Contributed Article in Neural Networks, Vol. 19, No. 1, pp. 62–75, 2005

The Asymptotic Equipartition Property in Reinforcement Learning and its Relation to Return Maximization

Kazunori Iwata[1] Kazushi Ikeda[2] Hideaki Sakai[2]

[1] Faculty of Information Sciences, Hiroshima City University
3-4-1 Ozuka-Higashi, Asa-Minami-Ku, Hiroshima, 731-3194, JAPAN
Phone: +81-82-830-1679 Fax: +81-82-830-1679
Email: kiwata@im.hiroshima-cu.ac.jp

[2] Department of Systems Science, Graduate School of Informatics,
Kyoto University
Yoshida-Honmachi, Sakyo-ku, Kyoto, 606-8501, JAPAN
Phone: +81-75-753-5501 Fax: +81-75-753-4755
Email: kazushi@i.kyoto-u.ac.jp, hsakai@i.kyoto-u.ac.jp

Abstract

We discuss an important property called the asymptotic equipartition property on empirical sequences in reinforcement learning. This states that the typical set of empirical sequences has probability nearly one, that all elements in the typical set are nearly equi-probable, and that the number of elements in the typical set is an exponential function of the sum of conditional entropies if the number of time steps is sufficiently large. The sum is referred to as stochastic complexity. Using the property we elucidate the fact that the return maximization depends on two factors, the stochastic complexity and a quantity depending on the parameters of environment. Here, the return maximization means that the best sequences in terms of expected return have probability one. We also examine the sensitivity of stochastic complexity, which is a qualitative guide in tuning the parameters of action-selection strategy, and show a sufficient condition for return maximization in probability.

Keywords: reinforcement learning, Markov decision process, information theory, asymptotic equipartition property, stochastic complexity, return maximization

1 Introduction

In information theory the weak law of large numbers is known as the asymptotic equipartition property (AEP) which was first stated in (Shannon, 1948) and then developed by the type method in (Csiszár & Körner, 1997; Csiszár, 1998). When a sequence of random variables is drawn many times, independently and according to an identical probability distribution, the AEP states that there exists the typical set of the sequences with probability nearly one, that all elements in the typical set are nearly equi-probable, and that the number of elements in the typical set is given by an exponential function of the entropy of the probability distribution. In addition, the number of elements in the typical set is quite small compared to the number of possible sequences. If the AEP also holds on empirical sequences generated from a Markov decision process (MDP) in reinforcement learning (RL), it facilitates the analysis of the learning process since most of our attention can be focused on the typical set of the empirical sequences. This leads us to the question of whether or not the AEP holds for an empirical sequence. The fact is that a similar AEP holds but it is more complicated than the original AEP. Using the type method, first we introduce an information-theoretic formulation for almost stationary ergodic MDPs in general and then describe the AEP that holds on the empirical sequences. From the AEP, we indicate the existence of an important factor called the stochastic complexity which consists of the sum of conditional entropies and elucidate that the return maximization is characterized by two factors, the stochastic complexity and a quantity which depends on the parameters of environment. Here, the return maximization means that the probability of best sequences that yield the maximal expected return goes to probability one. Also, useful knowledge for tuning the parameters of action-selection strategy is described by examining the sensitivity of the stochastic complexity. Furthermore, we show that the stochastic complexity is derived from the algorithmic complexity which was explored by Chaitin (Chaitin, 1977, 1987).

The organization of this paper is as follows. We introduce some notation and the type of empirical sequence in Section 2. Section 3 shows the main theorems associated with the AEP. Using the AEP we analyze the RL process in Section 4. Finally, we give some conclusions in Section 5. Appendices A and B are the related theorems to the AEP and those proofs, respectively.

2 Preliminaries

We concentrate on the discrete-time MDP with discrete states and actions in this paper. Let $\mathcal{S} \stackrel{\text{def}}{=} \{s_1, s_2, \ldots, s_I\}$ be the finite set of states of the environment, $\mathcal{A} \stackrel{\text{def}}{=} \{a_1, a_2, \ldots, a_J\}$ be the finite set of actions, and $\mathbb{R}_0 \stackrel{\text{def}}{=} \{r_1, r_2, \ldots, r_K\} \subset \mathbb{R}$ be the finite set of rewards which are discrete real numbers. Notice that $|\mathcal{S}| = I$, $|\mathcal{A}| = J$, and $|\mathbb{R}_0| = K$. We assume that elements in these sets are recognized without error by the learner, hereinafter called the agent. We denote a time step by t. The stochastic variables of state, action, and reward at time step t ($t = 1, 2, \ldots$) are written as $s(t)$, $a(t)$, and $r(t)$, respectively. The agent improves the policy by observing one-by-one each element of the empirical sequence that is generated by the interactions between the agent and the environment, as shown in Figure 1.

[Figure 1 about here.]

