)} = \frac{1}{R(z)}. \tag{55}$$

Next, let the optimal noise transfer function (NTF) for low-pass filter $H(z)$ obtained by solving the optimization problem (51)-(54) is denoted as $R_{opt}(z) = \dfrac{b_r}{a_r}$, where $b_r$ and $a_r$ are the numerator and denominator of the noise transfer function, respectively. Then the optimal noise-shaping transfer function (NSF) is $F_{opt}(z) = 1 - R_{opt}(z) = \dfrac{a_r - b_r}{a_r}$. Finally, the optimal low pass filter for MPC implementation is $H_{mpc} = \frac{1}{R_{opt}(z)} = \frac{a_r}{b_r}$.

## 7.1 Simulation 1:

Let us consider a third-order butter-worth low pass filter $H(z)$ with $F_c = 100kHz$ and $F_s = 1Mhz$. Then, from the expression in the equation 55 the corresponding noise shaping filter $F(z)$ and the noise transfer function $R(z)$ can be obtained. Next, the optimization problem (51)-(54) with the constraint $\gamma_\eta < 1.5^2$ (Lee's condition) is solved to obtain the optimal noise transfer function $R_{opt}(z)$ and consequently the $F_{opt}(z)$ and $H_{mpc}(z)$ are obtained. The frequency responses of $H(z)$, $F(z)$ and $R(z)$ are shown in the Figure. 11 and that of $H_{opt}(z)$, $F_{opt}(z)$ and $R_{opt}(z)$ are shown in Figure 12.

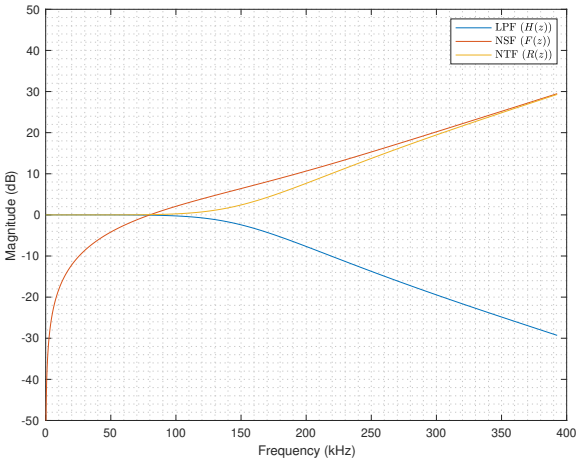


Figure 11: Frequency response: Butterworth filter $H(z)$ with $n = 3$ and $F_c = 100kHz$ and corresponding noise shaping filter $F(z)$ and noise transfer function $R(z)$.

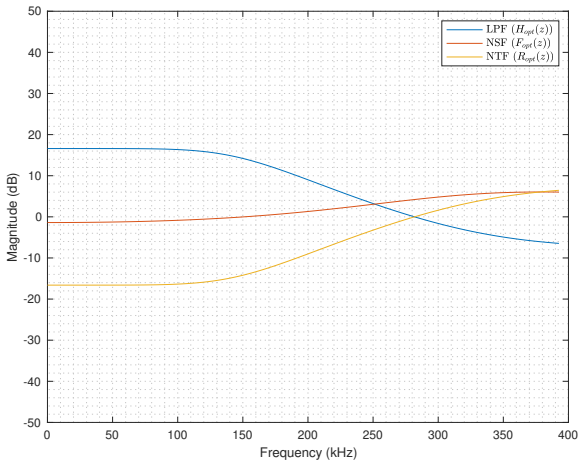


Figure 12: Frequency response: Optimal noise transfer function $R_{opt}(z)$ for $H(z)$ and corresponding noise shaping filter $F_{opt}(z)$ and low pass filter $H_{opt}(z)$.

Table 2: ENOB obtained for different methods with uniform quantisation models. Optimal NTF solved for second order butterworth filter with $Fc = 100kHz$ and $Fs = 1Mhz$.

| ENOB/ Methods | Direct | DSM | NSD (Optimal) | MPC ($N = 1$) | MPC ($N = 2$) | MPC ($N = 3$) |
|---|---|---|---|---|---|---|
| 6-bit | 6.936 | 8.291 | 9.30 | 9.30 | 9.303 | 9.361 |
| 8-bit | 9.650 | 10.366 | 11.357 | 11.357 | 11.326 | 11.380 |
| 12-bit | 13.285 | 14.365 | 15.372 | 15.372 | 15.290 | 15.281 |
| 16-bit | 17.148 | 18.390 | 19.373 | 19.373 | 19.247 | 19.134 |

