Estimation of mean squared errors of non-binary A/D-encoders through Fredholm determinants of piecewise-linear transformations

Katsutoshi Shinohara¹a) and Kenta Kobayashi¹b)

¹ Department of Commerce and Management, Hitotsubashi University, 2-1 Naka, Kunitachi, Tokyo 186-8601, Japan

a) ka.shinohara@r.hit-u.ac.jp
b) kenta.k@r(hit-u.ac.jp

Received July 22, 2017; Revised November 9, 2017; Published April 1, 2018

Abstract: Using the theory of Fredholm determinants of Perron-Frobenius operators of piecewise-linear maps, we derive a mathematically rigorous upper bounds of mean squared errors of analog-to-digital converters based on $\beta$-expansions. We also explain the technique of calculating the upper bound numerically by means of the numerical verification method.

Key Words: $\beta$-encoders, Fredholm determinant, numerical verification method

1. Introduction

Analog-to-digital (A/D) converters are electronic devices which convert given analog input into a digital sequence. Conventionally, A/D converters are designed based on the binary expansion of real numbers. Daubechies et al. proposed that $\beta$-encoders, which are A/D converters designed based on $\beta$-expansions, would show better performance ([1]). Their suggestions are examined experimentally (see for instance [2]).

Since $\beta$-encoders have advantages from the viewpoint of engineering, it is interesting and important to give theoretical, mathematically rigorous estimation of performances of $\beta$-encoders. On the other hand, their mathematical analysis is a challenging problem due to the chaotic behavior of the corresponding dynamical systems.

An analysis is given in [3] using a technique called segments. They derived a mathematically rigorous upper bound for the mean squared errors of $\beta$-encoders. The result was satisfactory in regard to the industrial applications. Meanwhile, there were several suggestions from researchers of dynamical systems that the result may be improved by combining the previous argument with classical results of one dimensional dynamical systems. The aim of this paper is to sharpen the previous result by using the spectral analysis of Perron-Frobenius operators of piecewise linear maps.

In the paper [3], one of the important steps to conclude the estimations was the counting of the
number of segments (see Section 2). In order to complete the counting they approximated the generation of segments by a Markovian process. This approximation was good enough to obtain the result. On the other hand, by using the theory of Perron-Frobenius operators, in other words, the theory of generating functions of segments, we can avoid taking the approximation and investigate the growth of the number of segments directly. As a result, we can reach the conclusion with more insightful argument and as a byproduct we can obtain an inequality which is valid in more general situations.

In [4], we announced this result without detailed explanation of the background mathematics due to the limitation of the pages. In this article, we elucidate how we can derive the formula necessary for the analysis and numerical techniques. We believe that the technique we propose would be useful for many other problems. On the other hand, the analysis of Perron-Frobenius operators of piecewise-linear maps are not so popular for researchers outside dynamical systems. Thus in this article we try to give a self-contained explanation of this theory.

We also discuss the techniques of numerical calculation of the upper bound. To obtain a mathematically rigorous upper bound we use the numerical verification method. The application of this method for our problem requires special care due to the presence of discontinuities of the systems. We explain how one can overcome this problem.

Let us explain the organization of this paper. In Section 2 we give the mathematical formulation of the problem. In Section 3 we give brief introduction of Perron-Frobenius operators of dynamical systems and summarize the known results which we will use afterward. In Section 4 we give the calculation of the Fredholm determinant of the Perron-Frobenius operators. In Section 5 we derive a theoretical upper bound of the MSEs using the results in Section 3. In Section 6 we explain how we can numerically calculate the upper bound we obtained in Section 5. Finally in Section 7 we give the numerical result we obtained.

2. Setting and precise statements

Let $\beta \in (1, 2]$. We consider the transformation of $I = [0, 1]$ defined as follows (notice that it is different from the conventional lazy $\beta$-transformation\(^1\) on the right branch):

\[
T_\beta(x) = T(x) = \begin{cases} 
\beta x & (x < 1/\beta) \\
\beta(x - 1) + 1 & (x \geq 1/\beta). 
\end{cases}
\]

For $x \in I$ and $i \geq 1$, we define the $i$-th digit $d_i(x)$ of $x$ by

\[
d_i(x) = \begin{cases} 
0 & (T^{i-1}(x) < 1/\beta) \\
1 & (T^{i-1}(x) \geq 1/\beta), 
\end{cases}
\]

where we put $T^0(x) = x$.

Then, the infinite sequence $(d_i(x)) \in \{0, 1\}^\mathbb{N}$ gives a $\beta$-expansion of $x$, namely, we have the following equality:

\[
x = (\beta - 1) \left( \sum_{i=1}^{\infty} \frac{d_i(x)}{\beta^i} \right). \tag{1}
\]

The coefficient $(\beta - 1)$ is a normalization constant put in order that the sequence $(1, 1, 1, \ldots)$ corresponds to $x = 1^2$. The sequence $(d_i(x))$ is a digital encoding of an input $x$. We can recover $x$ from the infinite $\{0, 1\}$-sequence $(d_i(x))$ by Eq. (1).

In the ideal situation where all the (infinitely many) digits are available, the encoding and decoding process does not give rise to any loss of information. However, in real situations only finitely many digits are available. We are interested in analyzing this loss of information. We denote the number of digits available by $L$. Then the resulted decoding for an input $x$ is no longer Eq. (1) but given by the formula below:

\(^1\)Recall that the conventional lazy $\beta$-transformation is given by $x \mapsto \beta x \ (	ext{mod} \ 1)$.
\(^2\)More precisely, the coefficient $(\beta - 1)$ is obtained as the reciprocal of the infinite series $\sum_{i=1}^{\infty} \beta^{-i}$.
\[ x_L(x) = (\beta - 1) \left( \sum_{i=1}^{L} \frac{d_i(x)}{\beta^i} \right) + \theta \beta^{-L}, \]  

(2)

where the term \( \theta \beta^{-L} \) is added in order to decrease the loss of information in the average. In our setting, a reasonable choice of \( \theta \) is \( \frac{1}{2}[(2 - \beta) + 1] = (3 - \beta)/2 \), since the interval \([2 - \beta, 1]\) is an invariant interval of \( T_\beta \) (namely, we have \( T_\beta([2 - \beta, 1]) \subseteq [2 - \beta, 1] \)). In the following discussion we fix \( \theta \) as such.