Now let us consider the empirical sequence of n time steps,

$$s(1), a(1), s(2), r(2), a(2), \ldots, s(n), r(n), a(n), r(n+1).$$

Let $r(n+1) = r(1)$ for notational convenience and let $\boldsymbol{x} = \{s(t), a(t), r(t)\}_{t=1}^{n}$ denote the empirical sequence of n time steps. The state sequence, action sequence, and reward sequence of the empirical sequence $\boldsymbol{x} \in (\mathcal{S} \times \mathcal{A} \times \mathbb{R}_0)^n$ are denoted by $\boldsymbol{s} = \{s(t)\}_{t=1}^{n}$, $\boldsymbol{a} = \{a(t)\}_{t=1}^{n}$, and $\boldsymbol{r} = \{r(t)\}_{t=1}^{n}$, respectively. We use the term return to express the sum of rewards.

Let $q_i \stackrel{\text{def}}{=} \Pr(s(1) = s_i)$ be the initial probability distribution and $\mathbf{q} \stackrel{\text{def}}{=} \{q_1, q_2, \ldots, q_I\}$ where $q_i > 0$ for all i. The empirical sequence is drawn according to an ergodic MDP specified by the following two conditional probability distribution matrices. Henceforth, the conditional probability distribution matrix is simply called the matrix. The policy matrix which the agent determines is an $I \times J$ matrix defined by

$$\mathbf{\Gamma}^{\pi} \stackrel{\text{def}}{=} \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1J} \\ p_{21} & p_{22} & \cdots & p_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ p_{I1} & p_{I2} & \cdots & p_{IJ} \end{pmatrix} = \begin{pmatrix} \mathbf{P}_{(1)} \\ \mathbf{P}_{(2)} \\ \vdots \\ \mathbf{P}_{(I)} \end{pmatrix}, \tag{1}$$

where $p_{ij} \stackrel{\text{def}}{=} \Pr(a(t) = a_j | s(t) = s_i)$. According to this matrix, the agent selects an action in a state at each time step. Note that $\mathbf{\Gamma}^{\pi}$ is actually time-varying because the agent improves the policy in the process of RL. However, $\mathbf{\Gamma}^{\pi}$ tends to be constant as the policy goes to be optimal by the learning. The state transition matrix of the environment is an $IJ \times IK$ matrix defined by

$$\mathbf{\Gamma}^{\mathrm{T}} \stackrel{\text{def}}{=} \begin{pmatrix} p_{1111} & p_{1112} & \cdots & p_{11IK} \\ p_{1211} & p_{1212} & \cdots & p_{12IK} \\ \vdots & \vdots & \ddots & \vdots \\ p_{IJ11} & p_{IJ12} & \cdots & p_{IJIK} \end{pmatrix} = \begin{pmatrix} \mathbf{P}_{(11)} \\ \mathbf{P}_{(12)} \\ \vdots \\ \mathbf{P}_{(IJ)} \end{pmatrix}, \tag{2}$$

where $p_{iji'k} \stackrel{\text{def}}{=} \Pr(s(t+1) = s_{i'}, r(t+1) = r_k | s(t) = s_i, a(t) = a_j)$. The agent does not know the matrix $\mathbf{\Gamma}^{\mathrm{T}}$ of the environment but can estimate it by observing the results for an action. We assume that $\mathbf{\Gamma}^{\mathrm{T}}$ is constant and that for simplicity of analysis $\mathbf{\Gamma}^{\pi}$ is temporarily fixed for n time steps where n is sufficiently large. For notational simplicity we define $\mathbf{\Gamma} \stackrel{\text{def}}{=} (\mathbf{\Gamma}^{\pi}, \mathbf{\Gamma}^{\mathrm{T}})$. Since MDPs are characterized by the finite sets, the initial probability distribution, and the matrices, we denote the MDP by $\mathrm{M}(\mathcal{S}, \mathcal{A}, \mathbb{R}_0, \mathbf{q}, \mathbf{\Gamma})$.

2.1 Type of Empirical Sequence

Let n_i ($n_i \leq n$) denote the number of times that a state $s_i \in \mathcal{S}$ occurs in the empirical sequence of n time steps, $\boldsymbol{x} = (\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{r}) \in (\mathcal{S} \times \mathcal{A} \times \mathbb{R}_0)^n$. In a similar manner, let n_{ij} ($n_{ij} \leq n_i$) be the number of occurrences of t such that $(s(t), a(t)) = (s_i, a_j) \in \mathcal{S} \times \mathcal{A}$ in the empirical sequence. With an additional "cyclic" convention that $s(n)$, $a(n)$, and $r(n+1) = r(1)$ precede $s(1)$, $a(1)$, and $r(2)$, let $n_{iji'k}$ ($n_{iji'k} \leq n_{ij}$) denote the number of occurrences of t such that $(s(t), a(t), s(t+1), r(t+1)) = (s_i, a_j, s_{i'}, r_k) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times \mathbb{R}_0$ in the empirical sequence. Note that the cyclic convention is for simplicity of development. The discussions in this paper strictly hold even if we do not assume this convention. The relationship among the non-negative numbers n, n_i, n_{ij}, and $n_{iji'k}$ is expressed as

$$n = \sum_{i=1}^{I} n_i = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{i'=1}^{I} \sum_{k=1}^{K} n_{iji'k}. \tag{3}$$

