Supremus Typicality

Sheng Huang, Mikael Skoglund
Communication Theory Lab, School of Electrical Engineering
KTH Royal Institute of Technology
Stockholm, Sweden
Email: sheng.huang@ee.kth.se, skoglund@ee.kth.se

***Abstract*—This paper investigates a new type of typicality for sequences, termed Supremus typical sequences, in both the strong and the weak senses. It is seen that Supremus typicality is a condition stronger than classic typicality in both the strong and the weak senses. Even though Supremus typical sequences form a (often strictly smaller) subset of classic typical sequences, the Asymptotic Equipartion Property is still valid for Supremus typical sequences. Furthermore, Supremus typicality leads to a generalized typicality lemma that is more accessible and easier to analyze than its classic counterpart.**

I. Introduction

The idea of typical sequence [1] has become an almost un-replaceable part of most asymptotical achievability results in information theory (in a narrow sense). This concept is mainly defined in either a strong sense or a weak sense (cf. [2]). However, regardless its presented form, the main features are captured by the Asymptotic Equipartion Property (AEP) and related lemmata, e.g. the typicality lemma and the packing lemma. These properties/lemmata illustrate the ergodic behaviours of the dynamical system (a random process is a dynamical system [3]) modelling a set of sources and/or several communication channels. For example, the AEP is in some sense the Shannon–McMillan–Breiman (SMB) Theorem [4], [5], [6] (the original SMB Theorem is for stationary systems, please refer to [7] for the one for asymptotically mean stationary (a.m.s.) systems which include stationary systems as special cases).

A classic typical sequence is usually defined to be a sequence reflecting the ergodic behaviour of an a.m.s. random process. For instance, a typical sequence admits an empirical distribution that is "close" enough to the genuine distribution. By the SMB Theorem, a randomly generated sequence by the source(s) is typical in probability close to 1. However, we will see that this classic definition is not strong enough in the sense that it includes a class of sequences that will occur in probability close to 0. Hence, it makes sense to exclude those sequences from consideration.

This paper studies a new concept that we call Supremus typical sequence, in both a strong sense and a weak sense. Roughly speaking, a sequence is Supremus typical if all of its reduced subsequences admit empirical distributions resembling the genuine distributions of the corresponding reduced random processes[1], respectively. It is easy to see that Supremus typical sequences are classic typical. However, the converse statement is not true as seen later. Supremus typical sequences form a strictly smaller subset in general. Nevertheless, the AEP still holds for Supremus typical sequences if the random process is a.m.s.. This follows from the fact that the SMB Theorem holds simultaneously for all the reduced random processes of an a.m.s. random process [8]. Other than possessing AEP, advantages have been seen when comparing the generalized typicality lemma of Supremus typicality, Lemma II.6, to the one of classic typicality, Lemma II.7. It is seen that the one of Supremus typical sequence is more accessible, while the one of classic typical sequence is usually hard to analyze (see Subsection II-C for more details).

The organization of this paper is as follows. We will introduce Supremus typicality in the strong sense in Section II. In order to clearly present the main idea, we will assume that the source is irreducible Markov but not necessarily stationary in this section. Supremus typicality in the weak sense is studied in Section III after a short introduction on related results from ergodic theory. Because of space limitation, jointly typical sequence will not be considered in this paper.

II. Supremus Typicality in the Strong Sense

In this section, we will focus on investigating Supremus typicality in the strong sense in the setting of Markov sources. Unless otherwise specified, all Markov chains/processes considered in this work are homogeneous, finite-state and irreducible, while their initial distributions are unknown. Assume that $\mathbf{P} = [p_{i,j}]_{i,j \in \mathscr{X}}$ is the transition matrix of some Markov process $\mathscr{M}$ with state space $\mathscr{X}$, then one can prove that $\mathscr{M}$ has a unique invariant distribution $\pi = [p_j]_{j \in \mathscr{X}}$, i.e. $\pi \mathbf{P} = \pi$, since $\mathscr{M}$ is finite-state and irreducible. However, π is not necessarily the initial distribution of $\mathscr{M}$, thus $\mathscr{M}$ is not necessarily stationary[2]. For convenience, defined $H(\mathbf{P}|\pi) = -\sum_{i,j \in \mathscr{X}} p_i p_{i,j} \log p_{i,j}$. It is easily seen that $H(\mathbf{P}|\pi)$ is the entropy rate of $\mathscr{M}$.

Recall that the classic Markov typical sequence in the strong sense is defined as follows.

Definition II.1 (Strong Markov Typicality (cf. [9], [10])). Let $\mathscr{M} = \left\{ X^{(n)} \right\}$ be an irreducible Markov chain with state space

[1] In some literature, a reduced random process is called an induced random process, because it is described by an induced transformation [8].

[2] In most literature, random processes are often assumed to be stationary, while irreducibility is a weaker condition.

This work was funded in part by the Swedish Research Council.

