Paper

Invariant set of two-dimensional dynamics of golden ratio encoders

Yutaka Jitsumatsu$^{1a}$

$^1$ Department of Informatics, Kyushu University
744 Motooka, Nishi-ku, Fukuoka 819-0395, Japan

$a$) jitumatu@inf.kyushu-u.ac.jp

Received April 11, 2020; Revised July 21, 2020; Published January 1, 2021

Abstract: A discrete-time two-dimensional dynamical system appears in a Golden Ratio Encoder (GRE), a type of analog-to-digital converter. One of the essential elements in analyzing a given dynamical system is identifying the invariant set of that system. The invariant set of dynamics of GREs is not known, except in special cases. We herein determine the invariant set of the dynamics of GREs with an amplification factor $\alpha$ and a threshold $\theta$ for a wide range of parameters ($\alpha, \theta$). The invariant set is separated into six sub-regions and the transition probabilities between the sub-regions are defined. We show that the uniform distribution on the invariant set is an invariant density for this dynamical system.

Key Words: golden ratio encoder (GRE), $\beta$-encoders, analog-digital converter (ADC), two-dimensional dynamical systems

1. Introduction

A $\beta$ encoder is an analog-to-digital converter (ADC) as proposed by Daubechies et al. [1, 2] and is based on $\beta$-expansion [3]. $\beta$-expansion is a typical example of a discrete-time one-dimensional dynamical system, and its properties have been reported [4, 5]. A $\beta$-encoder is robust to the uncertainty of its quantizer and also has exponential accuracy [6–10]. Such a property makes a $\beta$ encoder an attractive candidate for an ADC.

$\beta$-coders, however, have a drawback in that they are not robust to the fluctuation of the $\beta$ value. Therefore, a $\beta$ encoder has to be equipped with a precise estimation of $\beta$ value [2, 6]. Circuit implementations of $\beta$ encoders with $\beta$ value estimation were reported in [11, 12]. These studies showed that the benefit of a $\beta$-encoder is limited by the cost of $\beta$-value estimation.

A Golden Ratio Encoder (GRE) is another class of ADC [13] that is based on $\beta$ expansion, but the radix is fixed to the golden ratio and therefore the $\beta$ value estimation process is not needed. This implies that GREs can overcome the drawbacks of $\beta$ encoders. Unlike $\beta$-encoders, the dynamics of GREs are described by a discrete-time two-dimensional map. Such a map is called a GRE map. The basic robustness property of a GRE map was analyzed in [6, 13], but the accuracy analysis of the GRE was not sufficient. Signal to Quantization Noise Ratio (SQNR) is one of the important quantities that measures the accuracy of an ADC [14]. In order to do theoretical evaluation of the SQNR of GREs, it is necessary to determine the invariant set of the GRE map. In [13], an invariant set of a GRE map was given, but its specification was not sufficient to analyze the quantization error accurately.
The assumptions in this paper are different from those in [13] in the following points: (i) We assume the quantizer is ideal, while in [13] a flaky quantizer is employed. (ii) We assume no additive noise for the GRE map, while it is allowed in [13]. These assumptions make the GRE map deterministic and hence make the analysis tractable. The details of the definition of the GRE map are described in Section 2.

The contributions of this paper are as follows: We determine the invariant set of GREs and show that the invariant set is given by the union of two trapezoids and a parallelogram. We separate the invariant set into six sub-regions. Explicit forms of the transition probabilities between sub-regions are determined. Finally, we show that the uniform distribution on the invariant set is the invariant density of the GRE map.

2. Review of golden ratio encoders

In this section, we give a formal definition of a golden ratio encoder (GRE). A GRE generates the binary codes for the sample invariants.

\[ x = \sum_{n=1}^{\infty} b_n^* \beta^{-n} \quad \text{and} \quad \hat{x}_N = \sum_{n=1}^{N} b_n^* \beta^{-n}, \]  

where \( \beta = (1 + \sqrt{5})/2 \) is the golden ratio. Hereafter, we denote the golden ratio by \( \phi \) instead of \( \beta \) to avoid confusion with a \( \beta \) encoder in which \( \beta \) can take any real number in \([1, 2]\).

For the analysis of robustness and accuracy of GREs, it is beneficial to employ a \{+1, −1\}-valued bipolar code. In this case, the input range of \( x \) is \([-V_{FS}/2, V_{FS}/2]\), where \( V_{FS} \) is called a full-scale voltage. The reason we choose such a bipolar code is that the symmetry of this formulation makes the analysis of the dynamics easier than the normal \{0, 1\}-valued binary code with input range \([0, V_{FS}]\).

One can easily obtain the binary code \( b_n^* \) from the \{−1, +1\}-valued code \( b_n \) by \( b_n^* = (b_n + 1)/2 \).

In this paper, we consider a deterministic dynamical system so that the parameters \( \theta \) and \( \alpha \) are assumed to be fixed during AD conversion. We also assume that the encoder and decoder do not know the exact value of \( \alpha \) and \( \theta \). A formal definition of a GRE in this paper is as follows:

**Definition 1** An \( N \)-bit golden ratio encoder (GRE) with amplification factor \( \alpha \), threshold \( \theta \), and a full scale voltage \( V_{FS} \) is an analog-to-digital converter (ADC) with input value \( x \in [-\frac{V_{FS}}{2}, \frac{V_{FS}}{2}] \). Its output binary code \( b^N(x) = b_N = (b_1b_2...b_N), b_n \in \{-1, +1\} \) is recursively determined by

\[ b_n = Q(\alpha u_n + u_{n-1} - \theta), \quad n = 1, 2, \ldots, N, \]  

\[ u_{n+1} = u_n + u_{n-1} - \frac{V_{FS}}{2\phi} b_n, \]  

\[ (u_0, u_1) = (0, x) \]  

where \( Q(x) \) is a 1 bit quantizer defined by \( Q(x) = +1 \) for \( x \geq 0 \) and \( Q(x) = -1 \) for \( x < 0 \).