Now we define the type of $s_i \in \mathcal{S}$ by

$$f_i = \frac{n_i}{n}. \tag{4}$$

The type is generally called the empirical distribution (Han & Kobayashi, 2002, p. 42) because we can regard each sequence as a sample from a stochastic process. Also, the joint type of $(s_i, a_j) \in \mathcal{S} \times \mathcal{A}$ is defined as

$$f_{ij} = \frac{n_{ij}}{n}. \tag{5}$$

Let us denote all the types and the joint types by

$$\boldsymbol{F}_{\mathcal{S}} \stackrel{\text{def}}{=} (f_1, f_2, \ldots, f_I), \tag{6}$$

and

$$\boldsymbol{F}_{\mathcal{SA}} \stackrel{\text{def}}{=} \begin{pmatrix} f_{11} & f_{12} & \cdots & f_{1J} \\ f_{21} & f_{22} & \cdots & f_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ f_{I1} & f_{I2} & \cdots & f_{IJ} \end{pmatrix}, \tag{7}$$

respectively. In this case we say that the state sequence $\boldsymbol{s}$ and the state-action sequence $(\boldsymbol{s}, \boldsymbol{a})$ have the type $\boldsymbol{F}_{\mathcal{S}}$ and the joint type $\boldsymbol{F}_{\mathcal{SA}}$, respectively.

Conditional Type Relative to Policy If $n_i > 0$ for all i, then the conditional type g_{ij} of $(s_i, a_j) \in \mathcal{S} \times \mathcal{A}$ given a state sequence $\boldsymbol{s} \in \mathcal{S}^n$ is defined as

$$n_{ij} \stackrel{\text{def}}{=} g_{ij} n_i. \tag{8}$$

However, if there exists i such that $n_i = 0$, then we can not uniquely determine the conditional type (see Example 2.1). To avoid such a case, we consider the set of action sequences given any state sequence $\boldsymbol{s}$ having the type $\boldsymbol{F}_{\mathcal{S}}$ and an $I \times J$ matrix $\boldsymbol{\Phi}^\pi : \mathcal{S} \to \mathcal{A}$ expressed as

$$\boldsymbol{\Phi}^\pi \stackrel{\text{def}}{=} \begin{pmatrix} g_{11} & g_{12} & \cdots & g_{1J} \\ g_{21} & g_{22} & \cdots & g_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ g_{I1} & g_{I2} & \cdots & g_{IJ} \end{pmatrix} = \begin{pmatrix} \boldsymbol{G}_{(1)} \\ \boldsymbol{G}_{(2)} \\ \vdots \\ \boldsymbol{G}_{(I)} \end{pmatrix}. \tag{9}$$

In short, n_{ij} is decided by n_i and g_{ij} for every i, j. The set of action sequences, which is uniquely determined in this way, is referred to as $\boldsymbol{\Phi}^\pi$-shell (Csiszár & Körner, 1997, p. 31) of $\boldsymbol{s}$ and denoted by $\mathcal{C}^n(\boldsymbol{\Phi}^\pi, \boldsymbol{s})$. The entire set of possible matrices $\boldsymbol{\Phi}^\pi$ for any state sequence with the type $\boldsymbol{F}_{\mathcal{S}}$ is simply written as $\boldsymbol{\Lambda}_n^\pi$.

Example 2.1 *Let $I = J = 2$, the state sequence $\boldsymbol{s} = (s_1, s_1, s_1, s_1) \in \mathcal{S}^4$, and the action sequence $\boldsymbol{a} = (a_1, a_1, a_2, a_2) \in \mathcal{A}^4$. Then, from the definition of* (8) *we obtain $g_{11} = g_{12} = 1/2$. Also, because of $n_2 = 0$, letting $g_{21} = \omega$ where $0 \leq \omega \leq 1$ we have $g_{21} = \omega$ and $g_{22} = 1 - \omega$. Therefore, we can not uniquely determine the conditional type.*

Example 2.2 ($\boldsymbol{\Phi}^\pi$-shell) *Let $I = J = 2$, again. For the state sequence $\boldsymbol{s} = (s_1, s_1, s_1, s_2) \in \mathcal{S}^4$ with the type $\boldsymbol{F}_{\mathcal{S}} = (3/4, 1/4)$ and the matrix,*

$$\boldsymbol{\Phi}^\pi = \begin{pmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{pmatrix} = \begin{pmatrix} 2/3 & 1/3 \\ 0 & 1 \end{pmatrix}, \tag{10}$$

the $\boldsymbol{\Phi}^\pi$-shell of $\boldsymbol{s}$ is $\mathcal{C}^4(\boldsymbol{\Phi}^\pi, \boldsymbol{s}) = \{(a_1, a_1, a_2, a_2), (a_1, a_2, a_1, a_2), (a_2, a_1, a_1, a_2)\}$.

