Reducing the Error of an Averaged Quantiser in the Presence of a Small-Scale Dither Signal

Arnfinn A. Eielsen, John Leth, and Andrew J. Fleming

***Abstract*—A method to improve the reconstruction of a signal after quantisation by compensation of a systematic error is presented. The result is that the effective resolution of a quantiser can be increased. This is possible due to the smoothing effect of adding a dither signal before quantisation, and subsequent averaging. The method requires that the type of probability distribution of the dither signal and the step-size of the quantiser are known. An analysis of the error in the recovered signal is presented, and optimal variances for a normally distributed dither are presented for signals that have either large or small magnitudes compared to the quantiser step-size. A method for determining the dither probability distribution parameters from the quantised signal is also presented. This makes it possible to determine the dither probability distribution parameters in systems where it is not possible to exert accurate control over the added dither signal and therefore enable to use of the presented reconstruction method under such circumstances.**

***Index Terms*—analog digital conversion, quantisation, nonlinear distortion, dither**

I. Introduction

Quantisation is the process of mapping a large set of values to a smaller set of values; thereby discarding some values and introducing a signal dependent error, as illustrated in Fig. 1. However, by applying a dithering signal as illustrated in Fig. 2, a bijective property in terms of the expected value of the quantised signal y and the reference signal x can be recovered, as illustrated in Fig. 4 (see also Fig. 5). That is, the expected value of the output of a dithered quantiser is a deterministic function $N(x)$ of the reference signal [1,2].

By applying averaging, a resolution below the step-size of the dithered quantiser can then be achieved [3]. Improving the resolution in this manner is particularly useful when a signal is sampled using coarse quantisation or in applications where a large dynamic range is required.

The function $N(x)$ is in general non-linear, and determined by the distribution of the dither signal and the step-size of the quantiser. If this functional relationship between the reference and the expected value of the quantised signal is known, the error in the expected value introduced by $N(x)$ compared to x can be compensated for by using the inverse of the function [4].

When the probability distribution for a dither signal fulfils certain criteria, the moments of the quantised signal can be made independent of the reference signal [5]–[10]. In terms of the first moment, the expected value, the quantiser can then be perfectly linearised by applying a uniformly distributed dither signal with a range that is exactly $\pm\Delta/2$, where Δ denotes the step-size of the quantiser.

However, dither distributions that render the moments of the quantised signal independent of the reference signal, e.g. a uniform distribution, are infeasible to produce in a physical measurement system as they require a high degree of precision in the physical dither signal reproduction in order to provide good performance. This is because any dynamics in the signal path, by design or due to parasitic effects, will cause the reproduced dither signal to tend to a normal distribution. This effect is explained by the central limit theorem [11]. If the realisation of the distribution is not exact, the linearising effect is significantly reduced [12]. In contrast, since introducing dynamics in a signal path can be done using e.g. linear time-invariant filters, for a physical system, a normally distributed dither signal is almost trivial to produce.

In the normally distributed case, the quantiser can only be perfectly linearised in terms of the expected value when the standard deviation, or variance, of the dither is infinite, as shown in Sec. III-D. When applying averaging, considering both the error introduced by $N(x)$ (a deterministic effect) as well as the variance of the dither (a stochastic effect), the minimum error of the dithered and averaged quantiser output signal is obtained when the standard deviation is approximately $\Delta/2$ [12,13]. This error is illustrated in Fig. 5 using a non-optimal dither variance to emphasise the effect of $N(x)$.

As noted above, if the dither distribution parameters and quantiser step-size are known, then applying the inverse of the non-linear function $N^{-1}(\hat{y})$ to the output of the dithered and averaged quantiser output reduces the overall error [4]. This is illustrated in Fig. 6.

By applying the inverse it might intuitively be expected that the standard deviation of the dither can be reduced significantly below $\Delta/2$, as $N(x)$ becomes bijective for any standard deviation greater than zero. A reduced dither standard deviation, or variance, might reduce the noise due to the dither in the inverted output signal, hence increasing the resolution. As the results in this paper show, this is not the case: The derivative of $N(x)$ approaches zero in the midpoints between the steps when the dither variance is reduced, implying that the derivative of $N^{-1}(\hat{y})$ at these points approaches infinity. A large derivative, or sensitivity, for $N^{-1}(\hat{y})$ amplifies noise, and this noise contrbutes to the overall error more than the removal of the systematic error due to $N(x)$ can reduce it. An optimal dither variance that minimise the overall error in the inverted signal therefore exists.

A. A. Eielsen (eielsen@ux.uis.no) is with the Dept. of Electrical Engineering and Computer Science, University of Stavanger, Norway, J.-J. Leth (jjl@es.aau.dk) is with the Dept. of Electronic Systems, Automation & Control, Aalborg University, Denmark, A. J. Fleming (andrew.fleming@newcastle.edu.au) is with the School of Electrical Engineering and Computer Science, The University of Newcastle, Australia.

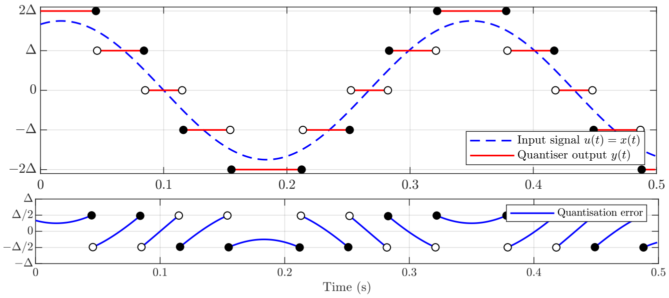

Figure 1: A quantised signal will have a signal dependent error, referred to as the quantisation error: $e(t) = y(t) - x(t)$.

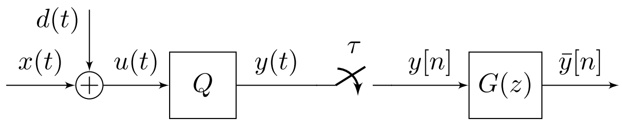

Figure 2: Quantiser dithering and averaging using an LTI-filter.

A. Contributions

The main contributions are an analysis of the overall error in the signal when applying $N^{-1}(\hat{y})$ and a method for identifying the parameters of the probability distribution using the quantised dither signal. The error analysis makes it possible to determine the performance limits when applying the inverse, and the identification method enables the application of the inverse to systems where a dither with a known type of probability distribution can be introduced, but where the exact distribution parameters are unknown.

The functional relationship between the reference signal and the expected value of the quantised signal as well as the expression for the variance of the output signal is developed directly using convolution integrals rather than via characteristic functions, such as in [4,12,13]. This approach makes it simpler to prove convergence to a linear function, as compared to the approaches in [5]–[7,9,10].

It is demonstrated that when using $N^{-1}(\hat{y})$ in combination with sufficient averaging, it is possible to reduce the error compared to the case when not using an inverse. It is also demonstrated that there exists an optimal dither variance, and that the optimal dither depends on assumptions made about the reference signal.

