Optimised Spectral Density Shaping of Quantisation Error Using Adaptive Dithering

Ahmad Faza, John Leth, and Arnfinn A. Eielsen

***Abstract*— A method for shaping the power spectral density (PSD) of the total error due to uniform quantisation is proposed. It utilises non-subtractive dithering, generated using a joint specification of the probability density function (PDF) and the PSD by way of stochastic minimisation (SM). The output of the quantiser can be made linear and continuous in the mean and the variance of the error can be made independent of the quantiser input by using a dither with a triangular PDF. However, shaping the error PSD to a desired form for reconstruction at the quantiser output remains input dependent. An adaptive dithering approach is implemented to address this dependency. By exploiting symmetry properties of uniform quantisation, it is possible to use SM to generate a limited number of dither sequences and reuse them to shape the total error PSD for arbitrary inputs. The approach is implemented using a look-up table (LUT). When optimised for reconstruction filtering, simulation results demonstrate an improved PSD shaping performance over the state-of-the-art feed-forward method of over two orders of magnitude within a given bandwidth.**

I. INTRODUCTION

Quantisation and re-quantisation are fundamental operations in digital signal processing, digital-analogue conversion, power electronics and measurement systems. Error is introduced since only a discrete subset of values can be represented [1,2].

Several methods exist to shape the power spectral density (PSD) of error due to quantisation. Combined with oversampling, PSD shaping reduces effective quantisation error by concentrating error power into specific frequency ranges for subsequent filtering. $\Delta\Sigma$-modulation achieves this by using quantiser feedback to shape the output PSD [3], but the discontinuous nature induces chaotic behaviour with input-dependent and empirically derived stability [4]. Model predictive control (MPC) offers higher performance and rigorous stability guarantees [5,6], though at the cost of significant computational effort.

Dithering is a feed-forward method that enables similar PSD shaping effects. As it is not reliant on feedback, it can be precomputed and provide guaranteed stability and determinism. An external signal is added to the input of the quantiser, and the resulting quantiser error can have a spectral distribution [7]–[9]. Dithering has been shown to mitigate the effect of static non-linearities such as element mismatch in digital-analogue converters [10] as well as dynamic non-linearities [11,12] such as glitches in digital-analogue converters [13]–[15]. The dither probability density function (PDF) is known to modify the characteristics of a non-linear system [16,17]. For uniform quantisation systems, the expected value of the quantiser output can be linearised and the variance of the error (the error power) can be made independent of the quantiser input using a dither of triangular PDF (TPDF) [7]. Hence, it is beneficial to generate dither of jointly specified PDF and PSD criteria for purposes of shaping the total error PSD via non-subtractive dither (NSD). This can be done by passing an initially white Gaussian noise through a linear noise colouring filter (PSD shaping) and a static non-linear transform (PDF shaping) as in [18,19]. This approach had been adapted analysed for uniform quantisation systems with arbitrary input signals [9]; where shaping the error PSD at the quantiser output was shown to be generally input dependent for any dither PDF choice of bounded variance. Another approach is to first generate an initially white noise with the desired PDF and then apply an iterative shuffling procedure (PSD shaping) until the shuffled dither sequence PSD approximates the desired one up to a defined margin [20,21]. In this paper, we study using NSD of TPDF generated via stochastic minimisation (SM) to shape the total error PSD. SM utilises the shuffling approach; where two elements of the dither sequence at the quantiser input are stochastically interchanged such that an adapted metric, defined to attain the desired PSD at the quantiser output, is minimised.

A. Contributions

Results from [7] and the SM method from [20] are used to provide a novel solution to shape the PSD of the total error due to uniform quantisation via adaptive non-subtractive dithering. The approach addresses the inherent dependency of shaping the error PSD on the quantiser input signal. SM is adapted to generate dither sequences attaining PSDs for optimal signal reconstruction given subsequent filtering. By uniformly segmenting a single quantisation level into several regions, SM computes dither sequences for a limited number of input values represented by the equidistant centres of said regions. These sequences are saved in a look-up table (LUT); where they are adaptively streamed each time the input value is best approximated by a centre value indexed in the LUT. Exploiting the symmetry of uniform quantisation, the use of an amplitude folding index function allows reusing the LUT to shape the error PSD for arbitrary input signals. Simulations show that, independent of the quantiser input, significantly improved error PSD shaping can be achieved compared to the PSDs realisable using the method in [9].

