ILLINOIS JOURNAL OF MATHEMATICS
Volume 36, Number 4, Winter 1992

NOTE ON CONTINUITY OF INFORMATION RATE

BY

NAOHUMI MURAKI AND MASANORI OHYA

1. Introduction

The mutual information is a measure describing how much information carried by an input probability measure of an input system X is correctly transmitted to an output system Y through a channel λ. This quantity plays an essential role in communication theory since it is the measure that quantitively expresses efficiency of information transmission [9].

L. Breiman [3] showed that the information rate, i.e., the mutual information per unit time is upper semicontinuous in the vague topology of input probability measures for stationary channels with finite-memory. This result has been used [3] to show that the stationary capacity of such a channel coincides with the ergodic capacity of the channel (conjecture by Khinchin [9]). The continuity property of the information rate has been discussed as a basic property of information function. In this paper, we study this continuity with respect to several topologies in the space of input probability measures. In §2, we briefly review finite-alphabet, discrete-time communication processes and fix terminologies used here. In §3, we explain several topologies for the space of input probability measures and show the continuity of the information rate in the set-wise convergence of the input measures. In §4, we show that the information rate is not continuous in the cylinder-wise convergence of the input measures. In §5, for finite-memory stationary channels, we show the continuity with respect to Ornstein's $\bar{d}$-distance for the input measures. The $\bar{d}$-distance topology is weaker than the set-wise convergence and is stronger than the cylinder-wise convergence (see §6). In §6, we present an example of a non-finite-memory channel for which the information rate is discontinuous in the $\bar{d}$-distance. We give, also in §6, some remarks for the results obtained in §4 and §5.

2. Preliminaries

In information theory, a finite-alphabet and discrete-time communication process is mathematically formulated as follows [9]: Let A and B be finite

Received June 20, 1989.
1991 Mathematics Subject Classification. Primary 94A40; Secondary 94A15.

alphabets, that is, non empty finite sets having more than two elements. Denote by Z the set of all integers. We take the input and output systems X, Y as the spaces $X = A^Z$ and $Y = B^Z$ of doubly infinite sequences of letters in A, B, respectively. The joint system $X \times Y = A^Z \times B^Z$, denoted by C^Z, is the space of doubly infinite sequences of letters in the alphabet $C = A \times B$. $\mathscr{F}_X$ is the σ-field of X generated from the field $\mathscr{M}_X$ consisting of all cylinder sets on X. The set X becomes a compact dynamical system with the product topology of the discrete topology for the alphabet A and with the shift T_A. Here the shift is defined as

$$T_A((x_n)_{n\in Z}) = (x_{n+1})_{n\in Z} \quad \text{for } x = (x_n)_{n\in Z} \in A^Z.$$

We denote by $\mathscr{P}(X)$ and $\mathscr{P}_T(X)$, the set of all probability measures on X and the set of all stationary probability measures on X with respect to the shift T_A, respectively. $\mathscr{F}_Y, \mathscr{M}_Y, T_B, \mathscr{P}(Y), \mathscr{P}_T(Y), \ldots$ are defined. We use the same symbol T for T_A, T_B, T_C when no confusion occurs. A *stationary channel* λ from X to Y is a mapping from X to $\mathscr{P}(Y)$ satisfying:

(1) $\lambda(x)(F)$ is a measurable function of $x \in X$ for any fixed $F \in \mathscr{F}_Y$,
(2) $\lambda(T_A x)(F) = \lambda(x)(T_B^{-1}F)$ for any $x \in X$ and any $F \in \mathscr{F}_Y$.

We often use the notation $\lambda(x, F)$ to denote $\lambda(x)(F)$. This channel λ induces the following two transformations of probability measures $\mu \to \lambda(\mu)$ and $\mu \to (\delta \otimes \lambda)(\mu)$. The first is from $\mathscr{P}(X)$ to $\mathscr{P}(Y)$ and the second is from $\mathscr{P}(X)$ to $\mathscr{P}(X \times Y)$ defined by

$$\lambda(\mu)(F) \equiv \int_X \lambda(x)(F)\, d\mu(x) \quad \text{for } F \in \mathscr{F}_Y,$$

$$((\delta \otimes \lambda)(\mu))(G) \equiv \int_X (\delta(x) \otimes \lambda(x))(G)\, d\mu(x) \quad \text{for } G \in \mathscr{F}_X \otimes \mathscr{F}_Y,$$

where $\delta(x)$ is the Dirac measure with respect to a point x. Note that if $\mu \in \mathscr{P}_T(X)$, then $\lambda(\mu) \in \mathscr{P}_T(Y)$ and $(\delta \otimes \lambda)(\mu) \in \mathscr{P}_T(X \times Y)$. For a pair of integers s and t with $s \leqq t$, let $[s, t]$ be the interval in Z consisting of all integer $n \in Z$ such that $s \leqq n \leqq t$. We canonically identify a finite sequence $\alpha = (\alpha_s, \ldots, \alpha_t)$ in $A^{[s,t]}$ with a uniquely determined measurable set

$$\{x \in A^Z | x_n = \alpha_n, n \in [s, t]\} \in \mathscr{F}_x.$$

Any element of $A^{[s,t]}$ is called a *message*. We often denote $A^{[0,n-1]}$ by A^n for short. The *entropy rate* $\tilde{S}(\mu)$ of a stationary probability measure μ and the information rate $\tilde{I}(\mu; \lambda)$ of a stationary probability measure μ with

respect to a stationary channel λ are defined as

$$\tilde{S}(\mu) = \lim_{n\to\infty} \frac{1}{n} S_n(\mu), \qquad \tilde{I}(\mu;\lambda) = \lim_{n\to\infty} \frac{1}{n} I_n(\mu;\lambda).$$

Here $S_n(\mu)$ and $I_n(\mu;\lambda)$ are the entropy and the mutual information for the finite spaces $A^{[0,n-1]}$ and $B^{[0,n-1]}$ generated by messages of length n, which are given by

$$S_n(\mu) \equiv -\sum_{\alpha\in A^{[0,n-1]}} \mu(\alpha)\log\mu(\alpha),$$

$$I_n(\mu;\lambda) \equiv \sum_{(\alpha,\beta)\in A^{[0,n-1]}\times B^{[0,n-1]}} \mu''(\alpha,\beta)\log\frac{\mu''(\alpha,\beta)}{\mu(\alpha)\mu'(\beta)},$$

where $\mu' = \lambda(\mu)$ and $\mu'' = (\delta\otimes\lambda)(\mu)$. The entropy rate $\tilde{s}(\mu)$ represents the averaged information generated by a stationary probability measure μ, and the information rate $\tilde{I}(\mu;\lambda)$ does the averaged information transmitted from μ to $\lambda(\mu)$ by a stationary channel λ. Using the Radon-Nikodym derivative $d\mu''/d(\mu\otimes\mu')$, the *mutual information* $I(\mu;\lambda)$ between μ and μ' is defined [14] by

$$I(\mu;\lambda) \equiv \int_{X\times Y}\left(\frac{d\mu''}{d\mu\otimes\mu'}\log\frac{d\mu''}{d\mu\otimes\mu'}\right)d\mu\otimes\mu'$$

when μ'' is absolutely continuous with respect to $\mu\otimes\mu'$, and otherwise $I(\mu;\lambda)\equiv\infty$.

