Rigorous Justification of Uniform White-noise Model for Roundoff Errors in Fixed-point Digital Filters

Phil Diamond [*], Igor Vladimirov [†]

Abstract. There are four sources of the roundoff noise arising in digital implementations of linear dicrete-time-invariant filters in fixed-point computer arithmetic with accuracy ε: overflow, coefficient quantization, input quantization and arithmetic errors. In the framework of the uniform white noise model, the last two kinds of roundoff errors are approximated by independent random vectors, distributed uniformly on appropriately dimensioned cubes. In its part concerning the input quantization errors, this heuristic model is justified by the well-known Quantization Theorem, which is not directly applicable to the arithmetic errors due to their complicated dependence on the input signal of the filter. In this paper, we provide a rigorous justification for the complete uniform white noise model in terms of weak convergence of probability measures as $\varepsilon \to 0$ under nonresonance conditions on the state-space realization matrices of the filter with generic random inputs.

1 Introduction

Quantization effects in digital systems have been a field of intensive research for decades (see [7] and references therein). These effects manifest themselves as the distortion of statistical properties of signals under analog-to-digital conversion and of functional properties of input-output systems under digitization of internal parameters and operations due to finiteness of machine arithmetic. When digitally implemented, linear systems become nonlinear.

The roundoff noise arising in digital implementations of linear filters in fixed-point computer arithmetic with accuracy ε comes from the following four sources: overflow, coefficient quantization, input quantization and arithmetic errors. In the framework of a commonly used paradigm, frequently called the uniform white noise model, the last two kinds of roundoff errors are approximated by independent random vectors, distributed uniformly on appropriately dimensioned cubes. In practice, the model is considered to be applicable if the dynamic range of the internal and external signals is reasonably large [10, 17].

In its part concerning the input quantization errors, this heuristic model is justified by the well-known Quantization Theorem which dates back to [2, 8, 15] and

[*]Department of Mathematics, University of Queensland, Brisbane, QLD 4072, Australia; e-mail: pmd@maths.uq.edu.au.

[†]Institute for Information Transmission Problems, Russian Academy of Sciences, 19 Bolshoi Karetny Lane, GSP–4, 101447 Moscow, Russia; current address: Department of Mathematics, University of Queensland, Brisbane, QLD 4072, Australia; e-mail: igv@maths.uq.edu.au.

describes the situation where a random signal undergoes simple analog-to-digital conversion. However, the theorem is not directly applicable to the arithmetic errors due to their extremely complicated dependence on the input signal of the filter. Note that the statistics of the arithmetic roundoff error in scalar multiplication without any dynamics was studied in [1, 12]. In this paper, we overcome the difficulty and provide a rigorous justification for the complete uniform white noise model both for the input quantization and arithmetic errors in digitally implemented linear discrete-time-invariant filters in terms of weak convergence of probability measures [3] as $\varepsilon \to 0$. The asymptotic validity of the model is shown to be essentially based on certain nonresonance conditions on the state-space realization matrices of the filter with generic random input.

The proof of the result exploits a quasiperiodic recursive representation of the roundoff errors in terms of the input signal, skew products of measure-preserving automorphisms of multidimensional torii and also a result on preservation of weak convergence of probability measures under continuously convergent mappings [13]. This machinery was developed in [14] for autonomous quantized linear systems and then extended to quantized nonlinear dynamical systems in [5].

The paper is organised as follows. In Section 2, digital multiplication in fixed-point arithmetic is described, and an auxiliry notion of a preproduct of matrices is defined. In Section 3, digital implementation of linear filters is described and the problem of justifying the uniform white noise model for the input quantization and arithmetic errors in terms of weak convergence of probability measures is stated. Section 4 formulates the main result of the paper on the asymptotic validity of the model, as well as defines generic input signals and nonresonance conditions on the state-space realization matrices of the filter. The complete proof of the main theorem is based on results of Sections 5–7 and is given in Section 8. In Section 5, skew products and continuous convergence of measure-preserving automorphisms of torii are considered. In Section 6, a quasiperiodic dependence of the input quantization and arithmetic errors on the input signal is established. In Section 7, a modification of the Quantization Theorem for linear mappings is proved. Section 9 contains concluding remarks which outline directions of further research.

2 Multiplication in fixed-point arithmetic

In fixed-point arithmetic with b digits after the binary point, a real number u is represented by the nearest node $R_\varepsilon(u)$ of the uniform grid $\varepsilon\mathbb{Z}$ with stepsize $\varepsilon = 2^{-b}$ (we do not take the overflow into account). With a convention to round up midpoints,

$$R_\varepsilon(u) = \varepsilon\lfloor u/\varepsilon + 1/2\rfloor, \tag{1}$$

where $\lfloor\cdot\rfloor$ is the floor function. Associated with the *roundoff operator* R_ε is the *magnified roundoff error*

$$E_\varepsilon(u) = (u - R_\varepsilon(u))/\varepsilon = E_1(u/\varepsilon) \tag{2}$$

which takes values in the interval $[-1/2, 1/2)$. The roundoff operator and magnified roundoff error extend in an obvious way to spaces of vectors and matrices, where they act component-wise. In particular, when extended to $\mathbb{R}^q$, the mapping E_ε takes values in the cube

$$V_q = [-1/2, 1/2)^q.$$

Suppose that the digital product $[u, v]_\varepsilon$ of two numbers $u, v \in \mathbb{R}$ is implemented as multiplication of their digital representations, performed with double precision and then followed by rounding the result off: $[u, v]_\varepsilon = R_\varepsilon(R_\varepsilon(u)R_\varepsilon(v))$. Similarly, multiplication of a column-vector $v \in \mathbb{R}^r$ with entries v_k by a matrix $M \in \mathbb{R}^{q\times r}$ with columns $m_{\bullet k}$ is implemented as

$$[M, v]_\varepsilon = \sum_{k=1}^{r} R_\varepsilon(R_\varepsilon(m_{\bullet k})R_\varepsilon(v_k)). \tag{3}$$

