

Optimal Quantiser Performance with a Small Dither

Arnfinn A. Eielsen^{a,*}, John Leth^b, Andrew J. Fleming^c

^a*Department of Electrical Engineering and Computer Science, University of Stavanger, Postboks 8600, 4036 Stavanger, Norway*

^b*Department of Electronic Systems, Automation & Control, Aalborg University, Denmark*

^c*School of Electrical Engineering and Computer Science, The University of Newcastle, Australia*

Abstract

A quantiser is a non-smooth function and no inverse function exist that can be applied to correct for the error it introduces. Applying a dither signal and averaging to a quantiser produces a smooth image for which an inverse function does exist. This article describes methods for minimising the error after inverse compensation. We show that there is an optimal dither variance that minimises the error after inversion. Simple rules for choosing the optimal dither variance are presented. The error after inversion can be made arbitrarily small by increasing the averaging length. This can be done by oversampling the signal by the same factor as the number of averages. Quantisation of a dither signal with a continuous probability distribution, produces a discrete probability mass function. We discuss a method for recovering an unknown continuous probability distribution from the empirical discrete probability mass function of the quantised dither signal. This enables inverse compensation in systems where exact control of the dither signal is not possible, as inverse compensation requires information about the continuous probability distribution of the dither signal and the step-size of the quantiser.

Keywords: Analog digital conversion, Quantisation error, Nonlinear distortion, Dither, Averaging, Gaussian noise

1. 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 correspondence between the expected value of the quantised signal and the reference signal can be obtained; that is, $E\{y\} = N(x)$, where $y = y(t)$ is the quantized version of $x = x(t)$, as illustrated in Fig. 4 (see also Fig. 5)

By applying averaging, a resolution below the step-size of the dithered quantiser can then be achieved [1]. 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.

In general, the function $N(x)$ is 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 by using the inverse of the function [2].

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 [3, 4, 5, 6, 7, 8]. 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 [7].

However, dither distributions that render the moments of the quantised signal independent of the reference, e.g. a uniform distribution, are infeasible to produce in a physical system as they require a high degree of precision in the dither signal reproduction. 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 [9]. If the realisation of the distribution is not exact, the linearising effect is significantly reduced [10]. In contrast, since introducing dynamics in a signal path can be done using e.g. linear filters, 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. 3.4. When applying averaging, considering the combined 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 $\Delta/2$ [10, 11]. This error is illustrated in Fig. 5 using a non-optimal dither variance to

*Corresponding author.

Email addresses: eielsen@ux.uis.no (Arnfinn A. Eielsen), jjl@es.aau.dk (John Leth), andrew.fleming@newcastle.edu.au (Andrew J. Fleming)

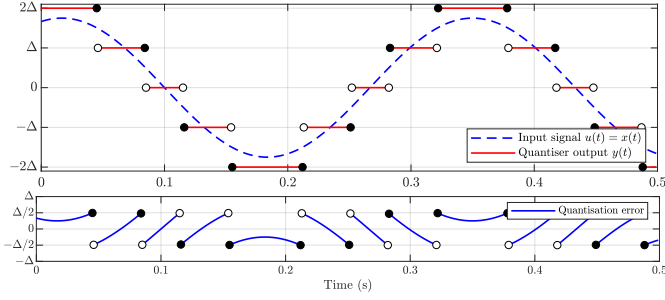


Figure 1: A quantised signal will have a signal dependent error, referred to as the quantisation error: $y(t) - x(t)$.

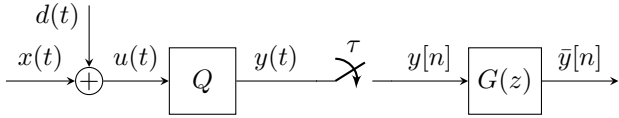


Figure 2: Quantiser dithering and averaging using an LTI-filter.

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 reduces the overall error [2]. This is illustrated in Fig. 6.

By using the inverse, it may be intuitively 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. 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 contributes to the overall error more than the removal of the systematic error due to $N(x)$ can reduce it. An optimal dither variance that minimises the overall error in the inverted signal therefore exists.

1.1. 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 [10, 2, 11]. This approach makes it simpler to prove convergence to a linear function, as compared to the approaches in [3, 4, 5, 7, 8].

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.

1.2. 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, $\mu = E\{d(t)\}$ denote the mean value, $C_d(t, s) \triangleq E\{d(t)d(s)\} - E\{d(t)\}E\{d(s)\}$ denote the auto-covariance and $C_d(t) \triangleq C_d(t, t)$ the variance. The indicator function for the set S is denoted $\chi_S(z)$ and the floor operator by $\lfloor \cdot \rfloor$.

2. 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 $\text{trunc}(u)$ as

$$\text{trunc}(u) \triangleq \left\lfloor \frac{u}{\Delta} + \frac{1}{2} \right\rfloor \in \mathbb{Z}. \quad (1)$$

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

$$y = n(u) \triangleq \Delta \text{trunc}(u). \quad (2)$$

The uniform quantiser is demonstrated in Fig. 1.

2.1. Alternative formulation

The uniform quantiser can equivalently be expressed as

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

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

$$n_k(u) \triangleq n_k^+(u) - n_k^-(u) \quad (4a)$$

in the quantiser, is described by the step functions

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

$$n_k^-(u) \triangleq \Delta H(-u - T_k), \quad (4c)$$

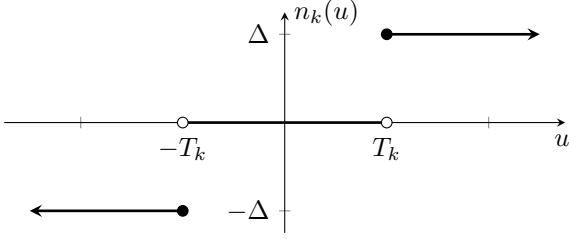


Figure 3: Step-function.