The continuity of the information rate $\tilde{I}(\cdot;\lambda)$ as a function of an input probability measure has been studied by Breiman as follows:

THEOREM 2.1 [3]. *The information rate $\tilde{I}(\cdot;\lambda)$ of a finite-memory stationary channel λ is upper semicontinuous with respect to the vague topology in the space of stationary probability measures.*

The finite-memory channel will be precisely defined in §5. Since the mutual information $I(\cdot;\lambda)$ is lower semicontinuous in general, it is interesting for us to ask under which conditions the information rate $\tilde{I}(\mu;\lambda)$ is continuous with respect to μ. In this note, we mainly study this question. Throughout this paper we use the notation $S(p)$ or $S(p_1,\ldots,p_n)$ to denote the entropy of a probability distribution $p=(p_1,\ldots,p_n)$: $S(p)=\sum_{i=1}^n\eta(p_i)$, where $\eta(t) = -t\log t$ for $t>0$ and $\eta(0)=0$. The symbol e is used to denote the base of the natural logarithm. The base of the logarithm is always assumed to be 2.

3. Several topologies in the space of input probability measures

In order to investigate the continuity of the function $\tilde{I}(\cdot;\lambda)$, we introduce several topologies for the space of stationary probability measures. There exist three natural topologies as follows:

(T1) Norm-topology defined by the total variation norm of real valued measures:

$$\mu_j \to \mu \quad \Leftrightarrow \quad \|\mu_j - \mu\| \to 0;$$

(T2) Topology defined by the "set-wise" convergence

$$\mu_j \to \mu \quad \Leftrightarrow \quad \mu_j(E) \to \mu(E) \quad (\forall E \in \mathscr{F}_X);$$

(T3) Topology defined by the "cylinder-wise" convergence

$$\mu_j \to \mu \quad \Leftrightarrow \quad \mu_j(E) \to \mu(E) \quad (\forall E \in \mathscr{M}_X).$$

Obviously T1 is stronger than T2, and T2 is stronger than T3. The topology T3 is often called the vague topology. We shall later show that the information rate is continuous in topology T1 and in topology T2 for every stationary channel and is discontinuous in topology T3 for almost stationary channels. Additionally, we also study the continuity in the following topology, i.e., Ornstein's $\bar{d}$-distance [11]:

(T4) Metric topology defined by the $\bar{d}$-distance:

$$\bar{d}(\mu,\mu') = \sup_n \inf_{\omega \in \mathscr{P}_n} \int_{A^{[0,n-1]} \times A^{[0,n-1]}} \frac{1}{n} \sum_{i=0}^{n-1} d(\alpha_i, \alpha_i')\, d\omega(\alpha, \alpha')$$

where $\mathscr{P}_n = \mathscr{P}_n(\mu,\mu')$ is the set of all probability measures ω on $A^{[0,n-1]} \times A^{[0,n-1]}$ such that the left marginal and the right marginal of ω coincide with the restrictions $\mu \restriction A^{[0,n-1]}$ and $\mu' \restriction A^{[0,n-1]}$, respectively. The $\bar{d}$-distance was used by Ornstein to prove his famous isomorphism theorem for Bernoulli shifts [10], [15], and it has been applied to several aspects in information theory [6], [7]. This $\bar{d}$-distance topology T4 is strictly weaker than the topology T2, and is strictly stronger than topology T3 (see §6). Thus the $\bar{d}$-distance fills the gap between the set-wise convergence and the cylinder-wise convergence. The information rate is not always continuous in the $\bar{d}$-distance (see §6). As will be shown in §5, however, it is continuous in the $\bar{d}$-distance for a special class of channels, i.e., the finite-memory channels.

It is easy to show the continuity of the function $\tilde{I}(\cdot;\lambda)$ for the total variation distance T1 as follows. We can extend the affine functional $\tilde{I}(\cdot;\lambda)$ on $\mathscr{P}_T(X)$ to a linear functional on the Banach space $M_T(X)$ (the space of all stationary real-valued measures on X) for which we use the same

notation $\tilde{I}(\cdot;\lambda)$ as above, and it is defined by

$$\tilde{I}(0;\lambda) \equiv 0,$$

$$\tilde{I}(\mu;\lambda) \equiv \|\mu\|\tilde{I}\left(\frac{\mu}{\|\mu\|};\lambda\right) \quad (\mu \geqq 0, \mu \neq 0),$$

$$\tilde{I}(\mu;\lambda) \equiv \tilde{I}(\mu^+;\lambda) - \tilde{I}(\mu^-;\lambda) \quad (\mu \in M_T(X)),$$

where μ^+ and μ^- are the positive and negative components in the Jordan decomposition of μ, respectively. This definition is consistent. Then it is easily shown that

$$C_s \equiv \sup_{\mu \in \mathscr{P}_T(X)} \tilde{I}(\mu;\lambda) = \sup_{\mu \in M_T(X),\, \|\mu\|=1} \tilde{I}(\mu;\lambda) = \|\tilde{I}(\cdot;\lambda)\|.$$

So the linear functional $\tilde{I}(\cdot;\lambda)$ is bounded and its norm is just C_S, where C_S is called the stationary capacity of a stationary channel λ. So the functional is continuous in the total variation distance.

The continuity of the information rate $\tilde{I}(\cdot;\lambda)$ for the set-wise convergence T2 is also easily shown by using the Parthasarathy-Umegaki ergodic decomposition theorem of the information rate [13], [16], as follows. Since any bounded measurable function is approximated by simple functions in the uniform norm, the topology T2 equals the topology T2′ defined by

$$\text{(T2}') \qquad \mu_j \to \mu \quad \Leftrightarrow \quad \int_X f d\mu_j \to \int_X f d\mu \quad (\forall f \in B(X)),$$

where $B(X)$ is the Banach space of all real-valued bounded measurable functions on X. By the Parthasarathy-Umegaki ergodic decomposition theorem of the information rate, there exists a bounded measurable function h on X such that

$$\tilde{I}(\mu;\lambda) = \int_X h(x)\, d\mu(x) \quad (\forall \mu \in \mathscr{P}_T(X)).$$

Therefore we obtained the continuity of the information rate $\tilde{I}(\cdot;\lambda)$ in the topology T2.

In the following two sections, we study the continuity of the information rate in the cylinder-wise convergence T3 and in the $\bar{d}$-distance T4.

4. Discontinuity of information rate in the cylinder-wise convergence

In this section we show that the information rate is discontinuous in the cylinder-wise convergence for almost all channels except those with stationary capacity 0.

THEOREM 4.1. *The information rate* $\tilde{I}(\cdot;\lambda)$ *of a stationary channel* λ *is discontinuous for the vague topology in* $\mathscr{P}_T(X)$ *unless its stationary capacity is* 0.

To prove Theorem 4.1, we prepare several notions. It is easily checked by the Stone-Weierstrass theorem and by the continuity of the characteristic functions of cylinder sets that the topology T3 equals the vague topology T3′ defined by

$$\text{(T3}') \qquad \mu_j \to \mu \quad \Leftrightarrow \quad \int_X f d\mu_j \to \int_X f d\mu \quad (\forall f \in C(X)),$$

where $C(X)$ is the Banach space of all real-valued continuous functions on X. In a compact dynamical system (X, T), a point $x \in X$ is called a *quasi-regular point* if

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} f(T^i x)$$

exists for any $f \in C(X)$. For any quasi-regular point x of (X, T), there exists a unique stationary probability measure ν_x such that

$$\int_X f d\nu_x = \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} f(T^i x) \quad (\forall f \in C(X))$$

in virtue of Riesz's theorem. A point $x \in X$ is called a *regular point* if it is a quasi-regular point and the associated stationary probability measure ν_x is ergodic. Let us denote the set of all quasi-regular points on (X, T) by Q and the set of all regular points on (X, T) by R. The Parthasarathy-Umegaki ergodic decomposition theorem of information rate [13, 17] asserts that

$$\tilde{I}(\mu;\lambda) = \int_R \tilde{I}(\nu_x;\lambda)\, d\mu(x) \quad (\forall \mu \in \mathscr{P}_T(X)),$$

where ν_x is the ergodic measure associated to each regular point $x \in R$. When the functional $\tilde{I}(\cdot;\lambda)$ is not identically 0, there exists some stationary probability measure μ such that $\tilde{I}(\mu;\lambda) > 0$, and hence some regular point $x \in R$ such that $\tilde{I}(\nu_x;\lambda) > 0$.