In the sequel, we shall often encounter expressions like (3). To write them in a matrix-operator form, denote by $A \odot B$ a *preproduct* of matrices $A \in \mathbb{R}^{q\times r}$ and $B \in \mathbb{R}^{r\times s}$ with columns $a_{\bullet k}$ and rows $b_{k\bullet}$, respectively, as the block $(qr \times s)$-dimensional matrix

$$A \odot B = \begin{bmatrix} a_{\bullet 1}b_{1\bullet} \\ \vdots \\ a_{\bullet r}b_{r\bullet} \end{bmatrix}.$$

The bilinear operation $\odot$ over compatibly dimensioned matrices is related with the standard multiplication as $AB = J_{q,r}(A \odot B)$, where $J_{q,r}$ is a $(q \times qr)$-dimensional matrix defined by

$$J_{q,r} = \underbrace{\begin{bmatrix} I_q & \dots & I_q \end{bmatrix}}_{r \text{ times}},$$

with I_q denoting the identity matrix of order q. On the other hand, $A \odot B = (A \odot I_r)B$. Moreover, the preproduct satisfies the property

$$A \odot (BC) = (A \odot B)C,$$

so that the brackets may be omitted. With these notations in mind, the implementation (3) reads

$$[M, v]_\varepsilon = J_{q,r}R_\varepsilon(R_\varepsilon(M) \odot R_\varepsilon(v)).$$

It is straightforward to verify that the magnified accumulated error of the implementation is

$$(Mv - [M, v]_\varepsilon)/\varepsilon = E_\varepsilon(M)v + R_\varepsilon(M)E_\varepsilon(v) + J_{q,r}E_\varepsilon(R_\varepsilon(M) \odot R_\varepsilon(v)). \tag{4}$$

Regarding $[M, v]_\varepsilon$ as output of a digitally implemented static linear filter with input vector v, the first term on the right of (4) is caused by coefficient quantization, the second comes from analog-to-digital conversion (or quantization) of the input and the last from digital arithmetic operations. In this paper, we shall be concerned with the two last sources of the round-off noise: input quantization and arithmetic errors.

3 Problem formulation

Consider a linear discrete-time-invariant system with m-dimensional input w, p-dimensional output z and n-dimensional internal state x governed by the equations

$$\begin{aligned} x_{k+1} &= Ax_k + Bw_k, \\ z_k &= Cx_k + Dw_k, \end{aligned} \tag{5}$$

with initial condition $x_0 = 0$, where A, B, C, D are appropriately dimensioned real matrices. The system will be interpreted as a linear filter to process an analog signal w. The filter may satisfy an optimization principle with the matrices A, B, C, D obtained as a numerical solution to an optimal filtering/control problem. Whilst the appropriate filter design is usually performed on high-precision computers (with potentially unlimited accuracy), the filter itself is then implemented on a computer (say, a microprocessor) with substantially lower precision ε, though supposed to operate in real time. Considering the computations once performed on those high-precision computers perfectly accurate, it seems reasonable to investigate asymptotic properties of the digitally implemented filter as $\varepsilon \to +0$.

The digital implementation of the filter (5) in fixed-point arithmetic with accuracy ε can be modeled as

$$\begin{aligned} x^\varepsilon_{k+1} &= [A, x^\varepsilon_k]_\varepsilon + [B, w_k]_\varepsilon, \\ z^\varepsilon_k &= [C, x^\varepsilon_k]_\varepsilon + [D, w_k]_\varepsilon, \end{aligned} \tag{6}$$

with the same initial condition $x^\varepsilon_0 = 0$. In particular, the accurate matrices A, B, C, D are replaced with their digital counterparts

$$A_\varepsilon = R_\varepsilon(A),\ B_\varepsilon = R_\varepsilon(B),\ C_\varepsilon = R_\varepsilon(C),\ D_\varepsilon = R_\varepsilon(D).$$

Clearly, these last converge to the accurate matrices as $\varepsilon \to +0$. For the sake of brevity in the sequel, it will be convenient to use the notations

$$\Phi_\varepsilon = \begin{bmatrix} A_\varepsilon \\ C_\varepsilon \end{bmatrix}, \quad \Phi = \begin{bmatrix} A \\ C \end{bmatrix}, \tag{7}$$

$$\Psi_\varepsilon = \begin{bmatrix} B_\varepsilon \\ D_\varepsilon \end{bmatrix}, \quad \Psi = \begin{bmatrix} B \\ D \end{bmatrix}, \tag{8}$$

$$\Theta_\varepsilon = \begin{bmatrix} \Phi_\varepsilon & \Psi_\varepsilon \end{bmatrix}, \quad \Theta = \begin{bmatrix} \Phi & \Psi \end{bmatrix}. \tag{9}$$

It is easy to see that

$$\|\Theta_\varepsilon - \Theta\| = O(\varepsilon) \quad \text{as } \varepsilon \to +0, \tag{10}$$

where $\|\cdot\|$ is a matrix norm. To isolate the input quantization and arithmetic effects from the coefficient quantization ones, rewrite (6) as

$$\begin{aligned} x^\varepsilon_{k+1} &= A_\varepsilon x^\varepsilon_k + B_\varepsilon w_k - \varepsilon\lambda^\varepsilon_k, \\ z^\varepsilon_k &= C_\varepsilon x^\varepsilon_k + D_\varepsilon w_k - \varepsilon\mu^\varepsilon_k, \end{aligned} \tag{11}$$

where

$$\begin{bmatrix} \lambda_k^\varepsilon \\ \mu_k^\varepsilon \end{bmatrix} = \Psi_\varepsilon \alpha_k^\varepsilon + \begin{cases} J_{n+p,m}\beta_0^\varepsilon & \text{for} \quad k=0 \\ J_{n+p,n+m}\beta_k^\varepsilon & \text{for} \quad k>0 \end{cases}. \tag{12}$$