Ahmad Faza' (ahmad.faza@uis.no), Arnfinn A. Eielsen (eielsen@ux.uis.no) are with the Dept. of Energy and Petroleum Engineering, University of Stavanger, Norway

J. Leth (jjl@es.aau.dk) is with the Dept. of Electronic Systems, Automation & Control, Aalborg University, Denmark

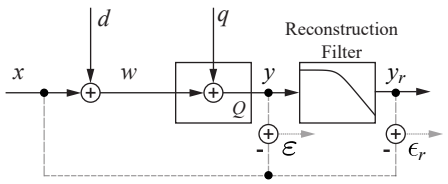

Fig. 1: Non-subtractively dithered quantiser with subsequent filtering, with input x, dither d, quantiser input w, quantiser output y, quantisation error q, total quantisation error ε, reconstructed output y_r and residual error ϵ_r.

II. ANALYTICAL FRAMEWORK

A. Uniform Quantisation

A uniform quantiser, Q in Fig. 1, with step-size $\delta \in \mathbb{R}_{>0}$ and mid-tread behaviour can be defined as $Q_\perp(w) \triangleq \delta \left\lfloor \frac{w}{\delta} + \frac{1}{2} \right\rfloor$; or with mid-rise behaviour and step-size $\Delta \in \mathbb{R}_{>0}$ which is defined as $Q_\top(w) \triangleq \Delta \left(\left\lfloor \frac{w}{\Delta} \right\rfloor + \frac{1}{2} \right)$ where $\lfloor \cdot \rfloor$ denotes the floor operator. The quantisation error $q(w)$ given an input w is defined as the function

$$q(w) \triangleq Q(w) - w \,. \tag{1}$$

The output can then be modelled as:

$$y = w + q(w) = Q(w) \,. \tag{2}$$

If we consider a mid-rise multi-bit quantiser with a word-size of B bits, it has 2^B output levels. With an output range restricted between V_{min} and V_{max} the step-size is

$$\Delta = \frac{V_{\text{max}} - V_{\text{min}}}{2^B - 1} \,, \tag{3}$$

and in this case $Q(w)$ can be expressed as

$$Q_\top(w) = \Delta \sum_{i=1}^{N_T} \left(\Gamma(w - T_i) - \frac{1}{2} \right) \,, \tag{4}$$

where $N_T = 2^B - 1$ is the number of quantisation levels, and the step-functions have the thresholds $T_i : i \in \{1, 2, 3, ...N_T\}$; where, $T_j = (j - i)\Delta + T_i$ for $j > i$ and $T_{2^{B-1}} = (V_{\text{max}} + V_{\text{min}})/2$. Where the Heaviside step-function denoted $\Gamma(u)$ is defined as

$$\Gamma(u) \triangleq \begin{cases} 0 \,, & u \leq 0 \\ 1 \,, & u > 0 \end{cases} \,. \tag{5}$$

B. Error PSD Shaping via Non-subtractive Dithering (NSD)

Consider the quantiser configuration in Fig. 1. The total error ε is defined as the difference between the output y and input x, $\varepsilon \triangleq y - x$, to distinguish it from the quantisation error q in (1). For NSD, $\varepsilon = Q(x + d) - x = q(x + d) + d$.

For given specifications of the PSD $S_\varepsilon(\omega)$ of the total error ε it is generally desired to generate a dither d that both linearises the uniform quantiser in mean, that is, $E[y]$ becomes a linear function of x, and induces a PSD $S_\varepsilon(\omega)$ having the required specifications.

As for the linearisation, first note that $E[y] = x + E[\varepsilon]$. Now from [7], we know that choosing a zero mean dither with a TPDF will result in the total error ε being a zero mean ($E[\varepsilon] = 0$) stationary process with $E[\varepsilon^2] = \Delta^2/4$. Hence, $E[y] = x$ (quantiser linearisation can be achieved), and $\text{Var}(y) = E[\varepsilon^2]$ is constant and independent of the value of x (noise modulation effect can be mitigated). Furthermore, it follows from the choice of a TPDF dither, that the PSD S_y of the output y is just the sum of the PSD S_ε of the total error ε and the PSD S_x of the input x [7,8].

When it comes to spectrally shaping ε, it is ideally desirable to attain S_ε that has minimal power content at frequencies where S_x has significant power content. For example, the reconstructed output y_r after the reconstruction filter in Fig. 1 contains the low-pass power content of y where S_x is concentrated; therefore, it is advantageous to shape S_ε so that it has a high-pass power content away from S_x and optimised for attenuation by the subsequent reconstruction filter ($y_r \approx x$).