$\mathscr{X}$, and $\mathbf{P} = [p_{i,j}]_{i,j \in \mathscr{X}}$ and $\pi = [p_j]_{j \in \mathscr{X}}$ be its transition matrix and invariant distribution, respectively. For any $\epsilon > 0$, a sequence $\mathbf{x} \in \mathscr{X}^n$ of length n (≥ 2) is said to be *strongly Markov ϵ-typical* with respect to $\mathbf{P}$ if

$$\begin{cases} \left| \dfrac{N(i,j;\mathbf{x})}{N(i;\mathbf{x})} - p_{i,j} \right| < \epsilon; \\ \left| \dfrac{N(i;\mathbf{x})}{n} - p_i \right| < \epsilon, \end{cases} \quad \forall\, i, j \in \mathscr{X}, \tag{1}$$

where $N(i,j;\mathbf{x})$ denotes the number of occurrences of subsequence $[i,j]$ in $\mathbf{x}$ and $N(i;\mathbf{x}) = \sum_{j \in \mathscr{X}} N(i,j;\mathbf{x})$. The set of all strongly Markov ϵ-typical sequences with respect to $\mathbf{P}$ in $\mathscr{X}^n$ is denoted by $\mathcal{T}_\epsilon(n, \mathbf{P})$.

It is well-known that the AEP holds for strongly Markov typical sequences. We soon will see that this definition can be strengthen to obtain refined properties.

Some background is needed before proceeding.

A. Stochastic Complement

Given a Markov chain $\mathscr{M} = \left\{X^{(n)}\right\}$ with state space $\mathscr{X}$ and a non-empty subset A of $\mathscr{X}$, let

$$T_{A,l} = \begin{cases} \inf\left\{n > 0 | X^{(n)} \in A\right\}; & l = 1, \\ \inf\left\{n > T_{A,l-1} | X^{(n)} \in A\right\}; & l > 1, \\ \sup\left\{n < T_{A,l+1} | X^{(n)} \in A\right\}; & l < 1. \end{cases}$$

It is well-known that $\mathscr{M}_A = \left\{X^{(T_{A,l})}\right\}$ is Markov by the strong Markov property [11, Theorem 1.4.2]. In particular, if $\mathscr{M}$ is irreducible, so is $\mathscr{M}_A$. To be more precise, if $\mathscr{M}$ is irreducible with invariant distribution $\pi = [p_i]_{i \in \mathscr{X}}$ and transition matrix

$$\mathbf{P} = \begin{bmatrix} \mathbf{P}_{A,A} & \mathbf{P}_{A,A^c} \\ \mathbf{P}_{A^c,A} & \mathbf{P}_{A^c,A^c} \end{bmatrix},$$

respectively, then

$$\mathbf{S}_A = \mathbf{P}_{A,A} + \mathbf{P}_{A,A^c}\left(\mathbf{1} - \mathbf{P}_{A^c,A^c}\right)^{-1}\mathbf{P}_{A^c,A},$$

is the transition matrix of $\mathscr{M}_A$ [12, Theorem 2.1 and Section 3]. $\pi_A = \left[\dfrac{p_i}{\sum_{j \in A} p_j}\right]_{i \in A}$ is an invariant distribution of $\mathscr{M}_A$, i.e. $\pi_A \mathbf{S}_A = \pi_A$ [12, Theorem 2.2]. Since $\mathscr{M}_A$ inherits irreducibility from $\mathscr{M}$ [12, Theorem 2.3], π_A is unique. The matrix $\mathbf{S}_A$ is termed the *stochastic complement* of $\mathbf{P}_{A,A}$ in $\mathbf{P}$, while $\mathscr{M}_A$ is named a *reduced Markov chain* (or *reduced Markov process*) of $\mathscr{M}$. It has state space A obviously.

B. Supremus Typicality in the Strong Sense and AEP

Definition II.2 (Supremus Typicality in the Strong Sense)**.** Following the notation defined above, given $\epsilon > 0$ and a sequence $\mathbf{x} = \left[x^{(1)}, x^{(2)}, \cdots, x^{(n)}\right] \in \mathscr{X}^n$ of length n ($\geq 2\,|\mathscr{X}|$), let $\mathbf{x}_A$ be the subsequence of $\mathbf{x}$ formed by all those $x^{(l)}$'s that belong to A in the original ordering. $\mathbf{x}$ is said to be *Supremus ϵ-typical* with respect to $\mathbf{P}$ if $\mathbf{x}_A$ is strongly Markov ϵ-typical with respect to $\mathbf{S}_A$ for all feasible non-empty subset A of $\mathscr{X}$.

In Definition II.2, the set of all Supremus ϵ-typical sequences with respect to $\mathbf{P}$ in $\mathscr{X}^n$ is denoted as $\mathcal{S}_\epsilon(n, \mathbf{P})$. $\mathbf{x}_A$ is called a *reduced subsequence* (with respect to A) of $\mathbf{x}$. It follows immediately from this definition that

Proposition II.3. *Every reduced subsequence of a Supremus ϵ-typical sequence is Supremus ϵ-typical.*

However, the above proposition does not hold for strongly Markov ϵ-typical sequences. In other words, *a reduced subsequence of a strongly Markov ϵ-typical sequence is not necessarily strongly Markov ϵ-typical.*

Example II.4. Let $\{\alpha, \beta, \gamma\}$ be the state space of an i.i.d. process with a uniform distribution, i.e.