B. Notation

The set of natural numbers $\{0, 1, 2, \dots\}$ is denoted $\mathbb{N}$. A definition is denoted by $\triangleq$ and $*$ indicate the convolution product. The Laplace operator is denoted $\mathcal{L}$ and the Z-transform $\mathcal{Z}$. Functions of time t (or discrete-time n) are usually denoted by lower case, e.g. $g(t)$ (or $g[n]$), and the Laplace transform (or Z-transform) by upper case, e.g. $G(s) = \mathcal{L}\{g\}(s)$ (or $G[z] = \mathcal{Z}\{g\}[z]$). The standard notation $L^p(\mathbb{R})$ and l^p indicate the L^p-space and l^p-space $p = 1, 2, \dots, \infty$ and $\|g\|_p$ the p-norm of $g(t)$ or $g[n]$. For a stochastic process $d(t, \omega)$ the dependency on the sample variable ω is omitted, $\mu = \mathrm{E}\{d(t)\}$ denote the mean value, $C_d(t, s) \triangleq \mathrm{E}\{d(t)d(s)\} - \mathrm{E}\{d(t)\}\mathrm{E}\{d(s)\}$ denote the auto-covariance and $C_d(t) \triangleq C_d(t, t)$ the variance.

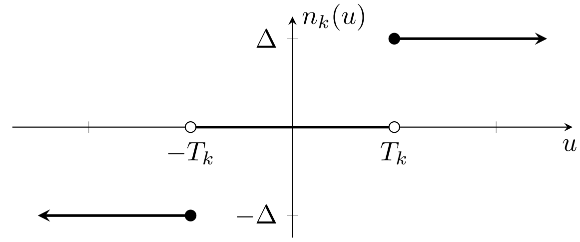

Figure 3: Step-function.

The indicator function for the set S is denoted $\chi_S(z)$ and the floor operator by $\lfloor \cdot \rfloor$.

II. The uniform quantiser

A uniform quantiser can represent values separated by the quantisation step-size $\Delta > 0$ and can be defined using the truncation operator $\mathrm{trunc}(u)$ as

$$q = \mathrm{trunc}(u) \triangleq \left\lfloor \frac{u}{\Delta} + \frac{1}{2} \right\rfloor \in \mathbb{Z}\,. \tag{1}$$

The quantised output $y(t)$ given input $u(t)$ is then

$$y(t) = n(u(t)) \triangleq \Delta\, \mathrm{trunc}(u(t)) = \Delta\, q(t)\,. \tag{2}$$

The uniform quantiser is demonstrated in Fig. 1.

A. Alternative formulation

The uniform quantiser can equivalently be expressed as

$$y(t) = n(u(t)) = \sum_{k \in \mathbb{N}} n_k(u(t)) \tag{3}$$

where each (odd) step-function, as shown in Fig.3,

$$n_k(u) \triangleq n_k^+(u) - n_k^-(u) \tag{4a}$$

in the quantiser is described by the step functions

$$n_k^+(u) \triangleq \Delta\, H(u - T_k)\,, \tag{4b}$$

$$n_k^-(u) \triangleq \Delta\, H(-u - T_k)\,, \tag{4c}$$

with the threshold of the step T_k given by

$$T_k \triangleq \Delta\,(k + 1/2) \tag{5}$$

and $H(u) \triangleq \chi_{[0,\infty)}(u)$ the Heaviside step-function. Note that the sum in (3) converges for every t since the summands become zero when $T_k \geq u(t) \geq -T_k$ or equivalently when $k \geq |u(t)|/\Delta - 1/2$.

III. Smoothing the quantiser

The input signal $u(t)$ when a stochastic dithering signal is present is the combination of two different signals

$$u(t) \triangleq x(t) + d(t)\,, \tag{6}$$

where $x(t)$ is the deterministic reference signal that should be recovered from a quantised and averaged time-series. It is assumed that the stochastic dither signal $d(t)$ is an identically distributed stochastic process having the strictly white noise

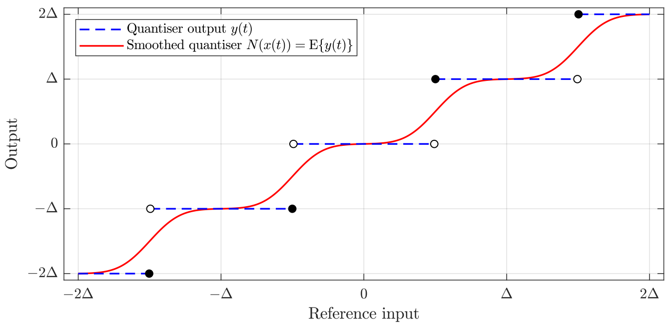

Figure 4: A quantiser is not bijective as it maps the continuum to a discrete set, but the smoothed quantiser (17) is; thus giving a one-to-one correspondence between reference and expected value. Here a zero-mean, normally distributed dither with ($\sigma_d = \Delta/6$) has been used.

property, i.e. $d(t_i)$ and $d(t_j)$ are independent for each pair t_i and t_j [14].

The dither signal $d(t)$ is either already present naturally due to sampling a noisy measurement, or it is added artificially. In either case, the effect is that by averaging the quantised signal $y(t)$, the effect of the discontinuous quantiser can be smoothed, imbuing the quantiser with continuity, as illustrated in Fig. 5. However, apart from the smoothing effect, the dither signal is otherwise unwanted in the recovered signal.

Let $F_d(z)$ be the cumulative distribution function (CDF) and $f_d(z)$ the probability density function (PDF) corresponding to $d(t)$, and $f_x(z,t) = \delta(z - x(t))$ the PDF corresponding to $x(t)$. Here, $\delta(\cdot)$ denotes the Dirac delta function. The signal $u(t)$ will then have the PDF

$$f_u(z,t) = \int_{\mathbb{R}} f_x(z-w,t) f_d(w)\,\mathrm{d}w = f_d(z - x(t))\,. \tag{7}$$

Consider the strictly white noise output for each step

$$y_k(t) \triangleq n_k(u(t))\,, \tag{8}$$

and the (strictly white noise) quantised output signal

$$y(t) = \sum_{k\in\mathbb{N}} y_k(t)\,. \tag{9}$$

The application of the dither signal makes it possible to define an averaged and smoothed step-function

$$N_k(z) \triangleq \int_{\mathbb{R}} n_k(z+w) f_d(w)\,\mathrm{d}w\,, \tag{10}$$

related to $n_k(u)$ and the stochastic dither signal $d(t)$.

Note that in general, when $n_k(z)$ is any arbitrary function of bounded variation,

$$\|N_k\|_\infty \le \|f_d\|_1 \|n_k\|_\infty = \|n_k\|_\infty \tag{11}$$

since $n_k(u) \in L^\infty(\mathbb{R})$, and that if $L_D \triangleq \sup_{z\in\mathbb{R}} |f_d(z)| < \infty$ then by [2, Lemma A.1]

$$L_{N_k} \le L_D TV(n_k) \tag{12}$$

with L_{N_k} a Lipshitz constant of $N_k(z)$ and $TV(n_k)$ the total variation of $n_k(u)$. In the case considered in this paper

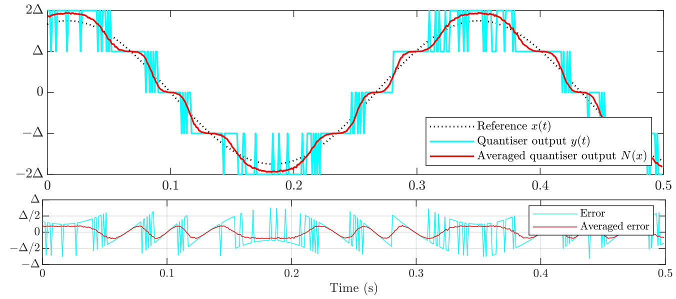

Figure 5: Applying dithering makes is possible to reduce the quantisation error if the output is averaged. Here a zero-mean, normally distributed dither (with $\sigma_d = \Delta/6$) has been used. The averaged signal was obtained with $M = 1000$ samples.