Here,

$$\alpha_k^\varepsilon = E_\varepsilon(w_k) \tag{13}$$

are the input quantization errors, and β_k^ε the arithmetic errors given by

$$\beta_k^\varepsilon = \begin{cases} E_\varepsilon(\Psi_\varepsilon \odot R_\varepsilon(w_0)) & \text{for} \quad k=0 \\ \\ E_\varepsilon(\Theta_\varepsilon \odot \begin{bmatrix} x_k^\varepsilon \\ R_\varepsilon(w_k) \end{bmatrix}) & \text{for} \quad k>0 \end{cases} \tag{14}$$

(we have taken the initial condition $x_0^\varepsilon = 0$ into account). For the case where the signal w is random, it is nowadays a wide-spread approach to approximate the input quantization and arithmetic errors by independent (mutually and of the input) random vectors distributed uniformly on the appropriately dimensioned cubes, which is frequently referred to as the uniform white noise model [7]. However, since each of the errors α_k^ε and β_k^ε is a deterministic function of $w_0, \ldots, w_k$, the validity of the model can only be conceived asymptotically, as $\varepsilon \to +0$. This can be formulated rigorously in terms of weak convergence of probability measures [3] as follows: for any $k \in \mathbb{Z}_+$ and any bounded continuous function $f : \mathbb{R}^{(k+1)m} \times V_{(n+p+1)m+k(m+(n+p)(n+m))} \to \mathbb{R}$,

$$\lim_{\varepsilon \to +0} \mathbf{E} f(W_k, \gamma_k^\varepsilon) = \mathbf{E} f(W_k, N_k). \tag{15}$$

Here, $\mathbf{E}(\,\cdot\,)$ denotes mathematical expectation; the vector γ_k^ε is formed by stacking the input quantization and arithmetic errors (13)–(14) upto the instant k,

$$\gamma_k^\varepsilon = \begin{bmatrix} \alpha_0^\varepsilon \\ \beta_0^\varepsilon \\ \vdots \\ \alpha_k^\varepsilon \\ \beta_k^\varepsilon \end{bmatrix}; \tag{16}$$

N_k is a random vector, distributed unifomly on the cube $V_{(n+p+1)m+k(m+(n+p)(n+m))}$ and independent of the k-th initial segment of the input signal

$$W_k = \begin{bmatrix} w_0 \\ \vdots \\ w_k \end{bmatrix}. \tag{17}$$

While part of the uniform white noise approximation concerning the input quantization errors is justified by the Quantization Theorem [2, 6, 8, 11, 15, 16], the whole property (15) does not hold automatically, and fulfilment of certain conditions on the state-space realization matrices A, B, C, D in (5) is essential for the model to be asymptotically valid.

4 Main result

The input signal w in (6) will be called *generic* if for every $k \in \mathbb{N}$, the probability distribution of the random vector (17) is absolutely continuous with respect to $(k+1)m$-dimensional Lebesgue measure.

Say that a matrix is *nonresonant* if its rows are linearly independent over the field of real rationals. The state-space realization (5) will be called *totally nonresonant* if for every $k \in \mathbb{N}$, the matrix

$$\Sigma_k = \begin{bmatrix} I_m \\ \Psi \odot I_m \\ (\Phi J_{n,k}) \odot \begin{bmatrix} I_n \\ A \\ \vdots \\ A^{k-1} \end{bmatrix} B \end{bmatrix} \tag{18}$$

is nonresonant (the notations (7)–(8) are used).

Theorem 1 *Suppose that the state-space realization (5) is totally nonresonant and the input signal w is generic. Then the uniform white noise model applies to the input quantization and arithmetic errors (13)–(14) in the sense of (15). More precisely, for every $k \in \mathbb{Z}_+$, the joint probability distribution of the vectors W_k and γ_k^ε weakly converges to that of W_k and N_k as $\varepsilon \to +0$.*

The proof of the theorem is based on lemmas of Sections 5–7 and will be given in Section 8.

5 Skew products and continuous convergence

Denote by $\mathcal{P}_{q,r}$ the class of measurable mappings $F : \mathbb{R}^q \to \mathbb{R}^r$, unit periodic in its q variables, and by $\mathcal{T}_q$ the class of mappings $F \in \mathcal{P}_{q,q}$ such that the restriction $F|_{V_q}$ is a mes_q-preserving bijection of the cube V_q onto itself (here, mes_q denotes q-dimensional Lebesgue measure). Clearly, any $F \in \mathcal{T}_q$ can be identified with a measure-preserving automorphism of the q-dimensional torus. For example, the mapping E_1 in (2), when considered on $\mathbb{R}^q$, belongs to the class $\mathcal{T}_q$ (note that $E_1|_{V_q}$ is the identity of V_q).

Given two mappings $F \in \mathcal{T}_q$ and $G \in \mathcal{P}_{q,r}$, define a mapping $F \star G : \mathbb{R}^{q+r} \to V_{q+r}$ by

$$(F \star G)(\begin{bmatrix} u \\ v \end{bmatrix}) = \begin{bmatrix} F(u) \\ E_1(v - G(u)) \end{bmatrix}, \quad u \in \mathbb{R}^q,\ v \in \mathbb{R}^r. \tag{19}$$

Slightly modifying the traditional definition, $G \star H$ will be called a *skew product* of F and G. The well-known properties of the standard skew product of mappings [4, 9] imply that $F \star G \in \mathcal{T}_{q+r}$.