Conditional Markov Type Relative to State Transition In a slightly different manner we need to deal with the conditional Markov type[1]. We consider the set of state-reward sequences such that the

[1]For Markov type, see (Davisson, Longo, & Sgarro, 1981).

joint type is $\boldsymbol{F}_{\mathcal{SA}}$ given any action sequence and an $IJ \times IK$ matrix $\boldsymbol{\Phi}^{\mathrm{T}} : \mathcal{S} \times \mathcal{A} \to \mathcal{S} \times \mathbb{R}_0$ designated by

$$\boldsymbol{\Phi}^{\mathrm{T}} \stackrel{\mathrm{def}}{=} \begin{pmatrix} g_{1111} & g_{1112} & \cdots & g_{11IK} \\ g_{1211} & g_{1212} & \cdots & g_{12IK} \\ \vdots & \vdots & \ddots & \vdots \\ g_{IJ11} & g_{IJ12} & \cdots & g_{IJIK} \end{pmatrix} = \begin{pmatrix} \boldsymbol{G}_{(11)} \\ \boldsymbol{G}_{(12)} \\ \vdots \\ \boldsymbol{G}_{(IJ)} \end{pmatrix}. \tag{11}$$

The set of state-reward sequences is referred to as $\boldsymbol{\Phi}^{\mathrm{T}}$-shell and denoted by $\mathcal{C}^n(\boldsymbol{\Phi}^{\mathrm{T}}, \boldsymbol{F}_{\mathcal{SA}})$. The entire set of possible matrices $\boldsymbol{\Phi}^{\mathrm{T}}$ such that the joint type is $\boldsymbol{F}_{\mathcal{SA}}$ for any action sequence is simply written as $\boldsymbol{\Lambda}_n^{\mathrm{T}}$.

For simplicity, we define $\boldsymbol{\Phi} \stackrel{\mathrm{def}}{=} (\boldsymbol{\Phi}^{\pi}, \boldsymbol{\Phi}^{\mathrm{T}})$ and $\boldsymbol{\Lambda}_n \stackrel{\mathrm{def}}{=} \boldsymbol{\Lambda}_n^{\pi} \times \boldsymbol{\Lambda}_n^{\mathrm{T}}$. The set of empirical sequences that consists of the $\boldsymbol{\Phi}^{\pi}$-shell and $\boldsymbol{\Phi}^{\mathrm{T}}$-shell is called the $\boldsymbol{\Phi}$-shell and denoted by $\mathcal{C}^n(\boldsymbol{\Phi}, \boldsymbol{F}_{\mathcal{S}}, \boldsymbol{F}_{\mathcal{SA}})$. The structure of the $\boldsymbol{\Phi}$-shell is depicted in Figure 2. When a joint-type $\boldsymbol{F}_{\mathcal{SA}}$ and a matrix $\boldsymbol{\Phi}^{\mathrm{T}}$ are given, the $\boldsymbol{\Phi}^{\mathrm{T}}$-shell having the type $\boldsymbol{F}_{\mathcal{S}}$ is uniquely determined and then the combination of each element in the $\boldsymbol{\Phi}^{\mathrm{T}}$-shell and a matrix $\boldsymbol{\Phi}^{\pi}$ produces the $\boldsymbol{\Phi}^{\pi}$-shell. Therefore, the $\boldsymbol{\Phi}$-shell is uniquely determined. Notice that

$$|\mathcal{C}^n(\boldsymbol{\Phi}, \boldsymbol{F}_{\mathcal{S}}, \boldsymbol{F}_{\mathcal{SA}})| = \sum_{(\boldsymbol{s}', \boldsymbol{r}') \in \mathcal{C}^n(\boldsymbol{\Phi}^{\mathrm{T}}, \boldsymbol{F}_{\mathcal{SA}})} |\mathcal{C}^n(\boldsymbol{\Phi}^{\pi}, \boldsymbol{s}')|. \tag{12}$$

In this case we write that the empirical sequence has the conditional type matrix $\boldsymbol{\Phi}$.

[Figure 2 about here.]

2.2 V-typical and W-typical sequences

In order to prove the AEP on empirical sequences, we have to introduce the $\mathbf{V}$-typical sequence with respect to the state sequences and the $\mathbf{W}$-typical sequences with respect to the state-action sequences.