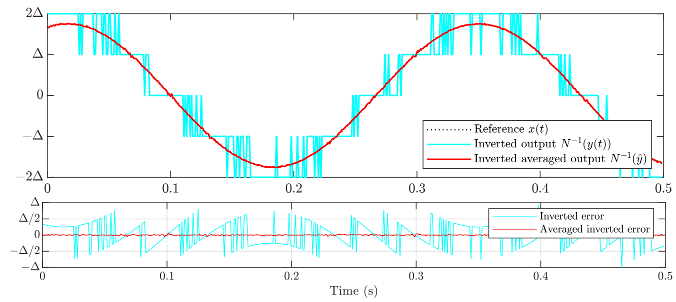

Figure 6: The inverse of the non-linear function relating reference and expected value can be used to reduce the quantisation error of the averaged signal. Here a zero-mean, normally distributed dither with ($\sigma_d = \Delta/6$) has been used. The averaged signal was obtained with $M = 1000$ samples.

$TV(n_k) = 2\Delta$, hence the dither PDF and quantisation step-size determines an upper bound on the smoothing effect on the quantiser. Moreover, using (4)

$$N_k(z) = \Delta\left(\int_{T_k - z}^{\infty} f_d(w)dw - \int_{-\infty}^{-T_k - z} f_d(w)dw\right) \tag{13a}$$
$$= \Delta\left(1 - F_d(T_k - z) - F_d(-T_k - z)\right)\,. \tag{13b}$$

It will be assumed throughout that the CDF F_d is such that

$$N(z) \triangleq \sum_{k\in\mathbb{N}} N_k(z) \tag{14a}$$
$$= \sum_{n\in\mathbb{Z}} \Delta\left(\frac{1}{2} - F_d(T_n - z)\right) \tag{14b}$$

is finite for each z. The summand k in (14a) correspond to the two summands $n = k$ and $n = -(k+1)$ in (14b).

In the special case where $f_d(z)$ is symmetric around zero[1] (13) can be written as

$$N_k(z) = \Delta \int_{-T_k - z}^{-T_k + z} f_d(w)dw \tag{15a}$$
$$= \Delta(F_d(-T_k + z) - F_d(-T_k - z))\,. \tag{15b}$$

Hence $\operatorname{sign}(N_k(z)) = \operatorname{sign}(z)$.

[1] $f_d(\mu + x) = f_d(\mu - x)$ with $\mu = E\{d(t)\} = 0$

A. First and Second Order Moments

Using (10), the expected value for each step is

$$\mathrm{E}\{y_k(t)\} = \int_{\mathbb{R}} n_k(u) f_d(u - x(t))\,\mathrm{d}u \tag{16a}$$
$$= \int_{\mathbb{R}} n_k(x(t) + w) f_d(w)\,\mathrm{d}w \tag{16b}$$
$$= N_k(x(t))\,. \tag{16c}$$

The expected value of the quantiser output is therefore

$$\mathrm{E}\{y(t)\} = \sum_{k\in\mathbb{N}} N_k(x(t)) = N(x(t))\,. \tag{17}$$

With (17), this leads to a output auto-covariance given by

$$C_y(t,s) = \mathrm{E}\{y(t)y(s)\} - \mathrm{E}\{y(t)\}\mathrm{E}\{y(s)\} \tag{18a}$$
$$= \begin{cases} 0 & t \neq s \\ \mathrm{E}\{y^2(t)\} - N^2(x(t)) & t = s \end{cases} \tag{18b}$$
$$\triangleq v(t)\chi_{\{t\}}(s) \tag{18c}$$

where the first case in (18b) follows by the strictly white noise property of $y(t)$ and the second case is by (17).

Now, consider the zero-mean strictly white noise error signal

$$\varepsilon(t) = y(t) - N(x(t))\,, \tag{19}$$

having auto-covariance $C_\varepsilon(t,s) = C_y(t,s)$. In order to find the time-varying variance

$$v(t) = C_\varepsilon(t) = \mathrm{E}\{\varepsilon^2(t)\}$$

of the error signal $\varepsilon(t)$, the output of the terms (4) in the summation (3) is considered individually as correlated variables. The variance and covariance of the terms can then be found individually and then summed using the formula

$$\mathrm{E}\{\varepsilon^2(t)\} = \sum_{k\in\mathbb{Z}} \mathrm{var}(y_k(t)) + 2\sum_{k\in\mathbb{Z}}\sum_{l>k} \mathrm{cov}(y_k(t), y_l(t))\,, \tag{20}$$

to find the variance of the error.

Note that the Heaviside step-function has the properties

$$H^2(u) = H(u)\,, \tag{21}$$

and (more generally)

$$H(u-a)H(u-b) = H(u-b) \tag{22}$$

if $a \leq b$. The variance for $y_k(t)$ can then be found as

$$\begin{aligned}
&\mathrm{var}(y_k(t)) \\
&\quad\triangleq \mathrm{E}\{y_k^2(t)\} - N_k(x(t))^2 \\
&\quad= \Delta^2\big(1 - F_d(T_k - x(t)) + F_d(-T_k - x(t))\big) \\
&\qquad\quad - N_k(x(t))^2 \qquad\qquad (23a) \\
&\quad= 2\Delta^2 F_d(-T_k - x(t)) + \Delta N_k(x(t)) \\
&\qquad\quad - N_k(x(t))^2 \qquad\qquad (23b)
\end{aligned}$$

using (21) to obtain (23a), and (13b) to obtain (23b). For $k < l$, the covariance between terms can be found as

$$\begin{aligned}
\mathrm{cov}(y_k(t), y_l(t)) &\triangleq \mathrm{E}\{y_k(t)y_l(t)\} - \mathrm{E}\{y_k(t)\}\mathrm{E}\{y_l(t)\} \\
&= \Delta^2\big(1 - F_d(T_l - x(t)) + F_d(-T_l - x(t))\big) \\
&\quad - N_k(x(t))N_l(x(t))
\end{aligned} \tag{24}$$

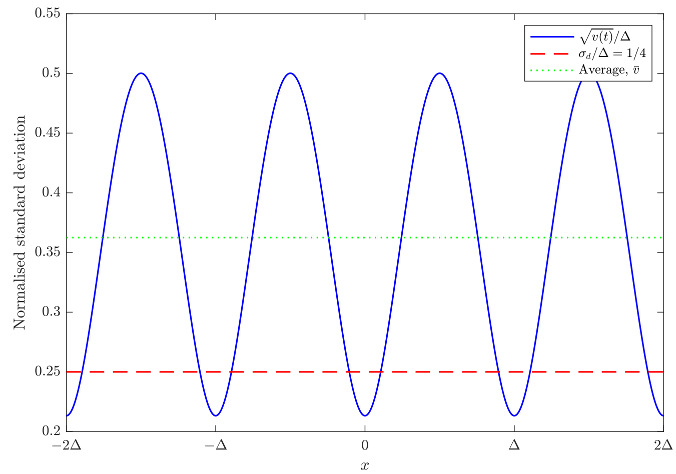

Figure 7: The variance $v(t) = \mathrm{E}\{\varepsilon^2(t)\}$ depends on $x(t)$. The plot shows the normalised standard deviation for the normally distributed case with $\sigma_d/\Delta = 1/4$.

using (22). An example of the variance of $\varepsilon(t)$ and its dependence of $x(t)$ is shown in Fig. 7.