Table 3: ENOB obtained for different methods with nonlinear quantisation models. Optimal NTF solved for second order butterworth filter with $Fc = 100kHz$ and $Fs = 1Mhz$.

| ENOB/ Methods | Direct | DSM | NSD (Optimal) | MPC ($N = 1$) | MPC ($N = 2$) | MPC ($N = 3$) |
|---|---|---|---|---|---|---|
| 6-bit | 4.177 | 7.977 | 7.022 | 7.022 | 6.99 | 7.007 |
| 8-bit | 6.526 | 9.968 | 9.228 | 9.228 | 9.251 | 9.241 |
| 12-bit | 10.830 | 13.679 | 13.524 | 13.524 | 13.503 | 13.509 |
| 16-bit | 13.577 | 15.77 | 16.156 | 16.156 | 16.157 | 16.145 |

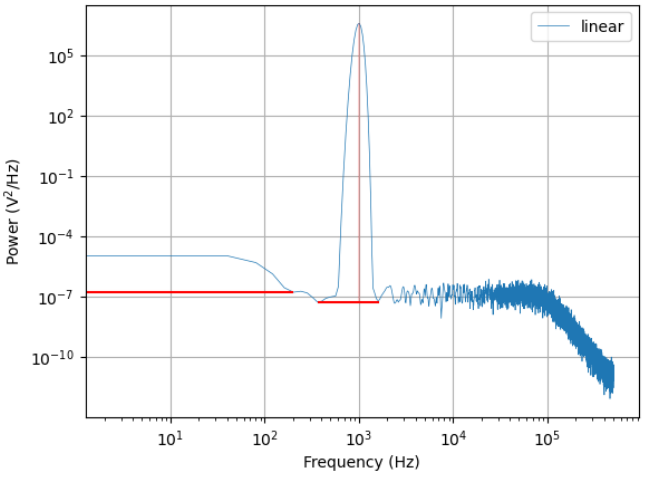


Figure 13: Direct Quantisation

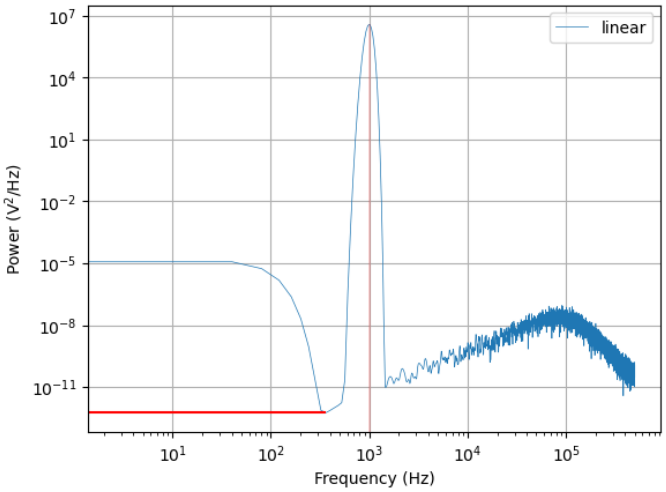


Figure 14: Delta sigma modulator

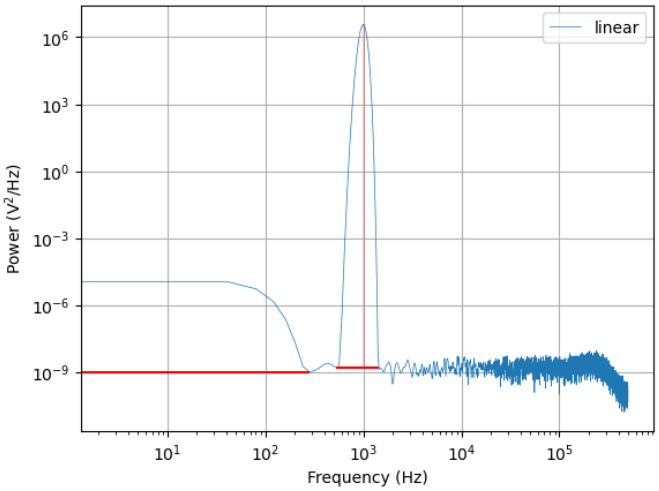


Figure 15: Noise shaping quantiser

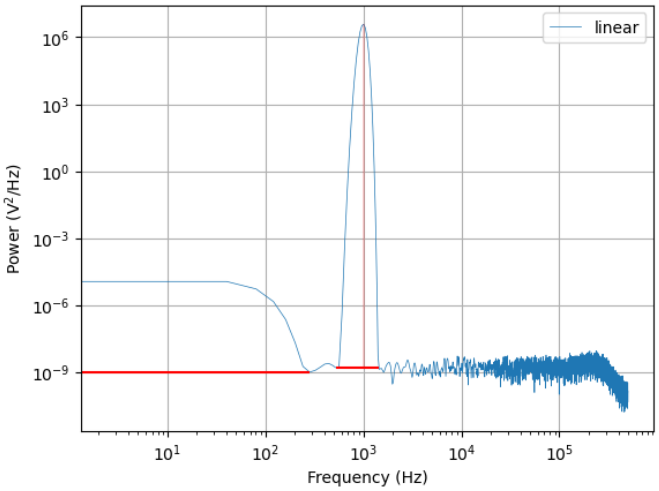


Figure 16: MPC with optimal noise shaping

Table 4: ENOB obtained for different methods with nonlinear quantisation models. Optimal NTF solved for second order butterworth filter with $Fc = 100kHz$ and $Fs = 10Mhz$.

| ENOB/ Methods | Direct | DSM | NSD (Optimal) | MPC ($N = 1$) | MPC ($N = 2$) | MPC ($N = 3$) |
|---|---|---|---|---|---|---|
| 6-bit | 5.041 | 12.355 | 15.576 | 15.576 | 15.548 | 15.577 |
| 8-bit | 7.392 | 13.950 | 16.988 | 16.988 | 16.406 | 17.008 |
| 12-bit | 11.210 | 17.368 | 20.952 | 20.952 | 20.950 | 20.931 |
| 16-bit | 13.877 | 19.413 | 23.369 | 23.369 | 23.372 | 23.372 |

Table 5: ENOB for different values of $\gamma_\eta$

| Methods/$\gamma_\eta$ | 1.5 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Direct | 17.148 | - | - | - | - | - | - | - | - | - | - | - |