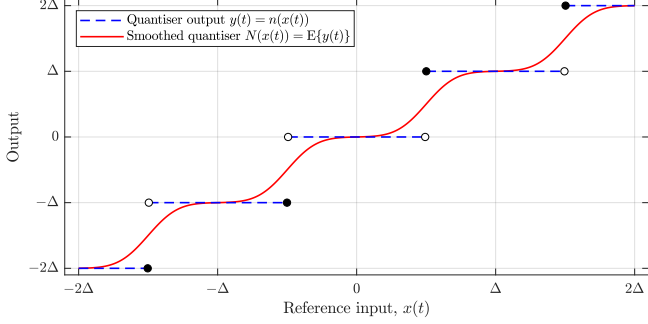


Figure 4: A quantiser is not bijective as it maps the continuum to a discrete set, but the smoothed quantiser (17) is; thus giving a one-to-one correspondence between reference and expected value. Here a zero-mean, normally distributed dither (with $\sigma_d = \Delta/6$) has been used.

with the threshold of the step T_k given by

$$T_k \triangleq \Delta(k + 1/2) \quad (5)$$

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 $T_k \geq u \geq -T_k$ or equivalently when $k \geq |u|/\Delta - 1/2$.

3. Smoothing the quantiser

The input signal u when a stochastic dithering signal is present is

$$u(t) \triangleq x(t) + d(t), \quad (6)$$

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 property, i.e. $d(t_i)$ and $d(t_j)$ are independent for each pair t_i and t_j [12]. In the sequel the input u is viewed as a function of x and t viz.,

$$u = u(x, t) = x + d(t),$$

when the time dependency of x is irrelevant. If this is not the case we write $u(t) = u(x(t), t)$.

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 $y(x, t) = n(u(x, t))$, 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 note that neither of these functions depend on time t by the i.d., property of $d(t)$. Moreover, let $\delta(\cdot)$ denotes the Dirac delta function and $x \in \mathbb{R}$. Then $f_x(z, x) = \delta(z - x)$ is the PDF corresponding to x which depends on time t if $x = x(t)$ does; $f_x(z, t) \triangleq f_x(z, x(t))$. The signal $u(x, t)$ will then have the PDF

$$f_u(z, x) = \int_{\mathbb{R}} f_x(z - w, x) f_d(w) dw = f_d(z - x). \quad (7)$$

with $f_u(z, t) \triangleq f_u(z, x(t))$ when $x = x(t)$.

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

$$N_k(z) \triangleq \int_{\mathbb{R}} n_k(z + w) f_d(w) dw, \quad (8)$$

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,

$$\|N_k\|_{\infty} \leq \|f_d\|_1 \|n_k\|_{\infty} = \|n_k\|_{\infty} \quad (9)$$

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 [13, Lemma A.1]

$$L_{N_k} \leq L_D TV(n_k) \quad (10)$$

with L_{N_k} a Lipschitz constant of $N_k(z)$ and $TV(n_k)$ the total variation of $n_k(u)$. In the case considered in this article $TV(n_k) = 2\Delta$, hence the dither PDF and quantisation step-size determines an upper bound on the smoothing effect on the quantiser. Moreover, using (4)

$$N_k(z) = \Delta \left(\int_{T_k - z}^{\infty} f_d(w) dw - \int_{-\infty}^{-T_k - z} f_d(w) dw \right) \quad (11a)$$

$$= \Delta (1 - F_d(T_k - z) - F_d(-T_k - z)). \quad (11b)$$

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

$$N(z) \triangleq \sum_{k \in \mathbb{N}} N_k(z) \quad (12a)$$

$$= \sum_{n \in \mathbb{Z}} \Delta \left(\frac{1}{2} - F_d(T_n - z) \right) \quad (12b)$$

is finite for each z . The summand k in (12a) correspond to the two summands $n = k$ and $n = -(k + 1)$ in (12b).

In the special case where $f_d(z)$ is symmetric around zero¹ (11) can be written as

$$N_k(z) = \Delta \int_{-T_k - z}^{-T_k + z} f_d(w) dw \quad (13a)$$

$$= \Delta (F_d(-T_k + z) - F_d(-T_k - z)). \quad (13b)$$

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

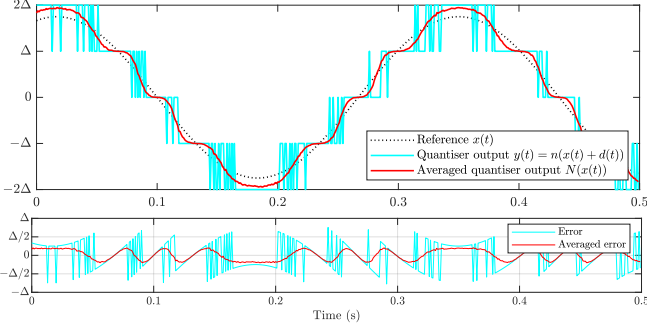


Figure 5: Applying dither makes it possible to reduce the quantisation error if the output is averaged. Here a zero-mean, normally distributed dither (with $\sigma_d = \Delta/6$) has been used. The averaged signal was obtained with $M = 1000$ samples.

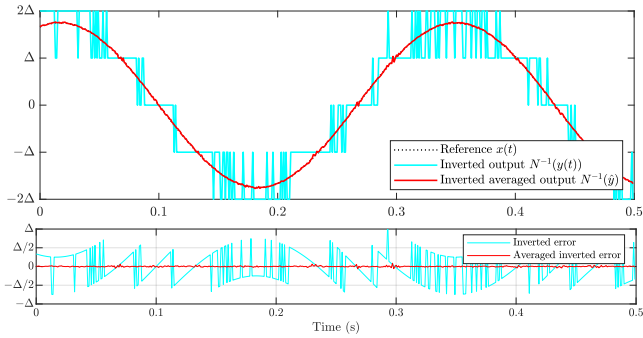


Figure 6: The inverse of the non-linear function relating the reference and expected value can be used to reduce the quantisation error of the averaged signal. Using a zero-mean, normally distributed dither with $(\sigma_d = \Delta/6)$, the averaged signal $\bar{y} = \langle y(t) \rangle_M$ was obtained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

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

3.1. First and Second Order Moments

Consider the strictly white noise output for each step

$$y_k(x, t) \triangleq n_k(u(x, t)), \quad (14)$$

and the (strictly white noise) quantised output signal

$$y(x, t) \triangleq \sum_{k \in \mathbb{N}} y_k(x, t), \quad (15)$$

with $y_k(t) \triangleq y_k(x(t), t)$ and $y(t) \triangleq y(x(t), t)$ if $x = x(t)$.