B. Convergence to a linear function

We show that $\frac{\mathrm{d}}{\mathrm{d}z}N(z)$ is constant, and therefore that $N(z)$ is affine linear, when the characteristic function $\varphi_d(v)$ of the PDF $f_d(z)$ has a specific property (see (29)).

From (14b) we get

$$\frac{\mathrm{d}}{\mathrm{d}z}N(z) = \sum_{n\in\mathbb{Z}} \Delta f_d(T_n - z) \tag{25a}$$
$$= \sum_{n\in\mathbb{Z}} \Delta f_d(\Delta/2 - z + n\Delta)\,, \tag{25b}$$

which by the Poisson sum formula [11] yields:

$$\frac{\mathrm{d}}{\mathrm{d}z}N(z) = \sum_{n\in\mathbb{Z}} \hat{f}_d\left(-\frac{2\pi n}{\Delta}\right) \mathrm{e}^{i\frac{2\pi n}{\Delta}(z-\Delta/2)} \tag{26}$$
$$= \sum_{n\in\mathbb{Z}} \varphi_d\left(\frac{2\pi n}{\Delta}\right) \mathrm{e}^{i\frac{2\pi n}{\Delta}(z-\Delta/2)} \tag{27}$$
$$= 1 + \sum_{n\in\mathbb{Z}\setminus\{0\}} \varphi_d\left(\frac{2\pi n}{\Delta}\right) \mathrm{e}^{i\frac{2\pi n}{\Delta}(z-\Delta/2)} \tag{28}$$

It can then be seen that if

$$\varphi_d\left(\frac{2\pi n}{\Delta}\right) = 0 \;\forall n \in \mathbb{Z}\setminus\{0\}\,, \tag{29}$$

then the derivative of $N(z)$ is constant, and hence $N(z)$ is affine linear, that is, when (29) is fulfilled, then the quantisation error

$$e_x(t) = N(x(t)) - x(t) = N(0). \tag{30}$$

Note that the quantisation error (30) is in terms of the reference $x(t)$, whereas the output error (19) is in terms of the output $y(t)$. If $N(x(t))$ intersects the origin, $N(0) = 0$. The above derivation leads to the same well-known result as found in [5]–[7,9,10]. Below are two examples of distributions that can fulfill the criterion (29) and in principle linearise the quantiser and remove the quantisation error. As will be seen, neither have practical use in measurement systems.

C. Application of a uniformly distributed dither

The simplest distribution that can fulfil (29) is the uniform distribution [9,10]. In this case

$$f_d(z) = \begin{cases} \frac{1}{\Delta}\,, & |z| \le \frac{\Delta}{2} \\ 0\,, & |z| > \frac{\Delta}{2} \end{cases} \tag{31}$$

which has characteristic function

$$\varphi_d(v) = \frac{\mathrm{e}^{iv\frac{\Delta}{2}} - \mathrm{e}^{-iv\frac{\Delta}{2}}}{iv\Delta} = \frac{\sin\left(\frac{\Delta}{2}v\right)}{\frac{\Delta}{2}v}\,,$$

hence

$$\varphi_d\left(\frac{2\pi n}{\Delta}\right) = \frac{\sin(\pi n)}{\pi n} = \operatorname{sinc}(n) = 0 \;\forall n \in \mathbb{Z}\setminus\{0\}\,.$$

If the dither signal has a uniform distribution, the dither signal reconstruction must be absolutely exact in order to linearise the quantiser. Engineering a device that can reproduce a physical white noise signal with a uniform distribution is in general infeasible due to the precision requirements. Hence, the uniformly distributed case has little practical use in most physical measurement systems.

D. Application of a normally distributed dither

Producing a normally distributed signal is comparatively trivial compared to producing a physical realisation of a uniformly distributed white noise signal. If the process d is normally distributed with mean μ_d and standard deviation σ_d, then

$$f_d(z) = \frac{1}{\sqrt{2\pi}\,\sigma_d}\mathrm{e}^{-\frac{(z-\mu_d)^2}{2\sigma_d^2}} \tag{32}$$

$$\varphi_d\,(v) = \mathrm{e}^{i\mu_d v}\mathrm{e}^{-\frac{1}{2}\sigma_d^2 v^2}, \tag{33}$$

hence

$$\varphi_d\left(\frac{2\pi n}{\Delta}\right) = \mathrm{e}^{i\mu_d\frac{2\pi n}{\Delta}}\,\mathrm{e}^{-\sigma_d^2\frac{2\pi^2 n^2}{\Delta^2}}\,, \tag{34}$$

and $N(z)$ will therefore approach an affine linear function as $\sigma_d \to \infty$, i.e. this distribution only fulfills (29) for an infinite standard deviation making it irrelevant for practical applications.

IV. Averaging

A. Sample-averaging

Considering (18b) and (19), the output $y(t)$ can be seen to be a linear combination of the effective non-linearity $N(x(t))$, and the white noise error term $\varepsilon(t)$, that is

$$y(t) = N(x(t)) + \varepsilon(t) = \mathrm{E}\{y(t)\} + \varepsilon(t)\,. \tag{35}$$

The expression $\mathrm{E}\{y(t)\} = N(x(t))$ represents the expected value in terms of an average over the sample space. We can in principle find this average as follows.

For $i = 1, 2, \ldots, M$ let $d_i(t)$ denote a stochastic dither signal and assume that $d_1(t), d_2(t), \ldots, d_M(t)$ are independent and identically distributed (i.i.d.) for each fixed time t. As above we construct, for each i, the (strictly white noise) quantised output signal

$$y^i(t) \triangleq \sum_{k\in\mathbb{N}} y_k^i(t) \triangleq \sum_{k\in\mathbb{N}} n_k(x(t) + d_i(t))\,. \tag{36}$$

Hence $y(t), y^1(t), \ldots, y^M(t)$ are i.i.d. and the sample average

$$\langle y(t)\rangle_M = \frac{1}{M}\sum_{i=1}^{M} y^i(t)\,, \tag{37}$$

converge[2] to $\mathrm{E}\{y(t)\}$ by the law of large numbers. That is,

$$\langle y(t)\rangle = \mathrm{E}\{y(t)\}$$

with $\langle y(t)\rangle \triangleq \lim_{M\to\infty}\langle y(t)\rangle_M$. Note that by direct calculations the variance of $\langle y(t)\rangle_M$ is

$$C_{\langle y\rangle_M}(t) = \frac{1}{M}C_y(t) = \frac{1}{M}v(t)\,, \tag{38}$$

which converge to zero, implying that $\langle y(t)\rangle_M$ converge to $E\{y(t)\}$ in the mean square sense since

$$\begin{aligned} &E\{(\langle y(t)\rangle_M - E\{y(t)\})^2\} \\ &\quad = E\{(\langle y(t)\rangle_M - E\{\langle y(t)\rangle_M\})^2\} = C_{\langle y\rangle_M}(t)\,. \end{aligned} \tag{39}$$

For the sample average

$$\langle\varepsilon(t)\rangle_M \triangleq \frac{1}{M}\sum_{i=1}^{M}\varepsilon^i(t) = \langle y(t)\rangle_M - E\{y(t)\}\,, \tag{40}$$

with $\varepsilon^i(t) \triangleq y^i(t) - E\{y^i(t)\} = y^i(t) - E\{y(t)\}$, we get

$$\langle\varepsilon(t)\rangle \triangleq \lim_{M\to\infty}\langle\varepsilon(t)\rangle_M = 0$$

and

$$C_{\langle\varepsilon\rangle_M}(t) = \frac{1}{M}C_\varepsilon(t) = \frac{1}{M}C_y(t)\,. \tag{41}$$

The sample averaging is often not implementable in applications since, in general, a large value of M is required (that is, a large number of physical channels is required). A common configuration in measurement systems is to have a single channel. Hence, time-averaging may therefore be the only option for producing an average value.