Therefore it is sufficient to prove the following two lemmas for the proof of Theorem 4.1.

LEMMA 4.2. *If* $x \in X$ *is periodic in the sense that* $|\{T^n x;\, n \in N\}| = P < \infty$, *then* $\tilde{s}(\nu_x) = 0$ *and* $\tilde{I}(\nu_x;\lambda) = 0$.

LEMMA 4.3. *For any quasi-regular point $x \in Q$, there exists a sequence $(y^n)_{n \in N}$ of periodic points y^n so that ν_{y^n} converges to ν_x in the vague topology.*

Proof of Lemma 4.2. Since $\tilde{S}(\nu_x)$, $\tilde{S}(\lambda(\nu_x))$ and $\tilde{S}((\delta \otimes \lambda)(\nu_x))$ are finite, we can decompose $\tilde{I}(\nu_x; \lambda)$ as $\tilde{I}(\nu_x; \lambda) = \tilde{S}(\nu_x) + \tilde{S}(\lambda(\nu_x)) - \tilde{S}((\delta \otimes \lambda)(\nu_x))$. Let us compute these entropies.

Calculation of $\tilde{S}(\nu_x)$. Since the period of x is P, for any message $\alpha \in A^{[0, n-1]}$ with sufficiently large length $n \geqq P$, we have

$$\nu_x(\alpha) = \frac{1}{P} \sum_{i=1}^{P} \delta(T^i x)(\alpha) = \begin{cases} \frac{1}{P} & (\exists i \in \{1, 2, \ldots, P\}, T^i x \in \alpha) \\ 0 & (\text{otherwise}). \end{cases}$$

So we have

$$S_n(\nu_x) = \sum_{\alpha} \eta(\nu_x(\alpha)) = P\eta\left(\frac{1}{P}\right) = \log P.$$

Hence we get

$$\tilde{S}(\nu_x) = \lim_{n \to \infty} \frac{S_n(\nu_x)}{n} = 0.$$

Calculation of $S_n((\delta \otimes \lambda)(\nu_x))$. $(\delta \otimes \lambda)(\nu_x)$ is written as

$$(\delta \otimes \lambda)(\nu_x) = \frac{1}{P} \sum_{i=1}^{P} (\delta \otimes \lambda)(\delta(T^i x)).$$

For any message $(\alpha, \beta) \in X^{[0, n-1]} \times Y^{[0, n-1]}$ with $n \geqq P$, we have

$$\begin{aligned} ((\delta \otimes \lambda)(\delta(T^i x)))(\alpha, \beta) &= (\delta(T^i x) \otimes \lambda(T^i x))(\alpha, \beta) \\ &= \begin{cases} \lambda(T^i x)(\beta) & (T^i x \in \alpha) \\ 0 & (\text{otherwise}). \end{cases} \end{aligned}$$

We therefore obtain

$$((\delta \otimes \lambda)(\nu_x))(\alpha, \beta) = \frac{1}{P} \sum_{i=1}^{P} \delta(T^i x, \alpha) \lambda(T^i x, \beta)$$

$$= \begin{cases} \frac{1}{P} \lambda(T^i x)(\beta) & (\exists i \in \{1, 2, \ldots, P\}, T^i x \in \alpha) \\ 0 & (\text{otherwise}), \end{cases}$$

which implies

$$S_n((\delta \otimes \lambda)(\nu_x)) = \sum_{\alpha, \beta} \eta(((\delta \otimes \lambda)(\nu_x))(\alpha, \beta))$$

$$= \sum_{\beta} \sum_{i=1}^{P} \eta\left(\frac{\lambda(T^i x)(\beta)}{P} \right).$$

$$= \sum_{\beta} \sum_{i=1}^{P} \left\{ \frac{\log P}{P} \lambda(T^i x)(\beta) + \frac{1}{P} \eta(\lambda(T^i x)(\beta)) \right\}$$

$$= \sum_{i=1}^{P} \left\{ \frac{\log P}{P} + \frac{1}{P} S_n(\lambda(T^i x)) \right\}$$

$$= \log P + \frac{1}{P} \sum_{i=1}^{P} S_n(\lambda(T^i x)).$$

Calculation of $\tilde{I}(\nu_x; \lambda)$: Since $S_n(\lambda(\nu_x))$ is written as

$$S_n(\lambda(\nu_x)) = S_n\left(\frac{1}{P} \sum_{i=1}^{P} \lambda(T^i x) \right) = \frac{1}{P} \sum_{i=1}^{P} S_n(\lambda(T^i x)) + C_n$$

with a bounded sequence C_n, we have

$$S_n(\lambda(\nu_x)) = (S_n((\delta \otimes \lambda)(\nu_x)) - \log P) + C_n. \tag{4.1}$$

By taking the limit as $n \to \infty$ in (4.1) $\times$ $(1/n)$, we get $\tilde{S}(\lambda(\nu_x)) = \tilde{S}((\delta \otimes \lambda)(\nu_x))$. Therefore we have

$$\tilde{I}(\nu_x; \lambda) = \tilde{S}(\nu_x) + (\tilde{S}(\lambda(\nu_x)) - \tilde{S}((\delta \otimes \lambda)(\nu_x)) = 0. \qquad \square$$

Proof of Lemma 4.3. The topology of X is metrizable by a distance

$$d(x, y) \equiv \sum_{m \in Z} \frac{d_m(x, y)}{2^{|m|}} \quad (x, y \in X),$$

where

$$d_m(x, y) \equiv \begin{cases} 1 & (x_m \neq y_m) \\ 0 & (x_m = y_m). \end{cases}$$

For a given quasi-regular point x, we define a sequence $(y^n)_{n \in N}$ of periodic points y^n by

$$y^n \equiv (y_k^n)_{k \in Z}, \quad y_k^n \equiv x_{(k \bmod n)},$$

where the expression "$k \bmod n$" means the only integer $j \in \{0, 1, \ldots, n-1\}$ such that $j \equiv k \pmod{n}$. Then, for any i such as $0 \leqq i \leqq n-1$, we have

$$\begin{aligned} d(T^i x, T^i y^n) &= \sum_{m \in Z} \frac{d_m(T^i x, T^i y^n)}{2^{|m|}} \\ &\leqq \sum_{m \leqq -i} \frac{1}{2^{|m|}} + \sum_{n-i \leqq m} \frac{1}{2^{|m|}} \\ &= \frac{2}{2^i} + \frac{2}{2^{n-i}}. \end{aligned}$$

For any fixed continuous function $f \in C(X)$ and any $\varepsilon > 0$, there exists $\delta > 0$ such as $|f(x) - f(y)| \leqq \varepsilon$ $(d(x, y) \leqq \delta)$, because of the compactness of X. Let M be an integer such as $4/2^M \leqq \delta$, and n be any sufficiently large integer such as $n \geqq 2M$. Then we have

$$d(T^i x, T^i y^n) \leqq \frac{2}{2^i} + \frac{2}{2^{n-i}} \leqq \frac{4}{2^M} \leqq \delta \quad \text{for } M \leqq i \leqq n - M,$$

and hence

$$|f(T^i x) - f(T^i y^n)| \leqq \varepsilon \quad \text{for } M \leqq i \leqq n - M,$$

by which we can evaluate the difference

$$\left| \frac{1}{n} \sum_{i=1}^{n} f(T^i x) - \int_X f d\nu_y \right|$$

for sufficiently large n:

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{n} f(T^i x) - \int_X f d\nu_{y^n}\right| \\
&\quad = \left|\frac{1}{n}\sum_{i=1}^{n} f(T^i x) - \frac{1}{n}\sum_{i=1}^{n} f(T^i y^n)\right| \\
&\quad \leqq \frac{1}{n}\sum_{i=1}^{n} |f(T^i x) - f(T^i y^n)| \\
&\quad = \frac{1}{n}\left\{\sum_{i=M}^{n-M} |f(T^i x) - f(T^i y^n)| + \sum_{i=1}^{M-1} |f(T^i x) - f(T^i y^n)|\right. \\
&\qquad\qquad \left. + \sum_{i=n-M+1}^{n} |f(T^i x) - f(T^i y^n)|\right\} \\
&\quad \leqq \frac{1}{n}\{(n - 2M + 1)\varepsilon + 2(M - 1)\|f\| + 2M\|f\|\} \\
&\quad \leqq 2\varepsilon.
\end{aligned}$$

On the other hand, since $(1/n)\sum_{i=1}^{n} f(T^i x)$ converges to $\int_X f d\nu_x$, we get

$$\int_X f d\nu_{y^n} \to \int_X f d\nu_x \quad (\forall f \in C(X)).$$

That is, the sequence of stationary probability measures ν_{y^n} converges to the stationary probability measure ν_x in the vague topology. □

So we obtain the discontinuity of the information rate in the cylinder-wise convergence. Furthermore we obtain the following as a consequence of Lemma 4.2 and Lemma 4.3.

THEOREM 4.4. *The set of zero points of the information rate function $\tilde{I}(\cdot; \lambda)$ is dense in $\mathscr{P}_T(X)$ in the vague topology.*

Proof. $\mathscr{P}_T(X)$ is the vague closure of the convex hull of the set of all ergodic probability measures on X. Any ergodic probbility measure μ is represented by some regular point $x \in R$ such that $\mu = \nu_x$ [12]. Therefore the set of all stationary probability measures having finite supports is dense in $\mathscr{P}_T(X)$ with respect to the vague topology according to Lemma 4.3. From Lemma 4.2, the set of zero points of $\tilde{I}(\cdot; \lambda)$ is dense in $\mathscr{P}_T(X)$ in the vague topology. □

5. Continuity of information rate in the $\bar{d}$-distance

As shown in the preceding two sections, the information rate is continuous in the set-wise convergence T2 and discontinuous in the cylinder-wise convergence T3 for almost all channels. So it is a natural question to ask whether the information rate is continuous or not in the Ornstein $\bar{d}$-distance defining the topology between the set-wise convergence topology and the cylinder-wise convergence topology. An answer to this question is that the information rate is continuous in the $\bar{d}$-distance for finite-memory channels (Theorem 5.1). Here, a stationary channel λ from A^Z to B^Z is said to have *finite-memory*, if there exists an integer $m \geqq 0$ such that

(1) for any pair of integers $i \leqq j$, any $\beta \in B^{[i,j]}$ and any $x, x' \in A^Z$,

$$x_k = x'_k (i - m \leqq k \leqq j) \quad \text{implies} \quad \lambda(x, \beta) = \lambda(x', \beta),$$

and

(2) for any integers $n \leqq r \leqq s \leqq t$,

$$s - r > m \quad \text{implies} \quad \lambda(x, \beta \cap \beta') = \lambda(x, \beta)\lambda(x, \beta')$$

for all $x \in A^Z$, all $\beta \in B^{[n,r]}$ and all $\beta' \in B^{[s,t]}$. An example of a non-finite-memory stationary channel for which the information rate is discontinuous in the $\bar{d}$-distance will be given in §6. In this section we prove the following.

THEOREM 5.1. *For a finite-memory stationary channel λ, the information rate $\tilde{I}(\cdot; \lambda)$ is uniformly continuous in the Ornstein $\bar{d}$-distance.*

Theorem 5.1 is essentially reduced to the continuity of the entropy rate $\tilde{S}(\mu)$ in the $\bar{d}$-distance (Theorem 5.2) and to the $\bar{d}$-continuity of the transformation of measures associated to a finite-memory stationary channel (Theorem 5.4). The $\bar{d}$-continuity of the entropy rate has been mentioned in [8, 11], but its proof has not appeared yet. So we give a proof of the continuity of the entropy rate in the $\bar{d}$-distance for completeness.

THEOREM 5.2. *The entropy rate is uniformly continuous in the Ornstein $\bar{d}$-distance.*

This theorem is a consequence of the following lemma 5.3. Let $d(\alpha, \alpha')$ be the Hamming distance between two finite sequences $\alpha, \alpha' \in A^{[0,n-1]}$, i.e.,

$$d(\alpha, \alpha') = \sum_{i=0}^{n-1} d(\alpha_i, \alpha'_i), \quad d(\alpha_i, \alpha'_i) = \begin{cases} 0 & (\alpha_i = \alpha'_i), \\ 1 & (\alpha_i \neq \alpha'_i). \end{cases}$$

For a probability measure ω on $X \times X'$ where $X = X' = A^n$, denote by p and p' the marginals of ω with respect to X and X', respectively. For any integer $d \geqq 0$, denote by Δ_d or Δ the set $\{(\alpha, \alpha') \in X \times X';\ d(\alpha, \alpha') \leqq d\}$.

LEMMA 5.3. *There exists an integer $L \geqq 0$ such that the inequality*

$$\left| \frac{S(p)}{n} - \frac{S(\omega)}{n} \right| \leqq \frac{6}{n} + 6h \log|A| + (\log|A| + 4)\frac{d}{n} + \eta\left(\frac{d}{n}\right)$$

holds for any probability measure ω on $X \times X'$, any integer d with $L \leqq d \leqq (n+1)/2$, and any real number h with $\omega(\Delta_d^C) \leqq h < 1/e$.

To prove Lemma 5.3, we use the following fundamental inequalities (5.1) and (5.2). Let $p_1, \ldots, p_N$ be positive real numbers such that $0 < \sum_{i=1}^N p_i \leqq 1$. Then the following hold:

(1) For $\sum_{i=1}^N p_i \leqq h \leqq 1/e$, we have

$$\sum_{i=1}^{N} \eta(p_i) \leqq h \log N + \eta(h) \tag{5.1}$$

(2) For $P = \sum_{i=1}^N p_i$, we have

$$\left| S\left(\frac{p_1}{P}, \ldots, \frac{p_N}{P}\right) - \sum_{i=1}^{N} \eta(p_i) \right| \leqq (1-P)\log N + \eta(P) \tag{5.2}$$

Proof of Lemma 5.3. Let $\bar{A}, \bar{B}, \bar{C}, \bar{D}, \bar{E}$ be finite families of subsets of $X \times X'$ given by

$$\bar{A} = \{\{\alpha\} \times A^n | \alpha \in A^n\},$$
$$\bar{B} = \{\{\alpha\} \times A^n \cap \Delta | \alpha \in A^n\} \cup \{\{\alpha\} \times A^n \cap \Delta^C | \alpha \in A^n\},$$
$$\bar{C} = \{\{\alpha\} \times A^n \cap \Delta | \alpha \in A^n\},$$
$$\bar{D} = \{\{(\alpha, \alpha')\} | (\alpha, \alpha') \in \Delta\},$$
$$\bar{E} = \{\{(\alpha, \alpha')\} | (\alpha, \alpha') \in A^n \times A^n\},$$

where $\{\alpha\}$ (resp. $\{(\alpha, \alpha')\}$) is the set of single element α (resp. (α, α').) Let $S(\bar{F}, \omega)$ be a function of a finite family $\bar{F}$ of subsets of $X \times X'$ and a probability measure ω on $X \times X'$, given by

$$S(\bar{F}, \omega) = \sum_{G \in \bar{F}} \eta(\omega(G)).$$

Let $\bar{\omega}$ be a probability measure on $X \times X'$ given by $\bar{\omega}(\alpha, \alpha') = \omega(\alpha, \alpha')/\omega(\Delta)$ for $(\alpha, \alpha') \in \Delta$ and $\bar{\omega}(\alpha, \alpha') = 0$ for $(\alpha, \alpha') \notin \Delta$. Then we have an inequality

$$|S(p) - S(\omega)| = |S(\bar{A}, \omega) - S(\bar{E}, \omega)| \leqq D_1 + D_2 + D_3 + D_4 + D_5 + D_6 \tag{5.3}$$

where

$$D_1 = |S(\bar{A}, \omega) - S(\bar{B}, \omega)|, \qquad D_2 = |S(\bar{B}, \omega) - S(\bar{C}, \omega)|,$$
$$D_3 = |S(\bar{C}, \omega) - S(\bar{C}, \bar{\omega})|, \qquad D_4 = |S(\bar{C}, \bar{\omega}) - S(\bar{D}, \bar{\omega})|,$$
$$D_5 = |S(\bar{D}, \bar{\omega}) - S(\bar{D}, \omega)|, \qquad D_6 = |S(\bar{D}, \omega) - S(\bar{E}, \omega)|.$$