The best error power spectral shaping performance achievable for uniform quantisation using TPDF NSD is presented in [9]. While analytical derivations illustrate how the shaping of S_ε depends on the input signal in uniform quantisation systems employing NSD, it is demonstrated that selecting a TPDF dither and designing the colouring filter to generate d as if x were fixed at a quantisation threshold T_i effectively decouples the filter synthesis step (necessary for shaping S_ε) from the dependence on x. However, this nonlinear filtering method (NFM) imposes inherent restrictions on the set of realisable PSDs. For applications that aim to achieve S_ε optimised for minimising the total power of ϵ_r ($\text{Var}(\epsilon_r) = E[\epsilon_r^2]$) at the output of a specific reconstruction filter, these limitations are shown to be quite restrictive.

C. Reconstruction Filters

Low-pass filtering subsequent to a dithered uniform quantiser is a fundamental step often used to reduce repeated spectra due to sampling, and tends to reduce quantisation error improving the fidelity of the reconstructed signal. E.g. for Digital-to-Analog Converters (DACs) applications, the quantised signal is often sampled at a high rate where high-frequency noise from quantisation folds back (aliases) into the base-band, introducing artefacts. Using a reconstruction filter (often a low-pass filter) as shown in Fig. 1, ensures that only the desired signal is retained, preventing aliasing. In [22], a $\Delta\Sigma$-modulation framework uses feedback from the output of the uniform quantiser (noise shaping filter driven by q) to generate d that optimally shapes S_ε to minimise $\text{Var}(\epsilon_r)$ for a given infinite impulse response (IIR) reconstruction filter $H(z)$. In this work we consider r-th order Butterworth filters of the form:

$$H(z) = \frac{1}{1 + \left(\frac{\omega_z}{\pi \omega_c} \frac{1 - z^{-1}}{1 + z^{-1}} \right)^{2r}} \,, \tag{6}$$

where ω_c is the cut-off angular frequency, often designed to match the bandwidth of the original signal to retain as much useful information as possible and ω_z is the throughput angular frequency. As a benchmark to evaluate the performance of the method proposed in Sec. III, we use [9] to produce a TPDF d with the target specification being the optimal S_ε to minimise $\text{Var}(\epsilon_r)$ for a given $H(z)$ as formulated by [22].

III. Stochastically Minimised Adaptive Dithering

A. Dither Generation via Stochastic Minimisation (SM)

The method in [20] uses an iterative shuffling approach to generate random sequences with jointly specified PSD and PDF. Starting with an initial sequence of desired PDF, two sequence indices are randomly chosen and then the sequence values associated with these indices are interchanged in each iteration. By advancing only the interchanges that minimise an error metric to a target PSD, this PSD-shaping PDF-preserving shuffling is repeated until the target PSD is attained. In this work, the metric is adapted to minimise $\mathrm{Var}(\epsilon_r)$ for a given IIR reconstruction filter $H(z)$ as discussed in Sec. II-B. Hence, by stochastically shuffling an initial spectrally white TPDF dither sequence d_0 to minimise $\mathrm{Var}(\epsilon_r)$ for a fixed $x = u$, SM (in Algorithm 1) returns the dither sequence d optimised for that u. Note how given a fixed $x = u$ and $H(z)$, SM indirectly obtains S_ε optimised for subsequent filtering.

Due to the symmetry of the uniform quantisation (e.g. $Q_\top(\cdot)$ in (4)), note how for all $x_i, x_j \in [V_{min} + \Delta, V_{max} - \Delta]$: $x_i \neq x_j$ that if $(x_i - Q_\top(x_i)) = (x_j - Q_\top(x_j))$ then $\varepsilon_i = \varepsilon_j$ for any TPDF dither sequence $d(:)$. Exploiting this property allows for reduced computation time, since the number of dither sequences required for generation via SM can be limited. By uniformly decimating the Δ-wide quantisation step into a finer grid of $M \in \mathbb{Z}_{>0}$ bands, the δ spaced band centres ($\Delta = M\delta$) can be used to approximate $(x - Q_\top(x)) \in [-\Delta/2, \Delta/2]$ into $Q_\perp(x - Q_\top(x))$ which has a space of only M values $\delta\{-M/2...M/2\}$ (if M is even) or $\delta\{(1-M)/2...(M+1)/2\}$ (if M is odd). This approximation is equivalent to representing x by $Q_\perp(x)$ (δ-wide quantisation step). Consequently, SM generates M unique dither sequences $d_k(:) : k \in \{1, 2...M\}$, optimised for a given $H(z)$, and saves them to a LUT.