| DSM | 18.390 | - | - | - | - | - | - | - | - | - | - | - |
| NSD | 19.152 | 19.370 | 19.532 | 19.522 | 19.472 | 19.339 | 19.222 | 19.127 | 18.976 | 18.705 | 2.786 | 2.2.458 |
| MPC(N=1) | 19.152 | 19.370 | 19.532 | 19.522 | 19.472 | 19.339 | 19.222 | 19.127 | 18.976 | 18.705 | 2.786 | 2.458 |
| MPC(N=2) | 19.098 | 19.237 | 19.533 | 19.402 | 19.301 | 19.191 | 19.023 | 18.909 | 18.787 | 18.738 | 18.641 | 18.577 |
| MPC(N=3) | 18.955 | 19.133 | 19.266 | 19.180 | 19.040 | 18.868 | 18.746 | 18.677 | 18.638 | 18.560 | 18.456 | 18.379 |

Table 6: ENOB for different values of $\gamma_\eta$ using 16 bit DAC.

| $\gamma_\eta$/ Headroom | 0 | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1.5 | 19.152 | 19.066 | 19.004 | 18.924 | 18.830 | 18.728 | 18.616 | 18.531 | 18.412 | 18.278 |
| 3 | 19.532 | 19.455 | 19.387 | 19.313 | 19.210 | 19.117 | 19.028 | 18.904 | 18.793 | 18.676 |
| 9 | 18.976 | 18.980 | 18.983 | 18.826 | 18.747 | 18.640 | 18.527 | 18.429 | 18.314 | 18.187 |
| 10 | 18.705 | 18.877 | 18.807 | 18.728 | 18.637 | 18.549 | 18.439 | 18.340 | 18.225 | 18.090 |
| 11 | 2.786 | 18.814 | 18.719 | 18.622 | 18.546 | 18.457 | 18.341 | 18.245 | 18.129 | 18.00 |
| 12 | 2.458 | 18.700 | 18.634 | 18.543 | 18.447 | 18.340 | 18.269 | 18.142 | 18.032 | 17.911 |
| 15 | 1.898 | 18.457 | 18.366 | 18.285 | 18.192 | 18.102 | 18.101 | 17.897 | 17.790 | 17.665 |
| 20 | 1.647 | 18.118 | | | | | | | | |
| 30 | 1.398 | 17.602 | | | | | | | | |
| 100 | 1.398 | 15.964 | | | | | | | | |
| 1000 | 0.828 | 12.642 | 12.545 | 12.484 | | | | | | |
| 10000 | 0.826 | 0.631 | 0.495 | 10.318 | 10.25 | 10.187 | 10.139 | 9.925 | 9.847 | 9.788 |