Using (8), the expected value for each step is

$$\mathbb{E}\{y_k(x, t)\} = \int_{\mathbb{R}} n_k(u) f_d(u - x) du \quad (16a)$$

$$= \int_{\mathbb{R}} n_k(x + w) f_d(w) dw \quad (16b)$$

$$= N_k(x). \quad (16c)$$

The expected value of the quantiser output is therefore

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

With (17) and $x = x(t)$, this leads to an output auto-covariance given by

$$C_y(t, s) = \mathbb{E}\{y(t)y(s)\} - \mathbb{E}\{y(t)\}\mathbb{E}\{y(s)\} \quad (18a)$$

$$= \begin{cases} 0 & t \neq s \\ \mathbb{E}\{y^2(t)\} - N^2(x(t)) & t = s \end{cases} \quad (18b)$$

$$\triangleq v(t)\chi_{\{t\}}(s) \quad (18c)$$

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

It is remarked that the right-hand side of (16) and (17) depend on time only when $x = x(t)$, implying the same for the left-hand sides. That is, the time evolution of the stochastic dither does not affect the expected value of the output. This is in general not the case for the auto-covariance (18). However, as shown below the variance $v = v(t)$ can be viewed as a function of x when time dependency of $x(t)$ is irrelevant.

Consider the zero-mean strictly white noise error signal

$$\varepsilon(x, t) = y(x, t) - N(x), \quad (19)$$

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

$$C_\varepsilon(t) = C_y(t) = v(t) \quad (20)$$

of the error signal $\varepsilon(t) \triangleq \varepsilon(x(t), t)$, 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

$$v(t) = \sum_{k \in \mathbb{Z}} \text{var}(y_k(t)) + 2 \sum_{k \in \mathbb{Z}} \sum_{l > k} \text{cov}(y_k(t), y_l(t)), \quad (21)$$

to find the variance of the error.

For this note first that the Heaviside step-function has the properties

$$H^2(u) = H(u), \quad (22)$$

and (more generally)

$$H(u - a)H(u - b) = H(u - b) \quad (23)$$

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

$$\begin{aligned} \text{var}(y_k(t)) &\triangleq \mathbb{E}\{y_k^2(t)\} - N_k(x(t))^2 \\ &= \Delta^2(1 - F_d(T_k - x(t)) + F_d(-T_k - x(t))) \\ &\quad - N_k(x(t))^2 \end{aligned} \quad (24a)$$

$$= 2\Delta^2 F_d(-T_k - x(t)) + \Delta N_k(x(t)) - N_k(x(t))^2 \quad (24b)$$

using (22) to obtain (24a), and (11b) to obtain (24b). For $k < l$, the covariance between terms can be found as

$$\begin{aligned} \text{cov}(y_k(t), y_l(t)) &\triangleq \mathbb{E}\{y_k(t)y_l(t)\} - \mathbb{E}\{y_k(t)\}\mathbb{E}\{y_l(t)\} \\ &= \Delta^2(1 - F_d(T_l - x(t)) + F_d(-T_l - x(t))) \\ &\quad - N_k(x(t))N_l(x(t)) \end{aligned} \quad (25)$$

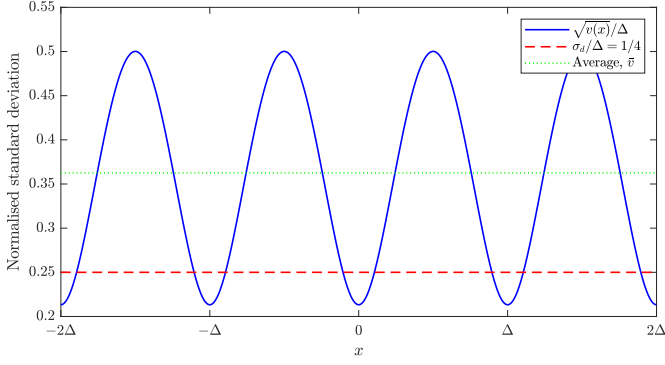


Figure 7: The normalised standard deviation for the normally distributed case with $\sigma_d/\Delta = 1/4$.

using (23). Note that when the time dependency of $x(t)$ is irrelevant then, from (24) and (25), the variance v may be viewed as a function of x viz., $v = v(x)$ with the special case $v(t) = v(x(t))$. For this reason (20) is generalised to

$$C_\varepsilon(x) = C_y(x) = v(x) \quad (26)$$

with (20) obtained from (26) by letting $x = x(t)$.

An example of the variance $v(x)$ and its dependence on x is shown in Fig. 7.

3.2. Convergence to a linear function

We show that $\frac{d}{dz}N(z)$ 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 (31)).

From (12b) we get

$$\frac{d}{dz}N(z) = \sum_{n \in \mathbb{Z}} \Delta f_d(T_n - z) \quad (27a)$$

$$= \sum_{n \in \mathbb{Z}} \Delta f_d(\Delta/2 - z + n\Delta), \quad (27b)$$

which by the Poisson sum formula [9] yields:

$$\frac{d}{dz}N(z) = \sum_{n \in \mathbb{Z}} \hat{f}_d\left(-\frac{2\pi n}{\Delta}\right) e^{i\frac{2\pi n}{\Delta}(z-\Delta/2)} \quad (28)$$

$$= \sum_{n \in \mathbb{Z}} \varphi_d\left(\frac{2\pi n}{\Delta}\right) e^{i\frac{2\pi n}{\Delta}(z-\Delta/2)} \quad (29)$$

$$= 1 + \sum_{n \in \mathbb{Z} \setminus \{0\}} \varphi_d\left(\frac{2\pi n}{\Delta}\right) e^{i\frac{2\pi n}{\Delta}(z-\Delta/2)} \quad (30)$$

It can then be seen that if

$$\varphi_d\left(\frac{2\pi n}{\Delta}\right) = 0 \quad \forall n \in \mathbb{Z} \setminus \{0\}, \quad (31)$$

then the derivative of $N(z)$ is constant, and hence $N(z)$ is affine linear, that is, when (31) is fulfilled, then the quantisation error

$$e_x(x(t)) \triangleq N(x(t)) - x(t) = N(0), \quad (32)$$

is constant. Note that the quantisation error (32) is in terms of the reference $x(t)$, whereas the output error (19), with $x = x(t)$ 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 [3, 4, 5, 7, 8]. Below are two examples of distributions that can fulfill the criterion (31) and in principle linearise the quantiser and remove the quantisation error. As will be seen, neither have practical use in measurement systems.