Let us evaluate each term D_i of the inequality (5.3). Since

$$S(\bar{A}, \omega) \leqq S(\bar{B}, \omega) = S(\bar{A}, \omega) + \sum_{G \in \bar{A}} \omega(G) S\left(\frac{\omega(G \cap \Delta)}{\omega(G)}, \frac{\omega(G \cap \Delta^C)}{\omega(G)} \right)$$
$$\leqq S(\bar{A}, \omega) + 1,$$

we have $D_1 \leqq 1$. Using (5.1) we have

$$D_2 = |S(\bar{B}, \omega) - S(\bar{C}, \omega)| = \sum_{\alpha \in A^n} \eta(\omega(\{\alpha\} \times A^n \cap \Delta^c))$$
$$\leqq \omega(\Delta^c) \log |A|^n + \eta(\omega(\Delta^C)) \leqq h \log |A|^n + 1$$

and

$$D_6 = |S(\bar{D}, \omega) - S(\bar{E}, \omega)| \leqq \omega(\Delta^C) \log |A|^{2n} + \eta(\omega(\Delta^c)) \leqq h \log |A|^{2n} + 1.$$

Using (5.2), we have

$$D_3 = |S(\bar{C}, \omega) - S(\bar{C}, \bar{\omega})| \leqq \omega(\Delta^C) \log |A|^n + \eta(\omega(\Delta)) \leqq h \log |A|^n + 1$$

and

$$D_5 = |S(\bar{D}, \bar{\omega}) - S(\bar{D}, \omega)| \leqq \omega(\Delta^c) \log |A|^{2n} + \eta(\omega(\Delta)) \leqq h \log |A|^{2n} + 1.$$

Let us evaluate the term D_4. We have

$$S(\overline{C},\overline{\omega}) = \sum_{G\in\overline{C}} \eta(\overline{\omega}(G)) \leqq \sum_{G\in\overline{D}} \eta(\overline{\omega}(G)) = \sum_{\alpha\in A^n} \sum_{\alpha'\in\Delta(\alpha)} \eta(\overline{\omega}(\alpha,\alpha'))$$

$$= S(\overline{C},\overline{\omega}) + \sum_{\alpha\in A^n} \overline{\omega}(\{\alpha\}\times A^n\cap\Delta) \sum_{\alpha'\in\Delta(\alpha)} \eta\left(\frac{\overline{\omega}(\alpha,\alpha')}{\overline{\omega}(\{\alpha\}\times A^n\cap\Delta)}\right)$$

$$\leqq S(\overline{C},\overline{\omega}) + \log|\Delta(\alpha)|, \tag{5.4}$$

$$|\Delta(\alpha)| = |\{\alpha' | d(\alpha,\alpha') \leqq d\}| = \sum_{k=0}^{d} \binom{n}{k} (|A|-1)^k \leqq (d+1)\binom{n}{d}|A|^d \tag{5.5}$$

for $d \leqq (n+1)/2$. By the Stirling formula $n! \sim \sqrt{2\pi n}\, n^n e^{-n}$, there exists an integer L such that

$$\sqrt{2\pi d}\, d^d e^{-d}/d! < 2 \quad (\forall d \geqq L).$$

Then, for $L \leqq d \leqq (n+1)/2$, we get

$$\binom{n}{d} \leqq \frac{n^d}{d!} \leqq 2\left(\frac{n}{d}\right)^d \frac{e^d}{\sqrt{2\pi d}} \leqq 2\left(\frac{n}{d}\right)^d e^d. \tag{5.6}$$

From (5.4), (5.5), (5.6), we get

$$D_4 \leqq \log|\Delta(\alpha)| \leqq \log\left\{2(d+1)\left(\frac{n}{d}\right)^d |A|^d e^d\right\}$$

$$\leqq 1 + (\log|A| + 4)d + d\log\frac{n}{d}.$$

Combining each evaluation of D_i, we get the desired inequality. □

Proof of Theorem 5.2. For any $\varepsilon > 0$, there exists a natural number q such that

$$\frac{\log|A| + 4}{q} + \eta\left(\frac{1}{q}\right) \leqq \varepsilon.$$

For a positive integer n, let d, r be a pair of intergers such that

$$n = d\cdot q + r \quad (0 \leqq r \leqq q-1).$$

Then, for sufficiently large $n \geqq N_0$, we have $d \geq L$ and $6/n \leqq \varepsilon$, where L is the same number as in Lemma 5.3. Let $\delta > 0$ be a sufficiently small number

δ such that

$$6(q+1)\delta \log|A| \leqq \varepsilon.$$

Let μ, μ' be probability measures such that $\bar{d}(\mu, \mu') < \delta$. Then, for any $n \geqq N_0$, there exists $\omega \in \mathscr{P}_n$ such that

$$\sum_{(\alpha,\alpha')} \frac{1}{n} \sum_{i=0}^{n-1} d(\alpha_i, \alpha_i')\omega(\alpha, \alpha') < \delta.$$

For $\Delta = \{(\alpha, \alpha') \in A^n \times A^n | d(\alpha, \alpha') \leqq d\}$, we have

$$\delta > \sum_{(\alpha,\alpha')} \frac{1}{n} \sum_{i=0}^{n-1} d(\alpha_i, \alpha_i')\omega(\alpha, \alpha') \geqq \sum_{d(\alpha,\alpha')>d} \frac{d}{n}\omega(\alpha, \alpha') = \frac{d}{n}\omega(\Delta^C),$$

and hence

$$\omega(\Delta^C) < \frac{n}{d}\delta \leqq (q+1)\delta.$$

By Lemma 5.3, we have

$$\left|\frac{S(\mu \upharpoonright A^n)}{n} - \frac{S(\omega)}{n}\right| \leqq \frac{6}{n} + 6(q+1)\delta \log|A| + (\log|A| + 4)\frac{d}{n} + \eta\left(\frac{d}{n}\right) \leqq 3\varepsilon,$$

and hence

$$\left|\frac{S(\mu \upharpoonright A^n)}{n} - \frac{S(\mu' \upharpoonright A^n)}{n}\right| \leqq 6\varepsilon,$$

and hence $|\tilde{S}(\mu) - \tilde{S}(\mu')| \leqq 6\varepsilon$. This shows that the entropy rate $\tilde{S}(\mu)$ is uniformly continuous with respect to the $\bar{d}$-distance. □

We next show the $\bar{d}$-continuity of the transformation $\mu \to \lambda(\mu)$.