Table 7: ENOB obtained for different methods with nonlinear quantisation models. Optimal NTF solved for second order butterworth filter with $Fc = 100kHz$ and $Fs = 1Mhz$.

| ENOB/ Methods | Direct | DSM | NSD (Optimal) | MPC ($N = 1$) [Time (s)] | MPC ($N = 2$)[Time (s)] | MPC ($N = 3$) |
|---|---|---|---|---|---|---|
| 6-bit (BINARY) | 6.693 | 8.322 | 9.051 | 9.051 [$\approx$ 880 s] | 9.135 [$\approx$ 3820 s] | - |
| 6-bit (SCALED) | 6.693 | 8.322 | 9.051 | 9.051 [$\approx$ 120 s] | 9.093 [$\approx$ 380 s] | 9.141 [$\approx$ 450 s] |

# 8 MPC with Switching Frequency Minimization

The optimization problem (6)-(9) can be reformulated as an optimization problem with the binary variables. Let $\mathcal{B}$ be the number of bits. $b_i = \{0,1\}$ and $Q_i$ , $i = \{0,1,\ldots,2^{\mathcal{B}}-1\}$, be the binary variables and quantisation levels, respectively.

$$y^*(t) = \arg\min_{y(t)} V_N = \sum_{t=k}^{k+N-1} e^2(t) + \beta\mathcal{N}_s \tag{56}$$

subject to

$$x(t+1) = Ax(t) + B(w(t) - y(t)) \tag{57}$$
$$e(t) = Cx(t) + (w(t) - y(t)) \tag{58}$$
$$y(t) = \sum_{i=0}^{2^{\mathcal{B}}-1} Q_i b_i(t), \quad \sum_{i=0}^{2^{\mathcal{B}}-1} b_i(t) = 1, \quad b_i = \{0,1\}. \tag{59}$$
$$\mathcal{N}_s = \sum_{i=0}^{2^{\mathcal{B}-1}} |b_i(t) - b_i(t-1)|^2 \tag{60}$$

where $b_i(t-1)$ is the state of the switch $i$ at the previous state and $b_i(t)$ is the sate of switch $i$ at the current state and $\beta$ is the weighting factor.

Table 8: Simulation Results: Carrier Frequency $999Hz$, Total number of samples: 21021.

| Weighting Factor ($\beta$) | 0 | 0.1 | 0.5 | 1 | 5 | 10 | 100 |
|---|---|---|---|---|---|---|---|
| ENOB | 9.392 | 9.374 | 8.694 | 8.036 | 6.574 | 6.080 | 4.000 |
| Total number of switches (% switching of total number of samples) | 16238 ($\approx 77\%$) | 15213 ($\approx 72\%$) | 13875 ($\approx 66\%$) | 13233 ($\approx 63\%$) | 11582 ($\approx 55\%$) | 10912 ($\approx 52\%$) | 9376 ($\approx 45\%$) |

Table 9: Simulation Results: Carrier Frequency $99Hz$, Total number of samples: 111111.

| Weighting Factor ($\beta$) | 0 | 0.1 | 0.5 | 1 | 5 | 10 | 100 |
|---|---|---|---|---|---|---|---|
| ENOB | 9.444 | 9.177 | 8.693 | 8.158 | 6.833 | 5.845 | 3.733 |
| Total number of switches (% switching of total number of samples) | 85548 ($\approx 77\%$) | 79474 ($\approx 72\%$) | 70175 ($\approx 63\%$) | 68803 ($\approx 62\%$) | 53833($\approx$ 48%) | 53319($\approx$ 48%) | 43703 ($\approx 39\%$) |