3.3. Application of a uniformly distributed dither

The simplest distribution that can fulfil (31) is the uniform distribution [7, 8]. In this case

$$f_d(z) = \begin{cases} \frac{1}{\Delta}, & |z| \leq \frac{\Delta}{2} \\ 0, & |z| > \frac{\Delta}{2} \end{cases} \quad (33)$$

which has characteristic function

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

hence

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

If the dither signal has a uniform distribution, the dither signal reconstruction must be 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.

3.4. Application of a normally distributed dither

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

$$f_d(z) = \frac{1}{\sqrt{2\pi}\sigma_d} e^{-\frac{(z-\mu_d)^2}{2\sigma_d^2}} \quad (34)$$

$$\varphi_d(v) = e^{i\mu_d v} e^{-\frac{1}{2}\sigma_d^2 v^2}, \quad (35)$$

hence

$$\varphi_d\left(\frac{2\pi n}{\Delta}\right) = e^{i\mu_d \frac{2\pi n}{\Delta}} e^{-\sigma_d^2 \frac{2\pi^2 n^2}{\Delta^2}}, \quad (36)$$

and $N(z)$ will therefore approach an affine linear function as $\sigma_d \rightarrow \infty$, i.e. this distribution only fulfills (31) for an infinite standard deviation. An infinite variance is of course impractical, and in general, dithering with a large variance has limited practical application.

4. Averaging

4.1. Sample-averaging

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

$$y(x, t) = N(x) + \varepsilon(x, t) = E\{y(x, t)\} + \varepsilon(x, t). \quad (37)$$

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

For $i = 1, 2, \dots, M$ let $d_i(t)$ denote a stochastic dither signal and assume that $d_1(t), d_2(t), \dots, d_M(t)$ 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

$$y^i(x, t) \triangleq \sum_{k \in \mathbb{N}} y_k^i(x, t) \triangleq \sum_{k \in \mathbb{N}} n_k(x + d_i(t)). \quad (38)$$

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

$$\langle y(x, t) \rangle_M = \frac{1}{M} \sum_{i=1}^M y^i(x, t), \quad (39)$$

converges almost surely to $E\{y(x, t)\} = N(x)$ by the law of large numbers. That is,

$$\langle y(x) \rangle = N(x)$$

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

$$C_{\langle y \rangle_M}(x) = \frac{1}{M} C_y(x) = \frac{1}{M} v(x), \quad (40)$$

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

$$\begin{aligned} & E\{(\langle y(x, t) \rangle_M - E\{y(x, t)\})^2\} \\ &= E\{(\langle y(x, t) \rangle_M - E\{\langle y(x, t) \rangle_M\})^2\} = C_{\langle y \rangle_M}(x). \end{aligned} \quad (41)$$

Moreover, for

$$\begin{aligned} \langle \varepsilon(x, t) \rangle_M &\triangleq \frac{1}{M} \sum_{i=1}^M \varepsilon^i(x, t) \\ &= \langle y(x, t) \rangle_M - E\{y(x, t)\}, \end{aligned} \quad (42)$$

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

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

and

$$C_{\langle \varepsilon \rangle_M}(x) = \frac{1}{M} C_\varepsilon(x) = \frac{1}{M} C_y(x), \quad (43)$$

with the time dependent case $x = x(t)$ as a special case.

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.

4.2. 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

$$\bar{y}[n] = (g * N(x))[n] + (g * \varepsilon)[n] \quad (44a)$$

$$= N(x[n]) + (h * N(x))[n] + (g * \varepsilon)[n] \quad (44b)$$

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

$$\bar{y}(z) = G(z)y(z) \quad (45a)$$

$$= \mathcal{Z}\{N(x)\}(z) + H(z)\mathcal{Z}\{N(x)\}(z) + G(z)\mathcal{E}(z), \quad (45b)$$

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 [14] do not apply. However, an LTI-filter will still reduce the variance of the error signal $\varepsilon(t)$ as will be shown next. First, the filtered error signal is defined as

$$\bar{\varepsilon}[n] \triangleq (g * \varepsilon)[n], \quad (46)$$

where $g[n]$ is the output filter and $\varepsilon[n]$ is the error signal. The variance of the filtered error signal is then

$$C_{\bar{\varepsilon}}[n, n] = E\{\bar{\varepsilon}^2[n]\} \quad (47a)$$

$$= (v * g^2)[n], \quad (47b)$$

where (47b) follows by the discrete-time version of [12, Theorem 9-3]. Hence if $v \in l^1$ and $g^2 \in l^2$ then

$$\|C_{\bar{\varepsilon}}\|_2 = \|v * g^2\|_2 \quad (48a)$$

$$\leq \|v\|_1 \|g^2\|_2 \quad (48b)$$

where (48b) follows by Young's convolution theorem for sequences. By (48) the variance of the filtered error signal will be attenuated by the energy of the impulse response squared $\|g^2\|_2$. 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 $\bar{\varepsilon}$.

4.2.1. 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

$$\mathbb{E}\{\bar{\varepsilon}^2[n]\} = \frac{1}{M^2} \sum_{i=0}^{M-1} v[n-i] \quad (49)$$

according to (47), and it will be attenuated by

$$\|g^2\|_2 = \frac{1}{M}, \quad (50)$$

according to (48).