$$\mathbf{P} = \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{bmatrix},$$

and $\mathbf{x} = (\alpha, \beta, \gamma, \alpha, \beta, \gamma, \alpha, \beta, \gamma)$. It is easy to verify that $\mathbf{x}$ is a strongly Markov $5/12$-typical sequence. However, the reduced subsequence $\mathbf{x}_{\{\alpha,\gamma\}} = (\alpha, \gamma, \alpha, \gamma, \alpha, \gamma)$ is no long a strongly Markov $5/12$-typical sequence, because $\mathbf{S}_{\{\alpha,\gamma\}} = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}$ and

$$\left| \frac{\text{the number of subsequence } (\alpha,\alpha)\text{'s in } \mathbf{x}_{\{\alpha,\gamma\}}}{6} - 0.5 \right|$$
$$= |0 - 0.5| > 5/12.$$

Proposition II.5 (AEP of Strong Supremus Typicality)**.** *Let $\mathscr{M} = \left\{X^{(n)}\right\}$ be an irreducible Markov chain with state space $\mathscr{X}$, and $\mathbf{P} = [p_{i,j}]_{i,j \in \mathscr{X}}$ and $\pi = [p_j]_{j \in \mathscr{X}}$ be its transition matrix and invariant distribution, respectively. For any $\eta > 0$, there exist $\epsilon_0 > 0$ and $N_0 \in \mathbb{N}^+$, such that, $\forall\ \epsilon_0 > \epsilon > 0$, $\forall\ n > N_0$ and $\forall\ \mathbf{x} = \left[x^{(1)}, x^{(2)}, \cdots, x^{(n)}\right] \in \mathcal{S}_\epsilon(n, \mathbf{P})$,*

1) $\exp_2\left[-n\left(H(\mathbf{P}|\pi) + \eta\right)\right] < \Pr\{[X^{(1)}, X^{(2)}, \cdots, X^{(n)}] = \mathbf{x}\} < \exp_2\left[-n\left(H(\mathbf{P}|\pi) - \eta\right)\right]$;
2) $\Pr\left\{\left[X^{(1)}, X^{(2)}, \cdots, X^{(n)}\right] \notin \mathcal{S}_\epsilon(n, \mathbf{P})\right\} < \eta$; *and*
3) $|\mathcal{S}_\epsilon(n, \mathbf{P})| < \exp_2\left[n\left(H(\mathbf{P}|\pi) + \eta\right)\right]$.

Sketch of the Proof [3]*:* 1) and 3) follow from the fact that a Supremus typical sequence (in the strong sense) is Strongly Markov typical. Let $\mathbf{X} = \left[X^{(1)}, X^{(2)}, \cdots, X^{(n)}\right]$. Since $\mathscr{M}_A$ is irreducible with transition matrix $\mathbf{S}_A$ for any non-empty subset A of the state space, $\Pr\left\{\mathbf{X}_A \notin \mathcal{T}_\epsilon(n, \mathbf{S}_A)\right\} < \dfrac{\eta}{2^{|\mathscr{X}|} - 1}$ if ϵ_0 is small enough and N_0 is large enough. Henceforth, 2) follows, because

$$\left\{\mathbf{X} \notin \mathcal{S}_\epsilon(n, \mathbf{P})\right\} = \bigcup_{\emptyset \neq A \subseteq \mathscr{X}} \left\{\mathbf{X}_A \notin \mathcal{T}_\epsilon(n, \mathbf{S}_A)\right\}$$

by Proposition II.3. ■

Remark 1. From Proposition II.5, one sees that $\mathcal{S}_\epsilon(n, \mathbf{P})$ has probability close to 1, as it is also the case for $\mathcal{T}_\epsilon(n, \mathbf{P})$ even though $\mathcal{S}_\epsilon(n, \mathbf{P})$ is always smaller than $\mathcal{T}_\epsilon(n, \mathbf{P})$. As a consequence, although a coding procedure makes mistakes in

[3] A complete proof is found in [13].

encoding/decoding sequences in $\mathcal{T}_\epsilon(n,\mathbf{P})\setminus\mathcal{S}_\epsilon(n,\mathbf{P})$, the error is negligible as for non-typical sequences.

The idea of "ignoring $\mathcal{T}_\epsilon(n,\mathbf{P})\setminus\mathcal{S}_\epsilon(n,\mathbf{P})$" has already been used in [14], [15], even though the concept of Supremus typicality was never mentioned in these works. In fact, these works consider only i.i.d. scenarios allowing some tricks to be played and sparing them the introduction of this concept. Unfortunately, the situation changes as soon as we move away from the i.i.d. condition (as many real communication systems do). More elaboration on this is given in the next subsection.

C. The Generalized Typicality Lemmata

Given a set $\mathscr{X}$, a *partition* $\coprod_{k\in\mathscr{K}} A_k$ of $\mathscr{X}$ is a disjoint union of $\mathscr{X}$, i.e. $A_{k'}\cap A_{k''}\neq\emptyset \Leftrightarrow k'=k''$, $\bigcup_{k\in\mathscr{K}} A_k=\mathscr{X}$ and A_k's are not empty.