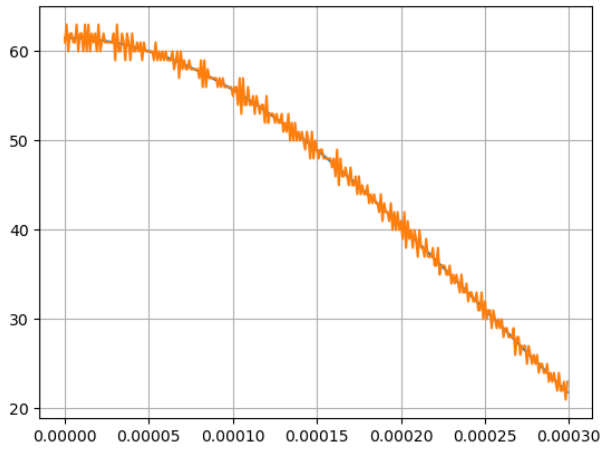


Figure 17: Carrier frequency: $999Hz, \beta = 0$

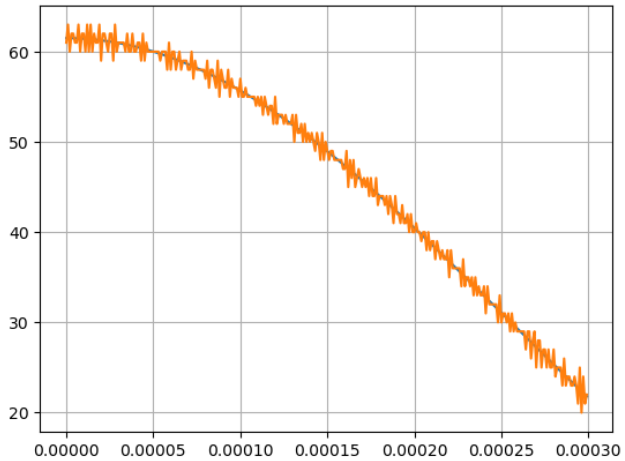


Figure 18: Carrier frequency: $999Hz, \beta = 0.1$

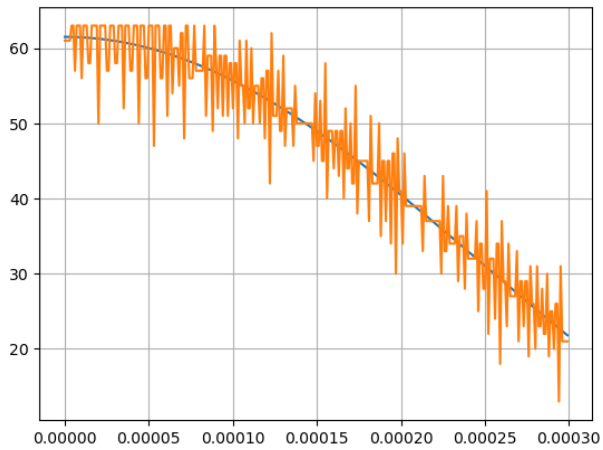


Figure 19: Carrier frequency: $999Hz, \beta = 10$

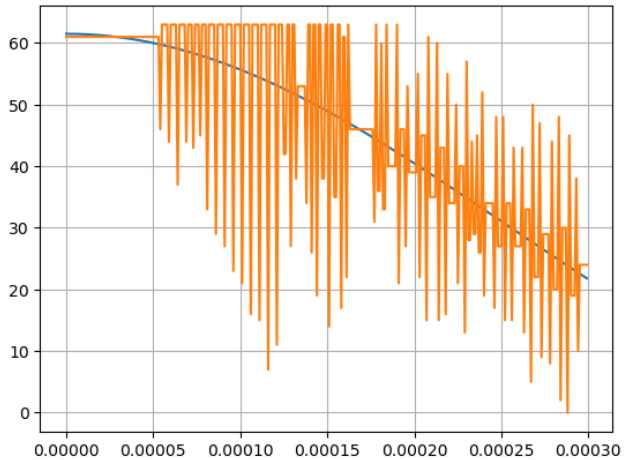


Figure 20: Carrier frequency: $999Hz, \beta = 100$

# 9 MPC with switching rate limitation:

The MPC formulation with the rate limitation is as follows:

$$y^*(t) = \arg\min_{y(t)} V_N = \sum_{t=k}^{k+N-1} e^2(t) \tag{61}$$

subject to

$$x(t+1) = Ax(t) + B(w(t) - y(t)) \tag{62}$$
$$e(t) = Cx(t) + w(t) - y(t) \tag{63}$$
$$-L_r\Delta t \leq \Delta y(t) \leq L_r\Delta t \tag{64}$$
$$y(t) \in \mathbb{U}. \tag{65}$$

The rate limitation is imposed by the constraint (64) where $\Delta y(t) = y(t) - y(t - \Delta t)$ and the constraint (64) can be further simplifed as

$$\Delta y(t) \leq L_r\Delta t$$
$$-\Delta y(t) \leq L_r\Delta t$$

Here $\Delta t = T_s$ and the rate limitation $L_r$ are assumed to constant for all sampling instants.