B. Time-averaging

Time-averaging the output of the quantiser is in practice done using discrete time-samples of $y(t)$, at the instances $t_n = \tau n$ where τ is the sampling-time. The samples will then be averaged using a linear time-invariant (LTI) low-pass filter. The discrete-time output is denoted $y[n] \triangleq y(t_n)$, and the LTI-filter impulse response is denoted $g[n]$. Denoting the time-averaged output $\bar{y}[n] \triangleq (g * y)[n]$, we have

$$\bar{y}[n] = (g * N(x))[n] + (g * \varepsilon)[n] \tag{42a}$$

$$= N(x[n]) + (h * N(x))[n] + (g * \varepsilon)[n] \tag{42b}$$

[2] almost surely or in probability

where $h[n] = g[n] - \delta[n]$. The z-domain representation is then

$$\bar{y}(z) = G(z)y(z) \tag{43a}$$
$$= \mathcal{Z}\{N(x)\}(z) + H(z)\mathcal{Z}\{N(x)\}(z) + G(z)\mathcal{E}(z)\,, \tag{43b}$$

where $H(z) = G(z) - 1$. The filter $H(z)$ is used to describe the error introduced by the time-averaging operation.

Since the error signal (19) is not stationary, common systems norms [15] do not apply. However, an LTI-filter will still reduce the variance of the error signal $\varepsilon(t)$ as will be shown next. Define the filtered error signal

$$\bar{\varepsilon}[n] \triangleq (g * \varepsilon)[n]\,, \tag{44}$$

as the output of the filter $g[n]$ due to the error signal $\varepsilon[n]$. The variance of the filtered error signal is then

$$C_{\bar{\varepsilon}}[n, n] = \mathrm{E}\{\bar{\varepsilon}^2[n]\} \tag{45a}$$
$$= (v * g^2)[n]\,, \tag{45b}$$

where (45b) follows by the discrete-time version of [14, Theorem 9-3]. Hence if $v \in l^1$ and $g^2 \in l^2$ then

$$\|C_{\bar{\varepsilon}}\|_2 = \|v * g^2\|_2 \tag{46a}$$
$$\leq \|v\|_1 \|g^2\|_2 \tag{46b}$$

where (46b) follows by Young's convolution theorem for sequences. By (46) the variance of the filtered error signal will be attenuated by the energy of the impulse response squared $\left\|g^2\right\|_2$. For strictly proper, unity gain low-pass filters, the energy of the impulse response is directly related to the cut-off frequency: A lower cut-off frequency yields lower energy, hence smaller variance for the filtered error signal $\bar{\varepsilon}$.

1) Moving average filter: Applying a moving average filter

$$g[n] = \frac{1}{M} \sum_{i=0}^{M-1} \delta[n - i]\,,$$

gives

$$\bar{y}[n] = (g * y)[n] = \frac{1}{M} \sum_{i=0}^{M-1} y[n - i]\,.$$

The variance of the filtered error signal will then be

$$\mathrm{E}\{\bar{\varepsilon}^2[n]\} = \frac{1}{M^2} \sum_{i=0}^{M-1} v[n - i] \tag{47}$$

according to (45), and it will be attenuated by

$$\|g^2\|_2 = \frac{1}{M}\,, \tag{48}$$

according to (46).

Note that a moving average filter in the z-domain is a uniformly weighted sum of delay elements, which can simplified using the geometric series sum formula to yield

$$G(z) = \frac{1}{M} \sum_{i=0}^{M-1} z^{-i} = \frac{1}{M} \frac{1 - z^{-M}}{1 - z^{-1}}\,. \tag{49}$$

By setting $z = \mathrm{e}^{i2\pi f}$ the frequency response of $G(z)$ can be found as

$$G(f) = \frac{\mathrm{sinc}(Mf)}{\mathrm{sinc}(f)} \mathrm{e}^{-i2\pi(\frac{M}{2} - 1)f}\,. \tag{50}$$

The magnitude response is shown in Fig. 8. The frequency response can be used to compensate for some of the effects the moving average filter has on the reference signal $x[n]$, as explained in Sec. V-B.

2) Accumulate-and-dump filter: A large number of samples is desireable in order to reduce the variace due to the dither signal. However, in terms of implementability, it may be necceary to decimate, or downsample, the output of the moving average filter in order to reduce the sampling rate. If the moving filter with M samples is subsequently downsampled by the same factor M, it is known as an accumulate-and-dump filter. This operation is attractive due to its simplicity. The effect of the decimation can be analysed by considering the interpolation formula

$$y'(t) = \sum_{n \in \mathbb{Z}} \bar{y}[n] \mathrm{sinc}(t - n)\,.$$

Decimating $\bar{y}[n]$ is achieved by evaluating $\bar{y}'[Mn] \triangleq y'(Mn)$.

C. Variance for Small Δ

If it can be assumed that Δ is small in the sense that $x(t)$ varies approximately linearly over the quantisation step-size, the variance $v(t)$ can be approximated by an average [4,13]; the limiting case being that $x(t)$ varies linearly across one step.

In the sample-averaged case the average variance is defined as

$$\bar{v} \triangleq \frac{1}{\Delta} \int_{-\Delta/2}^{\Delta/2} \frac{1}{M} v(t)\, \mathrm{d}t \tag{51}$$

and the variance (41) of $\langle \varepsilon(t) \rangle_M$ can be approximated by

$$\mathrm{E}\{\langle \varepsilon(t) \rangle_M^2\} \approx \frac{1}{M} \bar{v}\,. \tag{52}$$

using (38).

In the time-averaged case the average variance is similarly defined as

$$\bar{v}_d = \frac{\Delta}{M} \sum_{k=0}^{M-1} v\left(-\frac{\Delta}{2} + \frac{\Delta}{M} k\right) \tag{53}$$

and, assuming a moving average filter, the variance of $\bar{\varepsilon}[n]$ will then be

$$\mathrm{E}\{\bar{\varepsilon}^2[n]\} \approx \frac{1}{M} \bar{v}_d\,, \tag{54}$$

by (47).

Note that if Δ/M in (53) is interpreted as the sampling-time then $\bar{v} \approx \bar{v}_d$ for sufficiently small sampling-time (or equivalently for sufficiently large M). An example of the average variance is shown in Fig. 7, and compared to the case when it is dependent on $x(t)$. By assuming that the variance is contant, the process describing the error becomes stationary and infomation about the worst-case error variance is neglected.