This is detailed in (Algorithm 1), where an added selection refinement step primes d_0 (an initial TPDF sequence of length L) to ensure that $E[Q(u + d_0) - u] \approx 0$ for all $u = Q_\perp(x - Q_\top(x))$. This step mitigates noise modulation artefacts in the reference base-band.

B. Adaptive Dither Streaming

A block diagram for the proposed SMAD method is shown in Fig. 2. By approximating $x(t) \in [V_{min} + \Delta, V_{max} - \Delta]$ with the sampled finely quantised $Q_\perp(x(nT_s))$ ($n \in \mathbb{Z}_{\geq 0}$, $T_s = \frac{2\pi}{\omega_s}$ is the sampling period), the LUT of optimised dither sequences from Sec. III-A can be utilised to shape S_ε for an arbitrary input signal. Provided a sufficiently fine grid is considered ($M > 10$) and adequate reference oversampling is adopted ($T_s \leq \frac{\min(T_{ref})}{1000}$; $\frac{2\pi}{\min(T_{ref})}$ is the highest frequency component in the reference band), $Q_\perp(x(nT_s)) \approx x(t)$. This can be achieved by introducing an amplitude folding LUT indexing function as follows:

$$\digamma_M(u) \triangleq \left(\frac{Q_\perp(u - Q_\top(u))}{\delta} \bmod M \right) + 1 \qquad (7)$$

where $u \bmod M \triangleq u - M \left\lfloor \frac{u}{M} \right\rfloor$. The index choice of the k-th dither sequence d_k to be streamed from the LUT to the

Algorithm 1 LUT Generation via SM, MATLAB

```
B = 5; % Number of quantisation bits
V_max=10; V_min=-10; % Dynamic range - upper/lower limit
LSB = V_max-V_min/(2^B-1); % Least Significant Bit (Delta)
Th = V_min:LSB:V_max; % Quantised Output Levels (2^B)
M = 64; % Number of finer grid bands (Delta = M*delta)
v_delta = Th(2^(B-1)):LSB/M:Th(2^(B-1)+1);%delta-grid vals
TPDF=makedist('Triangular','A',-LSB,'B',0,'C',LSB);%(TPDF)
L = 1e3; % Streamed Dither Sequence Length
d_Mid = random(TPDF, 1, L); % Initial Dither Sequence
LUT = zeros(M,L); % Look-Up-Table
%% IIR Butterworth filter H(z) Synthesis
Fc=1e3;Fz=1e6; % cutoff/ throughput frequency
r = 3; % Filter order
[S_num,S_den]=butter(r,Fc/(Fz/2),"low");% H(z) Synthesis
%% Initialise SM-Alg Parameters
alpha=1e-4; beta=5e-3; NUM=39e1;
for k=1:M
%% Refine the selection of d_0 (best realisation out of L)
    d_0 = d_Mid;
    for i=1:L
    d_Mid = random(TPDF, 1, 1e3);
        if(abs(mean(v_delta(k)-Q(d_Mid+v_delta(k),V_min,
             V_max,B)))<abs(mean(v_delta(k)-Q(d_0+v_delta(
             k),V_min,V_max,B))))
        % See Q() in Eq(4)-Sec.II-A
            d_0 = d_Mid;
        end
    end
%% Stochastic Minimisation (Algorithm 1) for u=v_delta(k)
    d_SM = SM(d_0,v_delta(k),S_num,S_den,alpha,beta,NUM,
         V_min,V_max,B);
    LUT(k,:) = d_SM; %Save d_SM optimised for u=v_delta(k)
end
%% Generation of a Dither Sequence using Stochastic
     Minimization (SM)
function d = SM(d0, u, s_num, s_denum, alpha, beta, NUM,
     Vmin,Vmax,Bit)
    %% Inputs:
    % d0 - Initial sequence of desired PDF (e.g., TPDF)
    % u - Input level
    % alpha - Target normalized filtered error variance (0
         < alpha <= 1)
    % beta - Marginal decrease ratio (0 < beta <= 1)
    % NUM - Number of consecutive interchanges with
         marginal decrease ratio < beta before termination
    % s_num, s_denum - IIR H(z) (numerator, denumerator)
    %% Output:
    % d - Target sequence where SS1 < alpha
    %% Initialize variables
    SS1 = 1; d1 = d0; Var_d0=var(filter(s_num,s_denum,Q(d0
         +u,Vmin,Vmax,Bit)-u));
    SS = SS1; SuccessCount = 0; TerminateCount = 0;
    while (SS1 >= alpha) && (TerminateCount < NUM)
        SS1 = var(filter(s_num,s_denum,Q(d1+u,Vmin,Vmax,
             Bit)-u))/Var_d0;
        % Step 1: Randomly choose two indices for
             interchange
        idx = randperm(length(d1), 2);
        j1 = idx(1); j2 = idx(2);
        % Perform interchange while preserving PDF
        d2 = d1; d2([j1, j2]) = d2([j2, j1]);
        SS2 = var(filter(s_num,s_denum,Q(d2+u,Vmin,Vmax,
             Bit)-u))/Var_d0;
        % Step 2: Commit to the interchange if it
             minimizes SS1
        if (SS2 < SS1)
            d1 = d2; SS = [SS; SS2]; SS1 = SS2;
            SuccessCount = SuccessCount + 1;
        end
        % Step 3: Check termination condition
        if (SuccessCount > NUM)
            TerminateCount = 0;
            for i = NUM:-1:1
                if ((SS(end-i) - SS(end-i+1)) / SS(end-i)
                     < beta)
                    TerminateCount = TerminateCount + 1;
                end
            end
        end
    end
    d = d1; % Return the optimized dither sequence
end
```