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

$$G(z) = \frac{1}{M} \sum_{i=0}^{M-1} z^{-i} = \frac{1}{M} \frac{1 - z^{-M}}{1 - z^{-1}}. \quad (51)$$

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

$$G(f) = \frac{\text{sinc}(Mf)}{\text{sinc}(f)} e^{-i2\pi(\frac{M}{2}-1)f}. \quad (52)$$

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. 5.2.

4.2.2. Accumulate-and-dump filter

A large number of samples is desirable in order to reduce the variance due to the dither signal. However, in terms of implementability, it may be necessary 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

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

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

4.3. 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 $v(x)$ can be approximated by an average [2, 11]; the limiting case being that x varies linearly across one step.

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

$$\bar{v} \triangleq \frac{1}{\Delta} \int_{-\Delta/2}^{\Delta/2} v(x) dx \quad (53)$$

and the variance $C_{\langle \varepsilon \rangle_M}(x)$ from (43) can be approximated by

$$C_{\langle \varepsilon \rangle_M}(x) \approx \frac{1}{M} \bar{v}. \quad (54)$$

using (40).

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

$$\bar{v}_d = \frac{1}{M} \sum_{k=0}^{M-1} v\left(-\frac{\Delta}{2} + \frac{\Delta}{M}k\right) \quad (55)$$

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

$$\mathbb{E}\{\bar{\varepsilon}^2[n]\} \approx \frac{1}{M} \bar{v}_d, \quad (56)$$

by (49).

Note that if Δ/M in (55) 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 . By assuming that the variance is constant, the process describing the error becomes stationary and information about the worst-case error variance is neglected.

5. Inverting the Effective Non-linearity

Consider again the output of the quantiser (37) with $x = x(t)$. 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

$$x(t) = N^{-1}(y(t)). \quad (57)$$

Clearly (57) 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. $\langle y(t) \rangle_M \triangleq \langle y(x(t), t) \rangle_M$ 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

$$F(z) \triangleq N(z) - \hat{y}. \quad (58)$$

One such method is the bisection method [15], 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. See Fig. 6 for an example of $N^{-1}(y(t))$, solved using the bisection method.

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,

$$N^{-1}(y(t)) \approx x(t) + \alpha(x(t))\varepsilon(t), \quad (59)$$

where we have defined the sensitivity α as

$$\alpha(x) \triangleq \left. \frac{dN^{-1}}{dw} \right|_{w=N(x)} = \left(\left. \frac{dN}{dz} \right|_{z=x} \right)^{-1}. \quad (60)$$

Using (59) directly, or the delta-method [16], an approximation to the variance of $N^{-1}(y(t))$ can be found as

$$C_{N^{-1}(y)}(x) \approx \alpha^2(x)v(x), \quad (61)$$

with $x(t)$ replaced by x since time dependence is irrelevant.

5.1. 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 (59) and (42) we get

$$N^{-1}(\langle y(t) \rangle_M) \approx x(t) + \varepsilon_{e,M}(t), \quad (62)$$

where

$$\varepsilon_{e,M}(t) = \alpha(x(t))\langle \varepsilon(t) \rangle_M \quad (63)$$

denotes the stochastic error due to the dithering signal, see Fig. 6 for an example of $N^{-1}(\langle y(t) \rangle_M)$, found solving (58). Using (43) and (40), the variance of (62) is

$$\begin{aligned} C_{N^{-1}(\langle y \rangle_M)}(x) &\approx C_{\varepsilon_{e,M}}(x) \\ &= \frac{1}{M} \alpha^2(x) v(x) \approx \frac{1}{M} C_{N^{-1}(y)}(x), \end{aligned} \quad (64)$$

with $x(t)$ replaced by x since time dependence is irrelevant. Hence the sample-average reduces the variance (61) by approximately a factor of $1/M$.

5.2. Inverting the time-average

Time-averaging using an LTI low-pass filter $G(z)$ such as (51) 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 (45b). From (59) and (19), the inverse can be written as

$$N^{-1}(\bar{y}[n]) \approx x[n] + e_f[n] + \varepsilon_t[n], \quad (65)$$

where

$$e_f[n] = \alpha[n](h * N(x))[n], \quad (66)$$

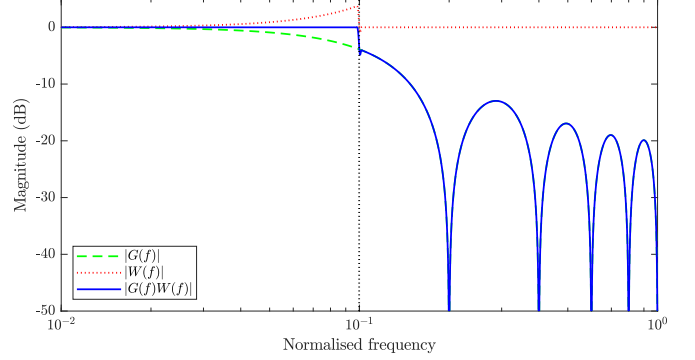


Figure 8: The frequency response of the moving average filter $G(z)$ in the passband can be compensated for by using a pre-filter $W(z)$. By using the pre-filter, the amplitude response can be made to be unity in the passband, compensating for the amplitude attenuation by the moving average filter. The filter $W(z)$ has been synthesised using the least-squares method [17].

denotes the deterministic error due to filtering, and

$$\varepsilon_t[n] = \alpha[n](g * \varepsilon)[n]. \quad (67)$$

denotes the stochastic error due to the dithering signal. Under the assumptions of Sec. 4.3, the variance of $\varepsilon_t[n]$ will be reduced by a factor of $1/M$.

Note that if $x(t)$ is bandwidth limited, it is straightforward to compensate for the error $e_f[n]$ by designing a finite impulse-response (FIR) compensating pre-filter $W(z)$ which minimises

$$\|1 - G(f)W(f)\|_2 \quad (68)$$

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

$$\|1 - W(f)\|_2 \quad (69)$$

elsewhere, as demonstrated in Fig. 8 for the moving average filter. Given (68), 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 $H(z)$ can be ignored (e.g. it was neglected for the results obtained in Fig. 11).