The aim of this paper is to give an estimation of the mean squared error of this imperfect encoding-decoding process, that is, to give an upper bound of the integral

\[ \text{MSE}(\beta, L) := \int_0^1 |x_L(x) - x|^2 \, dx. \]

Since this is an MSE for \( \beta \)-encoders, it is natural to guess \( \text{MSE}(\beta, L) = O(\beta^{-2L}) \). We are interested in estimating the coefficient of right hand side.

In [3], we introduced an important sequence of intervals called segments which are useful for the analysis of the MSE. Let us give a brief review of it. Given \( k \geq 1 \), the points of discontinuity of \( T_k \) divide the interval \( I \) into sub-intervals. Each point in such the same sub-interval has the same expansions up to \( k \)-th digits. The number of such intervals increases exponentially as \( k \) increases and this makes the analysis difficult. However, if we take the image of them under \( T_k \), the situation becomes simpler. More precisely, we have the following:

1. every image interval appearing in the graph of \( T_k \) appears among those of \( T_{k+1} \);
2. there are at most one new image interval in \( T_{k+1} \) which do not appear in \( T_k \). Hence the kind of image intervals appears at most linearly.

We label each image interval \( J_0 = I, J_1, \ldots J_k \) in such a way that \( J_k \) firstly appears in \( T_k \). We call them segments of \( T_k \) (see Fig. 1). Notice that some of \( J_i \) may coincide. We denote the population (number) of \( J_i \) in \( T_k \) by \( n_i^{(k)} \).

Then, by a change of variable argument (see [3] for the detail), we have the following equality:

\[ \text{MSE}(\beta, L) = \beta^{-3L} \sum_{i=0}^{L} n_i^{(L)} I_i, \]

where \( (I_i) \) are real numbers given as follows: We put \( J_i = [l_i, r_i] \). Then,

\[ I_i = \int_{l_i}^{r_i} |y - \theta|^2 \, dy = \frac{1}{3} \left[ (r_i - \theta)^3 - (l_i - \theta)^3 \right]. \]  

(3)

In our setting, we have \( r_i = 1 \) and for \( l_i \) we have the following: \( l_i = 0 \) if \( i = 0 \) and \( l_i = T^{i(1/\beta)} \) for \( i \geq 1 \).

Intuitively, we may guess that for fixed \( \beta \), MSE has the order of \( \beta^{-2L} \). Thus in the following we are interested in estimating the following constant:

\[ K_{\beta, L} := \sum_{i=0}^{L} \frac{n_i^{(L)}}{\beta^L} I_i, \]  

(4)

The following is the main result of this paper:

**Theorem 1** For \( \beta \in [1.6, 2.0] \) and \( L \geq 20 \), we have \( K_{\beta, L} \leq 0.0857(\beta - 1)^2 \).

Notice that an upper bound for \( K_{\beta, L} \) (\( L \leq 20 \)) can be obtained by the technique in [3]. The novelty here is that our result is valid for every \( L \geq 20 \). In section 7, we discuss this result further. From the next section we start the argument of how we prove this result.
The graphs of $T^i$ for $i = 1, 2, 3, 4$ and $\beta = 1.65$. For $T^1$ there are two branches. We label the image of the left branch as $J_0$ and that of the right one as $J_1$. For $T^2$ there are four branches. One branch has $J_0$ as its image and two branches have the image $J_1$. There is one branch which has new image and we call it $J_2$. For $T^3$ there are one $J_0$, four $J_1$, two $J_2$. The other one has new image and we call it $J_3$. Inductively we define $J_k$. We denote the number of branches of the graph of $T^k$ whose image is equal to $J_i$ by $n^{(k)}_i$ and call them populations of segments.

3. Perron-Frobenius operator and segments

3.1 Perron-Frobenius operators and their generating funtions

In this subsection, we give a brief introduction of the Perron-Frobenius operator of a transformation. As we will see later, this operator is tightly related to the generation of segments and plays an important role in our analysis.

For a piecewise $C^1$ transformation $S$ of an interval $I$, we can define the Perron-Frobenius operator $P$ acting on $L^1(I)$ as the (extension of the) adjoint operator of the Koopman operator $f(x) \mapsto f(S(x))$ with respect to the $L^2$ inner product. Namely, it is an operator which satisfies the following relation for every pair of continuous functions $f$ and $g$:

$$\int_{[0,1]} (P(f)(x)) g(x) dx = \int_{[0,1]} f(x) g(S(x)) dx.$$ 

It is not difficult to see that $P$ is an operator which pushes $f$ forward as a distribution function. Let us calculate $P(1_J)$ where $P$ is the Perron-Frobenius operator for $T$ (the transformation introduced in Section 2) and $1_J$ is the characteristic function of an interval $J \subset I$. By definition, together with a change of variable argument, we have

$$P(1_J) = (1/\beta)(1_{T(J_-)} + 1_{T(J_+)}) ,$$

where $J_- = J \cap [0, 1/\beta]$ and $J_+ = J \cap [1/\beta, 1]$. To be precise, we should put $J_- = [0, 1/\beta)$. However,
since \(1_{[0,1/\beta]}\) and \(1_{[0,1/\beta]}\) define the same function in \(L^1(I)\) and the latter form is more convenient for our purpose, we define \(J_+\) as \([0,1/\beta]\).

Notice that by definition we have

\[ P^L(1_I) = \beta^{-L} \sum_{i=0}^{L} n_i^{(L)} 1_{J_i}, \tag{5} \]

where \((J_i)\) are the segments in Section 2. Thus, the analysis of the MSE can be reduced to the study of corresponding Perron-Frobenius operator. In the following, we investigate the behavior of the sequence of functions \((P^L(1_I))\).