Say that a mapping $F_\varepsilon : X \to Y$ of a set $X \subset \mathbb{R}^q$ to a separable metric space Y, parameterized by $\varepsilon > 0$, *converges continuously* to a mapping F, as $\varepsilon \to +0$, if the set $\{x \in X:\ F_\varepsilon(u) \to F(x) \text{ as } (\varepsilon, u) \to (+0, x)\}$ is of full q-dimensional Lebesgue measure (this will be denoted as $F_\varepsilon \xrightarrow{\mathrm{c}} F$).

Lemma 1 *Suppose that mappings $F_\varepsilon \in \mathcal{T}_q$ and $G_\varepsilon \in \mathcal{P}_{q,r}$, parameterized by $\varepsilon > 0$, converge continuously to $F \in \mathcal{T}_q$ and $G \in \mathcal{P}_{q,r}$, respectively, as $\varepsilon \to +0$. Then $F_\varepsilon \star G_\varepsilon \xrightarrow{\mathrm{c}} F \star G$.*

Proof. Define the sets

$$\begin{aligned} S &= \{u \in V_q:\ F_\varepsilon(z) \to F(u),\ G_\varepsilon(z) \to G(u) \text{ as } (\varepsilon, z) \to (+0, u)\}, \\ T &= \{x \in V_{q+r}:\ (F_\varepsilon \star G_\varepsilon)(y) \to (F \star G)(x) \text{ as } (\varepsilon, y) \to (+0, x)\}. \end{aligned}$$

Note that the discontinuity set of the mapping E_1 on $\mathbb{R}^r$ is $\partial V_r + \mathbb{Z}^r$, where $\partial(\cdot)$ denotes the boundary of a set. Hence, the definition (19) easily implies the inclusion

$$T \supset \left\{\begin{bmatrix} u \\ v \end{bmatrix}:\ u \in S,\ v \in V_r \setminus (G(u) + \partial V_r + \mathbb{Z}^r)\right\}.$$

Since $\mathrm{mes}_r(\partial V_r + \mathbb{Z}^r) = 0$ (as for a union of denumerably many hyperplanes in $\mathbb{R}^r$) and $\mathrm{mes}_q S = 1$ (by the assumptions of the lemma), we have $\mathrm{mes}_{q+r} T = 1$, which yields the assertion of the lemma. ■

6 Quasiperiodic structure of roundoff errors

Lemma 2 *For every $k \in \mathbb{Z}_+$, the vector (16) can be expressed in terms of (17) as*

$$\gamma_{\varepsilon,k} = F_{\varepsilon,k}(\varepsilon^{-1} L_{\varepsilon,k} W_k). \tag{20}$$

Here, the matrices $L_{\varepsilon,k} \in \mathbb{R}^{(n+p+1)m + k(m+(n+p)(n+m))}$ are defined by the recursion

$$L_{\varepsilon,k+1} = \left[\begin{array}{c|c} L_{\varepsilon,k} & 0 \\ \hline 0 & I_m \\ \Phi_\varepsilon \odot \left[\, A_\varepsilon^k B_\varepsilon \mid \ldots \mid B_\varepsilon \,\right] & 0 \\ 0 & \Psi_\varepsilon \odot I_m \end{array}\right] \tag{21}$$

with initial condition

$$L_{\varepsilon,0} = \begin{bmatrix} I_m \\ \Psi_\varepsilon \odot I_m \end{bmatrix}, \tag{22}$$

where the notations (7)–(8) are used; each of the mappings $F_{\varepsilon,k}$ belongs to the class $\mathcal{T}_{(n+p+1)m+k(m+(n+p)(n+m))}$ and continuously converges to a mapping F_k from the same class, as $\varepsilon \to +0$.

Proof. The proof of the lemma will be carried out by induction on k. To verify the basis of induction, rewrite (13) and (14) for $k = 0$ as

$$\begin{aligned} \alpha_0^\varepsilon &= E_1(\varepsilon^{-1} w_0), \\ \beta_0^\varepsilon &= E_\varepsilon(\Psi_\varepsilon \odot (w_0 - \varepsilon E_\varepsilon(w_0))) \\ &= E_1(\varepsilon^{-1}(\Psi_\varepsilon \odot I_m) w_0 - \Psi_\varepsilon \odot E_1(\varepsilon^{-1} w_0)). \end{aligned}$$

Hence, (20) holds for $k=0$ with the matrix (22) and the mapping $F_{\varepsilon,0} : I\!R^{(n+p+1)m} \to V_{(n+p+1)m}$ defined by

$$F_{\varepsilon,0}(\left[\begin{array}{c} u \\ v \end{array}\right]) = E_1(\left[\begin{array}{c} u \\ v - \Psi_\varepsilon \odot E_1(u) \end{array}\right]), \quad u \in I\!R^m,\ v \in I\!R^{(n+p)m}.$$

It is easy to see that $F_{\varepsilon,0}$ is the skew product of the mappings $E_1 \in \mathcal{T}_m$ and $\Psi_\varepsilon \odot E_1 \in \mathcal{P}_{m,(n+p)m}$, and consequently, belongs to the class $\mathcal{T}_{(n+p+1)m}$. Hence, application of Lemma 1 yields the continuous convergence

$$F_{\varepsilon,0} \xrightarrow{\mathrm{c}} F_0 = E_1 \star (\Psi \odot E_1) \in \mathcal{T}_{(n+p+1)m} \quad \text{as } \varepsilon \to +0.$$

Now let the representation (20) and the continuous convergence described in the lemma hold for some $k \in Z\!\!\!Z_+$. Note that, by (12) and (16),

$$\lambda_j^\varepsilon = \left[\begin{array}{c|c} B_\varepsilon & \left[\begin{array}{cc} I_n & 0 \end{array}\right] \left\{\begin{array}{ll} J_{n+p,m} & \text{for } \ j=0 \\ J_{n+p,n+m} & \text{for } \ j>0 \end{array}\right. \end{array}\right] \left[\begin{array}{c} \alpha_j^\varepsilon \\ \beta_j^\varepsilon \end{array}\right].$$