Figure 1 shows a block diagram of a GRE. A continuous-time analog signal \( x(t) \) is sampled at \( t = kT_s \), where \( T_s \) is the sampling interval. At this moment, the statuses of the three switches in Fig. 1 are inverted and \( u_0 \) and \( u_1 \) are reset to 0 and \( x(kT_s) \). During the time interval \([kT_s, (k+1)T_s]\), \( b_1, b_2, \ldots, b_N \) are obtained for the sample \( x(kT_s) \).

In Daubechies et al.’s model [13], the above ideal 1 bit quantizer \( Q \) is replaced by a more realistic one called a flaky quantizer with two parameters \( \nu_1 \) and \( \nu_2(> \nu_1) \) that outputs a binary code by the following rule:

\[ Q_{\nu_1, \nu_2}(x) = \begin{cases} 
-1, & \text{if } x < \nu_1, \\
-1 \text{ or } +1, & \text{if } \nu_1 \leq x \leq \nu_2, \\
+1, & \text{if } x > \nu_2.
\]  

A flaky quantizer is a useful model that captures the uncertainty of the behavior of a quantizer. In this
paper, however, we restrict our attention to GREs with an ideal quantizer to simplify our analysis. The aim of our study is the evaluation of the SQNR of a GRE. To this aim, we analyze the invariant set and the invariant density of a GRE map with an ideal quantizer. The analysis of the SQNR of a GRE with a flaky quantizer is left for a future study.

Definition 2 The decoder of an $N$-bit GRE with a full scale voltage $V_{FS}$ is defined as a digital-to-analog converter with an input bit sequence $b^N = (b_1, b_2, \ldots, b_N)$, $b_n \in \{+1, -1\}$ and an output analog value $x$ defined by

$$\hat{x} (b^N) = \frac{V_{FS}}{2\phi} \sum_{n=1}^{N} b_n \phi^{-n}.$$  

(6)  

The quantization error is defined by the difference between $x$ and $\hat{x}(b^N(x))$. We have

$$x - \frac{V_{FS}}{2\phi} \sum_{n=1}^{N} b_n(x) \phi^{-n} = \phi^{-N+1} (u_N + \phi u_{N+1}).$$  

(7)  

See Eq. (25) in [13] for the derivation. See also Section 3.1.

The following theorem was proved in [13] (See also [6]):

Theorem 1 (Daubechies et al. [13]) If $\{u_n\}_{n \geq 0}$ is bounded, then we have

$$|x - \hat{x}(b^N(x))| \leq \frac{V_{FS}}{2\phi} \phi^{-N+1}.$$  

(8)  

This theorem assures that a GRE has exponential accuracy in $N$ if $\{u_n\}_{n \geq 0}$ is bounded. Hence, analysis of the dynamics of $u_n$ is necessary to guarantee the accuracy of the GRE.

Here we give a short remark. The initial condition in [13] was $(u_0, u_1) = (x, 0)$. We replaced it by $(u_0, u_1) = (0, x)$ because this change makes the upper bound in (8) smaller by a factor of $\phi$.

In order to assure the robustness of a GRE against an inevitable error, Daubechies et al. employed the following two-dimensional map [13]:

$$T_{\alpha, \theta} \left( \begin{bmatrix} u \\ v \end{bmatrix} \right) = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} - \frac{V_{FS}}{2\phi} \cdot Q \begin{bmatrix} u + \alpha v - \theta \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix},$$  

(9)  

where $v$ is a variable to represent $u_n$ while $u$ is the one to represent $u_{n-1}$. Then, we have 

$$T_{\alpha, \theta} \left( \begin{bmatrix} u_{n-1} \\ u_n \end{bmatrix} \right).$$  

We derive the invariant set and the invariant measure of the two dimensional dynamical system (9) to analyze the accuracy of the GRE.

Daubechies et al. introduced a margin parameter $\mu$ and found an invariant set that satisfies $T_{\alpha, \theta}(R)+$
Fig. 2. The set of parameters \((\alpha, \theta)\) for which \(u_n\) is bounded.

\[ B_\mu(0) \subset R, \text{ where } B_\mu(0) \text{ denotes the open ball around } 0 \text{ with radius } \mu. \] This condition implies the invariant set \(R\) is robust to unknown fluctuation on \(u_n\). Then, for a fixed \(\mu\) the parameter space \((\alpha, \theta)\) for which \(\{u_n\}_{n \geq 0}\) is bounded was identified. We, however, restrict our attention to the case \(\mu = 0\) to simplify the discussion. Accuracy analysis of GREs with a positive \(\mu\) is left for future work.

The parameter space for which \(\{u_n\}_{n \geq 0}\) is bounded has already been analyzed in [13]. Because we employ \(\{+1, -1\}\)-valued ADC outputs, unlike in [13], we have to rephrase the parameter space for which \(\{u_n\}_{n \geq 0}\) is bounded. When \(\mu = 0\), this is given by

\[ |\theta| \leq \min\{\alpha - 1, \phi - 3\alpha, \frac{3 - \alpha}{3 - \phi}\}, \tag{10} \]

which is illustrated by a dark-colored kite-shaped quadrilateral in Fig. 2. The two dashed line segments in the kite-shaped quadrilateral will be explained later.

The robustness of GREs were also analyzed by Ward [6] who argued that in a realistic implementation the delay element has leakage so that (3) should be replaced by

\[ u_{n+1} = \rho_1 u_n + \rho_2 u_{n-1} - \frac{V_{FS}}{2\phi} b_n, \]

where \(\rho_1, \rho_2 \in [0, 1]\) are parameters expressing what proportion of stored voltage can be kept. In this paper, we only consider the ideal case where \(\rho_1 = \rho_2 = 1\).