Definition 2.1 (V-typical and W-typical sequences) *We assume the existence of the following two unique stationary probability distributions,*

$$\mathbf{V} \stackrel{\mathrm{def}}{=} (v_1, v_2, \ldots, v_I), \tag{13}$$

$$\mathbf{W} \stackrel{\mathrm{def}}{=} \begin{pmatrix} w_{11} & w_{12} & \ldots & w_{1J} \\ w_{21} & w_{22} & \ldots & w_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ w_{I1} & w_{I2} & \ldots & w_{IJ} \end{pmatrix}, \tag{14}$$

and assume that $\boldsymbol{F}_{\mathcal{S}}$ and $\boldsymbol{F}_{\mathcal{SA}}$ tend to $\mathbf{V}$ and $\mathbf{W}$ as $n \to \infty$, respectively. The stationary probability distributions are uniquely determined by the MDP, $\mathrm{M}(\mathcal{S}, \mathcal{A}, \mathbb{R}_0, \mathbf{q}, \boldsymbol{\Gamma})$. In this case, there exists a sequence of positive κ_n such that $\kappa_n \to 0$ as $n \to \infty$, and if the type $\boldsymbol{F}_{\mathcal{S}}$ of a state sequence $\boldsymbol{s} \in \mathcal{S}^n$ satisfies

$$\mathrm{D}(\boldsymbol{F}_{\mathcal{S}} \| \mathbf{V}) = \sum_{i=1}^{I} f_i \log \frac{f_i}{v_i} \leq \kappa_n, \tag{15}$$

then we call the state sequence a $\mathbf{V}$-typical sequence. The set of $\mathbf{V}$-typical sequences is denoted by $\mathcal{C}_{\kappa_n}^n(\mathbf{V}) \stackrel{\mathrm{def}}{=} \{\boldsymbol{s} \in \mathcal{S}^n | \mathrm{D}(\boldsymbol{F}_{\mathcal{S}} \| \mathbf{V}) \leq \kappa_n\}$. In a similar manner, there exists a sequence of positive ξ_n such that $\xi_n \to 0$ as $n \to \infty$, and if

$$\mathrm{D}(\boldsymbol{F}_{\mathcal{SA}} \| \mathbf{W}) = \sum_{i=1}^{I} \sum_{j=1}^{J} f_{ij} \log \frac{f_{ij}}{w_{ij}} \leq \xi_n \tag{16}$$

holds, then the state-action sequences $(\boldsymbol{s}, \boldsymbol{a}) \in (\mathcal{S} \times \mathcal{A})^n$ *are referred to as* $\mathbf{W}$*-typical sequences. We define the set of* $\mathbf{W}$*-typical sequences as* $\mathcal{C}^n_{\xi_n}(\mathbf{W}) \stackrel{\text{def}}{=} \{(\boldsymbol{s}, \boldsymbol{a}) \in (\mathcal{S} \times \mathcal{A})^n | \mathrm{D}(\boldsymbol{F}_{\mathcal{SA}} \| \mathbf{W}) \leq \xi_n\}$.

In the rest of this section, we will introduce a few basic conventions in information theory (Cover & Thomas, 1991, Chapter 2). Let us use the convention that $0 \log 0 = 0$ henceforth. The function H indicates the entropy. For instance, we write the entropy of $\mathbf{P}_{(i)}$ in (1) for any i as

$$\mathrm{H}(\mathbf{P}_{(i)}) = -\sum_{j=1}^{J} p_{ij} \log p_{ij}, \tag{17}$$

and describe its conditional entropy given $\mathbf{V}$ as

$$\mathrm{H}(\mathbf{\Gamma}^\pi | \mathbf{V}) = \sum_{i=1}^{I} v_i \mathrm{H}(\mathbf{P}_{(i)}). \tag{18}$$

Also, as used in (15) and (16), the divergence is designated by the function D. The divergence between $\mathbf{\Phi}^\pi$ and $\mathbf{\Gamma}^\pi$ given $\boldsymbol{F}_{\mathcal{S}}$ is denoted as

$$\mathrm{D}(\mathbf{\Phi}^\pi \| \mathbf{\Gamma}^\pi | \boldsymbol{F}_{\mathcal{S}}) = \sum_{i=1}^{I} f_i \mathrm{D}(\boldsymbol{G}_{(i)} \| \mathbf{P}_{(i)}), \tag{19}$$

where

$$\mathrm{D}(\boldsymbol{G}_{(i)} \| \mathbf{P}_{(i)}) = \sum_{j=1}^{J} g_{ij} \log \frac{g_{ij}}{p_{ij}}. \tag{20}$$

3 Asymptotic Equipartition Property

In this section, it is elucidated that the empirical sequences generated from almost stationary ergodic MDPs have the AEP. Now we are in a position to give the definitions of the typical sequence and the typical set of empirical sequences, which will lead us to show that the AEP holds on empirical sequences.