5.3. Optimal dither variance when inverting

The variance $C_{N^{-1}(\langle y \rangle_M)}(x)$ of $N^{-1}(\langle y(t) \rangle_M)$ from (64) depends on the sensitivity $\alpha(x)$ from (60) and the variance of the error $v(x)$ from (26). Consider (27) for a normal distribution: The variance $v(x)$ will then have maxima at the thresholds (5), that is at $x = T_k$, as seen in Fig. 7. Moreover, the sensitivity $\alpha(x)$ will have maxima at $k\Delta$, as demonstrated in Fig. 10, with the maximum sensitivity tending to infinity as σ_d goes to zero (and $\alpha(x)$ tending to a constant as σ_d goes to ∞). It then follows that the variance $C_{N^{-1}(\langle y \rangle_M)}(x)$ will be dominated by $\alpha(x)$ when $\sigma_d/\Delta \ll 1$ and by $v(x)$ when σ_d/Δ grows larger.

As shown below it turns out that there is an optimal σ_d , balancing the sensitivity and the magnitude of σ_d . Two

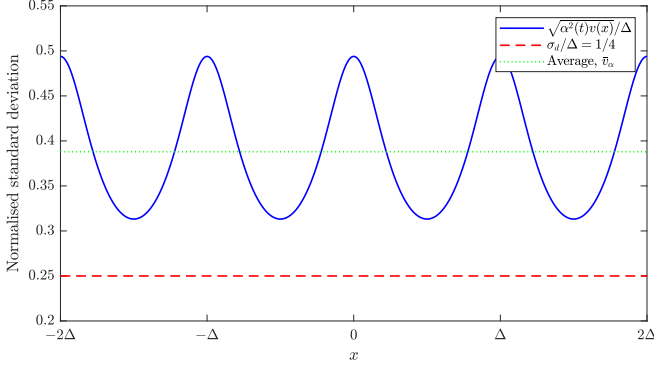


Figure 9: The variance $\alpha^2(x)v(x)$ from (61). As Fig. 7, the plot shows the normalised standard deviation for the normally distributed case with $\sigma_d/\Delta = 1/4$.

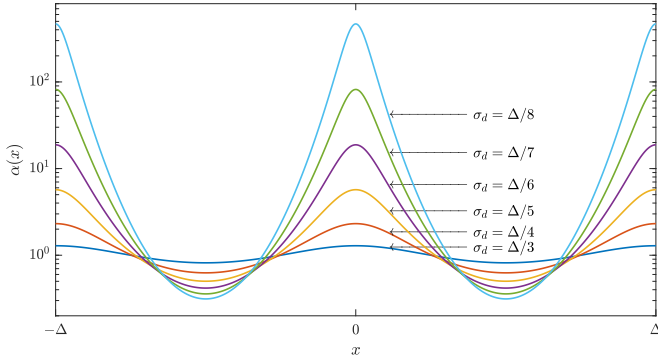


Figure 10: The sensitivity $\alpha(x)$ from (60) and its dependency on the variance σ_d for a normally distributed dither.

metrics for $C_{N^{-1}(\langle y \rangle_M)}(x)$ can be devised, suitable for different conditions on the reference signal: An averaged case where it is assumed that $x = x(t)$ varies linearly over the quantisation step-size (small Δ , as suggested in [2, 11]), and a worst case where the maximum variance is considered.

5.3.1. Small Δ

As in Sec. 4.3, it is assumed that Δ is small in the sense that $x(t)$ varies approximately linearly over the quantisation step-size. Then the variance $C_{N^{-1}(\langle y \rangle_M)}(x)$ can be approximated by an average variance $\bar{v}_\alpha$ with

$$C_{N^{-1}(\langle y \rangle_M)}(x) \approx \bar{v}_\alpha \triangleq \frac{1}{\Delta} \int_{-\Delta/2}^{\Delta/2} \frac{1}{M} \alpha^2(x) v(x) dx \quad (70)$$

$$\leq \frac{\bar{v}}{M} \int_{-\Delta/2}^{\Delta/2} \alpha^2(x) dx \triangleq a_C, \quad (71)$$

where the Cauchy-Schwarz inequality and (53) has been used to obtain (71). Note that $a_C = a_C(M, \sigma_d)$, so for each M an approximative smallest upper bound $a_C(M, \sigma_d^*)$ on the variance $C_{N^{-1}(\langle y \rangle_M)}(x)$ can be obtained through (71) by varying σ_d as in Fig. 10. This is shown in Fig. 11 for a scaled version of $a_C(M, \sigma_d/\Delta)$.

It is also informative to look at a mean square error (MSE) estimate of the quantisation error $e_q(x(t)) \triangleq y(t) - x(t)$, which is the case without inversion [11]. Consider the sample average

$$\begin{aligned} \langle e_q(x(t)) \rangle_M &\triangleq \frac{1}{M} \sum y^i(t) + x(t) \\ &= \langle y(t) \rangle_M - x(t) \\ &= e_x(x(t)) + \langle \varepsilon(t) \rangle_M \end{aligned} \quad (72)$$

where (72) follows from (42) and (32). The second moment of the sample average is therefore

$$\begin{aligned} E\{\langle e_q(x(t)) \rangle_M^2\} &= e_x^2(x(t)) + C_{\langle e_q \rangle_M}(x(t)) \\ &= e_x^2(x(t)) + C_{\langle \varepsilon \rangle_M}(x(t)) = e_x^2(x(t)) + \frac{1}{M} v(x(t)) \end{aligned} \quad (73)$$

Setting $x = x(t)$ and averaging yields an MSE

$$\text{mse} = \frac{1}{\Delta} \int_{-\Delta/2}^{\Delta/2} E\{\langle e_q(x) \rangle_M^2\} dx = \bar{e}_x^2 + \frac{1}{M} \bar{v} \quad (74)$$

where the MSE of the deterministic component is

$$\bar{e}_x^2 \triangleq \frac{1}{\Delta} \int_{-\Delta/2}^{\Delta/2} e_x^2(x) dx, \quad (75)$$

and is in addition to the average variance $\bar{v}$ representing the stochastic part of e_q . The MSE can be computed numerically, as in Fig. 11, or using the expressions presented in [11].