V. Inverting the Effective Non-linearity

Consider again the output of the quantiser (35). If there was no error signal, $\varepsilon(t) = 0$, then the reference signal $x(t)$ could directly be recovered by application of the inverse of $N(z)$, that is

$$x(t) = N^{-1}(y(t))\,. \tag{55}$$

Clearly (55) can not be true since we always have $\varepsilon(t) \neq 0$ due to the presence of the dither signal. However, we may obtain an estimate of $x(t)$ as follows. Let $\hat{y}$ denote an average of the measured output (e.g. $\langle y(t)\rangle_M$ or $\bar{y}[n]$). If the dither d is given, then $N(z)$ is known analytically and the estimate $\hat{x} \triangleq N^{-1}(\hat{y})$ can then be computed numerically by solving $F(\hat{x}) = 0$ where

$$F(z) \triangleq N(z) - \hat{y}\,. \tag{56}$$

One such method is the bisection method [16], which is guaranteed to converge if $F(z)$ is a continuous function in a domain $[z_l, z_h]$ and $F(z_l)$ and $F(z_h)$ have opposite signs. Since $|N(z) - \hat{y}| \leq \Delta$, it is always possible to choose a constant z_0 such that $w_l = \hat{y} - z_0$ and $z_h = \hat{y} + z_0$ will fulfil this encompassing condition.

It is clearly of interest to obtain an estimate of the variance of $\hat{x}$. This can be done using a first-order Taylor series of $N^{-1}(w)$, at $N(x(t))$, evaluated along the quantised output signal $y(t)$. That is,

$$N^{-1}(y(t)) \approx x(t) + \alpha(t)\varepsilon(t)\,, \tag{57}$$

where we have defined the sensitivity as

$$\alpha(t) \triangleq \left.\frac{\mathrm{d}N^{-1}}{\mathrm{d}w}\right|_{w=N(x(t))} \tag{58}$$

$$= \left(\left.\frac{\mathrm{d}N}{\mathrm{d}z}\right|_{z=N^{-1}(N(x(t)))}\right)^{-1}. \tag{59}$$

Using (57) directly, or the delta-method [17], an approximation to the variance of $N^{-1}(y(t))$ can be found as

$$C_{N^{-1}(y)}(t) \approx \alpha^2(t)\mathrm{E}\{\varepsilon^2(t)\} = \alpha^2(t)v(t)\,. \tag{60}$$

A. Inverting the sample-average

Sample-averaging reduces the variance of the quantised signal, and hence also the variance of the signal obtained by application of the inverse $N^{-1}(z)$. Using (57) and (40) we get

$$N^{-1}(\langle y(t)\rangle_M) \approx x(t) + \varepsilon_{e,M}(t)\,, \tag{61}$$

where

$$\varepsilon_{e,M}(t) = \alpha(t)\langle\varepsilon(t)\rangle_M \tag{62}$$

denotes the stochastic error due to the dithering signal. Using (41) and (38), the variance of this error after averaging is

$$\mathrm{E}\{\varepsilon^2_{e,M}(t)\} = \frac{1}{M}\alpha^2(t)\,v(t)\,. \tag{63}$$

B. Inverting the time-average

Time-averaging using an LTI low-pass filter $G(z)$ such as (49) reduces the variance of the stochastic error signal $\varepsilon[n]$ but introduces a frequency dependent error for the deterministic reference signal $x[n]$. At low frequencies the error for $x[n]$ is small but increases toward the Nyquist-frequency. This can be seen from the high-pass characteristic of the filter $H(z) = G(z) - 1$ in (43b). From (57) and (19), the inverse can be written as

$$N^{-1}(\bar{y}[n]) \approx x[n] + e_f[n] + \varepsilon_t[n]\,, \tag{64}$$

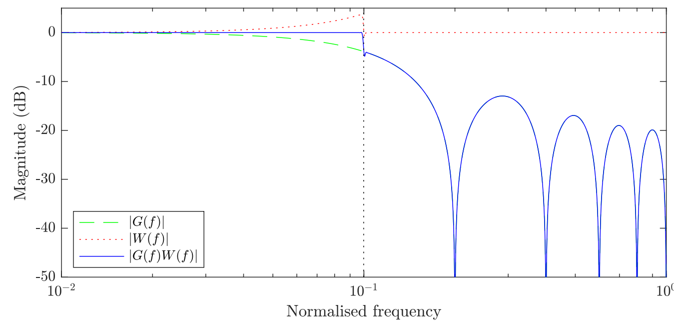

Figure 8: Compensating for the response of $G(z)$ (moving average) by using a pre-filter $W(z)$. The filter $W(z)$ has been synthesised using the least-squares method [18].

where

$$e_f[n] = \alpha[n](h * N(x))[n]\,, \tag{65}$$

denotes the deterministic error due to filtering, and

$$\varepsilon_t[n] = \alpha[n](g * \varepsilon)[n]\,. \tag{66}$$

denotes the stochastic error due to the dithering signal. Under the assumptions of Sec. IV-C, the variance of $\varepsilon_t[n]$ will be reduced by a factor of $1/M$.

If $x(t)$ is bandwidth limited, it is straight-forward to compensate for the error $e_f[n]$ by designing a finite impulse-response (FIR) compensating pre-filter $W(z)$ which mininize

$$\|1 - G(f)W(f)\|_2 \tag{67}$$

in the passband for $x(t)$, and minimize

$$\|1 - W(f)\|_2 \tag{68}$$

elsewhere, as demonstrated in Fig. 8 for the moving average filter. Given (67), the error $H(f) = G(f)W(f) - 1 \approx 0$ in the passband. If the passband is sufficiently small, the energy due to the difference $\|G(z) - G(z)W(z)\|_2$ will be small. Hence, the effect of $H(z)$ will be ignored in the remainder.

C. Optimal dither variance when inverting

The expression in (60) consists of the sensitivity (58) and the variance of the error (20), which both are dependent on $x(t)$. Consider (25), for a normal distribution. The variance of the error (20) will then have maxima at the thresholds (5), that is at $x(t) = T_k$, as seen in Fig. 7. In this case $N(z)$ and $N^{-1}(w)$ will intersect at Δk, and $N^{-1}(w)$ will therefore according to (59) have maxima at Δk. This is shown in Fig. 9.

As demonstrated in Fig. 10, the maximum sensitivity tends to infinity as σ_d goes to zero. This is the dominating factor for smaller σ_d. Hence, there is an optimal σ_d, balancing the sensitivity of $N^{-1}(z)$ and the magnitude of σ_d. Two metrics for the error after inversion can be devised, suitable for different conditions on the reference signal: An averaged case where it is assumed that $x(t)$ varies linearly over the quantisation step-size (small Δ), and a worst case where the maximum variance is considered.

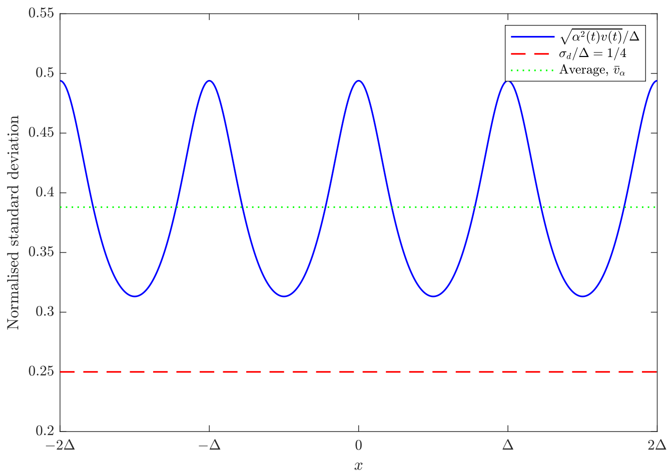

Figure 9: The variance (60) depends on $x(t)$. As Fig. 7, The plot shows the normalised standard deviation for the normally distributed case with $\sigma_d/\Delta = 1/4$.