To analyze \((P^L(1_I))\), we introduce a tool called \textit{generating function}. Consider the following formal power series

\[ s^L(z) := \sum_{i=0}^{\infty} (P^i(1_I))z^i, \]

where \(z\) is a formal variable. Notice that formally we have \(s^L(z) = (1 - zP)^{-1}(1_I)\). This suggests that the solution of the equation \(s^L(z) = 0\) is the reciprocal of the eigenvalue of the operator \(P\).

### 3.2 Fredholm determinant

Using the recursive relation of the sequence \((P^L(1_I))\), we can derive a closed formula of \(s^L(z)\). In section 4, we will derive the formula following the argument of Mori (see [5], where Fredholm determinants of wider class of piecewise linear maps are discussed). In this subsection, we admit the formula below and discuss its consequence.

In our setting, \(s^L(z)\) is given by the following formula:

\[ s^L(z) = \frac{1}{1 - z/\beta} 1_{J_0} + \frac{1}{(1 - z/\beta)(1 - E(z))} \left( \sum_{i=1}^{\infty} (z/\beta)^i 1_{J_i} \right), \tag{6} \]

where

\[ E(z) = \sum_{i=1}^{\infty} \phi(i-1) \left( \frac{z}{\beta} \right)^i, \quad \phi(i) = \begin{cases} 0 & (T^i(1/\beta) \geq 1/\beta), \\ 1 & \text{(otherwise)}. \end{cases} \]

Notice that the value \(\phi(i-1)\) is determined by the information of the left end point of \(J_i\). An interesting point about this closed formula is the following: while the domain of convergence of \(s^L(z)\) was initially \(|z| < 1\), \(s^L(z)\) converges for \(|z| < \beta\) in the new formula. Thus the formula above gives an analytic continuation of \(s^L(z)\).

In this formula, the part \((1 - z/\beta)(1 - E(z))\) is called the \textit{Fredholm determinant} of the Perron-Frobenius operator. As we will see in section 4, this part can be regarded as the characteristic equation of \(P\).

By taking the Taylor expansion of \(s^L(z)\), we can extract some information about \(n^{(L)}_i\). We put

\[ \frac{1}{(1 - z/\beta)(1 - E(z))} = \sum_{i=0}^{\infty} w_i z^i. \tag{7} \]

By expanding Eq. (6) and comparing the coefficients, we obtain

\[ P^L(1_I) = \frac{1}{(1/\beta)^L} 1_{J_0} + \sum_{i=1}^{L} w_{L-i} (1/\beta)^i 1_{J_i}. \tag{8} \]

Thus, by comparing the coefficients of \(1_{J_i}\), together with Eq. (5), we obtain

\[ n^{(L)}_0 = 1, \quad n^{(L)}_i = w_{L-i} \cdot \beta^{L-i} \quad (1 \leq i \leq L). \tag{9} \]
3.3 Taylor expansion of the coefficient function and MSE

In order to obtain the estimate of the MSE, we need to investigate the behavior of the sequence \((w_n)\). It can be deduced by the positions of zeros of \(1 - E(z)\). By definition of \(E(z)\), it is easy to see that \(|E(z)| < 1\) for \(|z| < 1\). Thus there is no zero of \(1 - E(z)\) in \(\{z \mid |z| < 1\}\).

We have the following.

**Lemma 1** Let
\[
R(z) = \sum_{i=0}^{\infty} a_i \left( \frac{z}{\beta} \right)^i, \quad a_i = \frac{1 - T^{i+1}(1/\beta)}{(\beta - 1)}. 
\]

1. We have \(1 - E(z) = (1 - z)R(z)\), \(\textup{(10)}\)
in particular, \(E(1) = 1\).
2. \(z = 1\) is a simple root of \(1 - E(z) = 0\).
3. There is no other root of \(1 - E(z) = 0\) on \(|z| = 1\).

**Proof.** (1.) By a direct calculation, it is easy to see that Eq. \((10)\) is true for the constant term. For \(n \geq 1\), the coefficient of \(z^n\) in the left hand side is \(-\phi(n - 1)/\beta^n\). The corresponding coefficient in the right hand side is
\[
\frac{1 - T^{n+1}(1/\beta)}{(\beta - 1)\beta^n} - \frac{1 - T^n(1/\beta)}{(\beta - 1)\beta^{n-1}} = \frac{(1 - \beta) - (T^{n+1}(1/\beta) - \beta T^n(1/\beta))}{(\beta - 1)\beta^n}.
\]
Then, by definition, we can check that \(T^{n+1}(1/\beta) - \beta T^n(1/\beta) = -(1 - \phi(n - 1))(\beta - 1)\) and this immediately gives the desired equality.

(2.) It follows because the coefficients of \(R(z)\) are non-negative.

(3.) For this item we just give an outline of the proof. It is known that the roots of \(1 - E(z) = 0\) on the unit circle \(|z| = 1\) form a finite group with respect to the multiplication (see for instance Proposition 3.5 of [6]). Meanwhile, we can see that \(z = e^{i\theta}\) where \(\theta \in 2\pi(Q \setminus \mathbb{Z})\) cannot be a root. Indeed, if it were, it corresponds to the existence of non-mixing measure, which is not possible in our setting. \(\square\)

This lemma tells us how the sequence \((w_n)\) grows. Since the rational function in Eq. \((10)\) has a simple root \(z = 1\) and there is no other root in the neighborhood of the disk \(|z| = 1\), we know that \(w_n = O(1) + o(\mu^{-n})\), where \(\mu\) is some constant greater than one. Furthermore, the term \(O(1)\) is computable as a residue of the Fredholm determinant at \(z = 1\). That is, we have
\[
w_n = \frac{\beta}{R(1)(\beta - 1)} + o(\mu^{-n}).
\]

3.4 Second eigenvalue and estimate of coefficient

In order to obtain the estimate of contribution of the second term \(o(\mu^{-n})\), we use the technique of contour integrals. Let \(\mu^1\) be the root of \(1 - E(z) = 0\) whose absolute value is second smallest (in other words, the root of \(R(z) = 0\) which has the smallest absolute value) and let \(\mu\) be any number in the open interval \(1, |\mu^1|\).