5.3.2. Worst case

If no assumptions can be made about $x(t)$, the worst-case variance

$$m_C = \max_x C_{N^{-1}(\langle y \rangle_M)}(x) \approx \frac{1}{M} \max_x \alpha^2(x) v(x). \quad (76)$$

will be a more reasonable estimate of $C_{N^{-1}(\langle y \rangle_M)}(x)$. Note that $m_C = m_C(M, \sigma_d)$, similar to a_C , which is shown in Fig. 11 for a scaled version.

5.3.3. Optimal variances

Upper bounds for the standard deviation $\sqrt{C_{N^{-1}(\langle y \rangle_M)}(x)}$ from (71) and (76), and the root mean squared error (RMSE) using (74) are shown in Fig. 11 for a normally distributed dither signal for different standard deviations σ_d and three values of M . Both axes have been normalised relative to the step-size Δ . The numerical values of the optimal cases are presented in Tab. 1.

Considering (74), it can be seen that increasing M decreases the error due to stochastic dither, $\bar{v}$. The contribution to the overall error due to the deterministic component $\bar{e}_x$ therefore becomes more significant relative to the stochastic component when using a large number of averages. This is the reason the minimal RMSE depends on M , in addition to the dither variance σ_d . When applying the inverse $N^{-1}(z)$, the deterministic component $\bar{e}_x$ is

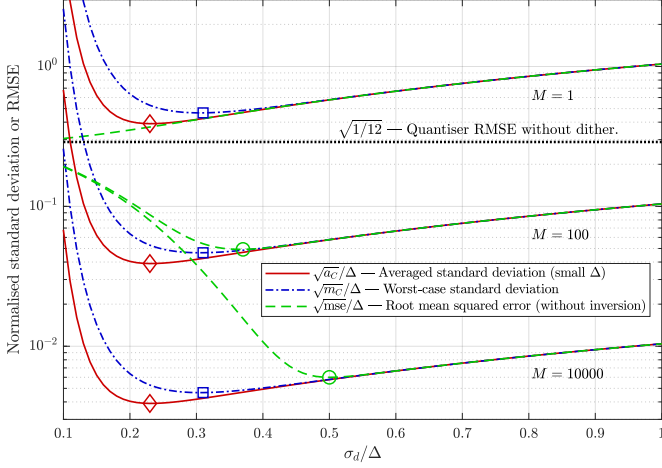


Figure 11: Upper bounds for the standard deviation $\sqrt{C_{N^{-1}(\langle y \rangle_M)}(x)}$, using normally distributed dither. Diamond, square and circle indicate optimal σ_d/Δ for the three cases.

Table 1: Optimal σ_d/Δ values for the three cases: The averaged standard deviation a_C , the worst-case standard deviation m_C and the root mean squared error (mse) without using inverse compensation.

	$M = 1$ (number of averages)
$\sqrt{a_C}/\Delta$	0.390 for $\sigma_d/\Delta = 0.230$
$\sqrt{m_C}/\Delta$	0.466 for $\sigma_d/\Delta = 0.310$
$\sqrt{\text{mse}}/\Delta$	0.289 for $\sigma_d/\Delta = 0.00$
	$M = 100$
$\sqrt{a_C}/\Delta$	0.0390 for $\sigma_d/\Delta = 0.230$
$\sqrt{m_C}/\Delta$	0.0466 for $\sigma_d/\Delta = 0.310$
$\sqrt{\text{mse}}/\Delta$	0.0493 for $\sigma_d/\Delta = 0.370$
	$M = 10000$
$\sqrt{a_C}/\Delta$	0.00390 for $\sigma_d/\Delta = 0.230$
$\sqrt{m_C}/\Delta$	0.00466 for $\sigma_d/\Delta = 0.310$
$\sqrt{\text{mse}}/\Delta$	0.00600 for $\sigma_d/\Delta = 0.500$

removed from the error, and the variances a_C and m_C are then only scaled by the factor $1/M$. The minimal error variance is determined by the relative significance of the size of σ_d and the sensitivity α which depends on σ_d .

By numerical evaluation, it was found that $a_C(M, \sigma_d)$ gave a smaller minimizer than $\text{mse}(M, \sigma_d)$ when $M \geq 4$, and that $m_C(M, \sigma_d)$ gave a smaller minimizer than $\text{mse}(M, \sigma_d)$ when $M \geq 38$. Hence, for $M \geq 4$ inverting the average output provides improved performance over only averaging an optimally dithered quantiser.

6. 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

$$f_u(z) = f_d(z - \mu_x). \quad (77)$$

In a similar manner as described in [7], the probability that the input signal $u(t) = \mu_x + d(t)$ will truncate to a given integer $k \in \mathbb{Z}$ can be found by considering the probability of $u(t)$ in the domain $\Delta[k - 1/2, k + 1/2]$

$$f(k) \triangleq \int_{\Delta(k-1/2)}^{\Delta(k+1/2)} f_d(z - \mu_x) dz \quad (78)$$

$$= F_d(\Delta(k + 1/2) - \mu_x) - F_d(\Delta(k - 1/2) - \mu_x) \quad (79)$$

with F_d the CDF of d . If $d(t)$ is normally distributed, then the input distribution is

$$f_u(z; \mu_u, \sigma_d) = \frac{1}{\sqrt{2\pi} \sigma_d} e^{-\frac{(z - \mu_u)^2}{2\sigma_d^2}} \quad (80)$$

with $\mu_u = \mu_x + \mu_d$, and f from (78) can be written as

$$f(k; \mu_u, \sigma_d) = \text{erf}\left(\frac{\Delta(k + \frac{1}{2}) - \mu_u}{\sqrt{2} \sigma_d}\right) - \text{erf}\left(\frac{\Delta(k - \frac{1}{2}) - \mu_u}{\sqrt{2} \sigma_d}\right). \quad (81)$$