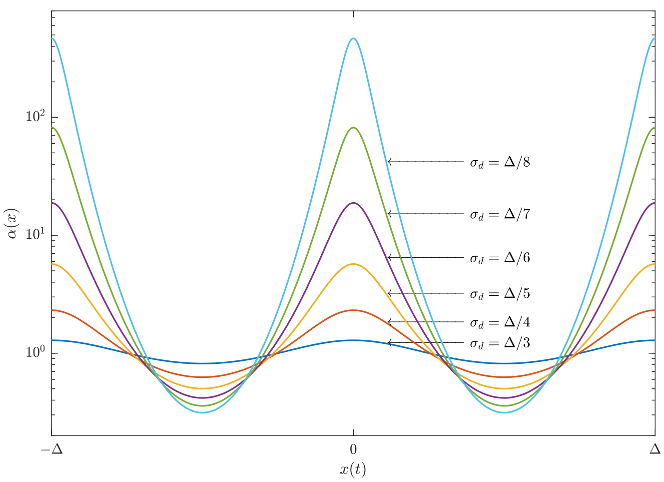

Figure 10: The sensitivity (58) depends on σ_d.

1) Small Δ*:* Under the same assumptions as in Sec. IV-C, then the average of (60) is a good approximation of the error variance after inversion, i.e.

$$\bar{v}_\alpha = \frac{1}{\Delta}\int_{-\Delta/2}^{\Delta/2} \alpha^2(x)v(x)\,\mathrm{d}x\,, \tag{69}$$

and using the Cauchy-Schwarz inequality and (51)

$$\bar{A}_{v_\alpha} = \mathrm{E}\left\{N^{-1}\left(\hat{y}\right)^2\right\} \leq \frac{1}{M}\bar{v}\,\frac{1}{\Delta}\int_{-\Delta/2}^{\Delta/2} \alpha^2(x)\,\mathrm{d}x\,, \tag{70}$$

it can be seen that the variance is reduced by a factor of $1/M$, and the expression provides a metric for the error variance after inversion with small Δ. By varying M and σ_d an optimal value for (70) can be found.

The above variance should be compared to the sample average of the quantisation error (when not inverting)

$$\langle e(t)\rangle_M = \langle y(t)\rangle_M - x(t) = \underbrace{N(x(t)) - x(t)}_{e_x(t)} + \langle\varepsilon(t)\rangle_M \tag{71}$$

which is discussed in detail in [13]. For $\langle e(t)\rangle_M$ the error is the deviation from $x(t)$ and not the function $N(x(t))$ as in (19), hence the deviation from a linear response $e_x(t)$ is deterministic and contributes to the error. The error (19) is the deviation from the function $N(x(t))$, and is therefore a zero-mean, stochastic process. A metric that includes the deterministic component $e_x(t)$ is the mean squared error (MSE) [13], and with the above error signal this becomes

$$\mathrm{MSE} = \frac{1}{\Delta}\int_{-\Delta/2}^{\Delta/2} \langle e(x)\rangle_M^2\,\mathrm{d}x = \bar{e}_x^2 + \frac{1}{M}\bar{v} \tag{72}$$

where the MSE of the deterministic component is

$$\bar{e}_x^2 = \frac{1}{\Delta}\int_{-\Delta/2}^{\Delta/2} e_x^2(x)\,\mathrm{d}x\,, \tag{73}$$

and comes in addition to the variance due to the stochastic component $\bar{v}$.

2) Worst case: If no assumptions can be made about $x(t)$, the worst-case variance will be a more reasonable estimate of the variance of the error after inversion, that is,

$$A^\star_{v_\alpha} = \max_x \mathrm{E}\{N^{-1}(\hat{y})^2\} = \frac{1}{M}\max_x\,\alpha^2(x)v(x)\,. \tag{74}$$

By varying M and σ_d an optimal value for (74) can be found.

3) Optimal variances: The standard deviation using the two expressions for the variance, (70) and (74), and the root mean squared error (RMSE) using (72) are shown in Fig. 11 for a normally distributed dither signal for different standard deviations σ_d and number for averages M. Both axes have been normalised relative to the step-size Δ. As increasing M decreases the variance in the error, the error due to the deterministic component $\bar{e}_x$ becomes more significant for larger M, hence the minimal RMSE will depend on M. This effect is in part counteracted by applying the inverse $N^{-1}(z)$, but eventually the increasing sensitivity α grows faster than the reduction of σ_d, and the overall standard deviation increases. In this case the minimal variance does not depend on M.

By numerical evaluation, inversion with $M \geq 4$ was found to yield a lower variance than the equivalent MSE for the averaged case, and for $M \geq 38$ the maximum variance of the inverted signal is smaller than the MSE. Hence, for $M \geq 4$ inversion provides improved performance over only averaging an optimally dithered quantiser.

VI. Probability mass function of the quantised dither

Consider the case when the reference signal is a constant, $x(t) = \mu_x$. The PDF for the input will then be

$$f_u(z) = f_d(z - \mu_x)\,. \tag{75}$$

In a similar manner as described in [9], the probability that the input signal $u(t) = \mu_x + d(t)$ will truncate to a given integer $q(t) = k \in \mathbb{Z}$ can be found by considering the probability of $u(t)$ in the domain $\Delta[k - 1/2, k + 1/2]$

$$f_q(k) = \Pr(q(t) = k) = \int_{\Delta(k-1/2)}^{\Delta(k+1/2)} f_d(z - \mu_x)\,\mathrm{d}z\,. \tag{76}$$

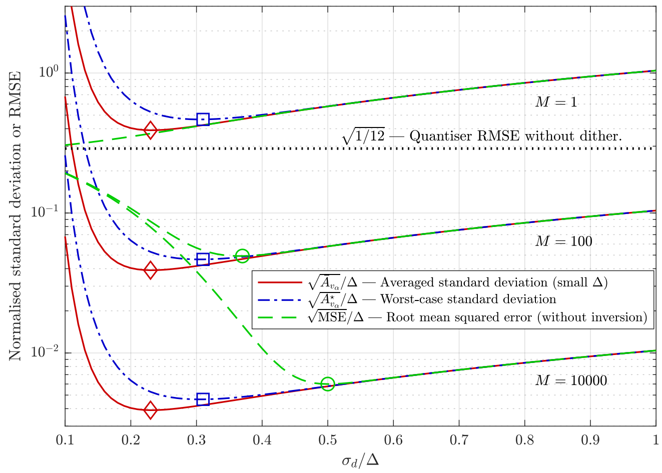

Figure 11: Optimal σ_d/Δ, using normally distributed dither.

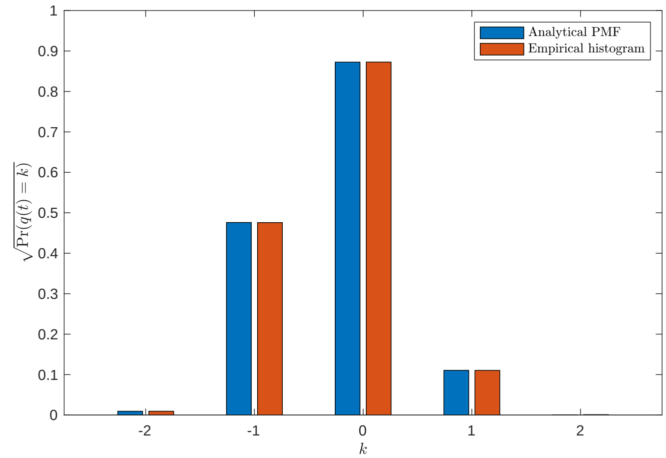

Figure 12: The analytical PDF compared to a numerical histogram (10^6 samples) for a quantiser with $\Delta = 1$ driven with a normally distributed signal with $\mu_u = -\Delta/4$ and $\sigma_d = \Delta/3$.