THEOREM 5.4. *For a finite-memory stationary channel λ, the output measure $\lambda(\mu)$ is a continuous function of an input measure μ in the $\bar{d}$-distance.*

For the proof of Theorem 5.4, we prepare the following Lemma 5.5. Let λ be a stationary channel from A^Z to B^Z with finite memory length m. Then, the channel λ canonically induces a channel $\alpha \to \lambda(\alpha)$ from a finite space

$A^{[-m,n-1]}$ to a finite space $B^{[0,n-1]}$ such as

$$\lambda(\alpha)(\beta) \equiv \lambda(x)(\beta) \quad \text{for } \alpha \in A^{[-m,n-1]}, \beta \in B^{[0,n-1]}, x \in A^Z, x \in \alpha.$$

LEMMA 5.5. *Let λ: $\alpha \to \lambda(\alpha)$ be a channel from $A^{[-m,n-1]}$ to $B^{[0,n-1]}$ given above. Then, for any $(\alpha, \alpha') \in A^{[-m,n-1]} \times A^{[-m,n-1]}$ with $d(\alpha, \alpha') \leqq d$, there exists a probability measure $\omega(\beta, \beta'|\alpha, \alpha')$ on $B^{[0,n-1]} \times B^{[0,n-1]}$ such that $\lambda(\alpha)(\beta)$ and $\lambda(\alpha')(\beta')$ are the marginals of $\omega(\beta, \beta'|\alpha, \alpha')$ and*

$$\sum_{(\beta,\beta')} \frac{1}{n} \sum_{i=0}^{n-1} d(\beta_i, \beta'_i)\omega(\beta, \beta'|\alpha, \alpha') \leqq \frac{d}{n}(m+1).$$

Proof. Let I, J be the subsets of the interval $[0, n-1]$ given by

$$I = \{t \in [0, n-1] | \alpha_s = \alpha'_s, \forall s \in [t-m, t]\},$$
$$J = \{t \in [0, n-1] | \alpha_s \neq \alpha'_s, \exists s \in [t-m, t]\}.$$

Let $\omega(\beta, \beta'|\alpha, \alpha')$ be a function of (β, β') given by

$$\omega(\beta, \beta'|\alpha, \alpha') \equiv (\lambda(\alpha)(\beta_I))\delta(\beta_I, \beta'_I)(\lambda(\alpha)(\beta_J|\beta_I))(\lambda(\alpha')(\beta'_J|\beta'_I)),$$

where $\beta_I = (\beta_s)_{s \in I}$, $\beta_J = (\beta_s)_{s \in J}$, $\delta(\beta_I, \beta'_I) = 1$ for $\beta_I = \beta'_I$ and $\delta(\beta_I, \beta'_I) = 0$ for $\beta_I \neq \beta'_I$, $\lambda(\alpha)(\beta_J|\beta_I) \equiv \lambda(\alpha)(\beta)/\lambda(\alpha)(\beta_I)$ and so on. The direct calculation shows that $\Sigma_{(\beta,\beta')}\omega(\beta, \beta'|\alpha, \alpha') = 1$, i.e., $\omega(\beta, \beta'|\alpha, \alpha')$ is a probability measure on $B^{[0,n-1]} \times B^{[0,n-1]}$. Let us show that $\lambda(\alpha')(\beta')$ is the marginal of $\omega(\beta, \beta'|\alpha, \alpha')$. The set I is represented as a union of disjoint intervals $I_j = [a_j, b_j]$ $(1 \leqq j \leqq k)$ such that

$$a_1 \leqq b_1 \leqq a_2 \leqq b_2 \leqq \cdots \leqq a_j \leqq b_j \leqq \cdots \leqq a_k \leqq b_k \quad \text{and} \quad a_{i+1} - b_i > m.$$

Since λ has finite memory length m, we have

$$\begin{aligned}
\lambda(\alpha)(\beta'_I) &= \lambda(\alpha)(\beta'_{I1} \cap \cdots \cap \beta'_{Ik}) \\
&= \lambda(\alpha)(\beta'_{I1}) \cdots \lambda(\alpha)(\beta'_{Ik}) \\
&= \lambda(\alpha')(\beta'_{I1}) \cdots \lambda(\alpha')(\beta'_{Ik}) \\
&= \lambda(\alpha')(\beta'_{I1} \cap \cdots \cap \beta'_{Ik}) \\
&= \lambda(\alpha')(\beta'_I).
\end{aligned}$$

So we have

$$\begin{aligned}\sum_{\beta}\omega(\beta,\beta'|\alpha,\alpha') &= \lambda(\alpha')(\beta'_J|\beta'_I)\sum_{(\beta_I,\beta_J)}\lambda(\alpha)(\beta_I)\delta(\beta_I,\beta'_I)\lambda(\alpha)(\beta_J|\beta_I)\\ &= \lambda(\alpha')(\beta'_J|\beta'_I)\sum_{\beta_I}\lambda(\alpha)(\beta_I)\delta(\beta_I,\beta'_I)\sum_{\beta_J}\lambda(\alpha)(\beta_J|\beta_I)\\ &= \lambda(\alpha')(\beta'_J|\beta'_I)\lambda(\alpha)(\beta'_I)\\ &= \lambda(\alpha')(\beta'_J|\beta'_I)\lambda(\alpha')(\beta'_I)\\ &= \lambda(\alpha')(\beta').\end{aligned}$$

This shows that $\lambda(\alpha')(\beta')$ is the marginal of $\omega(\beta,\beta'|\alpha,\alpha')$. In the same way, $\lambda(\alpha)(\beta)$ is shown to be the marginal of $\omega(\beta,\beta'|\alpha,\alpha')$. For fixed (β,β') and fixed $i\in I$, we have $d(\beta_i,\beta'_i)=0$ for $\beta_i=\beta'_i$ and $\delta(\beta_I,\beta'_I)=0$ for $\beta_i\neq\beta'_i$. So we have

$$\begin{aligned}&d(\beta_i,\beta'_i)\omega(\beta,\beta'|\alpha,\alpha')\\ &\quad= d(\beta_i,\beta'_i)(\lambda(\alpha)(\beta_I))\delta(\beta_I,\beta'_I)(\lambda(\alpha)(\beta_J|\beta_I))(\lambda(\alpha')(\beta'_J|\beta'_I)) = 0\end{aligned}$$

for $i\in I$. Hence we get

$$\begin{aligned}&\sum_{(\beta,\beta')}\frac{1}{n}\sum_{i=0}^{n-1}d(\beta_i,\beta'_i)\omega(\beta,\beta'|\alpha,\alpha')\\ &\quad=\sum_{(\beta,\beta')}\frac{1}{n}\sum_{i\in J}d(\beta_i,\beta'_i)\omega(\beta,\beta'|\alpha,\alpha')\leqq\frac{|J|}{n}\leqq\frac{d}{n}(m+1).\end{aligned}\qquad\square$$

Proof of Theorem 5.4. For any $\varepsilon>0$, let $q>0$ be an integer with $(m+1)/q<\varepsilon$, and d,r be integers with $n=dq+r$ $(0\leqq r\leqq q-1)$. Let $\delta>0$ be any real number such as $q\delta<\varepsilon$, and μ,μ' be any stationary probability measures on A^Z such that $\bar{d}(\mu,\mu')<\delta$. For each $(\alpha,\alpha')\in A^{[-m,n-1]}\times A^{[-m,n-1]}$, let $\omega(\beta,\beta'|\alpha,\alpha')$ be a probability measure on $B^{[0,n-1]}\times B^{[0,n-1]}$ constructed as in Lemma 5.5. For any probability measure $\overline{\mu}(\alpha,\alpha')$ on $A^{[-m,n-1]}\times A^{[-m,n-1]}$ having $\mu(\alpha)$ and $\mu'(\alpha')$ as its marginals,

$$\overline{\nu}(\beta,\beta')=\sum_{(\alpha,\alpha')}\omega(\beta,\beta'|\alpha,\alpha')\overline{\mu}(\alpha,\alpha')$$

is a probability measure on $B^{[0,n-1]}\times B^{[0,n-1]}$ having $(\lambda(\mu))(\beta)$ and