By the residue theorem, we can see that
\[
w_n = \frac{\beta}{R(1)(\beta - 1)} + v_n, \quad \text{where} \quad v_n = \frac{1}{2\pi i} \int_{|z| = \mu} \frac{dz}{(1 - z/\beta)(1 - z)z^{n+1}R(z)}. \quad (11)
\]
The second term decays exponentially as \(n \to +\infty\).

In section 5, we derive a formula which approximates \(K_{\beta,L}\) using these results.
4. Renewal of Perron-Frobenius operators

In this section, we derive Eq. (6). The result can be deduced from the general result by Mori [5]. Fortunately, in our setting a simpler and direct calculation is possible and we will give it in this section. We also refer to the article [7], where a similar calculation is carried out for conventional $\beta$-transformations. We obtain the closed formula by using the recursive relations of Perron-Frobenius operators repeatedly. This technique is called renewal method.

Let us start the calculation. We put $A = [0, 1/\beta)$ and $B = [1/\beta, 1]$. Then we define the sequence of intervals $K_i$ as follows ($X$ denotes the closure of $X$):

$$K_i = \begin{cases} J_i \setminus B & \text{(if } J_i \setminus B \neq \emptyset), \\ J_i & \text{(otherwise)}. \end{cases}$$

Notice that we have $T(K_i) = J_{i+1}$ and $\mathbb{1}_I = \mathbb{1}_A + \mathbb{1}_B$. Also $\mathbb{1}_{K_i}$ has the following simple formula (for the definition of $\phi(i)$, see the explanation of Eq. (6)):

$$\mathbb{1}_{K_i} + \phi(i-1)\mathbb{1}_B = 1_{J_i}.$$ 

Accordingly, we have the following:

$$P(1_{K_i}) = (1/\beta)1_{J_{i+1}} = (1/\beta)(1_{K_{i+1}} + \phi(i)1_B).$$

4.1 Renewal of $s^A$

For an interval $X \subset I$ we put $s^X(z) := \sum_{i=0}^{\infty} P^i(1_X)z^i$. We are interested in calculating $s^I$. Since $s^I = s^A + s^B$, we only need to calculate $s^A$ and $s^B$. First, we calculate $s^A(z)$. By definition, together with the formula $P(1_A) = (1/\beta)1_I = (1/\beta)(1_A + 1_B)$, we have

$$s^A(z) = \mathbb{1}_A + \sum_{i=0}^{\infty} (1/\beta)\mathbb{1}_AP^i(1_A + 1_B)z^i = \mathbb{1}_A + (z/\beta)(s^A(z) + s^B(z)). \quad (12)$$

This process is called renewal because we “renewed” the original equality using the recursive relation.

4.2 Renewal of $s^B$

Let us calculate $s^B$. First, using

$$P(1_B) = (1/\beta)1_{J_1} = (1/\beta)(1_{K_1} + \phi(0)1_B), = (1/\beta)(1_{K_1} + 1_B),$$

we have

$$s^B(z) = 1_B + \sum_{i=0}^{\infty} (1/\beta)\mathbb{1}_AP^i(1_K + \phi(0)1_B)z^i = 1_B + (z/\beta)s^{K_I}(z) + \phi(0)(z/\beta)s^B(z).$$

We continue the renewal process for $s^{K_I}(z)$. We have

$$s^{K_I}(z) = 1_{K_1} + \sum_{i=0}^{\infty} (1/\beta)\mathbb{1}_AP^i(1_K + \phi(1)1_B)z^i = 1_{K_1} + (z/\beta)s^{K_2}(z) + \phi(1)(z/\beta)s^B(z).$$

This calculation is valid for all $K_i$. Indeed, for every $i \geq 1$ we have

$$s^{K_i}(z) = 1_{K_i} + (z/\beta)s^{K_{i+1}}(z) + \phi(i)(z/\beta)s^B(z).$$

Now we are ready for completing the calculation. We have

$$s^B(z) = 1_B + \phi(0)(z/\beta)s^B(z) + (z/\beta)s^{K_1}(z)$$

$$= (1_B + (z/\beta)1_{K_1} + \phi(0) + \phi(1)(z/\beta))s^B(z) + (z/\beta)^2s^{K_2}(z)$$

$$= (1_B + (z/\beta)1_{K_1} + (z/\beta)^21_{K_2} + \phi(0) + \phi(1)(z/\beta) + \phi(2)(z/\beta)^2)s^B(z) + (z/\beta)^3s^{K_3}(z)$$

$$= \cdots = \left(1_B + \sum_{i=1}^{\infty} (z/\beta)^i1_{K_i} \right) + \sum_{i=1}^{\infty} \phi(i-1)(z/\beta)^i s^B(z).$$
We put
\[ \chi(z) := \mathbb{1}_B + \sum_{i=1}^{\infty} (z/\beta)^i \mathbb{1}_{K_i}, \quad E(z) := \sum_{i=1}^{\infty} \phi(i-1)(z/\beta)^i. \]
Then we have
\[ s^B(z) = \chi(z) + E(z)s^B(z). \] (13)