Hence, using (11) and (17), obtain that

$$\begin{aligned} x_{k+1}^\varepsilon &= \sum_{j=0}^{k} A_\varepsilon^{k-j}(B_\varepsilon w_j - \varepsilon\lambda_j^\varepsilon) \\ &= P_{\varepsilon,k}W_k - \varepsilon Q_{\varepsilon,k}F_{\varepsilon,k}(\varepsilon^{-1}L_{\varepsilon,k}W_k), \end{aligned} \tag{23}$$

where the matrices $P_{\varepsilon,k}$ and $Q_{\varepsilon,k}$ are governed by the recursions

$$P_{\varepsilon,k} = \left[\begin{array}{c|c} A_\varepsilon P_{\varepsilon,k-1} & B_\varepsilon \end{array}\right], \tag{24}$$

$$Q_{\varepsilon,k} = \left[\begin{array}{c|c} A_\varepsilon Q_{\varepsilon,k-1} & \left[\begin{array}{c|c} B_\varepsilon & \left[\begin{array}{cc} I_n & 0 \end{array}\right] J_{n+p,n+m} \end{array}\right] \end{array}\right], \tag{25}$$

with initial conditions

$$P_{\varepsilon,0} = B_\varepsilon, \tag{26}$$

$$Q_{\varepsilon,0} = \left[\begin{array}{c|c} B_\varepsilon & \left[\begin{array}{cc} I_n & 0 \end{array}\right] J_{n+p,m} \end{array}\right]. \tag{27}$$

Rewriting the definitions (13) and (14) for the $(k+1)$-th errors and taking (23) into account yield

$$\left[\begin{array}{c} \alpha_{k+1}^\varepsilon \\ \beta_{k+1}^\varepsilon \end{array}\right] = E_\varepsilon(\left[\begin{array}{c} w_{k+1} \\ \Theta_\varepsilon \odot \left[\begin{array}{c} P_{\varepsilon,k}W_k - \varepsilon Q_{\varepsilon,k}F_{\varepsilon,k}(\varepsilon^{-1}L_{\varepsilon,k}W_k) \\ w_{k+1} - \varepsilon E_\varepsilon(w_{k+1}) \end{array}\right] \end{array}\right]).$$

Therefore,

$$\gamma_{k+1}^\varepsilon = \left[\begin{array}{c} \gamma_k^\varepsilon \\ \alpha_{k+1}^\varepsilon \\ \beta_{k+1}^\varepsilon \end{array}\right] = F_{\varepsilon,k+1}(\varepsilon^{-1}L_{\varepsilon,k+1}W_{k+1}), \tag{28}$$

where the matrix $L_{\varepsilon,k+1}$ and the mapping $F_{\varepsilon,k+1}$ are given by

$$L_{\varepsilon,k+1} = \left[\begin{array}{c|c} L_{\varepsilon,k} & 0 \\ \hline 0 & I_m \\ \Phi_\varepsilon \odot P_{\varepsilon,k} & 0 \\ 0 & \Psi_\varepsilon \odot I_m \end{array}\right] \tag{29}$$

and

$$F_{\varepsilon,k+1}(\begin{bmatrix} u \\ v \\ s \end{bmatrix}) = \begin{bmatrix} F_{\varepsilon,k}(u) \\ v \\ E_1(\begin{bmatrix} s - \Theta_\varepsilon \odot \begin{bmatrix} Q_{\varepsilon,k} F_{\varepsilon,k}(u) \\ E_1(v) \end{bmatrix} \end{bmatrix}) \end{bmatrix} \tag{30}$$

for all $u \in I\!R^{(n+p+1)m+k(m+(n+p)(n+m))}$, $v \in I\!R^m$ and $s \in I\!R^{(n+p)(n+m)}$. Clearly, $F_{\varepsilon,k+1}$ is the skew product of the mappings $G_{\varepsilon,k} \in \mathcal{T}_{(n+p+2)m+k(m+(n+p)(n+m))}$ and $H_{\varepsilon,k} \in \mathcal{P}_{(n+p+2)m+k(m+(n+p)(n+m)),(n+p)(n+m)}$ defined by

$$G_{\varepsilon,k}(\begin{bmatrix} u \\ v \end{bmatrix}) = \begin{bmatrix} F_{\varepsilon,k}(u) \\ E_1(v) \end{bmatrix}, \tag{31}$$

$$H_{\varepsilon,k}(\begin{bmatrix} u \\ v \end{bmatrix}) = \Theta_\varepsilon \odot \begin{bmatrix} Q_{\varepsilon,k} F_{\varepsilon,k}(u) \\ E_1(v) \end{bmatrix}. \tag{32}$$

The inductive assumption $F_{\varepsilon,k} \xrightarrow{\mathrm{c}} F_k$ implies that $G_{\varepsilon,k} \xrightarrow{\mathrm{c}} G_k$ as $\varepsilon \to +0$, where the limiting mapping G_k is described by (31), with $F_{\varepsilon,k}$ replaced by F_k. Note that $Q_{\varepsilon,k} \to Q_k$, where the limiting matrices Q_k satisfy the equations (25) and (27) with the matrices $A_\varepsilon, B_\varepsilon$ replaced by their limits A, B, respectively. Hence, $H_{\varepsilon,k} \xrightarrow{\mathrm{c}} H_k$, where H_k is given by (32), with the matrices $\Theta_\varepsilon, Q_{\varepsilon,k}$ and the mapping $F_{\varepsilon,k}$ replaced by their limits. Consequently, applying Lemma 1 to the mapping (30), obtain that $F_{\varepsilon,k+1} \xrightarrow{\mathrm{c}} F_{k+1} = G_k \star H_k$. It now remains to notice that, by (24) and (26), the recursion (29) coincides with (21). Hence, (28) satisfies all the properties claimed by the lemma. This completes the inductive step, thereby finishing the proof. ■