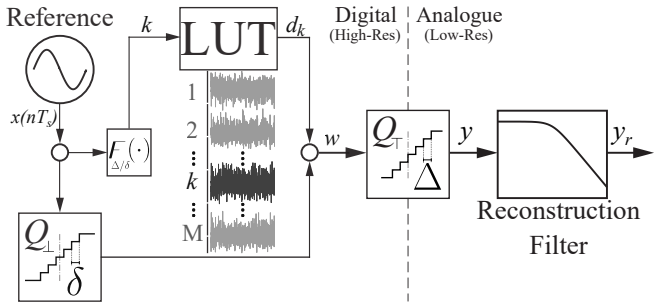

Fig. 2: Stochastically minimised adaptive dithering (SMAD) approach: sampled input $x(nT_s)$, input dependent LUT index k, streamed SM dither d_k, quantiser input w, quantiser output y and reconstructed output y_r.

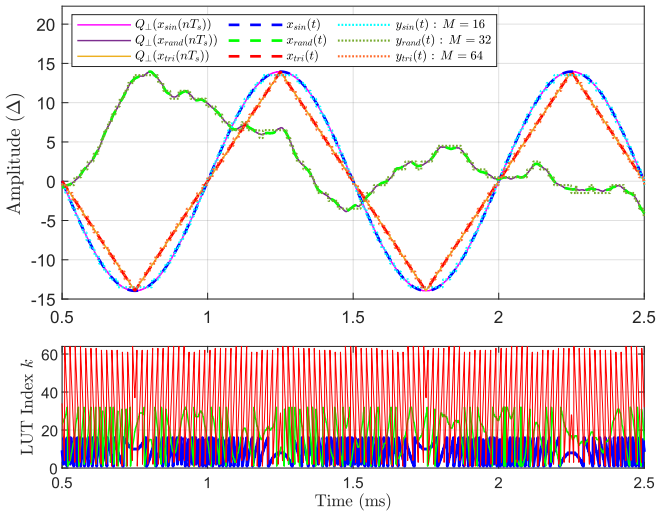

Fig. 3: Operating SMAD to reconstruct recorded signals: **A)** Various analogue realisations of 1 kHz band-limited input signals (sinusoidal, triangular, and an arbitrary random walk) $x(t) \in [V_{min} + \Delta, V_{max} - \Delta]$. Their digital recordings $Q_\perp(x(nT_s))$ when over-sampled at 1 MHz and finely quantised by a (9, 10, and 11)-Bit uniform quantiser, respectively. Re-quantised for reconstruction by a reduced resolution 5-Bit quantiser $\approx Q_\top(x(t))$. **B)** The LUT index $k = F_M(x(nT_s))$ (F_M from (7)) to stream d_k (in (8)) corresponding to $x(nT_s)$.

quantiser input w is adapted such that it is optimised for the current value of the sampled reference; $k = F_M(x(nT_s))$. Accordingly:

$$w\left(\left(n:\frac{1}{L}:n+\frac{(L-1)}{L}\right)T_s\right) = Q_\perp(x(nT_s))+d_k(1:L) \tag{8}$$

Note how this adaptive streaming approach effectively necessitates an angular frequency throughput capability of $\omega_z = L\omega_s$ (a minimum quantiser output update rate requirement).