If $d(t)$ is normally distributed and the reference signal is a constant, $x(t) = \mu_x$, then the input distribution is

$$f_u(z) = \frac{1}{\sqrt{2\pi}\,\sigma_d} \mathrm{e}^{-\frac{(z-\mu_u)^2}{2{\sigma_d}^2}} \tag{77}$$

with $\mu_u = \mu_x + \mu_d$ and then the probability mass function (PMF) becomes

$$f_q(k) = \mathrm{erf}\left(\frac{\Delta\left(k+\frac{1}{2}\right)-\mu_u}{\sqrt{2}\,\sigma_d}\right) - \mathrm{erf}\left(\frac{\Delta\left(k-\frac{1}{2}\right)-\mu_u}{\sqrt{2}\,\sigma_d}\right). \tag{78}$$

A. Identification of the dither probability density function

With the assumption that the reference signal can be held constant and that the dither signal has a PDF of a known type, the parameters of the input PDF can be determined from experiment, and used to construct an accurate inverse $N^{-1}(z)$. Sampling a quantised signal when the reference is constant, an empirical PMF can be produced using a histogram, denoted $\hat{f}_q(k)$, and is used to solve the parameter identification problem by solving, e.g.

$$\min_{\theta} \left\| f_q(k;\theta) - \hat{f}_q(k) \right\|_2 , \tag{79}$$

by using the analytic expression for the PMF (76). In the case of a normal distribution, robust estimates for $\hat{\mu}_u$ and $\hat{\sigma}_d$ are found solving a problem on the form

$$\min_{\mu_u,\sigma_d} \left\| \sqrt{f_q(k;\mu_u,\sigma_d)} - \sqrt{\hat{f}_q(k)} \right\|_2 , \tag{80}$$

where the problem has been scaled by the square root, as the masses associated with values away from k tend to be very small. In this case it is only possible to identify μ_u, which means that the bias μ_x can not be compensated for using $N^{-1}(z)$. The latter causes a constant bias to be introduced when applying $N^{-1}(z)$.

1) Numerical example: Using (80) to find the parameters for a normal distribution is illustrated in Fig. 12. In this case $\Delta = 1$, $\mu_u = -\Delta/4$ and $\sigma_d = \Delta/3$, and 10^6 samples have been used to produce the estimate $\hat{f}_q(k)$. Solving (80) with the data used in Fig. 12 yields $\hat{\mu}_u = -0.2502$ and $\hat{\sigma}_d = 0.3330$, which are very accurate estimates.

VII. Discussion and conclusions

When dithering a quantiser, the expected value of the output has a one-to-one correspondence to reference value, and this function can be determined if the probability density function of the dither and the step-size of the quantiser are known. The deterministic error introduced by the quantiser can then be compensated for using the inverse of this function, and the stochastic error due to the dither can be reduced by averaging. Sample and time-averaging achieves approximately the same performance gain. Time-averaging is more feasible in a physical system, but requires a compensation filter to reduce amplitude and phase distortion of the reference signal due to the averaging operation.

It was shown that there is a limit to how small the variance of the dither can be when applying inversion. The variance of the error grows large for very small dither variances, and similarity the variance grows large for larger dither variances. Optimal dither variances exists that minimise the variance of the error after inversion. One optimum exists when Δ can be assumed to be small compared to the reference, and another optimum exist when considering the maximum error variance. The variance of the error when using inversion scales with a factor $1/M$, where M is the number of averages. Using averaging and inversion with $M \geq 4$ and an optimal dither provides better performance than when only averaging and not inverting an optimally dithered quantiser.

A method for identifying the probability density function for the dither using empirical data was presented and demonstrated numerically. This enables compensation by inversion in systems where it is difficult to have exact control over the generated dither signal.

References

[1] G. Zames and N. Shneydor, "Dither in nonlinear systems," *IEEE Transactions on Automatic Control*, vol. 21, no. 5, pp. 660–667, 1976.

[2] L. Iannelli, K. H. Johansson, U. T. Jönsson, and F. Vasca, "Averaging of nonsmooth systems using dither," *Automatica*, vol. 42, no. 4, pp. 669–676, Apr. 2006.

[3] J. Vanderkooy and S. P. Lipshitz, "Resolution Below the Least Significant Bit in Digital Systems with Dither," *Journal of the Audio Engineering Society*, vol. 32, no. 3, pp. 106–113, Mar. 1984.

[4] P. Koeck, "Quantization errors in averaged digitized data," *Signal Processing*, vol. 81, no. 2, pp. 345–356, Feb. 2001.

[5] L. Schuchman, "Diher Signals and Their Effect on Quantization Noise," *IEEE Transactions on Communication Technology*, vol. 12, no. 4, Dec. 1964.

[6] A. Sripad and D. Snyder, "A necessary and sufficient condition for quantization errors to be uniform and white," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 25, no. 5, pp. 442–448, Oct. 1977.

[7] S. P. Lipshitz, R. A. Wannamaker, and J. Vanderkooy, "Quantization and dither: A theoretical survey," *Journal of the audio engineering society*, vol. 40, no. 5, pp. 355–375, 1992.

[8] R. M. Gray and T. G. Stockham, "Dithered Quantizers," *IEEE Transactions on Information Theory*, vol. 39, no. 3, pp. 805–812, May 1993.

[9] B. Widrow, I. Kollar, and M.-C. Liu, "Statistical theory of quantization," *IEEE Transactions on Instrumentation and Measurement*, vol. 45, no. 2, pp. 353–361, Apr. 1996.

[10] R. A. Wannamaker, S. P. Lipshitz, J. Vanderkooy, and J. N. Wright, "A Theory of Nonsubtractive Dither," *IEEE Transactions on Signal Processing*, vol. 48, no. 2, pp. 499–516, 2000.

[11] A. Papoulis, *The Fourier Integral and Its Applications*. McGraw-Hill, 1962.

[12] P. Carbone and D. Petri, "Effect of Additive Dither on the Resolution of Ideal Quantizers," *IEEE Transactions on Instrumentation and Measurement*, vol. 43, no. 3, pp. 389–396, 1994.

[13] R. Skartlien and L. Øyehaug, "Quantization Error and Resolution in Ensemble Averaged Data With Noise," *IEEE Transactions on Instrumentation and Measurement*, vol. 54, no. 3, pp. 1303–1312, Jun. 2005.

[14] A. Papoulis and S. U. Pillai, *Probability, random variables, and stochastic processes*, 4th ed. McGraw-Hill, 2002.

[15] S. Boyd and C. Barratt, *Linear Controller Design: Limits of Performance*. Prentice Hall, 1991.

[16] T. Sauer, *Numerical analysis*, 2nd ed. Pearson, 2012.

[17] G. Casella and R. L. Berger, *Statistical Inference*, 2nd ed. Cengage Learning, 2001.

[18] T. W. Parks and C. S. Burrus, *Digital Filter Design*. Wiley, 1987.