Definition 3.1 ($\mathbf{\Gamma}$-typical sequence and $\mathbf{\Gamma}$-typical set) *If the matrix* $\mathbf{\Phi} \in \mathbf{\Lambda}_n$ *of the conditional types with respect to an empirical sequence* $\boldsymbol{x} = (\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{r}) \in (\mathcal{S} \times \mathcal{A} \times \mathbb{R}_0)^n$ *satisfies*

$$\mathrm{D}(\mathbf{\Phi}^\pi \| \mathbf{\Gamma}^\pi | \boldsymbol{F}_{\mathcal{S}}) + \mathrm{D}(\mathbf{\Phi}^{\mathrm{T}} \| \mathbf{\Gamma}^{\mathrm{T}} | \boldsymbol{F}_{\mathcal{SA}}) \leq \lambda_n, \tag{21}$$

for any matrix $\mathbf{\Gamma}$ *and positive number* λ_n*, then the empirical sequence is called a* $\mathbf{\Gamma}$*-typical sequence. The set of such empirical sequences is also called the* $\mathbf{\Gamma}$*-typical set and denoted by* $\mathcal{C}^n_{\lambda_n}(\mathbf{\Gamma})$*. That is,* $\mathcal{C}^n_{\lambda_n}(\mathbf{\Gamma})$ *is given by*

$$\mathcal{C}^n_{\lambda_n}(\mathbf{\Gamma}) \stackrel{\text{def}}{=} \bigcup_{\substack{\mathbf{\Phi} \in \mathbf{\Lambda}_n: \\ \mathrm{D}(\mathbf{\Phi}^\pi \| \mathbf{\Gamma}^\pi | \boldsymbol{F}_{\mathcal{S}}) + \mathrm{D}(\mathbf{\Phi}^{\mathrm{T}} \| \mathbf{\Gamma}^{\mathrm{T}} | \boldsymbol{F}_{\mathcal{SA}}) \leq \lambda_n}} \mathcal{C}^n(\mathbf{\Phi}, \boldsymbol{F}_{\mathcal{S}}, \boldsymbol{F}_{\mathcal{SA}}). \tag{22}$$

Figure 3 illustrates the concept of Definition 3.1. The matrix $\mathbf{\Phi}$ of the $\mathbf{\Gamma}$-typical sequence exists in the neighborhood of $\mathbf{\Gamma}$, shown by the shaded circle on the manifold spanned by $\mathbf{\Gamma}$.

[Figure 3 about here.]

From the theorems, presented in Appendix A, we can derive the following three theorems regarding the AEP on empirical sequences. We begin with the theorem similar to (Wolfowitz, 1978).

Theorem 3.1 (Probability of the $\mathbf{\Gamma}$-typical set) *If $\lambda_n \to 0$ as $n \to \infty$ and λ_n satisfies*

$$\lambda_n - \frac{(IJ + I^2JK)\log(n+1) + \log I - \log \nu}{n} > 0, \tag{23}$$

where

$$\nu \stackrel{\text{def}}{=} \min_{1 \le i,i' \le I, 1 \le j \le J, 1 \le k \le K : p_{iji'k} > 0} p_{iji'k}, \tag{24}$$

there exists a sequence $\{\varepsilon_n(I,J,K,\lambda_n)\}$ such that $\varepsilon_n(I,J,K,\lambda_n) \to 0$ and then

$$\Pr\left(\mathcal{C}^n_{\lambda_n}(\mathbf{\Gamma})\right) = 1 - \varepsilon_n(I,J,K,\lambda_n). \tag{25}$$

Note that $n\lambda_n \to \infty$ because of (23). The proof is given in Appendix B.4. This theorem implies that the probability of the $\mathbf{\Gamma}$-typical set asymptotically goes to one independently of the underlying probabilistic structures, $\mathbf{\Gamma}^\pi$ and $\mathbf{\Gamma}^{\mathrm{T}}$. Next, the following theorem indicates the fact that all elements in the $\mathbf{\Gamma}$-typical set are nearly equi-probable.

Theorem 3.2 (Equi-probability of the $\mathbf{\Gamma}$-typical sequence) *If $\boldsymbol{s} \in \mathcal{C}^n_{\kappa_n}(\mathbf{V})$, $(\boldsymbol{s},\boldsymbol{a}) \in \mathcal{C}^n_{\xi_n}(\mathbf{W})$, $\boldsymbol{x} \in \mathcal{C}^n_{\lambda_n}(\mathbf{\Gamma})$ such that $\kappa_n \to 0$, $\xi_n \to 0$, $\lambda_n \to 0$ as $n \to \infty$, then there exists a sequence $\{\rho_n(I,J,K,\kappa_n,\xi_n,\lambda_n)\}$ such that*

$$\rho_n(I,J,K,\kappa_n,\xi_n,\lambda_n) \to 0.$$

Then,

$$\frac{\log \nu}{n} - \rho_n \le -\frac{1}{n}\log \Pr(\boldsymbol{x}) - \left\{\mathrm{H}(\mathbf{\Gamma}^\pi|\mathbf{V}) + \mathrm{H}(\mathbf{\Gamma}^{\mathrm{T}}|\mathbf{W})\right\} \le -\frac{\log \mu}{n} + \lambda_n + \rho_n, \tag{26}$$

where ν is given in (24) and

$$\mu \stackrel{\text{def}}{=} \min_{1 \le i \le I : q_i > 0} q_i. \tag{27}$$

This theorem is proved in Appendix B.5. Finally, we present the theorem which implies that the number of elements in the $\mathbf{\Gamma}$-typical set is written as an exponential function of the sum of the conditional entropies.