IV. Simulations

In the simulations, a mid-rise 5-bit uniform quantiser of the form in (4) is used (see Fig. 2) where the output range is set to $V_{max} = -V_{min} = 10$. Fig. 3 illustrates how SMAD

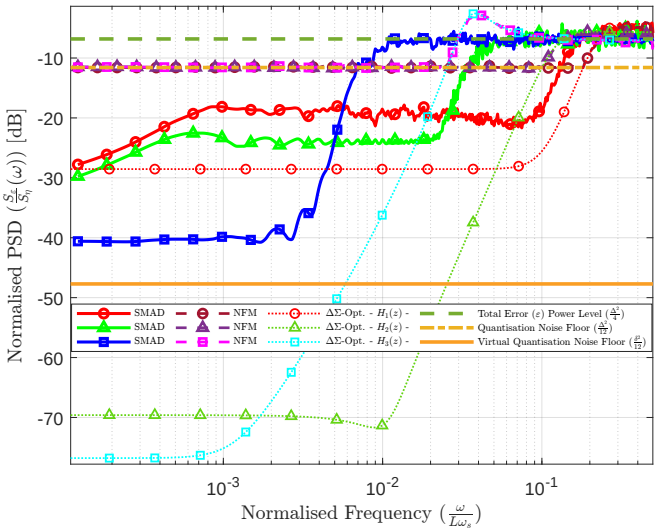

Fig. 4: Normalised single-sided total error PSD $S_\varepsilon(\omega)/S_\eta$: SMAD vs. NFM and the theoretical target $\Delta\Sigma$-Opt. PSDs (required spectral specification necessary to use NFM) as calculated in [22] for the filters (H_1, H_2, H_3).

TABLE I: Summary of simulation results NFM vs SMAD.

Case	Parameters		Filters		Method [**Var**(ϵ_r)/**Var**(ε)]	
#	**L**	**M**	**Shaping**	**Recon.**	**NFM**	**SMAD**
1	1000	64	H_1	H_1	3.49×10^{-2}	8.07×10^{-3}
2	1000	64	H_1	R	3.50×10^{-6}	1.22×10^{-10}
3	1000	32	H_1	H_1	3.49×10^{-2}	8.86×10^{-3}
4	1000	32	H_1	R	4.83×10^{-6}	9.19×10^{-10}
5	1000	64	H_2	H_2	3.54×10^{-3}	2.13×10^{-4}
6	1000	64	H_2	R	3.27×10^{-6}	4.45×10^{-9}
7	1000	16	H_2	H_2	3.55×10^{-3}	2.17×10^{-4}
8	1000	16	H_2	R	3.29×10^{-6}	5.45×10^{-9}
9	1000	64	H_3	H_3	3.54×10^{-4}	3.64×10^{-6}
10	1000	64	H_3	R	4.64×10^{-6}	3.86×10^{-9}
11	300	64	H_3	H_3	3.64×10^{-4}	2.97×10^{-6}
12	300	64	H_3	R	3.25×10^{-6}	2.60×10^{-9}
13	300	4	H_2	H_2	3.53×10^{-3}	1.23×10^{-4}
14	300	4	H_2	R	3.41×10^{-6}	2.28×10^{-11}