6.1. Identification of the dither probability density function

Assuming that the reference signal can be held constant and that the dither signal has a PDF of a known type, the parameters $\theta \in \mathbb{R}^p$ of the input PDF $f_u(k; \theta)$ can be determined from experiment and used to construct an accurate inverse $N^{-1}(z)$. By sampling a quantised signal, say $\{s_1, s_2, \dots, s_n\}$, when the reference is constant, an empirical probability mass function (PMF)

$$\hat{f}_n(k) \triangleq \frac{1}{n} \sum_{i=1}^n \delta(s_i - k)$$

can be produced. This can then be used to find the parameters θ , by solving the parameter identification problem

$$\min_{\theta} \|f(k; \theta) - \hat{f}_n(k)\|_2, \quad (82)$$

using the analytic expression (78). In the case of a normal distribution, robust estimates for $\hat{\mu}_u$ and $\hat{\sigma}_d$ are found by solving a problem of the form

$$\min_{\mu_u, \sigma_d} \left\| \sqrt{f(k; \mu_u, \sigma_d)} - \sqrt{\hat{f}_n(k)} \right\|_2, \quad (83)$$

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 $N^{-1}(z)$.

6.1.1. Numerical example

Using (83) 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 $n = 10^6$ samples have been used to produce the estimate $\hat{f}_n(k)$. Solving (83) with the data used in Fig. 12 yields $\hat{\mu}_u = -0.2502$ and $\hat{\sigma}_d = 0.3330$, which are very accurate estimates.

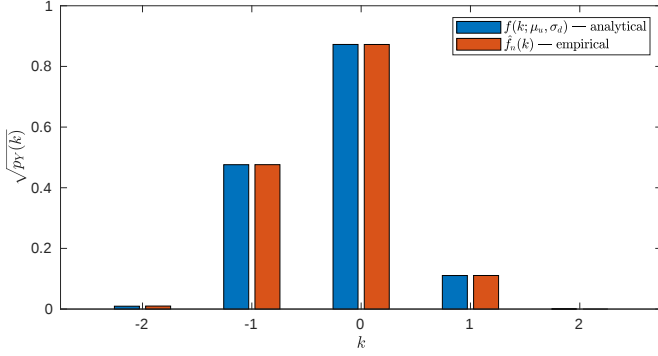


Figure 12: Histogram plot of $f(k)$ and $\hat{f}_n(k)$ for $n = 10^6$ samples of a quantiser with $\Delta = 1$ driven with a normally distributed signal with $\mu_u = -\Delta/4$ and $\sigma_d = \Delta/3$.

7. Conclusions

When dithering a quantiser, the expected value of the output has a one-to-one correspondence with the 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 by 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 may require a compensation filter to reduce amplitude and phase distortion of the reference signal due to the averaging operation.

It was demonstrated that when applying inversion there is an optimal dither variance that minimises the error after averaging. The optimal variance depends on assumptions about the reference signal: Either the step-size Δ is assumed to be small compared to the reference and an average error variance is considered, or the worst-case is considered by finding 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. For $M \geq 4$, using an optimal dither, inverting the average output provides improved performance over only averaging an optimally dithered quantiser. With a very small amount of averaging, applying inversion is better than mere averaging.

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] J. Vanderkooy, S. P. Lipshitz, Resolution Below the Least Significant Bit in Digital Systems with Dither, *Journal of the Audio Engineering Society* 32 (3) (1984) 106–113.

- [2] P. Koeck, Quantization errors in averaged digitized data, *Signal Processing* 81 (2) (2001) 345–356.
- [3] L. Schuchman, Dither Signals and Their Effect on Quantization Noise, *IEEE Transactions on Communication Technology* 12 (4) (Dec. 1964).
- [4] A. Sripad, D. Snyder, A necessary and sufficient condition for quantization errors to be uniform and white, *IEEE Transactions on Acoustics, Speech, and Signal Processing* 25 (5) (1977) 442–448.
- [5] S. P. Lipshitz, R. A. Wannamaker, J. Vanderkooy, Quantization and dither: A theoretical survey, *Journal of the audio engineering society* 40 (5) (1992) 355–375.
- [6] R. M. Gray, T. G. Stockham, Dithered Quantizers, *IEEE Transactions on Information Theory* 39 (3) (1993) 805–812.
- [7] B. Widrow, I. Kollar, M.-C. Liu, Statistical theory of quantization, *IEEE Transactions on Instrumentation and Measurement* 45 (2) (1996) 353–361.
- [8] R. A. Wannamaker, S. P. Lipshitz, J. Vanderkooy, J. N. Wright, A Theory of Nonsubtractive Dither, *IEEE Transactions on Signal Processing* 48 (2) (2000) 499–516.
- [9] A. Papoulis, *The Fourier Integral and Its Applications*, McGraw-Hill, 1962.
- [10] P. Carbone, D. Petri, Effect of Additive Dither on the Resolution of Ideal Quantizers, *IEEE Transactions on Instrumentation and Measurement* 43 (3) (1994) 389–396.
- [11] R. Skartlien, L. Øyehaug, Quantization Error and Resolution in Ensemble Averaged Data With Noise, *IEEE Transactions on Instrumentation and Measurement* 54 (3) (2005) 1303–1312.
- [12] A. Papoulis, S. U. Pillai, *Probability, random variables, and stochastic processes*, 4th Edition, McGraw-Hill, 2002.
- [13] L. Iannelli, K. H. Johansson, U. T. Jönsson, F. Vasca, Averaging of nonsmooth systems using dither, *Automatica* 42 (4) (2006) 669–676.
- [14] S. Boyd, C. Barratt, *Linear Controller Design: Limits of Performance*, Prentice Hall, 1991.
- [15] T. Sauer, *Numerical analysis*, 2nd Edition, Pearson, 2012.
- [16] G. Casella, R. L. Berger, *Statistical Inference*, 2nd Edition, Cengage Learning, 2001.
- [17] T. W. Parks, C. S. Burrus, *Digital Filter Design*, Wiley, 1987.