Lemma 3 *Let the state-space realization (5) be totally nonresonant. Then each of the matrices $L_{\varepsilon,k}$ in (21)–(22) converges to a nonresonant matrix L_k as $\varepsilon \to +0$.*

Proof. That the matrices $L_{\varepsilon,k}$ converge, follows immediately from (21)–(22). The limiting matrices L_k satisfy a similar recursion

$$L_{k+1} = \left[\begin{array}{c|c} L_k & 0 \\ \hline 0 & I_m \\ \Phi \odot \left[\begin{array}{c|c|c} A^k B & \ldots & B \end{array}\right] & 0 \\ 0 & \Psi \odot I_m \end{array}\right],$$

with initial condition

$$L_0 = \begin{bmatrix} I_m \\ \Psi \odot I_m \end{bmatrix}.$$

It is straightforward to show inductively that for every $k \in \mathbb{N}$, there is a permutation of the rows of L_k which reduces the matrix to

$$M_k = \left[\begin{array}{ccccc} \multicolumn{5}{c}{I_{(k+1)m}} \\ \hline \Psi \odot I_m & 0 & \cdots & 0 & 0 \\ 0 & \Psi \odot I_m & & & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & & & \Psi \odot I_m & 0 \\ 0 & 0 & \cdots & 0 & \Psi \odot I_m \\ \hline \Phi \odot B & 0 & \cdots & 0 & 0 \\ \vdots & \ddots & & & \vdots \\ \Phi \odot A^{k-2}B & & & 0 & 0 \\ \Phi \odot A^{k-1}B & \Phi \odot A^{k-2}B & \cdots & \Phi \odot B & 0 \end{array}\right].$$

The structure of this last implies that, if the matrix Σ_k in (18) is nonresonant, then so is M_k, and consequently, the matrix L_k. Indeed, let t be a vector obtained by stacking vectors $u_0, \ldots, u_k \in \mathbb{Z}^m$, $v_0, \ldots, v_k \in \mathbb{Z}^{(n+p)m}$, $s_0, \ldots, s_{k-1} \in \mathbb{Z}^{(n+p)n}$ and satisfying

$$t^T M_k = 0, \tag{33}$$

or equivalently,

$$u_j^T + v_j^T(\Psi \odot I_m) + \sum_{r=j}^{k-1} s_r^T(\Phi \odot A^{r-j}B) = 0 \quad \text{for all } 0 \le j \le k. \tag{34}$$

The row-vector on the left of each of these equalities is an integer linear combination of the rows of the matrix Σ_k. Hence, if Σ_k is nonresonant, then (34), considered for $j = 0$, implies that $u_0 = 0$, $v_0 = 0$, $s_0 = \ldots = s_{k-1} = 0$. By the same reasoning, from the other equations (34) it follows that $u_j = 0$, $v_j = 0$ for all $1 \le j \le k$, and consequently, the integer vector t satisfying (33) can only be zero. This completes the proof of the lemma. ■

7 Quantization Theorem for linear mappings

Lemma 4 *Suppose that a matrix $L_\varepsilon \in \mathbb{R}^{q\times r}$, parameterized by $\varepsilon > 0$, converges to a nonresonant matrix L as $\varepsilon \to +0$. Let W be an absolutely continuously distributed random vector with values in $\mathbb{R}^r$. Then the joint distribution of W and $E_\varepsilon(L_\varepsilon W)$ weakly converges to that of W and N, with N a random vector, independent of W and distributed uniformly on the cube V_q.*

Proof. The probability density function p of the vector W can be approximated by a piece-wise constant function with any given accuracy in the integral metric. More precisely, the quantity

$$\int_{\mathbb{R}^r} \left| p(w) - \sum_k p_k \mathcal{I}_{S_k}(w) \right| dw$$

can be made arbitrarily small by an appropriate choice of a finite collection of cubes S_k in $\mathbb{R}^r$ and numbers p_k (here, $\mathcal{I}(\cdot)$ denotes the indicator function of a set). Therefore, by the Weyl criterion [3], the lemma will be proved if we show that for any cube $S = g + \delta V_r$, with $g \in \mathbb{R}^r$ and $\delta > 0$, and for any $\sigma \in \mathbb{Z}^q \setminus \{0\}$,

$$\lim_{\varepsilon \to +0} \int_S \exp(2\pi i \sigma^T E_\varepsilon(L_\varepsilon w)) dw = 0. \tag{35}$$

Assuming the cube S and the vector σ fixed and using (2), write the absolute value of the last integral as

$$\begin{aligned} \varphi_\varepsilon &= \left| \int_S \exp(2\pi i \sigma^T E_1(L_\varepsilon w/\varepsilon)) dw \right| \\ &= \delta^{-r} \left| \int_{V_r} \exp(2\pi i \tau_\varepsilon^T w \delta/\varepsilon) dw \right| = \delta^{-r} \theta(\delta \tau_\varepsilon/\varepsilon), \end{aligned} \tag{36}$$

where $\tau_\varepsilon = L_\varepsilon^T \sigma$, and the function $\theta : \mathbb{R}^r \to [0,1]$ maps a vector $u = (u_k)_{1 \le k \le r}$ to

$$\theta(u) = \prod_{1 \le k \le r} \left| \frac{\sin(\pi u_k)}{\pi u_k} \right|,$$

with the convention $\left. \frac{\sin v}{v} \right|_{v=0} = 1$. The convergence $L_\varepsilon \to L$ and the nonresonance condition on L imply that

$$\lim_{\varepsilon \to +0} \tau_\varepsilon = L^T \sigma \neq 0,$$

and consequently, $|\tau_\varepsilon|/\varepsilon \to +\infty$ as $\varepsilon \to +0$. Therefore, since $\lim_{|u| \to +\infty} \theta(u) = 0$, obtain that the quantity (36) satisfies $\lim_{\varepsilon \to +0} \varphi_\varepsilon = 0$. This yields (35), thereby finishing the proof of the lemma. ■