4.3 Solving the renewal equation
Now we are ready to deduce the closed formula of \( s^I = s^A + s^B \). By Eq. (12) and Eq. (13), we have
\[ \begin{pmatrix} s^A \\ s^B \end{pmatrix} = \begin{pmatrix} \mathbb{1}_A \\ \chi(z) \end{pmatrix} + \begin{pmatrix} z/\beta \\ z/\beta \\ 0 \\ E(z) \end{pmatrix} \begin{pmatrix} s^A \\ s^B \end{pmatrix}. \]
Then, by solving this equation, we have
\[ \begin{pmatrix} s^A \\ s^B \end{pmatrix} = \frac{1}{(1 - z/\beta)(1 - E(z))} \begin{pmatrix} 1 - E(z) \\ 0 \\ 1 - z/\beta \end{pmatrix} \begin{pmatrix} \mathbb{1}_A \\ \chi(z) \end{pmatrix}. \]
Thus we have
\[ s^I = s^A + s^B = \frac{1}{(1 - z/\beta)(1 - E(z))}((1 - E(z))\mathbb{1}_A + \chi(z)). \]
Now, let us prove that this formula is equivalent to Eq. (6). We have \( \mathbb{1}_A = 1_I - 1_B \). So, the above formula is equal to
\[ \frac{1}{(1 - z/\beta)(1 - E(z))}((1 - E(z))1_I - (1 - E(z))1_B + \chi(z)) = \frac{1}{(1 - z/\beta)}1_I + \frac{1}{(1 - z/\beta)(1 - E(z))}(-(1 - E(z))1_B + \chi(z)). \]
Let us see that
\[ -(1 - E(z))1_B + \chi(z) = \sum_{i=1}^{\infty} (z/\beta)^i \mathbb{1}_{J_i}. \]
Indeed, one can see the equality of the constant term easily. In the left hand side the coefficient of \( z^n \) is equal to
\[ (\phi(n-1)1_B + \mathbb{1}_{K_n})/\beta^n = (1/\beta^n)1_{J_n}. \]
This concludes the proof.

5. Estimations
In this section, we numerically calculate the constant in Eq. (4). By virtue of Eq. (9) and Eq. (11), we already have theoretical estimations of the MSEs in Eq. (4). Let us make several estimations which is useful to obtain numerical upper bounds of \( K_{\beta,L} \).

5.1 The contour integrals
First, let us estimate the decaying term in Eq. (11). Suppose that we have some lower bound of \( |\mu^I| \). We denote it by \( \bar{\mu} \). As we will see later, given a small interval of \( \beta \) values, we can obtain some \( \bar{\mu} \) by numerical verification method. We fix \( \mu \in (1, \bar{\mu}) \) and put
\[ V_\mu = \frac{1}{2\pi} \int_0^{2\pi} \frac{d\theta}{|(1 - \mu e^{\sqrt{-1} \theta}/\beta)(1 - \mu e^{\sqrt{-1} \theta})R(\mu e^{\sqrt{-1} \theta})|}. \] (14)
Then we can see that \( |\nu_n| \leq V_\mu/\mu^n \), where \( \nu_n \) is the term defined in Eq. (11). Notice that in Eq. (14) the variable of the integration is \( \theta \), while it is \( z \) in Eq. (11). Hence, the upper bound of \( |\nu_n| \) is not \( V_\mu/\mu^{n+1} \) but \( V_\mu/\mu^n \). Thus, fixing \( \mu \) and calculating the integral numerically, we can obtain the upper bound of \( V_\mu \) and this gives the upper bound of \( \nu_n \) for \( n \geq 1 \).
Let us estimate $N$. We fix some integer $\mu$.

5.2 Estimation of $I_i$

Let us find some upper bound of $I_i$. By a direct calculation (using the definitions of $I_i$ in Eq. (3), $\theta = (3 - \beta)/2$ and $a_i$ in Lemma 1), we can see that

$$I_i = \frac{(\beta - 1)^3a_i(4a_i^2 - 6a_i + 3)}{12} \quad (i \geq 1).$$

(15)

Remember that $a_i$ is the sequence defined in Lemma 1. Also, for $I_i$ where $i \geq 1$, there is a trivial upper bound: $I_i$ is maximal when $i = 2 - \beta$ (i.e., the left end of $I_i$) and in that case $I_i = (\beta - 3)/12$. Thus we have

$$0 \leq I_i \leq \frac{(\beta - 1)^3}{12}.$$

The calculation of $I_i$ is easy for small $i$. Meanwhile, for larger $i$ it is difficult due to the sensitive dependency of $a_i$ on $\beta$. Fortunately, the contributions of $I_i$ for large $i$ are smaller in Eq. (4). Thus in the actual calculation we give rigorous estimation of $I_i$ for smaller $i$ and use the rough upper bound above for larger $i$.

5.3 Principal parts

We fix some integer $N$ (in the actual calculation we set $N = 20$) and put

$$R_N = \sum_{i=0}^{N} \frac{a_i}{\beta^i}, \quad M_N = \frac{1}{R_N} \sum_{i=1}^{N} \frac{I_i}{(\beta - 1)^{\beta i - 1}} = \frac{1}{R_N} \sum_{i=1}^{N} \frac{(\beta - 1)^2a_i(4a_i^2 - 6a_i + 3)}{12\beta^{i-1}}.$$

(16)

The number $R_N$ converges monotonically to $R(1)$ as $N \to \infty$ and $M_N$ converges to $\lim_{L \to \infty} K_{\beta,L}$ as $N \to \infty$. Thus these two values are good approximations of corresponding constants when $N$ is large.

5.4 Estimation of MSE

Let us estimate $K_{\beta,L}$ using $R_N$ and $M_N$ for $L \geq N$. First, using Eq. (4), Eq. (9), Eq. (11) and Eq. (16), we have the following:

$$K_{\beta,L} - M_N = \frac{I_0}{\beta^L} + \frac{1}{R(1)} \sum_{i=1}^{L} \frac{I_i}{(\beta - 1)^{\beta i - 1}} + \sum_{i=1}^{L} v_{L-i} \cdot \frac{I_i}{\beta^i} - \frac{1}{R_N} \sum_{i=1}^{N} \frac{I_i}{(\beta - 1)^{\beta i - 1}}$$

$$\leq \frac{I_0}{\beta^L} + \frac{1}{R_N} \sum_{i=N+1}^{\infty} \frac{I_i}{(\beta - 1)^{\beta i - 1}} + \sum_{i=1}^{L} v_{L-i} \cdot \frac{I_i}{\beta^i} - \frac{1}{R_N} \sum_{i=1}^{N} \frac{I_i}{(\beta - 1)^{\beta i - 1}}$$

$$= \frac{I_0}{\beta^L} + \frac{1}{R_N} \sum_{i=N+1}^{\infty} \frac{I_i}{(\beta - 1)^{\beta i - 1}} + \sum_{i=1}^{L} v_{L-i} \cdot \frac{I_i}{\beta^i}$$

$$\leq \frac{3\beta^2 - 12\beta + 13}{12\beta^N} + \frac{1}{R_N} \sum_{i=N+1}^{\infty} \frac{1}{(\beta - 1)^{\beta i - 1}} \cdot \frac{(\beta - 1)^3}{12} + \sum_{i=1}^{\infty} \frac{v_i}{\mu^{i-1}} \cdot \frac{(\beta - 1)^3}{12\beta^i}$$

$$= \frac{3\beta^2 - 12\beta + 13}{12\beta^N} + \frac{1}{R_N} \cdot \frac{(\beta - 1)}{12\beta^{N-1}} + \frac{(\beta - 1)^3}{12(\beta - \mu)\mu^{N-1}} =: E_N^\mu.$$