For the accuracy of ADCs, the most frequently used criteria is the signal to quantization noise ratio (SQNR), which is defined as the power of a sinusoidal waveform with the full-scale voltage \(V_{FS}\) divided by the mean square quantization error (MSQE) [14]. The well-known formula for the theoretical SQNR of an ideal \(N\)-bit ADC is 6.02\(N + 1.76\) dB. The MSQE and SQNR of a GRE are defined as follows:

**Definition 3** The MSQE of an \(N\)-bit GRE with an amplification factor \(\alpha\), a threshold value \(\theta\), and a full scale voltage \(V_{FS}\) is defined as

\[ \text{MSQE} = \frac{1}{V_{FS}} \int_{-V_{FS}/2}^{V_{FS}/2} |x - \hat{x}(b^N(x))|^2 dx. \tag{11} \]

The SQNR of the GRE is defined as

\[ \text{SQNR} = \frac{\frac{1}{8} V_{FS}^2}{\text{MSQE}}. \tag{12} \]

Because the SQNR is independent of \(V_{FS}\), we can set the \(V_{FS}\) at any positive number to evaluate SQNR. We chose \(V_{FS} = 2\phi\) to simplify the expression. It is easily found that the linear term of the SQNR of the GRE in decibel is 20\(N \log_{10} \phi = 4.12N\) because the radix of GREs is the golden ratio \(\phi\). Our target is the constant term.

The purpose of our study is to give a theoretical evaluation of SQNR for a GRE. In [13, Section III-E], Daubechies et al. gave only slight mention to the 2-norm (i.e., MSQE) of a GRE. The focus was on the computation of the bias term \(\xi_N\), which is an extra term for the right hand side of (1) to make the \(N\)-bit approximation closer to \(x\). They stated that when \(\alpha = \phi\), the invariant set \(\Gamma\) is the union of at most three rectangles. However, the details of this point were omitted in [13]. In this paper, we analyze the invariant set \(\Gamma\) for a wide range of \(\alpha\) and \(\theta\). We show that \(\Gamma\) is the union of two trapezoids and a parallelogram for any \(\alpha\) and \(\theta\) satisfying

\[ |\theta| + |\alpha - \phi| \leq \phi^{-1}. \tag{13} \]
When \( \alpha = \phi \), the two trapezoids and the parallelogram reduce to the above mentioned three rectangles. Examples of the three rectangles will be shown in Figs. 5(a) and (c) of Section 5.

The first step for the analysis of the two-dimensional dynamics of (9) is to consider the orthogonal transformation of the matrix \( \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \). Put

\[
P = \frac{1}{\sqrt{2 + \phi}} \begin{bmatrix} \phi & 1 \\ -1 & \phi \end{bmatrix}, \quad \Lambda = \begin{bmatrix} -\phi^{-1} & 0 \\ 0 & \phi \end{bmatrix}.
\]

(14)

Let \( r_n = \begin{bmatrix} s_n \\ t_n \end{bmatrix} = P^{-1} \begin{bmatrix} u_{n-1} \\ u_n \end{bmatrix} \). Then (2), (3) to (4) are replaced by

\[
b_n = Q \left( \frac{1}{\sqrt{2 + \phi}} \left[ \phi - \alpha, 1 + \alpha \phi \right] r_n - \theta \right),
\]

(15)

\[
r_{n+1} = \Lambda r_n - \frac{1}{\sqrt{2 + \phi}} \begin{bmatrix} -1 \\ \phi \end{bmatrix} b_n,
\]

(16)

\[
r_1 = \frac{1}{\sqrt{2 + \phi}} \begin{bmatrix} -1 \\ \phi \end{bmatrix} r.
\]

(17)

Define \( \hat{T}(r) = \hat{T}\alpha,\phi(r) = P^{-1}T\alpha,\phi(Pr) \) as a two-dimensional map from \( r_n \) to \( r_{n+1} \) so that we have \( r_{n+1} = \hat{T}(r_n) \). Substituting \( V_{FS} = 2\phi \) into this equation, we have

\[
\hat{T}(r) = \begin{bmatrix} -\phi^{-1} & 0 \\ 0 & \phi \end{bmatrix} r - \frac{1}{\sqrt{2 + \phi}} \begin{bmatrix} -1 \\ \phi \end{bmatrix} Q \left( \frac{1}{\sqrt{2 + \phi}} \left[ \phi - \alpha, 1 + \alpha \phi \right] r - \theta \right).
\]

(18)

The first term of Eq. (18) acts as reflection followed by contraction by a factor of \(-\phi^{-1}\) along the direction of \( s \) and acts as expansion by a factor of \( \phi \) along the direction of \( t \), while the second term is a translation. We will analyze the dynamics of \( \hat{T} \).

3. Main results

In this section, we state our results. The invariant set of \( \hat{T} \) under an ideal model with \( \mu = 0 \) is analyzed. We do not consider the leakage of the charge and use \( \rho_1 = \rho_2 = 1 \).

We first review the definition of invariant sets [15].

**Definition 4** For a given \( D \)-dimensional discrete dynamical system \( x_{n+1} = \tau(x_n), S \in \mathbb{R}^D \) is said to be an invariant set if \( x_n \in S \) implies \( x_{n+1} \in S \) for all \( n = 1, 2, \ldots \).

Let us define the following six regions.

\[
R_1 = \{(s, t) : f_0(s) \leq t \leq f_{1+}(s), -\phi^2/\sqrt{2 + \phi} \leq s \leq -\phi^{-1}/\sqrt{2 + \phi}\}
\]

(19)

\[
R_2 = \{(s, t) : f_0(s) \leq t \leq f_{1+}(s), -\phi^{-1}/\sqrt{2 + \phi} \leq s \leq -\phi^2/\sqrt{2 + \phi}\}
\]

(20)

\[
R_3 = \{(s, t) : f_0(s) \leq t \leq f_{2+}(s), \phi^{-1}/\sqrt{2 + \phi} \leq s \leq \phi^2/\sqrt{2 + \phi}\}
\]

(21)

\[
R_4 = \{(s, t) : f_{2-}(s) \leq t \leq f_0(s), -\phi^2/\sqrt{2 + \phi} \leq s \leq -\phi^{-1}/\sqrt{2 + \phi}\}
\]

(22)

\[
R_5 = \{(s, t) : f_{1-}(s) \leq t \leq f_0(s), -\phi^{-1}/\sqrt{2 + \phi} \leq s \leq -\phi^2/\sqrt{2 + \phi}\}
\]