Remark 1. A slight modification of the proof of Lemma 4 shows that the nonresonance condition on the limiting matrix L is essential in the following sense. Suppose, in addition to the assumption on W being absolutely continuously distributed, that $\|L_\varepsilon - L\| = O(\varepsilon)$ as $\varepsilon \to +0$. Then the weak convergence, asserted by the lemma, holds if and only if L is nonresonant.

8 Proof of the main theorem

For any $k \in \mathbb{Z}_+$, denote by Ω_k the probability distribution of the random vector W_k in (17), and by $\Pi_{\varepsilon,k}$ the joint distribution of W_k and

$$U_{\varepsilon,k} = E_\varepsilon(L_{\varepsilon,k} W_k), \tag{37}$$

where the matrices $L_{\varepsilon,k}$ are defined by (21)–(22). Using the suppositions of the theorem and applying Lemmas 3 and 4, obtain that the probability measure $\Pi_{\varepsilon,k}$, defined on Borel subsets of $\mathbb{R}^{(k+1)m} \times V_{(n+p+1)m+k(m+(n+p)(n+m))}$, converges weakly to an absolutely continuous measure

$$\Pi_k = \Omega_k \times \mathrm{mes}_{(n+p+1)m+k(m+(n+p)(n+m))}, \tag{38}$$

the direct product of Ω_k and the uniform distribution on the cube. By the properties of the mapping $F_{\varepsilon,k}$ described in Lemma 2, the representation (20) can be rewritten with the aid of (37) as

$$\gamma_{\varepsilon,k} = F_{\varepsilon,k}(U_{\varepsilon,k}).$$

Hence,

$$\left[\begin{array}{c} W_k \\ \gamma_k^\varepsilon \end{array}\right] = \Xi_{\varepsilon,k}\left(\left[\begin{array}{c} W_k \\ U_{\varepsilon,k} \end{array}\right]\right), \tag{39}$$

where the mapping $\Xi_{\varepsilon,k}$ is formed in an obvious way by the identity of $\mathbb{R}^{(k+1)m}$ and the mes-preserving bijection $F_{\varepsilon,k}|_{V_{(n+p+1)m+k(m+(n+p)(n+m))}}$. Clearly, $\Xi_{\varepsilon,k}$ continuously converges to a Π_k-preserving bijection Ξ_k as $\varepsilon \to +0$. Therefore, application of the results of [13] (see also [3, Theorem 5.5 on p. 34]) yields that the distribution of the random vector on the left of (39) weakly converges to the probability measure (38),

$$\Pi_{\varepsilon,k} \circ \Xi_{\varepsilon,k}^{-1} \xrightarrow{\mathrm{w}} \Pi_k \circ \Xi_k^{-1} = \Pi_k \quad \text{as } \varepsilon \to +0.$$

To complete the proof of the theorem, it remains to note that Π_k coincides with the joint distribution of W_k and N_k, where N_k is a random vector, independent of W_k and distributed uniformly on the cube $V_{(n+p+1)m+k(m+(n+p)(n+m))}$. ∎

Remark 2. Using (10), the definition (21)–(22) and the proof of Lemma 3, it is easy to verify inductively that for any given $k \in \mathbb{Z}_+$, the rate of convergence of the matrix $L_{\varepsilon,k}$ to its limit is $\|L_{\varepsilon,k} - L_k\| = O(\varepsilon)$ as $\varepsilon \to +0$. Consequently, by Remark 1, the nonresonance condition on L_k is not only sufficient but also necessary for the weak convergence $\Pi_{\varepsilon,k} \xrightarrow{\mathrm{w}} \Pi_k$, substantially used in the proof of the main theorem.

Remark 3. It may happen that the state-space realization (5) is not totally but still partially nonresonant in the sense that there exists a number $k_0 \in \mathbb{Z}_+$ so that the matrix Σ_{k_0} in (18) is nonresonant while Σ_{k_0+1} is not. As the proof of Theorem 1 shows, in that case, the assertion of the theorem remains valid for $k \leq k_0$. This means that the uniform white noise model applies until the critical instant k_0, after which it can break down. For $k > k_0$, persistent statistical cross-correlations may occur between the entries of the vector γ_k^ε in (16) and/or between them and the input signal.

9 Concluding remarks

We have rigorously justified the asymptotic validity of the uniform white noise model for the input quantization and arithmetic errors in digital implementations (6) of linear filters (5) in fixed-point arithmetic with accuracy $\varepsilon \to +0$. It was done under the nonresonance conditions on the state-space realization matrices of the filters and for generic input signals.

Not surprisingly, the fulfilment of the nonresonance conditions is sensitive to a particular state-space realization and is not invariant with respect to its equivalence transformations. To this end, it would be useful to show that the conditions are met for mes $_{(n+p)(n+m)}$-almost all matrices Θ in (9). Hence, of interest is the study of the structure of the appropriate resonance set in $\mathbb{R}^{(n+p)\times(n+m)}$.