In other words, we have the following upper bound:

$$K_{\beta,L} \leq M_N + E_N^\mu.$$

(17)

The quantities appearing the last formula can be calculated numerically and together with the numerical verification method, one can get the upper bound of it. Notice that $K_{\beta,L}$ can be bounded from above by $M_N + E_N^\mu$ for every $L \geq N$.

By a similar calculation, we can also find a lower bound of $K_{\beta,L}$:

$$K_{\beta,L} - M_N = \frac{I_0}{\beta^L} + \frac{1}{R(1)} \sum_{i=1}^{L} \frac{I_i}{(\beta - 1)^{\beta i - 1}} + \sum_{i=1}^{L} v_{L-i} \cdot \frac{I_i}{\beta^i} - \frac{1}{R_N} \sum_{i=1}^{N} \frac{I_i}{(\beta - 1)^{\beta i - 1}}$$

251
\[ K_{\beta,L} \geq M_N - E_N^f. \]

### 5.5 Estimation of the accuracy

Now we have an upper bound of \( K_{\beta,L} \). In the next section we calculate the upper bound numerically. Using the lower bound we obtained, we can estimate how accurate the upper bound is.

We are interested in estimating the error

\[ \Delta_N = \frac{M_N + E_N^u - K_{\beta,L}}{K_{\beta,L}}. \]

Notice that

\[ \Delta_N = \frac{M_N + E_N^u - K_{\beta,L}}{K_{\beta,L}} \leq \frac{M_N + E_N^u}{M_N} - \frac{K_{\beta,L}}{M_N} = \frac{E_N^u + E_N^f}{M_N} =: \tilde{\Delta}_N. \]

Thus \( \tilde{\Delta}_N \) gives the upper bound of \( \Delta_N \).

### 6. Numerical techniques

By the calculations in section 5, now we have theoretical upper bounds of \( K_{\beta,L} \). By means of numerical verification method, we can obtain mathematically rigorous upper bounds of these constants numerically. In this section, we explain the numerical techniques we used.

#### 6.1 Interval arithmetic

For the estimation of \( K_{\beta,L} \), we need to have rigorous estimates of the constants appearing in Eq. (17) for wide range of \( \beta \) values. The method we will use is so-called interval arithmetic (see [9]). This is a method of calculating real numbers keeping the information of their possible inaccuracy in each arithmetic operation. We deal with real numbers which in general have infinite number of digits. Since the number of digits are infinite, we cannot do the exact calculation of them in reality. However, it is possible to keep their upper and lower bounds given by numbers of finitely many digits. Thus, instead of treating a real number, we consider an interval which contains the real number we are interested in. In each calculation, we calculate the worst lower bound and the upper bound of the result, which gives us another interval. Then we continue the calculation. This calculation is not difficult to implement by computers and the result allows us to obtain rigorous conclusions for specific problems.

Let us explain in more precise way. Let \( \mathbb{IR} \) be the set of closed intervals of \( \mathbb{R} \). For \( X,Y \in \mathbb{IR} \) and any one of arithmetic operation \( \ast \) (where the symbol \( \ast \) denotes +, -, \( \times \) or \( \div \)), we define a binary operation

\[ X \ast Y = \{ x \ast y \mid x \in X, y \in Y \}. \]

In the actual interval arithmetic, we also take care the rounding errors resulted from the computer calculations.

As long as the calculations are formulated by finitely many arithmetic operations, we can obtain the upper and lower bounds of the result. Notice that the upper bound we obtain at last may have
big errors. However, by taking the initial intervals very close to the true numbers, we can decrease the errors.

Notice that the use of interval arithmetic enables us to obtain the upper bound of $K_{\beta,L}$ not only for a single, specific $\beta$ value but also for $\beta$ values contained in an (possibly small) interval, since in the interval arithmetics we do not deal with single real number but an interval containing the $\beta$ value we are interested in.

There are several softwares for the implementation of numerical verification method such as INTLAB, which is available in the toolbox of MATLAB ([8]). In this research, we natively implemented interval arithmetic by floating-point arithmetic with rounding mode (rounding direction) control of C++.

In the following, we explain how we applied this idea for the calculation of $K_{\beta,L}$.

6.2 Estimation of $a_i$ and bifurcation of intervals
For the estimation of $K_{\beta,L}$ we need to calculate the value $a_i$ given by the formula in Lemma 1. It is not difficult to check that $a_i$ satisfies the following recursive relation:

$$a_i = \begin{cases} \beta a_{i-1} - 1 & (a_{i-1} \leq 1/\beta) \\ \beta a_{i-1} - 1 - 1 & (a_{i-1} > 1/\beta) \end{cases}.$$  

Using this relation, one can calculate the possible range of $a_i$ for given an interval of $\beta$ values.

In the calculation of $a_i$, a direct application of interval arithmetic may bring large errors due to the discontinuity of the recursive relation of $a_i$. To avoid this problem, we used a technique which we call the bifurcation of intervals.

Let us explain this. For the calculation of the interval of $a_i$, we use the above formula. However, since it contains a point of discontinuity, the resulted interval may experience an unexpectedly large increase of interval width (this phenomenon is called wrapping effect) which cannot be overcome by shrinking the $\beta$-interval. To solve this problem, in the course of the calculation of $a_i$, when the interval touches the point of discontinuity we split the interval into two sub-intervals at the point of discontinuity and continue the calculation for each interval. This division technique allows us to avoid the wrapping effect.