(23)

\[
R_6 = \{(s, t) : f_{1+}(s) \leq t \leq f_0(s), \phi^{-1}/\sqrt{2 + \phi} \leq s \leq \phi^2/\sqrt{2 + \phi}\}
\]

(24)

where

\[
f_0(s) = \frac{1}{1 + \alpha \phi} \{\theta \sqrt{2 + \phi} + (\alpha - \phi) s\}
\]

(25)

\[
f_{1\pm}(s) = \frac{\phi}{1 + \alpha \phi} \{(\theta \pm 1) \sqrt{2 + \phi} - \phi (\alpha - \phi) s\},
\]

(26)
Fig. 3. Regions $R_i$ and lines $f_0(s)$, $f_1(s)$, and $f_2(s)$ for $\alpha = 1.9$ and $\theta = 0.1$. $s_1$ and $s_2$ imply $\frac{\phi - 1}{\sqrt{2 + \phi}}$ and $\frac{\phi^2}{\sqrt{2 + \phi}}$. The invariant set of $\hat{T}(r)$ is the union of $R_i$s.

$f_2(s) = \frac{\phi^2}{1 + \alpha\phi}\left\{ (\theta + (2 - \alpha))\sqrt{2 + \phi} + \phi^2(\alpha - \phi)s \right\}$, \hspace{1cm} (27)

where $\pm$ takes $+$ or $-$. The regions $R_i$ are depicted in Fig. 3. By definition, the union of $R_2$ and $R_5$ is a parallelogram and the union of $R_1$ and $R_4$ and the union of $R_3$ and $R_6$ are trapezoids.

We have the following lemma.

Lemma 1 If the parameters $\alpha$ and $\theta$ satisfy the condition

$$|\theta| + |\alpha - \phi| \leq \phi^{-1},$$ \hspace{1cm} (28)

then we have

$$\hat{T}(R_1 \cup R_2) = R_3 \cup R_6,$$ \hspace{1cm} (29)

$$\hat{T}(R_3 \cup R_4) = R_2 \cup R_5,$$ \hspace{1cm} (30)

$$\hat{T}(R_5 \cup R_6) = R_1 \cup R_4.$$ \hspace{1cm} (31)

The proof of Lemma 1 is given in the next section.

We give the following theorem.

Theorem 2 Let $R_i$ be regions defined by (19)–(24). If $\alpha$ and $\theta$ satisfy (28) then the invariant set $\Gamma$ of the two-dimensional map $\hat{T}$ is given by the union of $R_i$s, i.e.,

$$\Gamma = \bigcup_{i=1}^{6} R_i.$$ \hspace{1cm} (32)

Proof: Put $\Gamma = \bigcup_{i=1}^{6} R_i$. From Lemma 1, for any $r \in \Gamma$ we have $\hat{T}(r) \in \Gamma$. Therefore, by Definition 4, we say $\Gamma$ is an invariant set of $\hat{T}$. $\square$

In order to evaluate the MSQE of the GRE, we must obtain not only the invariant set $\Gamma$ but also the invariant measure on $\Gamma$. It directly follows from (18) that the determinant of $A$ is 1 and therefore the two-dimensional map $\hat{T}$ preserves the measure. In order to analyze the dynamics of $\hat{T}$, we assume $r_1$ is randomly distributed with a certain probability density function $h_1(r)$ and introduce random variables $Z_n$, $n = 1, 2, \ldots$ with

$$Z_n = i \quad \text{if} \quad r_n \in R_i, \quad i = 1, 2, \ldots, 6.$$ \hspace{1cm} (33)

Let $\pi_i^{(n)} = \text{Prob}\{Z_n = i\}$. We give the following two lemmas.

Lemma 2 The random variables $Z_1, Z_2, \ldots$ of (33) form a Markov chain with a state transition
matrix $P = (p_{ij})$, where $p_{ij} = \text{Prob}(Z_{n+1} = j | Z_n = i)$. If $\alpha$ and $\theta$ satisfy (28), the matrix $P$ is explicitly given by

$$P = \begin{bmatrix}
0 & 0 & p_{13} & 0 & 0 & p_{16} \\
0 & 0 & p_{23} & 0 & 0 & p_{26} \\
0 & p_{32} & 0 & 0 & p_{35} & 0 \\
0 & p_{42} & 0 & 0 & p_{45} & 0 \\
p_{51} & 0 & 0 & p_{54} & 0 & 0 \\
p_{61} & 0 & 0 & p_{64} & 0 & 0
\end{bmatrix},$$

(34)

where

$$p_{13} = \frac{1}{\phi} \frac{\phi^{-\alpha} + \theta}{\phi^{-\alpha} + \theta + (\alpha - \phi)},$$

$$p_{16} = 1 \frac{2\alpha \phi^{-1} - 1 - \phi^{-2}\theta}{\phi \phi^{-\alpha} + \theta + (\alpha - \phi)},$$

$$p_{23} = \frac{1}{\phi^2} \frac{\phi - \alpha + \phi \theta}{1 + \phi^{-2}\theta},$$

$$p_{26} = \frac{1}{\phi^2} \frac{\phi - \alpha - \phi \theta}{1 + \phi^{-2}\theta},$$

$$p_{32} = \begin{cases}
0, & \text{if } \theta \leq -|\alpha - \phi| \\
\frac{2\phi - 1 - \phi \theta}{2\phi - 1 + \phi \theta}, & \text{if } -|\alpha - \phi| < \theta < |\alpha - \phi| \\
|\alpha - \phi|, & \text{if } \theta \geq |\alpha - \phi|
\end{cases},$$

$$p_{35} = 1 - p_{32},$$

$$p_{45} = \begin{cases}
\frac{2\phi - 1 + \phi \theta}{2\phi - 1 + \phi \theta}, & \text{if } \theta \leq -|\alpha - \phi| \\
1, & \text{if } -|\alpha - \phi| < \theta < |\alpha - \phi| \\
0 & \text{if } \theta \geq |\alpha - \phi|
\end{cases},$$

$$p_{42} = 1 - p_{45},$$

$$p_{51} = \frac{1}{\phi^2} \frac{\phi - \alpha - \phi \theta}{1 + \phi^{-2}\theta},$$

$$p_{54} = \frac{1}{\phi^2} \frac{\phi - \alpha + \phi \theta}{1 + \phi^{-2}\theta},$$

$$p_{61} = \frac{2\alpha \phi^{-1} - 1 + \phi^{-2}\theta}{\phi \phi^{-\alpha} - \theta + (\alpha - \phi)},$$

$$p_{64} = \frac{1}{\phi} \frac{\phi^{-\alpha} + \theta}{\phi^{-\alpha} + \theta + (\alpha - \phi)}.$$

The proof will be given in the next section.