By the AEPs, an efficient coding system must ensure that all the Supremus/classic typical sequences are decoded correctly. An error occurs mainly because a typical sequence is confused with other typical sequences. In some specially-designed coding systems, the decoders will only confuse the correct message, say $\mathbf{x}$, with a subset of typical sequences of certain sequential pattern, say

$$S_\epsilon(\mathbf{x}) =\Big\{ \left[y^{(1)},y^{(2)},\cdots,y^{(n)}\right]\in\mathcal{S}_\epsilon(n,\mathbf{P})\Big| \\ y^{(l)}\in A_k\Leftrightarrow x^{(l)}\in A_k, \forall\, 1\leq l\leq n, \forall\, 1\leq k\leq m\Big\}$$

or

$$T_\epsilon(\mathbf{x}) =\Big\{ \left[y^{(1)},y^{(2)},\cdots,y^{(n)}\right]\in\mathcal{T}_\epsilon(n,\mathbf{P})\Big| \\ y^{(l)}\in A_k\Leftrightarrow x^{(l)}\in A_k, \forall\, 1\leq l\leq n, \forall\, 1\leq k\leq m\Big\},$$

where $\coprod_{k=1}^m A_k$ is some partition. Therefore, determining the sizes and upper bounds of $S_\epsilon(\mathbf{x})$ and $T_\epsilon(\mathbf{x})$ becomes important. An example of such a specially-designed system is the ring linear coding scheme from [14], [15], [13], with A_k's being cosets of some left (right) ideal in a ring.

Lemma II.6. *Given an irreducible Markov chain $\mathscr{M}=\left\{X^{(n)}\right\}$ with finite state space $\mathscr{X}$, transition matrix $\mathbf{P}$ and invariant distribution $\pi=[p_j]_{j\in\mathscr{X}}$. Let $\coprod_{k=1}^m A_k$ be any partition of $\mathscr{X}$. For any $\eta>0$, there exist $\epsilon_0>0$ and $N_0\in\mathbb{N}^+$, such that, $\forall\ \epsilon_0>\epsilon>0$, $\forall\ n>N_0$ and $\forall\ \mathbf{x}=\left[x^{(1)},x^{(2)},\cdots,x^{(n)}\right]\in\mathcal{S}_\epsilon(n,\mathbf{P})$,*

$$\begin{aligned}|S_\epsilon(\mathbf{x})| &< \exp_2\left\{n\left[\sum_{k=1}^m\sum_{j\in A_k}p_jH(\mathbf{S}_k|\pi_k)+\eta\right]\right\} &(2)\\ &=\exp_2\left\{n\left[H(\mathbf{S}|\pi)+\eta\right]\right\} &(3)\end{aligned}$$

where $\mathbf{S}_k$ is the stochastic complement of $\mathbf{P}_{A_k,A_k}$ in $\mathbf{P}$, $\pi_k=\dfrac{[p_i]_{i\in A_k}}{\sum_{j\in A_k}p_j}$ is the invariant distribution of $\mathbf{S}_k$ and

$$\mathbf{S}=\operatorname{diag}\left\{\{\mathbf{S}_k\}_{1\leq k\leq m}\right\}.$$

Proof: Let

$$\mathbf{x}_{A_k}=\left[x^{(n_1)},x^{(n_2)},x^{(n_{m_k})}\right]$$

be the subsequence of $\mathbf{x}$ formed by all those $x^{(l)}$'s belonging to A_k in the original ordering. Obviously, $\sum_{k=1}^m m_k=n$ and $\left|\dfrac{m_k}{n}-\sum_{j\in A_k}p_j\right|<|A_k|\,\epsilon+\dfrac{1}{n}$. For any $\mathbf{y}=\left[y^{(1)},y^{(2)},\cdots,y^{(n)}\right]\in S_\epsilon(\mathbf{x})$,

$$\mathbf{y}_{A_k}=\left[y^{(n_1)},y^{(n_2)},y^{(n_{m_k})}\right]\in A_k^{m_k}$$

is a strongly Markov ϵ-typical sequence of length m_k with respect to $\mathbf{S}_k$ by Proposition II.3, since $\mathbf{y}$ is Supremus ϵ-typical. Additionally, by the AEP of strongly Markov typicality, there exist $\epsilon_k>0$ and positive integer M_k such that the number of strongly Markov ϵ-typical sequences of length m_k is upper bounded by $\exp_2\{m_k\left[H(\mathbf{S}_k|\pi_k)+\eta/2\right]\}$ if $0<\epsilon<\epsilon_k$ and $m_k>M_k$. Therefore, if $0<\epsilon<\min_{1\leq k\leq m}\epsilon_k$ and $n>M=\max_{1\leq k\leq m}\left\{\dfrac{1+M_k}{\left|\sum_{j\in A_k}p_j-|A_k|\,\epsilon\right|}\right\}$ (this guarantees that $m_k>M_k$ for all $1\leq k\leq m$), then

$$\begin{aligned}|S_\epsilon(\mathbf{x})| &\leq\exp_2\left\{\sum_{k=1}^m m_k\left[H(\mathbf{S}_k|\pi_k)+\eta/2\right]\right\}\\ &=\exp_2\left\{n\left[\sum_{k=1}^m\frac{m_k}{n}H(\mathbf{S}_k|\pi_k)+\eta/2\right]\right\}.\end{aligned}$$