## Simulations results:

Table 10: Simulation Results: Carrier Frequency $999Hz$, Total number of samples: 21021.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Rate limit ($L_r\Delta t$) | N/A | 1e6 | - | - | - | - | - |
| ENOB (99Hz) | 9.444 | 4.18 | - | - | - | - | - |

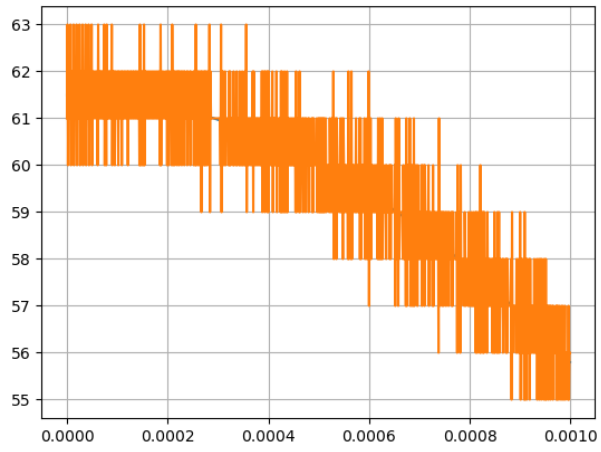


Figure 21: No rate limit. Carrier frequency: $99Hz$.

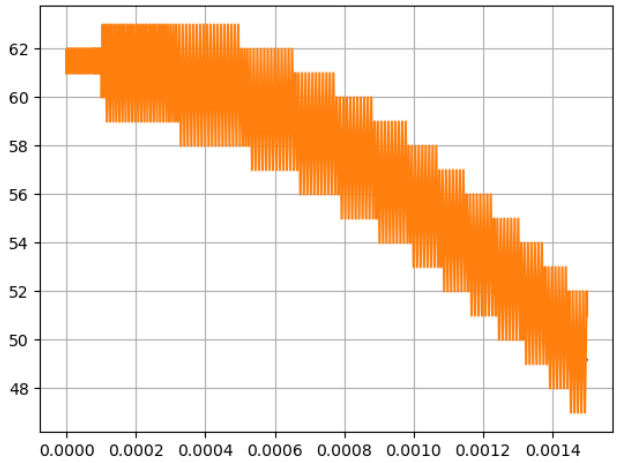


Figure 22: Carrier frequency: $999Hz, \beta = 100$

# 10 Closed Form Solution:

Let $N$ be the prediction horizon of the moving horizon implementation and $\mathbb{U}$ the quantisation levels. The $N$th power cartesian product of the set $\mathbb{U}$ for $N$ is defined as $\mathbb{U}^N := \mathbb{U} \ldots \times \ldots \mathbb{U}$. Let $n_{\mathbb{U}}$ represent the number of element in the set $\mathbb{U}$ i.e., number of quantisation levels, then the number of elements on the set $\mathbb{U}^N$ is $n_{\mathbb{U}}^N$. Also, let $v_i, i = \{1, 2, \ldots, n_{\mathbb{U}}^N\}$ represent the elements of the set $\mathbb{U}^N$. Then, the closed form solution of the moving horizon implementation of the optimisation problem (6)-(9) is

$$\mathbf{y}^*(k) = \Psi^{-1} q_{\tilde{\mathbb{U}}^N}\big[\Psi \mathbf{w}(k) + \Gamma x(k)\big] \tag{66}$$

where

$$\mathbf{y}^*(k) = \begin{bmatrix} y(k) \\ y(k+1) \\ \vdots \\ y(k+N-1) \end{bmatrix}, \quad \mathbf{w}(k) = \begin{bmatrix} w(k) \\ w(k+1) \\ \vdots \\ w(k+N-1) \end{bmatrix}, \quad \Gamma = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{N-1} \end{bmatrix}, \quad \Psi = \begin{bmatrix} h_0 & 0 & \cdots & 0 \\ h_1 & h_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ h_{N-1} & \ldots & h_1 & h_0 \end{bmatrix},$$

and $q_{\tilde{\mathbb{U}}^N}[.]$ represent the nearest neighbor vector quantisation (direct quantisation) and image of this mapping is the set

$$\tilde{\mathbb{U}}^N := \{\tilde{v}_1, \tilde{v}_2, \ldots, \tilde{v}_r\} \subset \mathbb{R}^N, \qquad r = n_{\mathbb{U}}^N,$$

with $\tilde{v}_i = \Psi v_i$ and $v_i \in \mathbb{U}^N = \{v_1, v_2, \ldots, v_{n_{\mathbb{U}}}\}$. The elements of the matrices $\Psi$, $\{h_0, h_1, \ldots, h_{N-1}\}$ are the impulse response of the filter $H(z) = 1 + C(zI - A)^{-1}B$ and $h_0 = 1$.