**Lemma 3** Suppose $\alpha$ and $\theta$ satisfy condition (28). Then, the stationary distribution of the Markov chain of $Z_n$ is proportional to the area of the regions $R_i$ ($i = 1, \ldots, 6$), given by

$$|R_1| = \frac{2}{1 + \alpha \phi} (\alpha \phi - 1 + \phi^{-1} \theta),$$

(35)

$$|R_2| = \frac{2}{1 + \alpha \phi} (1 + \phi^{-2} \theta),$$

(36)

$$|R_3| = \frac{2}{1 + \alpha \phi} (1 + \phi \theta),$$

(37)

$$|R_4| = \frac{2}{1 + \alpha \phi} (1 - \phi \theta),$$

(38)

$$|R_5| = \frac{2}{1 + \alpha \phi} (1 - \phi^{-2} \theta),$$

(39)
Fig. 4. An inverse image of $R_i$s under $\hat{T}$ with $\alpha = 1.9$ and $\theta = 0.1$. $s_2$ implies $\phi_1 = \sqrt{2+\phi}$ and $s_1$ implies $\frac{\phi^{-1}}{\sqrt{2+\phi}}$.

$|R_6| = \frac{2}{1+\alpha\phi}(\alpha\phi - 1 - \phi^{-1}\theta)$.  

We have $\sum_{i=1}^{6} |R_i| = 4$ irrespective of $\alpha, \theta$. Thus, the stationary distribution is $\pi^* = (|R_1|/4, |R_2|/4, \ldots, |R_6|/4)$.

**Proof:** One can easily check that $\pi^* P = \pi^*$ holds. This shows that $\pi^*$ is the stationary distribution of the Markov chain. Computation of $R_i$s is straightforward from its definition but tedious and thus omitted.

Finally we have the following theorem.

**Theorem 3** Suppose $\alpha$ and $\theta$ satisfy the condition (28). Then, the uniform distribution on the invariant set $\Gamma$ is the invariant density of the two dimensional map $\hat{T}$.

**Proof:** By Definition 4, $R_i$s are mutually disjoint. Assume that $r$ follows the uniform distribution on $\Gamma$. Then, the probability of $r$ belonging to $R_i$ is $|R_i|/|\Gamma|$. The map $\hat{T}$ restricted on $R_i$ preserves the measure. From (29) in Lemma 1, we have $R_1 \cup R_2 = \hat{T}^{-1}(R_3 \cup R_6) = \hat{T}^{-1}(R_3) \cup \hat{T}^{-1}(R_6)$. Similarly from (30) and (31), we have $R_3 \cup R_4 = \hat{T}^{-1}(R_2) \cup \hat{T}^{-1}(R_5)$ and $R_5 \cup R_6 = \hat{T}^{-1}(R_1) \cup \hat{T}^{-1}(R_4)$. Then, $\Gamma \equiv \bigcup_{i=1}^{6} R_i = \bigcup_{i=1}^{6} \hat{T}^{-1}(R_i)$. Here (a) follows from Theorem 2. An example of an inverse image of $R_i$s is shown in Fig. 4. By Lemma 3, the probability of $\hat{T}(r) \in R_i$ is $|R_i|/4$. This implies that $\hat{T}(r)$ also follows the uniform distribution on $\Gamma$. This completes the proof.

### 3.1 Evaluation of the MSQE of the GRE

In this section, we briefly analyze the MSQE of the GRE and give an approximation for it. By applying (18) recursively, we have

$$t_{N+1} = \phi^N t_1 - \frac{1}{\sqrt{2+\phi}} \sum_{n=1}^{N} b_n \phi^{N+1-n}. \quad (41)$$

From its definition, we have $t_1 = \frac{\phi}{\sqrt{2+\phi}} x$. Substituting this into (41) and dividing both sides of (41) by $\phi^{N+1}/\sqrt{2+\phi}$, we obtain

$$x - \sum_{n=1}^{N} b_n \phi^{-n} = \sqrt{2+\phi} \phi^{-N-1} t_{N+1}. \quad (42)$$

This equation is equivalent to (7) if $V_{FS} = 2\phi$ because $t_{N+1} = \frac{1}{\sqrt{2+\phi}} (u_N + u_{N+1})$. Therefore from (11) we have
For a rigorous evaluation, we have to evaluate the distribution of \( t_{N+1}(X) \) for a uniformly distributed \( X \in [-\phi, \phi] \). For \( \beta \)-encoders, such a rigorous evaluation can be found in [9, 10]. We cannot directly apply their analysis method because the dynamics of a \( \beta \) encoder is one-dimensional and that of GREs is two-dimensional.

This paper does not provide a rigorous evaluation of the MSQE. Instead we give its reasonable approximation by adopting the following heuristic. In [13], because the exact evaluation of \( \int u_X(x)dx \) is difficult, this term was replaced by \( \int_{(u,v)\in \Gamma} ududv \), where \( \Gamma \) is the invariant set for \( T_{\alpha,\theta} \) in (9).