Theorem 3.3 (Bound of the number of the $\mathbf{\Gamma}$-typical sequences) *If $\boldsymbol{s} \in \mathcal{C}^n_{\kappa_n}(\mathbf{V})$, $(\boldsymbol{s},\boldsymbol{a}) \in \mathcal{C}^n_{\xi_n}(\mathbf{W})$, $\boldsymbol{x} \in \mathcal{C}^n_{\lambda_n}(\mathbf{\Gamma})$ such that $\kappa_n \to 0$, $\xi_n \to 0$, $\lambda_n \to 0$ as $n \to \infty$, then there exist two sequences, $\{\zeta_n(I,J,K,\kappa_n,\xi_n,\lambda_n)\}$ and $\{\eta_n(I,J,K,\kappa_n,\xi_n,\lambda_n)\}$, such that*

$$\zeta_n(I,J,K,\kappa_n,\xi_n,\lambda_n) \to 0, \qquad \eta_n(I,J,K,\kappa_n,\xi_n,\lambda_n) \to 0,$$

respectively. Then, the number of elements in the $\mathbf{\Gamma}$-typical set is bounded by

$$\exp\left[n\left\{\mathrm{H}(\mathbf{\Gamma}^\pi|\mathbf{V}) + \mathrm{H}(\mathbf{\Gamma}^{\mathrm{T}}|\mathbf{W}) - \zeta_n\right\}\right] \le |\mathcal{C}^n_{\lambda_n}(\mathbf{\Gamma})| \le \exp\left[n\left\{\mathrm{H}(\mathbf{\Gamma}^\pi|\mathbf{V}) + \mathrm{H}(\mathbf{\Gamma}^{\mathrm{T}}|\mathbf{W}) + \eta_n\right\}\right]. \tag{28}$$

The proof is given in Appendix B.6. The ratio of the number of $\mathbf{\Gamma}$-typical sequences to that of all empirical sequences $\boldsymbol{x} \in (\mathcal{S} \times \mathcal{A} \times \mathbb{R}_0)^n$ of n time steps is

$$\frac{|\mathcal{C}^n_{\lambda_n}(\mathbf{\Gamma})|}{(IJK)^n} \le \exp\left[n\left\{\mathrm{H}(\mathbf{\Gamma}^\pi|\mathbf{V}) + \mathrm{H}(\mathbf{\Gamma}^{\mathrm{T}}|\mathbf{W}) + \eta_n - \log I - \log J - \log K\right\}\right] \to 0, \tag{29}$$

as $n \to \infty$, when the probability distributions of $\mathbf{\Gamma}^\pi$ and $\mathbf{\Gamma}^{\mathrm{T}}$ are not uniform distributions, that is,

$$\mathrm{H}(\mathbf{\Gamma}^\pi|\mathbf{V}) < \log I, \tag{30}$$

$$\mathrm{H}(\mathbf{\Gamma}^{\mathrm{T}}|\mathbf{W}) < \log J + \log K. \tag{31}$$

Hence, we can say that the $\mathbf{\Gamma}$-typical set is quite small in comparison to the set of all empirical sequences. Nonetheless, their existence is important enough because the total probability is almost one.

Remark 3.1 *The equation* (28) *shows*

$$|\mathcal{C}_{\lambda_n}^n(\mathbf{\Gamma})| \doteq \exp\left[n\left\{\mathrm{H}(\mathbf{\Gamma}^\pi|\mathbf{V}) + \mathrm{H}(\mathbf{\Gamma}^{\mathrm{T}}|\mathbf{W})\right\}\right], \tag{32}$$

where the notation $\doteq$ *indicates that both sides are equal to the first order in the exponent, namely,*

$$\lim_{n\to\infty}\frac{1}{n}\log|\mathcal{C}_{\lambda_n}^n(\mathbf{\Gamma})| = \lim_{n\to\infty}\frac{1}{n}\log\exp\left[n\left\{\mathrm{H}(\mathbf{\Gamma}^\pi|\mathbf{V}) + \mathrm{H}(\mathbf{\Gamma}^{\mathrm{T}}|\mathbf{W})\right\}\right], \tag{33}$$

(Cover & Thomas, 1991, p. 55).