Once we have obtained the intervals for $a_i$, then it is straightforward to calculate the intervals of $I_i$, $R_N$, $M_N$ and other constants appearing in Eq. (17) except $V_\mu$.

6.3 Estimation of $\mu$
In order to obtain the upper bound of the constant $V_\mu$, we need to have some estimate about the root of the equation $R(z) = 0$ having the smallest absolute value. This value is important for determining the radius of the integral $V_\mu$. In order to obtain a lower bound of it, we prove the non existence of roots of the equation $R(z) = 0$ by the numerical verification method. Let us see the idea of the calculation.

We divide the complex plane $\mathbb{C}$ into small squares whose edges are parallel to the real and the imaginary axes. Let us take one of these squares. We denote it by $D$. Suppose that $D$ has the form

$$D = \left\{ c + (a + b\sqrt{-1}) \mid c \in \mathbb{C}, |a|, |b| \leq h \right\},$$  

that is, $D$ is the square whose center is $c$ and the length of the edges are $2h$. Then, we calculate a lower bound of $|f(z)|$ where $z \in D$ by interval arithmetic. If this value were positive, then we can guarantee that there is no zero in $D$. We perform the same calculation for each $D$ and this process increases the size of the region where there is no zero.

It may be that for some of the squares the lower bounds we obtain are negative. Indeed, this could naturally happen for squares which are near some of the roots. In such a case, we divide the square under consideration into smaller ones and do the similar calculations for them.

Let us see precisely the calculation we have done. We are interested in estimating the value $|R(z)|$ in $D$ where $D$ is some square in the complex plane. We fix some integer $N$ and put $R_N(z) := \sum_{i=0}^{N-1} a_i(z/\beta)^i$.
(in the actual calculation, we set $N = 20$). We have (using the inequality $0 \leq a_i \leq 1$, which can be obtained by the definition of $a_i$ and the inequality $2 - \beta \leq T^i(1/\beta) \leq 1$ for $i \geq 1$)

$$|R(z) - R_N(z)| = \left| \sum_{i=N}^{\infty} a_i \left( \frac{z}{\beta} \right)^i \right| \leq \sum_{i=N}^{\infty} \left( \frac{z}{\beta} \right)^i \leq \frac{(|z|/\beta)^N}{1 - |z|/\beta}.$$  

Thus, the calculation of $R(z)$ can be approximated well by $R_N(z)$ when $|z/\beta|$ is small and $N$ is large. We put $r := \max_{z \in D} |z|$. Then we have

$$|R'(z)| = \left| \sum_{i=1}^{\infty} \frac{i a_i}{\beta} \left( \frac{z}{\beta} \right)^{i-1} \right| \leq \sum_{i=1}^{\infty} \frac{i r^{i-1}}{\beta^i} = \frac{\beta}{(\beta - r)^2}.$$  

Hence, for $z \in D$, we have

$$|R(z)| \geq |R(c)| - \sqrt{2} h \max_{z \in D} |R'(z)| \geq |R_N(c)| - \frac{(|c|/\beta)^N}{1 - |c|/\beta} - \frac{\sqrt{2}/\beta h}{(\beta - r)^2}.$$  

Notice that the number in the last formula can be calculated by finite number of arithmetic operation once we fix $c$, $h$, $\beta$ and $N$. Thus, by calculating this value, if the lower bound of the result is positive, then we can rigorously guarantee the non-existence of the root in $D$.

Figure 2 shows the result of the numerical verification of the position of the smallest root.

---

**Fig. 2.** Verification of non-zero regions of $R(z)$. $\beta = [1.6, 1.60001]$ (left-top), $\beta = [1.7, 1.70001]$ (right-top), $\beta = [1.8, 1.80001]$ (left-down), $\beta = [1.9, 1.90001]$ (right-down). Horizontal axis is the real axis and the vertical axis is the imaginary axis of the complex plane.
6.4 Estimation of $V_\mu$

By doing the calculation in the previous subsection for small $\beta$ value interval, we can find a lower bound of the absolute value of the smallest root of $R(z) = 0$. We denote it by $\tilde{\mu}$. Then, for every $\mu \in (1, \tilde{\mu})$, we have the value $V_\mu$ and the inequality $|\nu_n| \leq V_\mu/\mu^n$.

We have the freedom of the choice of $\mu$. What choice of $\mu$ is advantageous for our purpose? At a first glance, it seems that choosing $\mu$ as large as possible would be convenient because larger $\mu$ value seems to make the last term of Eq. (17) smaller. However, choosing $\mu$ close to the root is not always a good idea. For $\mu$ which is close to the root, the path of the integration passes close to the pole of the integrand and it makes the value of $V_\mu$ larger and it makes the estimation bad.

On the other hand, it is true that we need to choose $\mu$ large enough so that the last term in Eq. (17) becomes small. In the actual calculation, for each $\beta$ value, we used the upper bound obtained by the calculation multiplied by 0.95. We found this coefficient 0.95 empirically. The value $0.95\tilde{\mu}$ seems to be small enough so that we can avoid the explosion of the value of the integration $V_\mu$.

Once the $\mu$ value is fixed, then the calculation of $V_\mu$ can be done by numerical verification method. Indeed, for $h > 0$ and $\theta_0 \in \mathbb{R}$, put $G = [\theta_0 - h, \theta_0 + h]$. Then by a similar argument as we did for $R(z)$, for every $\theta \in G$, we have

$$\left| R(\mu e^{\sqrt{-1} \theta}) \right| \geq \left| R_N(\mu e^{\sqrt{-1} \theta_0}) \right| - \frac{(\mu/\beta)^N}{1 - \mu/\beta} - \frac{\beta \mu h}{(\beta - \mu)^2}.$$

Using this estimation and dividing the integral path into small pieces, we can obtain a rigorous upper bound of $V_\mu$ numerically.

7. Numerical results

In this section, we give the numerical result we obtained. We considered $\beta$-values in the interval $[1.6, 2.0]$.