can be adapted to reconstruct arbitrary reference signals for various choices of the finer grid ($\Delta = M\delta$) as discussed in Sec. III-A by adapting F_M (from (7)) accordingly, as discussed in Sec. III-B. Let $\eta(t)$ be a white, unity variance Gaussian process sampled at $L\omega_s$ rad/s (i.e. its single-sided PSD $S_\eta(\omega) = 2\pi/L\omega_s$); then the normalised single-sided PSD responses $S_\varepsilon/S_\eta(\omega)$ for various $x(t)$ (from Fig. 3-A) are shown in Fig. 4. The reference records (sampled at $\omega_s = 2\pi$ Mrad/s, with $M = 64$) are dithered (NFM vs. SMAD) prior to reconstruction where each record utilises a LUT ($L = 1 \times 10^3$) optimised for filters of the form in (6): H_1, H_2, H_3 with $r = 3$, $\omega_z = L\omega_s$ and $\omega_c/L\omega_s = 10^{-1}, 10^{-2}, 10^{-3}$; respectively. Tab. I provides case simulations of these reference signals designed to investigate the effect of varying parameters L, M, and the filtering choice for spectral shaping and reconstruction. Note how since all $x(t) \in [V_{min} + \Delta, V_{max} - \Delta]$ considered here are 1 kHz band-limited, a reconstruction filter R with $r = 3$, $\omega_z = L\omega_s$ and $\omega_c/L\omega_s = 10^{-5}$ is suitable to restore the base-band content. However, seeing how SMAD requires

a significantly larger bandwidth for operation (due to both "true" ($T_s \leq \min(T_{\text{ref}})/1000$) and artificial ($\omega_z = L\omega_s$) oversampling requirements), using SMAD "as if" H_1, H_2, H_3 is used for reconstruction, while it is R in reality gives rise to the opportunity of shaping the error content beyond the reference base-band as well. This may be interesting for applications where minimising the noise content even at certain frequency bands higher than reference base-band is beneficial.

V. Results and Discussion

Note that all subsequent discussions and comparisons are viewed in light of the approximation $\hat{x} \triangleq Q_\perp(x(nT_s)) \approx x(t)$ inherent to the operation of SMAD as explained in Sec. III-B (i.e. $\varepsilon = y - \hat{x}$). This shifts the appropriateness of using SMAD over NFM to applications where the reference signal is already recorded digitally (discretised) or applications where replacing x with $\hat{x}$ is either inevitable or an acceptable compromise in the application context. Otherwise, it is evident that given an output update rate of ω_z, the available sampling budget to use SMAD is restricted to ω_z/L whereas it is $\approx \omega_z$ for NFM. This has significance in terms of physical limitations to sampling-rate in applications.

Fig. 4 illustrates how (for H_1, H_2, H_3) NFM fails to meet its target for all ω where the $\Delta\Sigma$-Opt.-specification $S_\varepsilon/S_\eta(\omega) < \Delta^2/12$. However, SMAD performance is not bounded by the quantisation noise floor associated with the resolution reduction at the interface (SMAD achieves almost three orders of magnitude of improved attenuation in the case of H_3 within the filter's pass-band compared to NFM).

From Tab. I, note how $\text{Var}(\epsilon_r)/\text{Var}(\varepsilon)$ is proportional to $\omega_c/L\omega_s$ for any given choice of the reconstruction filter using both methods (this is expected since $\text{Var}(\varepsilon)$ is constant and the residual noise should be proportional to the filter pass-band). However, factoring out this aspect (normalising by $\omega_c/L\omega_s$) in cases 1, 5, and 9 indicates that SMAD exhibits improved attenuation with narrower filter pass-bands while NFM performance is bounded to $\approx 33\%$ (i.e. $\approx 33\%$ of the total error power $\Delta^2/4$ which is the quantisation noise floor $\Delta^2/12$). This is in fact corroborated by the results shown in Fig. 4. Comparing cases 9-12 (similar $\hat{x}$) suggests that reducing L (from 1000 to 300) does not deteriorate SMAD's performance (this is advantageous as it improves spectrum utilisation). Unfortunately, choosing $L < 300$ will deteriorate the performance (from attempts unreported in the table). Now considering the effect of varying M (e.g. cases 1-4 or 5-8), one should have in mind that this by default changes $\hat{x}$. Hence, it can be interpreted as shifting the virtual quantisation noise floor level $\delta^2/12$ (highlighted in Fig. 4 for $M = 64$) reflecting the new SMAD performance upper bound. E.g. the SMAD performance in case 2 means that the 5-bit quantised, R-filtered output ($R(z)y$) will be perceived as if it were reconstructed by a 10-bit quantiser ($y_r \approx \hat{x}$).

VI. Conclusions

It was demonstrated by way of a numerical algorithm that the proposed adaptive dithering method can, for arbitrary references, shape the spectral distribution of the total error power for a uniform quantiser. In practice, the method could be limited by the maximum available switching frequency in a hardware implementation. However, for applications of reconstructing recorded signals with high perceived fidelity, the method provides significantly improved performance compared to existing non-subtractive dithering methods.