$(\lambda(\mu'))(\beta')$ as its marginals. Then, for some $\bar{\mu}$, we have

$$\sum_{(\beta,\beta')} \frac{1}{n} \sum_{i=0}^{n-1} d(\beta_i, \beta_i')\bar{\nu}(\beta, \beta')$$

$$= \sum_{\substack{(\alpha,\alpha') \\ d(\alpha,\alpha') \leqq d}} \sum_{(\beta,\beta')} \frac{1}{n} \sum_{i=0}^{n-1} d(\beta_i, \beta_i')\omega(\beta, \beta' | \alpha, \alpha')\bar{\mu}(\alpha, \alpha')$$

$$+ \sum_{\substack{(\alpha,\alpha') \\ d(\alpha,\alpha') > d}} \sum_{(\beta,\beta')} \frac{1}{n} \sum_{i=0}^{n-1} d(\beta_i, \beta_i')\omega(\beta, \beta' | \alpha, \alpha')\bar{\mu}(\alpha, \alpha')$$

$$\leqq \frac{d}{n}(m+1) + \bar{\mu}(\Delta_d^C) \leqq \frac{d}{n}(m+1) + \frac{n}{d}\delta \leqq \frac{m+1}{q} + 2q\delta \leqq 3\varepsilon.$$

This shows that the transformation $\mu \to \lambda(\mu)$ is uniformly continuous in the $\bar{d}$-distance. □

Proof of Theorem 5.1. It is easily verified that the channel $\delta \otimes \lambda$ from A^Z to $A^Z \times B^Z$ also has finite memory length m. By Theorem 5.4, the convergence $\mu_j \to \mu$ in the $\bar{d}$-distance implies the convergence $\lambda(\mu_j) \to \lambda(\mu)$ and $(\delta \otimes \lambda)(\mu_j) \to (\delta \otimes \lambda)(\mu)$ in the $\bar{d}$-distance. Since the entropy rate $\tilde{S}(\cdot)$ is uniformly continuous in the $\bar{d}$-distance, so is the information rate $\tilde{I}(\cdot; \lambda) = \tilde{S}(\cdot) + \tilde{S}(\lambda(\cdot)) - \tilde{S}((\delta \otimes \lambda)(\cdot))$. □

6. Discontinuity of information rate in the $\bar{d}$-distance

In this section, we first show that (1) $\bar{d}$-distance topology is weaker than the set-wise convergence topology (Theorem 6.1) and that (2) $\bar{d}$-distance topology is stronger than the cylinder-wise convergence topology (Theorem 6.2). Then we construct a non-finite-memory channel for which the information rate is discontinuous in the $\bar{d}$-distance (Theorem 6.3). At the end of this section, we remark some continuity properties of the information rate in the case that the input and output alphabets A and B are standard Borel spaces.

THEOREM 6.1. *The $\bar{d}$-distance topology is weaker than the set-wise convergence topology.*

To prove Theorem 6.1, we need the following second definition [11] of the $\bar{d}$-distance which is equivalent to the first one given in §3. The $\bar{d}$-distance $\bar{d}(\mu, \mu')$ is the sup of the α satisfying the following: Given $\varepsilon > 0$ there is an integer N such that if $K > N$, then we can find two collections C_1 and C_2 of sequences in $A^{[0,K-1]}$ such that $\mu(C_1) > 1 - \varepsilon$ and $\mu'(C_2) > 1 - \varepsilon$, and any sequence in C_1 differs from any sequence in C_2 in more than αK places.

Proof of Theorem 6.1. Let $\{\mu_j\}$ be a net of stationary probability measures converging to a stationary probability measure μ in the set-wise convergence. Suppose that $\{\mu_j\}$ does not converge to μ in the $\bar{d}$-distance. Then there exist $\alpha > 0$ and a subsequence $\{\mu_{j(k)}\}$ of $\{\mu_j\}$ such that $\bar{d}(\mu_{j(k)}, \mu) > \alpha$ for all k. By the second definition of the $\bar{d}$-distance, for any $\varepsilon > 0$ and each $\mu_{j(k)}$, there exists a natural number $N^{(k)}$ such that, for any $K > N^{(k)}$, there exists subsets $C_1^{(k)}, C_2^{(k)}$ of A^K and the following hold:

$$\mu\big(C_1^{(k)}\big) > 1 - \varepsilon/2^k, \qquad \mu_{j(k)}\big(C_2^{(k)}\big) > 1 - \varepsilon/2^k, \qquad C_1^{(k)} \cap C_2^{(k)} = \varnothing.$$

We have

$$\mu\Big(\bigcup_k C_2^{(k)}\Big) \leqq \sum_k \mu\big(C_2^{(k)}\big) \leqq \sum_k \varepsilon/2^k = \varepsilon.$$

Since $\{\mu_{j(k)}\}$ converges to μ in the set-wise convergence, we have

$$\mu_{j(k)}(C_2) \to \mu(C_2),$$

where $C_2 \equiv \bigcup_k C_2^{(k)}$. So, for sufficiently large any $k \geqq k_0$, we have

$$\mu_{j(k)}(C_2) \leqq 2\varepsilon.$$

This contradicts $\mu_{j(k)}(C_2^{(k)}) > 1 - \varepsilon/2^k$. Hence $\{\mu_j\}$ converges to μ in the $\bar{d}$-distance. □

There exists a sequence $\{\mu_j\}$ of stationary probability measures converging to a stationary probability measure μ in the $\bar{d}$-distance, but not converging in the set-wise convergence, as follows: Let $\{x^{(j)};\ j = 1, 2, \dots\}$ be a sequence of doubly-infinite sequences $x^{(j)} = \{x_n^{(j)};\ n \in Z\}$ in A^Z with $A = \{0, 1\}$, given by

$$x_n^{(j)} = \begin{cases} 1 & \text{for } n \equiv j - 1 \pmod{j} \\ 0 & \text{otherwise.} \end{cases}$$

Let (μ_j) be a sequence of stationary measures given by

$$\mu_j = \frac{1}{j} \sum_{k=0}^{j-1} \delta(T^k x^{(j)})$$

and μ_j be a stationary measure $\delta(x^{(0)})$, where $x^{(0)} = (\dots, 0, 0, 0, \dots)$. Then it is easily shown that $\{\mu_j\}$ converges to μ in the $\bar{d}$-distance, but $\{\mu_j(E)\}$ does not converge to $\mu(E)$ for a set of single element $E = \{x^{(0)}\}$.

THEOREM 6.2. *The $\bar{d}$-distance topology is stronger than the cylinder-wise convergence topology.*

Proof. Let $\{\mu_j\}$ be a net of stationary probability measures converging to a stationary probability measure μ in the $\bar{d}$-distance. By stationarity of μ_j and μ, it suffices to show that $\mu_j(\alpha) \to \mu(\alpha)$ for cylinders $\alpha \in A^{[0,n-1]}$ in order to prove that $\{\mu_j\}$ converges to μ in the cylinder-wise convergence. By the $\bar{d}$-convergence of $\{\mu_j\}$ to μ, for any $\varepsilon > 0$, there exists j_0 such that $\bar{d}(\mu, \mu_j) \leqq \varepsilon$ $(j \geqq j_0)$. For μ_j $(j \geqq j_0)$ and any n there exists $\omega \in \mathscr{P}_n(\mu, \mu_j)$ such that

$$\int_{A^{[0,n-1]} \times A^{[0,n-1]}} \frac{1}{n} \sum_{i=0}^{n-1} d(\alpha_i, \alpha_i')\, d\omega(\alpha, \alpha') \leqq \varepsilon,$$

which implies $\omega\{(\alpha, \alpha') \in A^n \times A^n | \alpha \neq \alpha'\} \leqq n\varepsilon$. We have

$$\mu(\alpha) = \sum_{\alpha'} \omega(\alpha, \alpha') = \omega(\alpha, \alpha) + \sum_{\alpha' \neq \alpha} \omega(\alpha, \alpha') \leqq \omega(\alpha, \alpha) + n\varepsilon.$$