7.1 Second eigenvalues

We divided the interval $[1.6, 2.0]$ into sub-intervals of length $10^{-5}$ and on each sub-interval we calculated the lower bound $\tilde{\mu}$. The picture below shows the behavior of $\tilde{\mu}$ (see Fig. 3).

![Fig. 3. The behavior of $\tilde{\mu}$ as $\beta$ varies. Horizontal: $\beta$ value, vertical: $\tilde{\mu}$ value.](image-url)
The behavior of $V_\mu$ as $\beta$ varies. We calculated the upper bound of $V_\mu$ setting $\mu = 0.95 \tilde{\mu}$.

Table I. Examples of numerical results.

| $\beta$     | $\tilde{\mu}$ | $V_\mu$  | $M_N + E_N^u$ | $\tilde{\Delta}_N$ |
|-------------|---------------|----------|---------------|---------------------|
| [1.6, 1.60001] | 1.36557286    | 1.18722316 | 2.85593222 $\times 10^{-2}$ | 1.79246814 $\times 10^{-2}$ |
| [1.7, 1.70001] | 1.44949992    | 1.18158313 | 4.05269587 $\times 10^{-2}$ | 5.87183185 $\times 10^{-3}$ |
| [1.8, 1.80001] | 1.47558593    | 1.00272116 | 4.98729333 $\times 10^{-2}$ | 3.29438400 $\times 10^{-3}$ |
| [1.9, 1.90001] | 1.60520856    | 9.03527264 $\times 10^{-1}$ | 6.34505242 $\times 10^{-2}$ | 7.37710286 $\times 10^{-4}$ |

7.2 Calculation of $K_{\beta,L}$

Then, using these lower bounds of $\tilde{\mu}$, we calculated the upper bound of $K_{\beta,L}$. On each $\beta$ value sub-interval, we first calculated the upper bound of $V_\mu$. For the calculation we set $\mu$ to be the $\tilde{\mu}$ value obtained in the previous subsection multiplied by 0.95. For the calculation of the integration, we divided the integral path into $10^6$ pieces of curves and calculated the upper bound of $V_\mu$. For the result of the calculation of $V_\mu$, see Fig. 4. Using this $V_\mu$ value, on each sub-interval we calculated an upper bound of $K_{\beta,L}$, that is, $M_N + E_N^u$. Notice that this value is a upper bound valid for $K_{\beta,L}$ for every $L \geq N$. The graph below shows the result (see Fig. 5). The table below shows examples of values we calculated (see Table I).

7.3 Fitting with a quadratic function

Then, we calculated the quadratic function of the form $C(\beta - 1)^2$ which bounds the constant $K_{\beta,L} = M_N + E_N^u$ from above. We chose quadratic function as a fitting function because $M_N$ in Eq. (16) is quadratic with respect to $(\beta - 1)$.

As a result, we rigorously obtained the inequality $K_{\beta,L} \leq 0.0857(\beta - 1)^2$ for every $\beta \in [1.6, 2.0]$ and $L \geq 20$ (see Fig. 5).

7.4 Comparison with numerical result

We also calculated $\Delta_N$, the upper bound of the accuracy of the MSEs we obtained. The graph below (see Fig. 6) shows the graph of $\Delta_N$. For the $\beta$ values we have calculated, the percentage of the error between our result and the true MSEs value did not exceed 2%.
8. Conclusion
In this article, we gave a method of giving a mathematically rigorous upper bound of MSEs of non-binary A/D-converters. The method is based on the spectral analysis of Perron-Frobenius operators of dynamical systems. By means of numerical verification method, we obtained rigorous lower bounds of (reciprocals of) second eigenvalues of Perron-Frobenius operators. Using these bounds, we gave
the upper bound of MSEs. The result we obtained is valid for encodings of every sufficiently large number of digits.

For the sake of simplicity, we restricted our attention for lazy type piecewise linear maps. However, it is possible to extend our method for more general piecewise linear maps and in the future work we will investigate MSEs of such encoders.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 16K00333, JP16H03950 and CREST Program “Development of Verified Numerical Computations for Mathematical Modeling” of Japan Science and Technology Agency (JST).

References

[1] I. Daubechies, R.A. DeVore, C.S. Güntürk, and V.A. Vaishampayan, “Beta expansions: A new approach to digitally corrected A/D conversions,” Proc. IEEE Int. Symp. Circuits Syst., vol. 2, pp. 784–787, 2002.

[2] R. Suzuki, T. Maruyama, H. San, and M. Hotta, “Robust Cyclic ADC Architecture Based on β-Expansion,” IEICE TRANS. ELECTRON., vol.E96-C, no. 4, pp. 553–559, 2013.

[3] T. Makino, Y. Iwata, K. Shinohara, Y. Jitsumatsu, M. Hotta, H. San, and K. Aihara, “Rigorous estimates of quantization error for A/D converters based on beta-map,” NOLTA J., vol. 6, no. 1, pp. 130–141, 2015.

[4] K. Shinohara and K. Kobayashi, “Fredholm Determinants of Generalized β-Transformations and MSE Estimates of Corresponding AD-Converters,” Proc. NOLTA’16, paper ID 1054, pp. 514–517, November 2016.

[5] M. Mori, “Fredholm determinant for piecewise linear transformations,” Osaka J. Math., vol. 27, no. 1, pp. 81–116, 1990.

[6] V. Baladi, “Positive Transfer Operators and Decay of Correlations,” World Scientific, 2000.

[7] M. Mori, “Ichi-jigen rikigakukei to Cantor sets (Japanese),” Suuron to Ergodic Theory (Seminar on Probability, lecture note, 1996), vol. 61, 1998.

[8] S.M. Rump, “INTLAB - INTerval LABoratory”. T. Csendes (Ed), “Developments in Reliable Computing,” pp. 77–104, Kluwer Academic Publishers, Dordrecht, 1999.

[9] R.E. Moore, R.B. Kearfott, and M.J. Cloud, “Introduction To Interval Analysis,” Cambridge University Press, 2009.