To evaluate the MSQE, we replaced \( \frac{1}{2\phi} \int |t_{N+1}(x)|^2 dx \) by \( \int_{(s,t)\in \Gamma} t^2 h^*(s,t)dtds \), where \( h^*(s,t) \) is the invariant measure of \( \hat{T} \). We proved that \( h^* \) is a uniform distribution on the invariant set \( \Gamma \). Therefore, we give the following approximation of MSQE.

\[
\text{MSQE} = (2 + \phi)\phi^{-2N-2} \frac{1}{2\phi} \int_{-\phi}^{\phi} \{t_{N+1}(x)\}^2 dx. 
\]

(43)

Define \( s_1 = \phi^{-1}/\sqrt{2 + \phi} \) and \( s_2 = \phi^2/\sqrt{2 + \phi} \). By integrating \( t^2 \) with respect to \( t \) in the region defined by \( R_s \), we have

\[
\int_{(s,t)\in \Gamma} t^2 dtds = \frac{1}{3} \left[ \int_{s_2}^{s_1} f_{\phi}^3(s)ds + \int_{s_1}^{s_2} f_{\phi}^3(s)ds - \int_{-s_2}^{-s_1} f_{\phi}^3(s)ds - \int_{-s_1}^{-s_2} f_{\phi}^3(s)ds \right]. 
\]

(45)

Because the functions in the above integrals are linear, an explicit form of (45) can be obtained by elementary calculations. By substituting the explicit form of (45) and \( |\Gamma| = 4 \) into (44), we obtain

\[
\text{MSQE} = \frac{2 + \phi}{5} \phi^{-2N-2} \left\{ 1 + \frac{2\phi^4}{(1 + \alpha \phi)^2} \left( 5\phi^2 + \frac{\phi^2}{3} (\alpha - \phi)^2 \right) \right\}. 
\]

(46)

Details of the derivation are omitted. The right hand side of (46) is a function of \( \alpha \) and \( \theta \). Define \( d(\alpha, \theta) = \frac{2\phi^4}{(1 + \alpha \phi)^2} \left( 5\phi^2 + \frac{\phi^2}{3} (\alpha - \phi)^2 \right) \). This term takes zero if and only if \( \alpha = \phi \) and \( \theta = 0 \), otherwise it takes positive values. Therefore \( d(\alpha, \theta) \) expresses the increase of MSQE. We finally obtain the approximate SQNR in decibel expression as

\[
\text{SQNR}[\text{dB}] = 10 \log_{10} \frac{\frac{1}{5} V^2_{\text{FS}}}{\text{MSQE}} \\
= 20N \log_{10} \phi + 10 \log_{10} \frac{5\phi^4}{2(2 + \phi)} - 10 \log_{10}(1 + d(\alpha, \theta)) \\
\approx 4.12N + 6.75 - 10 \log_{10}(1 + d(\alpha, \theta)). 
\]

(47)

The last approximation is due to errors in rounding off the decimals. This expression clearly shows that the SQNR is degraded by a positive \( d(\alpha, \theta) \), and it is desirable to make \( d(\alpha, \theta) \) smaller than a prescribed level. The derivation of such an explicit form of the approximated SQNR is one of the contributions of this paper. Numerical simulations show that (47) is a good approximation for any \( \alpha \) and \( \theta \) satisfying (28) if \( N \) is large. The accuracy of this approximation will be analyzed in future research.

3.2 A special case of \( \alpha = \phi \)

In this subsection, we show that the case \( \alpha = \phi \) is special. Equation (15) is simplified to

\[
b_n = Q(\sqrt{2 + \phi} t_n - \theta) 
\]

(48)

Hence, output binary codes are decided solely by \( t_n \). Then, the two-dimensional dynamics becomes

\[
\begin{bmatrix} s_{n+1} \\ t_{n+1} \end{bmatrix} = \begin{bmatrix} -\phi^{-1} s_n \\ \phi t_n \end{bmatrix} - \frac{1}{\sqrt{\phi}} \begin{bmatrix} -1 \\ \phi \end{bmatrix} Q(\sqrt{2 + \phi} t_n - \theta). 
\]

(49)
Therefore the sequence \{t_n\}_{n \geq 0} is governed by one-dimensional dynamics. The lines \( f_0, f_{1+}, f_{2+} \) are all parallel to the \( s \)-axis. Either \( p_{23} \) or \( p_{15} \) of the transition probability matrix \( P \) is zero. Moreover we observe that two of the eigenvalues of \( P \) are zero in this case. The other eigenvalues are explicitly expressed as \( 1, -\phi^{-2}, \) and \( \frac{1}{2\phi} \pm \sqrt{\frac{1}{4\phi^2} - \frac{1 - \phi^2}{\phi^2}} \). In the case of \( \alpha \neq \phi \), we numerically found that \( P \) has six isolated non-zero eigenvalues. We have not found an explicit formula for the eigenvalues of \( P \) for such general cases.

4. Proofs of lemmas

In this section we give the proofs of lemmas and theorems. As a preliminary for the proofs, we give some definitions.

Figure 3 shows the invariant set of \( \hat{T} \). Two linear maps associated with \( \hat{T} \) are defined as [13]

\[
\hat{T}_+(r) = \Lambda r - \frac{1}{\sqrt{2 + \phi}} \begin{bmatrix} 1 \\ \phi \end{bmatrix}, \tag{50}
\]

\[
\hat{T}_-(r) = \Lambda r + \frac{1}{\sqrt{2 + \phi}} \begin{bmatrix} 1 \\ \phi \end{bmatrix}. \tag{51}
\]

Note that linear maps \( \hat{T}_+ \) and \( \hat{T}_- \) have fixed points at \( r = (\phi^{-1}, \phi^2)/\sqrt{2 + \phi} \) and \( r = (-\phi^{-1}, -\phi^2)/\sqrt{2 + \phi} \).