Furthermore, choose $0<\epsilon_0\leq\min_{1\leq k\leq m}\epsilon_k$ and $N_0\geq M$ such that $\dfrac{m_k}{n}<\sum_{j\in A_k}p_j+\dfrac{\eta}{2\sum_{k=1}^m H(\mathbf{S}_k|\pi_k)}$ for all $0<\epsilon<\epsilon_0$ and $n>N_0$ and $1\leq k\leq m$, we have

$$|S_\epsilon(\mathbf{x})|<\exp_2\left\{n\left[\sum_{k=1}^m\sum_{j\in A_k}p_jH(\mathbf{S}_k|\pi_k)+\eta\right]\right\},$$

(2) is established. Direct calculation yields (3). ■

Lemma II.7. *In Lemma II.6, define $\Gamma(x)=l\Leftrightarrow x\in A_l$. We have that*

$$\begin{aligned}|S_\epsilon(\mathbf{x})|\leq|T_\epsilon(\mathbf{x})|<\exp_2\Bigg\{n\Bigg[&H(\mathbf{P}|\pi)\\ &-\lim_{w\to\infty}\frac{1}{w}H\left(Y^{(w)},Y^{(w-1)},\cdots,Y^{(1)}\right)+\eta\Bigg]\Bigg\},\end{aligned}\qquad(4)$$

where $Y^{(w)}=\Gamma\left(X^{(w)}\right)$.

Proof: $|S_\epsilon(\mathbf{x})|\leq|T_\epsilon(\mathbf{x})|$ is trivial. Let

$$\overline{\mathbf{y}}=\left[\Gamma\left(x^{(1)}\right),\Gamma\left(x^{(2)}\right),\cdots,\Gamma\left(x^{(n)}\right)\right].$$

By definition,

$$\left[\Gamma\left(y^{(1)}\right),\Gamma\left(y^{(2)}\right),\cdots,\Gamma\left(y^{(n)}\right)\right]=\overline{\mathbf{y}},$$

for any $\mathbf{y}=\left[y^{(1)},y^{(2)},\cdots,y^{(n)}\right]\in S_\epsilon(\mathbf{x})$. $\mathbf{y}$ is *jointly typical* [16] with $\overline{\mathbf{y}}$ with respect to the process

$$\cdots,\begin{pmatrix}X^{(1)}\\Y^{(1)}\end{pmatrix},\begin{pmatrix}X^{(2)}\\Y^{(2)}\end{pmatrix},\cdots,\begin{pmatrix}X^{(n)}\\Y^{(n)}\end{pmatrix},\cdots$$

Therefore, there exist $\epsilon_0 > 0$ and $N_0 \in \mathbb{N}^+$, such that, $\forall\ \epsilon_0 > \epsilon > 0$ and $\forall\ n > N_0$,

$$\begin{aligned}|S_\epsilon(\mathbf{x})| <& \exp_2\left\{n\left[\lim_{w\to\infty}\frac{1}{w}H\left(X^{(w)}, X^{(w-1)}, \cdots, X^{(1)}\right)\right.\right.\\ &\left.\left.- \lim_{w\to\infty}\frac{1}{w}H\left(Y^{(w)}, Y^{(w-1)}, \cdots, Y^{(1)}\right) + \eta\right]\right\}\\ =& \exp_2\left\{n\left[H\left(\mathbf{P}|\pi\right)\right.\right.\\ &\left.\left.- \lim_{w\to\infty}\frac{1}{w}H\left(Y^{(w)}, Y^{(w-1)}, \cdots, Y^{(1)}\right) + \eta\right]\right\},\end{aligned}$$

where the equality follows from the fact that $\lim_{w\to\infty}\dfrac{1}{w}H\left(X^{(w)}, X^{(w-1)}, \cdots, X^{(1)}\right) = H\left(\mathbf{P}\,|\pi\right)$ because $\mathscr{M}$ is irreducible Markov. ■

Remark 2. From the proof, it is seen that Lemma II.6 (Lemma II.7) establishes an upper bound for $S_\epsilon(\mathbf{x})$ ($T_\epsilon(\mathbf{x})$) based on the Supremus typicality argument (classic typicality argument). One sees that evaluating the bound (2) is rather straightforward. However, (4) is in general very hard to assess, because the entropy rate of $\{Y^{(w)}\}$ is hard to evaluate. In particular, given the initial distribution unknown, it is likely that $\{Y^{(w)}\}$ is no longer Markov [17]. This is why [13] is able to draw the optimal conclusion of the ring linear coding scheme based on Lemma II.6, but not Lemma II.7, for compressing Markov sources with unknown initial distributions.

III. Supremus Typicality in the Weak Sense

From Section II, we can see that irreducibility is a recursive property of irreducible Markov process [12, Theorem 2.1]. This is the observation that leads to the discovery of the concept Supremus typicality in the strong sense. In fact, this recursive phenomenon is only a special realization of some more universal results in ergodic theory [8]. We will investigate Supremus typicality in the weak sense based on these results.