The moving horizon implementation of the above solution is

$$y(k) = [1, 0, \ldots, 0]\mathbf{y}^*(k) = [1, 0, \ldots, 0]\Psi^{-1} q_{\tilde{\mathbb{U}}^N}\big[\Psi \mathbf{w}(k) + \Gamma x(k)\big]. \tag{67}$$

Replacing $x(k) = (zI - A)^{-1}B(a(k) - u(k))$ and with further manipulation, we get,

$$y(k) = [1, 0, \ldots, 0]\Psi^{-1} q_{\tilde{\mathbb{U}}^N}\big[\Psi\big\{\mathcal{H}(z)\mathbf{w}(k) - \big[\mathcal{H}(z) - I\big]\mathbf{y}^*(k)\big\}\big]. \tag{68}$$

where

$$\mathcal{H}(z) = \begin{bmatrix} 1 + \mathcal{H}_1'(z) & 0 & \ldots & \ldots & 0 \\ \mathcal{H}_2'(z) & 1 & 0 & \ldots & 0 \\ \mathcal{H}_3'(z) & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \mathcal{H}_N'(z) & 0 & 0 & \ldots & 1 \end{bmatrix} \quad \text{and} \quad \mathcal{H}_i'(z) = \big[\Psi^{-1}\big]_{i,*}\Gamma(zI - A)^{-1}B \tag{69}$$

and $\big[\Psi^{-1}\big]_{i,*}$ denotes the $i$ th row of $\Psi^{-1}$.

## 10.1 Prediction horizon $N = 1$:

When the prediction horizon is $N = 1$, $\Psi = 1$ since $h_0 = 1$ and

$$\mathcal{H}(z) = 1 + \mathcal{H}_1'(z) = 1 + C(zI - A)^{-1}B = H(z)$$

and the moving horizon implementation is

$$y(k) = q_{\tilde{\mathbb{U}}^N}\big[H(z)w(k) - \big[H(z) - I\big]y^*(k)\big]. \tag{70}$$

and $q_{\tilde{\mathbb{U}}^N}[.] = q_{\mathbb{U}^N}[.]$ for $\Psi = 1$ and it becomes scalar (direct) quantisation.

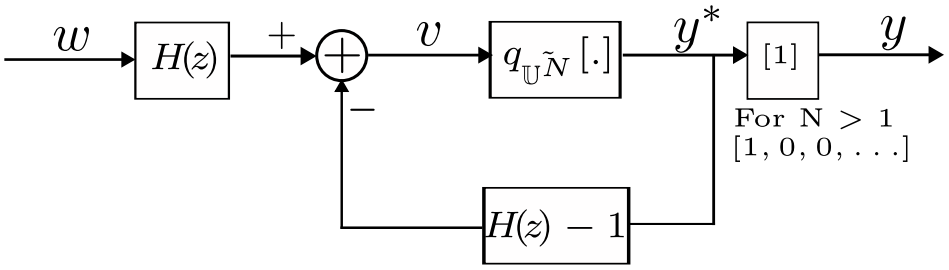


Figure 23: NSD equivalent of MHOQ implementation for $N = 1$.

### 10.2 Prediction horizon $N = 2$:

When the prediction horizon is $N = 1$, $\Psi = \begin{bmatrix} h_0 & 0 \\ h_1 & h_0 \end{bmatrix}$ and thus $\Psi^{-1} = \frac{1}{h_0^2}\begin{bmatrix} h_0 & -h_1 \\ 0 & h_0 \end{bmatrix} = \begin{bmatrix} 1 & -h_1 \\ 0 & 1 \end{bmatrix}$ since $h_0 = 1$. We get,

$$\mathcal{H}_1'(z) = \begin{bmatrix} 1 & -h1 \end{bmatrix} \Gamma(zI - A)^{-1}B = (C - h_1CA)(zI - A)^{-1}B = H(z) - h_1CA(zI - A)^{-1}B - 1$$
$$\mathcal{H}_2'(z) = \begin{bmatrix} 0 & h_0 \end{bmatrix} \Gamma(zI - A)^{-1}B = CA(zI - A)^{-1}B$$