In the same way, we have $\mu_j(\alpha) \leqq \omega(\alpha, \alpha) + n\varepsilon$, and hence

$$|\mu(\alpha) - \mu_j(\alpha)| \leqq 2n\varepsilon \qquad (j \geqq j_0).$$

Therefore $\{\mu_j\}$ converges to μ in the cylinder-wise convergence. □

There exists a sequence $\{\mu_n\}$ of stationary probability measures converging to a stationary probability measure μ in the cylinder-wise convergence, but not converging in the $\bar{d}$-distance, as follows: Let μ be a stationary probability measure of positive entropy rate $\tilde{s}(\mu) > 0$. There exists a sequence $\{\mu_n\}$ of stationary probability measures of entropy rate 0, which converge to μ in the cylinder-wise convergence. Since the entropy rate is continuous in the $\bar{d}$-distance, $\{\mu_n\}$ does not converge to μ in the $\bar{d}$-distance.

THEOREM 6.3. *There exists a stationary channel for which the information rate is discontinuous in the $\bar{d}$-distance.*

Proof. Put $A = B = \{0, 1, *\}$. Let us construct stationary probability measures $\{\mu_p;\ p \in N\}$ and μ on A^Z and a stationary channel λ from A^Z to B^Z such that μ_p converges to μ in the $\bar{d}$-distance but $\tilde{I}(\mu_p; \lambda)$ does not converge to $\tilde{I}(\mu; \lambda)$. Let $f_i^{(p)}\colon A^Z \to A^Z$ be a mapping defined by

$$y = f_i^{(p)}(x) \Leftrightarrow y_n = \begin{cases} * & \text{for } n \equiv i \pmod{p}, \\ x_n & \text{otherwise,} \end{cases}$$

for $x = \{x_n\}$, $y = \{y_n\} \in A^Z$. Let μ be a probability measure on A^Z given as the infinite direct product of an identical probability distribution

$$\begin{bmatrix} A \\ P \end{bmatrix} = \begin{bmatrix} 0 & 1 & * \\ 1/2 & 1/2 & 0 \end{bmatrix}$$

on A, and μ_p be a probability measure on A^Z given by

$$\mu_p = \frac{1}{p} \sum_{i=0}^{p-1} \int_{A^Z} \delta\left(f_i^{(p)}(x)\right) d\mu(x).$$

Considering a probability measure ω_p

$$\omega_p = \frac{1}{p} \sum_{i=0}^{p-1} \int_{A^Z} \delta(x) \otimes \delta\left(f_i^{(p)}(x)\right) d\mu(x)$$

on $A^Z \times A^Z$, for which μ and μ_p are the marginals, one can prove that $\bar{d}(\mu, \mu_p) \leqq 1/p$ and hence $\mu_p \to \mu$ in the $\bar{d}$-distance. Let λ be a stationary channel from A^Z to B^Z given by

$$\lambda(x) = \begin{cases} \delta(x) & \text{for } x \in \bigcup_{i=0}^{p-1} \left\{y = f_i^{(p)}(x);\ x \in A^Z\right\}, \\ \mu & \text{otherwise.} \end{cases}$$

After some calculation we get $\tilde{I}(\mu_p; \lambda) = (p - 1)/p$ and $\tilde{I}(\mu; \lambda) = 0$. Hence the information rate $\tilde{I}(\mu_p; \lambda)$ does not converge to $\tilde{I}(\mu; \lambda)$. □

As a final remark, let us generalize our results in the finite alphabet situation to those in the standard alphabet situation. Let A, B be standard Borel spaces, namely, Borel subsets of complete separable metric spaces. We call them standard alphabets. The concepts of entropy rate and information rate can be suitably defined for communication processes with standard alphabets (see [4] and [14]). We can naturally define the set-wise convergence topology and the cylinder-wise convergence topology for the set of all stationary probability measures on A^z. As proved in [4] the ergodic decomposition of the information rate still holds in the standard alphabet situation. So we can prove the continuity of the information rate in the set-wise convergence of input probability measures for a stationary channel λ with finite capacity, in the same way as in §3. We can also prove the that the set of zero points of the information rate is dense in $\mathscr{P}_T(A^Z)$ in the cylinder-wise convergence topology by the direct approximation of any measure $\mu \in \mathscr{P}_T(A^Z)$ with the measures of information rate 0.

We are now studying the continuity of the information rate in the $\bar{\rho}$-distance topology, where $\bar{\rho}$-distance is given in [6] and [7] as a generalization of $\bar{d}$-distance to the standard alphabet situation. We are also studying the continuity of the information rate for asymptotically mean stationary probability measures [5], using the results of [1] and [2].

Acknowledgements. The authors would like to thank the referee for a number of helpful suggestions.

References

1. P.H. Algoet and T.M. Cover, *A sandwich proof of the Shannon-McMillan-Breiman theorem*, Ann. Probab., vol. 16 (1988), pp. 899–909.
2. A.R. Barron, *The strong ergodic theorem for densities: Generalized Shannon-McMillan-Breiman theorem*, Ann. Probab., vol. 13 (1985), pp. 1292–1303.
3. L. Breiman, *On achieving channel capacity in finite-memory channels*, Illinois J. Math., vol. 4 (1960), pp. 246–252.
4. R.M. Gray and J.C. Kieffer, *Mutual information rate, distortion, and quantization in metric spaces*, IEEE Transactions on Information Theory, vol. 26 (1980), pp. 412–422.
5. _____, *Asymptotically mean stationary measures*, Ann. Probab., vol. 8 (1980), 962–973.
6. R.M. Gray, D.L. Neuhoff and D.S. Ornstein, *Nonblock source coding with a fidelity criterion*, Ann. Probab., vol. 3 (1975), pp. 478–491.
7. R.M. Gray, D.L. Neuhoff and P.C. Shields, *A generalization of Ornstein's $\bar{d}$ distance with applications to information theory*, Ann. Probab., vol. 3 (1975), pp. 315–328.
8. R.M. Gray and P.C. Shields, *The maximum mutual information between two random processes*, Information and Control, vol. 33 (1977), pp. 273–280.
9. A.I. Khinchin, *Mathematical foundations of information theory*, Dover, New York, 1957.
10. D.S. Ornstein, *An application of ergodic theory to probability theory*, Ann. Probab., vol. 1 (1973), pp. 43–65.
11. _____, *Ergodic theory, randomness, and dynamical systems*, Yale University Press, New Haven, 1974.
12. J.C. Oxtoby, *Ergodic sets*. Bull. Amer. Math. Soc., vol. 58 (1952), pp. 116–136.
13. K.R. Parthasarathy, *On the integral representation of the rate of transmission of a stationary channel*, Illinois J. Math., vol. 5 (1961), pp. 299–305.
14. M.S. Pinsker, *Information and information stability of random variables*, Holden-Day, San Francisco, 1964.
15. P. Shields, *The theory of Bernoulli shifts*, The University of Chicago Press, Chicago, 1973.
16. H. Umegaki, *General treatment of alphabet-message space and integral representation of entropy*, Kôdai Math. Sem. Rep., vol. 16 (1964), pp. 18–26.
17. _____, *A functional method for stationary channels*, Kôdai Math. Sem. Rep., vol. 16 (1964), pp. 27–39.

Tokyo Denki University
Hatoyama-Machi, Hiki-Gun, Saitama 350-03, Japan

Science University of Tokyo
Noda-City, Chiba 278, Japan