Careful readers might have foreseen the definition of Supremus typicality in the weak sense, but it is the proof of the related AEP that involves certain amount of background from ergodic theory.

A. Asymptotically Mean Stationary

Given a probability space $(\Omega, \mathscr{F}, \mu)$ and a measurable transformation $T : \Omega \to \Omega$ (not necessarily probability preserving), the dynamical system $(\Omega, \mathscr{F}, \mu, T)$ is said to be *asymptotically mean stationary* (*a.m.s.*) [4] [7] if there exists a measure $\overline{\mu}$ on $(\Omega, \mathscr{F})$ satisfying

$$\overline{\mu}(B) = \lim_{m\to\infty}\frac{1}{m}\sum_{i=0}^{m}\mu(T^{-i}B), \forall\ B \in \mathscr{F}.$$

Obviously, if $(\Omega, \mathscr{F}, \mu, T)$ is stationary, i.e. $\mu(B) = \mu(T^{-1}B)$, then it is a.m.s.. In addition, $(\Omega, \mathscr{F}, \mu, T)$ is said to be ergodic if $T^{-1}B = B \implies \mu(B) = 0$ or $\mu(B) = 1$.

[4] The a.m.s. condition is interesting because it is a sufficient and necessary condition for the Point-wise Ergodic Theorem to hold [7, Theorem 1].

Let $X : \Omega \to \mathscr{X}$ ($\mathscr{X}$ is always assumed to be finite from now on) be a measurable function. Then $\{X^{(n)}\} = \{X(T^n)\}$ defines a random process with state space $\mathscr{X}$. Actually, every random process can be realized as a process so defined. For instance, T can be taken as time shift, $\Omega = \prod_{i=-\infty}^{\infty}\mathscr{X}$, while X is the coordinate function

$$X : (\cdots, x^{(-1)}, x^{(0)}, x^{(1)}, \cdots) \mapsto x^{(0)}.$$

The joint distribution p is defined as

$$p\left(x^{(0)}, x^{(1)}, \cdots, x^{(n-1)}\right) = \mu\left(\bigcap_{i=0}^{n-1}T^{-i}\left(X^{-1}\left(x^{(i)}\right)\right)\right).$$

The random process $\{X^{(n)}\} = \{X(T^n)\}$ is said to be *a.m.s.* (stationary/ergodic) if $(\Omega, \mathscr{F}, \mu, T)$ is a.m.s. (stationary/ergodic).

By [7], if $\{X^{(n)}\} = \{X(T^n)\}$ is a.m.s. and ergodic, then the SMB Theorem holds. In exact terms,

$$-\frac{1}{n}\log p(X^{(0)}, X^{(1)}, \cdots, X^{(n-1)}) \to H \text{ with probability } 1,$$

where H is the entropy rate of $\{X^{(n)}\}$. Let

$$\begin{aligned}\mathcal{T}_\epsilon(n, \{X^{(n)}\}) = \Big\{ &\left(x^{(0)}, x^{(1)}, \cdots, x^{(n-1)}\right) \in \mathscr{X}^n \Big|\\ &n(H - \epsilon) < -\log p\left(x^{(0)}, x^{(1)}, \cdots, x^{(n-1)}\right) < n(H + \epsilon)\Big\}.\end{aligned}$$

$\mathcal{T}_\epsilon(n, \{X^{(n)}\})$ is known to be the set of classic typical sequences in the weak sense. Moreover, the AEP holds for these sequences.

B. Induced Transformation

Definition III.1. A dynamical system $(\Omega, \mathscr{F}, \mu, T)$ is said to be *recurrent* (*conservative*) if $\mu\left(B - \bigcap_{i=0}^{\infty}\bigcup_{j=i}^{\infty}T^{-j}B\right) = 0, \forall\ B \in \mathscr{F}$.

Remark 3. Poincaré's Recurrence Theorem says that a stationary system is always recurrent. However, generally even an a.m.s. system is not necessarily recurrent [3].

The physical interpretation of recurrence (conservativeness) states that an event of positive probability is expected to repeat itself infinitely often during the lifetime of the dynamical system. Because of this physical meaning, recurrence is often assumed for ergodic systems in literature [18].

Given a recurrent system $(\Omega, \mathscr{F}, \mu, T)$ and $A \in \mathscr{F}$ ($\mu(A) > 0$), one can define a new transformation T_A on $(A_0, \mathscr{A}, \mu|_{\mathscr{A}})$, where $A_0 = A \cap \bigcap_{i=0}^{\infty}\bigcup_{j=i}^{\infty}T^{-j}A$ and $\mathscr{A} = \{A_0 \cap B | B \in \mathscr{F}\}$, such that

$$T_A(x) = T^{\psi_A^{(1)}(x)}(x), \forall\ x \in A_0,$$

where

$$\psi_A^{(1)}(x) = \min\left\{i \in \mathbb{N}^+ | T^i(x) \in A_0\right\}$$

is the first return time function. It is easy to see that $(A_0, \mathscr{A}, \mu|_{\mathscr{A}}, T_A)$ forms a new dynamical system. Such a transformation T_A is called an *induced transformation* of $(\Omega, \mathscr{F}, \mu, T)$ with respect to A [19].