The lines in Fig. 3 are defined as follows: First, we define the following two horizontal lines:

\[
t = \phi^2/\sqrt{2 + \phi},
\]

\[
t = -\phi^2/\sqrt{2 + \phi}. \tag{53}
\]

It is easily verified that if \( r_n \) is above the line \( t = \phi^2/\sqrt{2 + \phi} \) at some time instance \( n \), then \( s_{n+k}, k \geq 0 \) diverges to \( +\infty \). Similarly, if \( r_n \) is below the line \( t = \phi^2/\sqrt{2 + \phi} \) at some \( n \), then \( s_{n+k}, k \geq 0 \) diverges to \( -\infty \). We also define the following two vertical lines:

\[
s = \phi^2/\sqrt{2 + \phi},
\]

\[
s = -\phi^2/\sqrt{2 + \phi}. \tag{55}
\]

There lines are shown to form the boundary of the invariant set of \( \hat{T} \).

In order to identify the invariant set, it is important to take care of the behavior of the set of points very close to the threshold. The threshold is defined by the following line segment.

\[
\ell_0 = \{(s,t) : t = f_0(s), -\phi^2/\sqrt{2 + \phi} \leq s \leq \phi^2/\sqrt{2 + \phi}\} \tag{56}
\]

Then, we define \( \ell_{1+} = \hat{T}_- (\ell_0), \ell_{1-} = \hat{T}_+ (\ell_0), \ell_{2+} = \hat{T}_+ (\ell_{1+}), \) and \( \ell_{2-} = \hat{T}_- (\ell_{1-}). \) Note that we put the positive and negative signs for \( \ell_{1\pm} \) and \( f_{1\pm} \) opposite the signs of the maps \( \hat{T}_{\pm} \). Calculating the image of the line segments, we find that

\[
\ell_{1+} = \{(s,t) : t = f_{1+}(s), -\phi^2/\sqrt{2 + \phi} \leq s \leq \phi^{-1}/\sqrt{2 + \phi}\} \tag{57}
\]

\[
\ell_{1-} = \{(s,t) : t = f_{1-}(s), -\phi^{-1}/\sqrt{2 + \phi} \leq s \leq \phi^2/\sqrt{2 + \phi}\} \tag{58}
\]

\[
\ell_{2+} = \{(s,t) : t = f_{2+}(s), \phi^{-1}/\sqrt{2 + \phi} \leq s \leq \phi^2/\sqrt{2 + \phi}\} \tag{59}
\]

\[
\ell_{2-} = \{(s,t) : t = f_{2-}(s), -\phi^2/\sqrt{2 + \phi} \leq s \leq -\phi^{-1}/\sqrt{2 + \phi}\}. \tag{60}
\]

We observe that a point just below the threshold line segment is mapped by \( \hat{T} \) into a point just below \( \ell_{1+} \) and a point just above the threshold line is mapped by \( \hat{T} \) into a point just above \( \ell_{1-} \). Similarly, a point just below \( \ell_{1+} \) is mapped by \( \hat{T} \) into a point just below \( \ell_{2+} \) and a point just above \( \ell_{1-} \) is mapped by \( \hat{T} \) into a point just above \( \ell_{2-} \).

It is important to note that \( \hat{T}_+ (\ell_{2+}) = \hat{T}_- (\ell_{2-}) \) holds for any \( \alpha \) and any \( \theta \) satisfying (28). We denote this line segment as \( \ell_3 \). One can find that \( \ell_3 \) is given by
These conditions guarantee that lines $R$ have mapped into $s$.

Moreover if (28) is satisfied, we have $f$ is equal to $s$.

Proof of Lemma 1: If Eq. (10) is satisfied, then we have $f_1(-s_2) \leq s_2$, $f_1(s_1) \leq s_2$, $f_1(-s_1) \geq -s_2$, $f_1(s_2) \geq -s_2$.

These conditions assure that $\{u_n\}_{n \geq 0}$ is bounded. Moreover if (28) is satisfied, we have $f_2(-s_1) \leq f_0(s_1)$, $f_2(s_1) \geq f_0(s_1)$.

These conditions guarantee that lines $\ell_2 -$ and $\ell_2 +$ do not cross line $\ell_0$.

As shown in Fig. 4, we have $\hat{T}^{-1}(R_3) \cup \hat{T}^{-1}(R_6) = R_1 \cup R_2$. The left hand side is equal to $\hat{T}^{-1}(R_3 \cup R_6)$. Thus we have $R_3 \cup R_6 = \hat{T}(R_1 \cup R_2)$. Similarly we have $R_1 \cup R_4 = \hat{T}(R_5 \cup R_6)$.

As shown in Fig. 4, the union of the left sub-region of $\hat{T}^{-1}(R_5)$ and the left sub-region of $\hat{T}^{-1}(R_2)$ is equal to $R_4$. The points on line $f_{(-1)}$ in $R_4$ are mapped into the points on line $f_0$. Line $f_{2-}$ is mapped into $f_3$. Similarly, the points on line $f_{(-1)}$ in $R_3$ are mapped to the points on line segment $f_0$ and line $f_{2+}$ is mapped to $f_3$. This shows that $\hat{T}^{-1}(R_2) \cup \hat{T}^{-1}(R_5) = R_3 \cup R_4$ holds. Therefore we have $R_2 \cup R_5 = \hat{T}(R_3 \cup R_4)$.

Proof of Lemma 2: As stated in the proof of Lemma 1, if (28) is satisfied, we have $f_2(-s_1) \leq f_0(s_1)$ and $f_2(s_1) \geq f_0(s_1)$, where $s_1 = \phi^{-1}/\sqrt{2\phi + 1}$. Then, we obtain $p_{ij}$ by the formula

$$p_{ij} = \frac{\hat{T}^{-1}(R_j) \cap R_i}{|R_i|}.$$
where $\hat{T}^{-1}(R_j)$ is the inverse image of $R_j$ under $\hat{T}$. We draw the inverse image of $R_j$s in Fig. 4, where the line segments $t = f_{(-1)+}(s)$ and $t = f_{(-1)-}(s)$ are the inverse image of $t = f_0(s)$ under $\hat{T}_+$ and $\hat{T}_-$. In this figure, $\hat{T}^{-1}(R_2)$ and $\hat{T}^{-1}(R_5)$ are separated into two parts, but one of the separated parts may vanish depending on $\alpha$ and $\theta$. Because the computation is elementary but tedious, we omit the details.