Then

$$\mathcal{H}(z) = \begin{bmatrix} 1 + \mathcal{H}_1'(z) & 0 \\ \mathcal{H}_2'(z) & 1 \end{bmatrix} = \begin{bmatrix} 1 + \mathcal{H}_1'(z) & 0 \\ \mathcal{H}_2'(z) & 1 \end{bmatrix} = \begin{bmatrix} H(z) - h_1CA(zI - A)^{-1}B & 0 \\ CA(zI - A)^{-1}B & 1 \end{bmatrix}$$

and the moving horizon implementation is

$$\begin{aligned}
y(k) &= [1,0]\Psi^{-1} q_{\tilde{\mathbb{U}}N}\left[\Psi\left\{\mathcal{H}(z)\mathbf{w}(k) - \left[\mathcal{H}(z) - I\right]\mathbf{y}^*(k)\right\}\right] \\
&= [1,0]\Psi^{-1} q_{\tilde{\mathbb{U}}N}\left[\Psi\mathcal{H}(z)\mathbf{w}(k) - \Psi\left[\mathcal{H}(z) - I\right]\mathbf{y}^*(k)\right] \\
&= [1,0]\Psi^{-1} q_{\tilde{\mathbb{U}}N}\left[\Psi\mathcal{H}(z)\big(\mathbf{w}(k) - \mathbf{y}^*(k)\big) + \Psi\mathbf{y}^*(k)\right] \\
&= [1,0]\Psi^{-1} q_{\tilde{\mathbb{U}}N}\left[\begin{bmatrix} H(z) - h_1CA(zI - A)^{-1}B & 0 \\ h_1H(z) - (h_1^2 - 1)CA(zI - A)^{-1}B & 1 \end{bmatrix}\mathbf{w}(k) - \begin{bmatrix} H(z) - 1 - h_1CA(zI - A)^{-1}B & 0 \\ h_1(H(z) - 1) - (h_1^2 - 1)CA(zI - A)^{-1}B & 0 \end{bmatrix}\mathbf{y}^*(k)\right] \\
&= [1,0]\Psi^{-1} q_{\tilde{\mathbb{U}}N}\left[\begin{bmatrix} H(z) & 0 \\ 0 & 0 \end{bmatrix}\mathbf{w}(k) - \begin{bmatrix} H(z) - 1 & 0 \\ 0 & 0 \end{bmatrix}\mathbf{y}^*(k)\right. \\
&\quad + \left.\begin{bmatrix} -h_1CA(zI - A)^{-1}B & 0 \\ h_1H(z) - (h_1^2 - 1)CA(zI - A)^{-1}B & 1 \end{bmatrix}\mathbf{w}(k) - \begin{bmatrix} -h_1CA(zI - A)^{-1}B & 0 \\ h_1(H(z) - 1) - (h_1^2 - 1)CA(zI - A)^{-1}B & 0 \end{bmatrix}\mathbf{y}^*(k)\right]
\end{aligned}$$

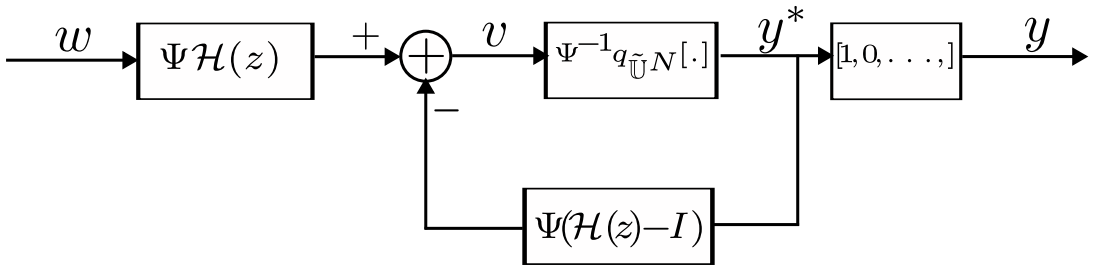


Figure 24: NSD equivalent of MHOQ implementation for $N > 2$.

# References


[1] Shuichi Ohno and M. Rizwan 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.

[2] Izumi Masubuchi, Atsumi Ohara, and Nobuhide Suda. Lmi-based controller synthesis: A unified formulation and solution. *International Journal of Robust and Nonlinear Control*, 8(8):669–686, 1998.

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


# 11 Scribble

We want to minimise this error: $\bar{e} = Hy - Hw = Hv + H\epsilon - Hw$
$\bar{e} = Hy - Hw = Hv + H\epsilon - Hw$
What is $\bar{e}$ when

- the feedback filter is a (double) delay (standard delta-sigma)
- the feedback filter is optimal
- the feedback filter is implemented in the MPC formulation with horizon length 1, 2, 3, ...

Things to study

done Need to handle saturation (no need as optimal gain appears to be 3)

done What happens when there is a constant noise floor, unaffected by the noise-shaping (no need as small headroom is sufficient for gain of 3)

- Can we limit the rate
- Adding in INL

Total noise and distortion:

$$\sigma_t = \sigma_n + \sigma_q$$

$$\frac{\sigma_s}{\sigma_n + \sigma_q}$$