Another line of further research is related to the use of the uniform white noise model in the asymptotic covariance analysis of the digital filter output in (6). More precisely, by (11)–(14) and by (16)–(17), the output is linearly expressed in terms of the input signal and of the input quantization and arithmetic errors as

$$\begin{aligned} z_k^\varepsilon &= C_\varepsilon \sum_{j=0}^{k-1} A_\varepsilon^{k-1-j}(B_\varepsilon w_j - \varepsilon\lambda_j^\varepsilon) + D_\varepsilon w_k - \varepsilon\mu_k^\varepsilon \\ &= S_{\varepsilon,k}W_k - \varepsilon T_{\varepsilon,k}\gamma_k^\varepsilon, \end{aligned}$$

where the deterministic matrices $S_{\varepsilon,k}$ and $T_{\varepsilon,k}$, as well their limits S_k and T_k, are computed in an obvious way. Clearly,

$$\begin{aligned} \mathbf{E}(z_k^\varepsilon z_k^{\varepsilon T}) &= S_{\varepsilon,k}\mathbf{E}(W_k W_k^T)S_{\varepsilon,k}^T \\ &\quad -\varepsilon(S_{\varepsilon,k}\mathbf{E}(W_k\gamma_k^{\varepsilon T})T_{\varepsilon,k}^T + T_{\varepsilon,k}\mathbf{E}(\gamma_k^\varepsilon W_k^T)S_{\varepsilon,k}^T) \\ &\quad +\varepsilon^2 T_{\varepsilon,k}\mathbf{E}(\gamma_k^\varepsilon\gamma_k^{\varepsilon T})T_{\varepsilon,k}^T. \end{aligned} \tag{40}$$

Since, according to the uniform white noise model justified by Theorem 1,

$$\lim_{\varepsilon\to+0}\mathbf{E}(\gamma_k^\varepsilon\gamma_k^{\varepsilon T}) = \frac{1}{12}I_{(n+p+1)m+k(m+(n+p)(n+m))},$$

from (40) it is easy to see that computation of principal terms upto order $O(\varepsilon^2)$ in the asymptotic expansion for the second-order moments of the output would be substantially simplified if

$$\mathbf{E}(W_k^T\gamma_k^\varepsilon) = o(\varepsilon) \quad \text{as } \varepsilon \to +0. \tag{41}$$

This last relationship could probably be established effectively in the case where the input signal w is a zero-mean Gaussian white noise and some enhanced nonresonance conditions are met. Note that even under the Gaussian assumption, (41) does not follow from Theorem 1 which yields just $o(1)$, nor it can be obtained from the convergence rate estimates [11] in the classic Quantization Theorem because of the complicated dependence of γ_k^ε on W_k. These issues are under investigation.

Acknowledgements

The work was supported by the Australian Research Council Grant A 4970 2246. The authors are grateful to Prof. Nikolai Kuznetsov for helpful discussions on quantization effects in digital systems. One of the authors (I.V.) thanks Prof. Flemming Topsøe for clarifying a terminological question on continuous convergence of mappings.

References

[1] Barnes C.W., Tran B.N., and Leung S.H. [1985] "On the statistics of fixed-point roundoff error", *IEEE Transactions on Acoustics, Speech and Signal Processing*, **ASSP–33**, 3, 595–606.

[2] Bennett W.R. [1948] "Spectra of quantized signals", *Bell Syst. Tech. J.*, **27**, 446–472.

[3] Billingsley P. [1968] *Convergence of Probability Measures* (John Wiley & Sons, New York).

[4] Cornfeld I.P., Fomin S.V., and Sinai Ya.G. [1982] *Ergodic Theory* (Springer-Verlag, New York etc.).

[5] Diamond P., and Vladimirov I. [1998] "Asymptotic independence and uniform distribution of quantization errors for spatially discretized dynamical systems", *International Journal of Bifurcation and Chaos*, **8**, 7, 1479–1490.

[6] Gray R.M. [1990] "Quantization noise spectra", *IEEE Trans. Informat. Theory*, **36**, 1220–1244.

[7] Gray R.M., and Neuhoff D.L. [1998] "Quantization", *IEEE Trans. Informat. Theory*, **44**, 6, 2325–2383.

[8] Kosyakin A.A. [1961] "The statistical theory of amplitude quantization", *Automation and Remote Control*, **22**, 624–630.

[9] Martin N., and England J. [1984] *Mathematical Theory of Entropy* (Cambridge University Press, London etc.).

[10] Oppenheim A.V., and Schafer R.W. [1989] *Digital Signal Processing* (Prentice-Hall, Englewood-Cliffs, New Jersey).

[11] Sripad A.B., and Snyder D.L. [1977] "A necessary and sufficient condition for quantization errors to be uniform and white", *IEEE Transactions on Acoustics, Speech and Signal Processing*, **ASSP–25**, 5, 442–449.

[12] Tokaji I., and Barnes C.W. [1987] "Roundoff error statistics for a continuous range of multiplier coefficients", *IEEE Transactions on Curcuits and Systems*, **CAS–34**, 1, 52–59.

[13] Topsøe F. [1967] "Preservation of weak convergence under mappings", *Annals of Mathematical Statistics*, **38**, 1661–1665.

[14] Vladimirov I.G. [1996] "Quantized linear systems on integer lattices: frequency-based approach, Part I, II", *Center for Applied Dynamical Systems and Environmental Modeling*, CADSEM Reports **96-032**, **96-033**, Deakin University, Geelong, Australia.

[15] Widrow B. [1956] "A study of rough amplitude quantization by means of Nyquist Sampling Theory", *IRE Transactions on Curcuit Theory*, **CT–3**, 266–276.

[16] Widrow B. [1960] "Statistical analysis of amplitude quantized sampled-data systems", *AIEE Transactions, Part II (Applications and Industry)*, **79**, 555–568.

[17] Williamson D. [1991] *Digital Control and Implementation: Finite Wordlength Considerations* (Prentice Hall, Englewood Cliffs, New Jersey).