Theorem III.2 ([8]). *If $(\Omega, \mathscr{F}, \mu, T)$ is recurrent a.m.s., then $(A_0, \mathscr{A}, \mu|_{\mathscr{A}}, T_A)$ is a.m.s. for all $A \in \mathscr{F}$ ($\mu(A) > 0$).*

Let $\{X^{(n)}\}$ be a random process with state space $\mathscr{X}$. A reduced process $\left\{X_{\mathscr{Y}}^{(k)}\right\}$ of $\{X^{(n)}\}$ with sub-state space $\mathscr{Y} \subseteq \mathscr{X}$ is defined to be $\left\{X_{\mathscr{Y}}^{(k)}\right\} = \left\{X^{(n_k)}\right\}$, where

$$n_k = \begin{cases} \min\{n \geq 0 | X^{(n)} \in \mathscr{Y}\}; & k = 0, \\ \min\{n > n_{k-1} | X^{(n)} \in \mathscr{Y}\}; & k > 0. \end{cases}$$

Assume that $\{X^{(n)}\} = \{X(T^n)\}$, where $(\Omega, \mathscr{F}, \mu, T)$ is recurrent a.m.s. ergodic and $X : \Omega \to \mathscr{X}$ is measurable, and let $A = X^{-1}(\mathscr{Y})$. It is easily seen that $\left\{X_{\mathscr{Y}}^{(k)}\right\}$ is essentially the random process $\left\{X\left(T_A^k\right)\right\}$ defined on the system

$$\left(A_0, A_0 \cap \mathscr{F}, \frac{1}{\mu(A)}\mu|_{A_0 \cap \mathscr{F}}, T_A\right),$$

which is also a.m.s. (by Theorem III.2) and ergodic (by [18, Proposition 1.5.2]). As a conclusion, the SMB Theorem holds for the reduced process $\left\{X_{\mathscr{Y}}^{(k)}\right\}$ as well [7].

C. Supremus Typicality in the Weak Sense and AEP

Based on the preliminaries introduced above, we state the definition of Supremus typicality in the weak sense as follows.

Definition III.3 (Supremus Typicality in the Weak Sense)**.** Let $\{X^{(n)}\}$ be a recurrent a.m.s. ergodic process with state space $\mathscr{X}$. A sequence $\mathbf{x} \in \mathscr{X}^n$ is said to be *Supremus ϵ-typical* with respect to $\{X^{(n)}\}$ for some $\epsilon > 0$, if $\forall\, \emptyset \neq \mathscr{Y} \subseteq \mathscr{X}$,

$$|\mathbf{x}_{\mathscr{Y}}| \left(H_{\mathscr{Y}} - \epsilon\right) < -\log p_{\mathscr{Y}}\left(\mathbf{x}_{\mathscr{Y}}\right) < |\mathbf{x}_{\mathscr{Y}}| \left(H_{\mathscr{Y}} + \epsilon\right),$$

where $p_{\mathscr{Y}}$ and $H_{\mathscr{Y}}$ are the joint distribution and entropy rate of the reduced process $\left\{X_{\mathscr{Y}}^{(k)}\right\}$ of $\{X^{(n)}\}$ with sub-state space $\mathscr{Y}$, respectively. The set of all Supremus ϵ-typical sequences with respect to $\{X^{(n)}\}$ in $\mathscr{X}^n$ is denoted by $\mathcal{S}_\epsilon(n, \{X^{(n)}\})$.

Obviously, Proposition II.3 is also valid for Supremus typical sequences in the weak sense. In addition,

$$\mathcal{S}_\epsilon(n, \{X^{(n)}\}) \subseteq \mathcal{T}_\epsilon(n, \{X^{(n)}\}).$$

Proposition III.4 (AEP of Weak Supremus Typicality)**.** *In Definition III.3,*

1) $\left|\mathcal{S}_\epsilon(n, \{X^{(n)}\})\right| < \exp_2\left[n\left(H_{\mathscr{X}} + \epsilon\right)\right]$*; and*
2) $\forall\, \eta > 0$, *there exists some positive integer N_0, such that*
$$\Pr\left\{\left[X^{(1)}, X^{(2)}, \cdots, X^{(n)}\right] \notin \mathcal{S}_\epsilon(n, \{X^{(n)}\})\right\} < \eta,$$
for all $n > N_0$.