If line segment $\ell_{2-}$ crosses line segment $\ell_0$, the invariant set $\Gamma$ is not determined by the union of the six sub-regions. $\Gamma$ consists of many small regions in this case. This happens also when line segment $\ell_{2+}$ crosses $\ell_0$. See Section 5 for the numerical results of $\Gamma$ of this case.

5. Numerical evaluation

Numerical simulations were performed to verify our results. We computed a long sequence of $u_n$ from a single $x$ according to Definition 1 with $V_{FS} = 2\phi$ and plotted $(u_{n-1}, u_n)$s, which gives the approximate invariant set, where $n = 1, 2, \ldots, 10^5$. The shape of the trajectory of $(u_{n-1}, u_n)$ largely depends on $\alpha$ and $\theta$.

Figures 5(a), (b), and (c) are the case of $\alpha = \phi$. Red lines show the threshold, i.e., $u_{n-1} + \alpha u_n = \theta$. The invariant set is composed of three rectangles in this case. Figures 5(d), (e), and (f) show the case $\alpha < \phi$. We observe that the $(u_{n-1}, u_n)$ are in $\Gamma$, as defined by Theorem 3. Figure 6 shows the case where $\alpha > \phi$. The set of plotted points in Figs. 6(a) and (b) form an invariant set that is exactly the same as in Theorem 3. The points of $(u_{n-1}, u_n)$ are uniformly distributed on $\Gamma$. This matches Theorem 3.

It is worth noting that if condition (28) is not satisfied, the invariant set has a complicated structure. Figures 6, (c), (d), (e), and (f) show different shapes than the union of $R_i$s. In these cases, (10) is satisfied so that $\{u_n\}_{n \geq 1}$ is bounded but (28) is not satisfied. Figure 6(c) clearly shows that the condition $f_{2-}(-s_1) \leq f_0(-s_1)$ stated in the proof of Lemma 1 is violated. We observe that (29) and
(31) hold, but (30) does not hold in this case. Fractal structures appear in Figs. 6(c), (d), (e), and (f). Because such a structure is quite complicated we did not analyze it in this paper.

6. Conclusion
We analyzed the invariant set and invariant density of two-dimensional GRE maps. For parameter $\alpha$ and $\theta$ with $|\theta| + |\alpha - \phi| \leq \phi^{-1}$, the invariant set has been explicitly identified. We have shown that the uniform distribution on the invariant set is the invariant density. This result contributes to the evaluation of MSQE and SQNR.

A rigorous valuation of the MSQE and SQNR of GREs for finite $N$ and $\mu > 0$ and $\rho_1, \rho_2 < 1$ are left for future work.

Acknowledgments
The author would like to thank Professor Yasutada Oohama and Professor Hiroshi Fujisaki for giving valuable comments for this research. This work was supported by JSPS KAKENHI Grant Numbers JP16K00333 and JP19K12156.

References
[1] I. Daubechies, R. DeVore, C. Gunturk, and V. Vaishampayan, “Beta expansions: a new approach to digitally corrected A/D conversion,” IEEE 2002 Int. Sympo. Circ. and Syst., ISCAS 2002, pp. II–II, IEEE, 2002.
[2] I. Daubechies and O. Yilmaz, “Robust and practical analog-to-digital conversion with exponential precision,” IEEE Transactions on Information Theory, vol. 52, no. 8, pp. 3533–3545, 2006.
[3] W. Parry, “On the $\beta$-expansions of real numbers,” Acta Math. Acad. Sci. Hung., vol. 11, pp. 401–416, 1960.
[4] W. Parry, “Representations for real numbers,” Acta Math. Acad. Sci. Hung., vol. 15, pp. 95–105, 1964.
[5] P. Erdös, I. Joó, and V. Komornik, “Characterization of the unique expansions $1 = \sum_{i=1}^{\infty} q^{n_i}$ and related problems,” Bull. Soc. Math. France, vol. 118, pp. 377–390, 1990.
[6] R. Ward, “On robustness properties of beta encoders and golden ratio encoders,” IEEE Transactions on Information Theory, vol. 54, no. 9, pp. 4324–4334, 2008.
[7] T. Kohda, Y. Horio, and K. Aihara, “$\beta$-expansion attractors observed in A/D converters,” Chaos: An Interdisciplinary Journal of Nonlinear Science, vol. 22, no. 4, pp. 047512, 2012.
[8] T. Kohda, Y. Horio, Y. Takahashi, and K. Aihara, “Beta encoders: Symbolic dynamics and electronic implementation,” Int. J. Bifurcation and Chaos, vol. 22, pp. 1230031, 2012.
[9] 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, vol. 6, no. 1, pp. 99–111, January 2015.
[10] K. Shinohara and K. Kobayashi, “Estimation of mean squared errors of non-binary A/D-encoders through Fredholm determinants of piecewise-linear transformations,” NOLTA, pp. 243–258, 4 2018.
[11] H. San, T. Kato, T. Maruyama, K. Aihara, and M. Hotta, “Non-binary pipeline analog-to-digital converter based on $\beta$-expansion,” IEICE Trans. Fundamentals of Electronics, Communications and Computer Sciences, vol. 96, no. 2, pp. 415–421, 2013.
[12] R. Suzuki, T. Maruyama, H. San, K. Aihara, and M. Hotta, “Robust cyclic ADC architecture based on $\beta$-expansion,” IEICE Transactions on Electronics, vol. 96, no. 4, pp. 553–559, 2013.
[13] I. Daubechies, C.S. Gunturk, Y. Wang, and Ö. Yilmaz, “The golden ratio encoder,” IEEE Trans. Inform. Theory, vol. 56, no. 10, pp. 5097–5110, 2010.
[14] K.L. Lin, A. Kemna, and B.J. Hosticka, “Modular Low-Power, High-Speed CMOS Analog-to-digital Converter for Embedded Systems,” Kluwer Academic Publishers, 2003.
[15] F. Blanchini, “Set invariance in control,” Automatica, vol. 35, no. 11, pp. 1747–1767, 1999.