References

[1] R. E. Crochiere and L. Rabiner, "Interpolation and decimation of digital signals—A tutorial review," in *Proc. IEEE*. IEEE, 1981, pp. 300–331.

[2] B. Widrow, I. Kollar, and M.-C. Liu, "Statistical Theory of Quantization," *IEEE Trans. Instrum. Meas.*, vol. 45, no. 2, pp. 353–361, 1996.

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

[4] M. Neitola, "Lee's Rule Extended," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 64, no. 4, pp. 382–386, 2017.

[5] G. C. Goodwin, D. E. Quevedo, and D. McGrath, "Moving-horizon optimal quantizer for audio signals," *Journal of the Audio Engineering Society*, vol. 51, no. 3, pp. 138–149, 2003.

[6] B. Adhikari, R. van der Rots, J. Leth, and A. Eielsen, "Linearisation of digital-to-analog converters by model predictive control," in *IFAC Conference on Nonlinear Model Predictive Control*, 2024.

[7] R. A. Wannamaker, S. Lipshitz, J. Vanderkooy, and J. N. Wright, "A Theory of Nonsubtractive Dither," *IEEE Trans. Signal Process.*, vol. 48, no. 2, pp. 499–516, 2000.

[8] B. Widrow and I. Kollar, *Quantization noise roundoff error in digital computation, signal processing, control, and communications*. Cambridge University Press, 2008.

[9] A. Faza, J. Leth, and A. A. Eielsen, "Spectral density shaping of quantisation error using dithering," in *2024 32nd European Signal Processing Conference (EUSIPCO)*. IEEE, 2024, pp. 2582–2586.

[10] A. A. Eielsen and A. J. Fleming, "Improving Digital-to-Analog Converter Linearity by Large High-Frequency Dithering," *IEEE Trans. Circuits Syst. I*, vol. 64, no. 6, pp. 1409–1420, 2017.

[11] C. A. Desoer and S. M. Shahruz, "Stability of dithered non-linear systems with backlash or hysteresis," *International Journal of Control*, vol. 43, no. 4, pp. 1045–1060, 1986.

[12] B. Armstrong-Hélouvry, P. Dupont, and C. C. De Wit, "A survey of models, analysis tools and compensation methods for the control of machines with friction," *Automatica*, vol. 30, no. 7, 1994.

[13] A. A. Eielsen, J. Leth, A. J. Fleming, A. G. Wills, and B. Ninness, "Large-amplitude dithering mitigates glitches in digital-to-analogue converters," *IEEE Transactions on Signal Processing*, vol. 68, pp. 1950–1963, 2020.

[14] A. Faza, J. Leth, and A. A. Eielsen, "Mitigating non-linear dac glitches using dither in closed-loop nano-positioning applications," in *2023 American Control Conference (ACC)*. IEEE, 2023, pp. 685–691.

[15] ——, "Criterion for sufficiently large dither amplitude to mitigate non-linear glitches," in *IEEE Conference on Control Technology and Applications*. IEEE, 2023, pp. 970–977.

[16] G. Zames and N. A. Shneydor, "Dither in Nonlinear Systems," *IEEE Trans. Autom. Control*, vol. 21, no. 5, pp. 660–667, 1976.

[17] L. Iannelli, K. H. Johansson, U. T. Jönsson, and F. Vasca, "Averaging of nonsmooth systems using dither," *Automatica*, vol. 42, no. 4, 2006.

[18] M. M. Sondhi, "Random processes with specified spectral density and first-order probability density," *Bell System Technical Journal*, vol. 62, no. 3, pp. 679–701, 1983.

[19] U. Gujar and R. Kavanagh, "Generation of random signals with specified probability density functions and power density spectra," *IEEE Transactions on Automatic Control*, vol. 13, no. 6, pp. 716–719, 1968.

[20] I. Hunter and R. Kearney, "Generation of random sequences with jointly specified probability density and autocorrelation functions," *Biological cybernetics*, vol. 47, no. 2, pp. 141–146, 1983.

[21] J. Nichols, C. Olson, J. Michalowicz, and F. Bucholtz, "A simple algorithm for generating spectrally colored, non-gaussian signals," *Probabilistic Engineering Mechanics*, vol. 25, no. 3, pp. 315–322, 2010.

[22] 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*, vol. 64, no. 4, pp. 918–930, 2016.