Proof: First of all, $\left|\mathcal{S}_\epsilon(n, \{X^{(n)}\})\right| \leq \left|\mathcal{T}_\epsilon(n, \{X^{(n)}\})\right| < \exp_2\left[n\left(H_{\mathscr{X}} + \epsilon\right)\right]$. Let $\mathbf{X} = \left[X^{(1)}, X^{(2)}, \cdots, X^{(n)}\right]$. Then

$$\left\{\mathbf{X} \notin \mathcal{S}_\epsilon(n, \{X^{(n)}\})\right\} = \bigcup_{\emptyset \neq \mathscr{Y} \subseteq \mathscr{X}} \left\{\mathbf{X}_{\mathscr{Y}} \notin \mathcal{T}_\epsilon(n, \{X_{\mathscr{Y}}^{(k)}\})\right\}.$$

Assume that $(\Omega, \mathscr{F}, \mu, T)$ and X are the recurrent a.m.s. ergodic system and the measurable function define $\{X^{(n)}\}$, i.e. $\{X^{(n)}\} = \{X(T^n)\}$. For any non-empty $\mathscr{Y} \subseteq \mathscr{X}$, we have that $\left\{X_{\mathscr{Y}}^{(k)}\right\} = \left\{X\left(T_A^k\right)\right\}$, where $A = X^{-1}(\mathscr{Y})$ and T_A is an induced transformation of $(\Omega, \mathscr{F}, \mu, T)$ with respect to A. Furthermore, Theorem III.2 and [18, Proposition 1.5.2] guarantee that

$$\left(A_0, A_0 \cap \mathscr{F}, \frac{1}{\mu(A)}\mu|_{A_0 \cap \mathscr{F}}, T_A\right),$$

is a.m.s. ergodic. Consequently, the SMB Theorem holds [7]. This says, with probability 1,

$$-\frac{1}{n}\log p_{\mathscr{Y}}\left(X_{\mathscr{Y}}^{(0)}, X_{\mathscr{Y}}^{(1)}, \cdots, X_{\mathscr{Y}}^{(n-1)}\right) \to H_{\mathscr{Y}}.$$

This implies that there exists a positive integer $N_{\mathscr{Y}}$ such that

$$\Pr\left\{\mathbf{X}_{\mathscr{Y}} \notin \mathcal{T}_\epsilon(n, \{X_{\mathscr{Y}}^{(k)}\})\right\} < \frac{\eta}{2^{|\mathscr{X}|} - 1}, \forall\, n > N_{\mathscr{Y}}.$$

Let $N_0 = \max_{\emptyset \neq \mathscr{Y} \subseteq \mathscr{X}} N_{\mathscr{Y}}$. One easily concludes that

$$\Pr\left\{\mathbf{X} \notin \mathcal{S}_\epsilon(n, \{X^{(n)}\})\right\} < \eta, \forall\, n > N_0.$$

The statement is proved. ■

References

[1] C. E. Shannon and W. Weaver, *The mathematical theory of communication*. Urbana: University of Illinois Press, 1949.

[2] R. W. Yeung, *A First Course in Information Theory*. Springer, Mar. 2002.

[3] R. M. Gray, *Probability, Random Processes, and Ergodic Properties*. Springer, 2nd ed., Aug. 2009.

[4] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1984.

[5] B. McMillan, "The basic theorems of information theory," *The Annals of Mathematical Statistics*, vol. 24, pp. 196–219, 1953.

[6] L. Breiman, "The individual ergodic theorem of information theory," *The Annals of Mathematical Statistics*, vol. 31, pp. 809–801, 1960.

[7] R. M. Gray and J. C. Kieffer, "Asymptotically mean stationary measures," *The Annals of Probability*, vol. 8, pp. 962–973, Oct. 1980.

[8] S. Huang and M. Skoglund, *Induced Transformations of Recurrent A.M.S. Dynamical Systems*. KTH Royal Institute of Technology, October 2013. Available: http://people.kth.se/∼sheng11

[9] L. D. Davisson, G. Longo, and A. Sgarro, "The error exponent for the noiseless encoding of finite ergodic Markov sources," *IEEE Transactions on Information Theory*, vol. 27, pp. 431–438, July 1981.

[10] I. Csiszár, "The method of types," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2505–2523, 1998.

[11] J. R. Norris, *Markov Chains*. Cambridge University Press, July 1998.

[12] C. D. Meyer, "Stochastic complementation, uncoupling Markov chains, and the theory of nearly reducible systems," *SIAM Rev.*, vol. 31, pp. 240–272, June 1989.

[13] S. Huang and M. Skoglund, *Encoding Irreducible Markovian Functions of Sources: An Application of Supremus Typicality*. KTH Royal Institute of Technology, May 2013. Available: http://people.kth.se/∼sheng11

[14] S. Huang and M. Skoglund, *On Linear Coding over Finite Rings and Applications to Computing*. KTH Royal Institute of Technology, October 2012. Available: http://people.kth.se/∼sheng11

[15] S. Huang and M. Skoglund, "On achievability of linear source coding over finite rings," in *2013 IEEE International Symposium on Information Theory Proceedings (ISIT)*, pp. 1984–1988, 2013.

[16] T. M. Cover, "A proof of the data compression theorem of slepian and wolf for ergodic sources," *IEEE Transactions on Information Theory*, vol. 21, pp. 226–228, Mar. 1975.

[17] C. J. Burke and M. Rosenblatt, "A Markovian function of a Markov chain," *The Annals of Mathematical Statistics*, vol. 29, pp. 1112–1122, Dec. 1958.

[18] J. Aaronson, *An Introduction to Infinite Ergodic Theory*. Providence, R.I.: American Mathematical Society, 1997.

[19] S. Kakutani, "Induced measure preserving transformations," *Proceedings of the Imperial Academy*, vol. 19, no. 10, pp. 635–641, 1943.