4 The Role of Stochastic Complexity in Reinforcement Learning

The agent learns the optimal policy via return maximization (RM) in RL. A number of studies have been made on the analysis of the process of RM (Jaakkola, Jordan, & Singh, 1994; Kushner & Yin, 1997; Singh, Jaakkola, Littman, & Szepesvári, 2000), but most of the studies focus on concrete stochastic approximation methods such as temporal difference (TD) learning. The aim here is to explore a more general mechanism of RM, how the probability of the subset of best sequences in terms of expected return is maximized, from a viewpoint of Shannon's ideas. In this section, we state the existence of an important factor called the stochastic complexity and show new insights about the role of the stochastic complexity in RM. We also discuss a sensitivity helpful in tuning the parameters of action-selection (AS) strategy and exhibit a sufficient condition for RM. We first give a review of the TD learning and typical AS strategies.

4.1 Temporal Difference Learning and Action Selection Strategy

Let Q_{ij} denote the estimate of an action-value function (Sutton & Barto, 1998, Chapter 3) with respect to a state-action pair $(s_i, a_j) \in \mathcal{S}\times\mathcal{A}$. Let $\mathcal{A}_i$ be the set of indices of actions available in a state $s_i \in \mathcal{S}$. The TD learning is an iterative approximation method to directly update the estimate of the action-value function from an observed event, without explicitly treating the matrix $\mathbf{\Gamma}^{\mathrm{T}}$ of the environment. We introduce the one-step version of Q-learning (Watkins & Dayan, 1992), here. At each time step t, for an observed one-step event $(s(t), a(t), s(t+1), r(t+1)) = (s_i, a_j, s_{i'}, r_k)$, the estimate Q_{ij} is updated by

$$Q_{ij} \leftarrow Q_{ij} + \alpha_t \delta Q_{iji'}, \tag{34}$$

where α_t is a learning rate at time step t and

$$\delta Q_{iji'} = r_k + \gamma \max_{j'\in\mathcal{A}_{i'}} Q_{i'j'} - Q_{ij}, \tag{35}$$

where γ denotes the discount factor that controls the relative importance of an immediate reward and delayed rewards. The learning rate α_t, where $0 \le \alpha_t \le 1$, is gradually decreased with respect to t such that the trajectory of the mean ordinary differential equation of Q_{ij} has a limit point. Under certain conditions (Dayan, 1992), it was proved that all Q_{ij} converge to the expected values with probability one. The convergence theorem was extended to more general versions using the stochastic approximation method in (Jaakkola et al., 1994; Tsitsiklis, 1994).

Next, we review the following two AS strategies that have been employed in many cases.

Softmax Method The softmax method (Sutton & Barto, 1998, Chapter 2) is the most popular strategy and is also termed the Boltzmann method when the exponential function is used. Recall that p_{ij} denotes the probability that the agent chooses an action a_j in a state s_i. The policy probability is defined as

$$p_{ij} \stackrel{\text{def}}{=} \pi\left(\beta, Q_{ij}\right) = \frac{\exp(\beta Q_{ij})}{\mathrm{Z}_i(\beta)}, \tag{36}$$

where the partition function is

$$\mathrm{Z}_i(\beta) \stackrel{\text{def}}{=} \sum_{j' \in \mathcal{A}_i} \exp(\beta Q_{ij'}). \tag{37}$$

The parameter β is gradually increased as $n \to \infty$ to promote the acceptance of actions which may produce a good return. Let us denote the value of β at time step n by β_n.

ϵ-greedy Method In the ϵ-greedy method (Sutton & Barto, 1998, Chapter 2), with probability ϵ, the agent randomly chooses an action from the possible ones. On the other hand, the agent chooses the best action with the largest estimated value with probability $1 - \epsilon$. That is, p_{ij} is given by

$$p_{ij} \stackrel{\text{def}}{=} \pi\left(\epsilon, Q_{ij}\right) = \frac{\epsilon}{J_i} + (1 - \epsilon)\theta_{ij}, \tag{38}$$

where $J_i \stackrel{\text{def}}{=} |\mathcal{A}_i|$ and

$$\theta_{ij} \stackrel{\text{def}}{=} \begin{cases} 1 & \text{if } j = \arg\max_{j' \in \mathcal{A}_i} Q_{ij'} \\ 0 & \text{if } j \neq \arg\max_{j' \in \mathcal{A}_i} Q_{ij'} \end{cases}. \tag{39}$$

The parameter ϵ is gradually decreased such that $\epsilon \to 0$ as $n \to \infty$. We denote the value of ϵ at time step n by ϵ_n.

Whether the softmax AS or the ϵ-greedy AS is better is unclear and it may depend on the task and on human factors (Sutton & Barto, 1998, p. 31). Added to this, the explicit role of the parameters β and ϵ is also unknown. In the rest of this section, we elucidate the mathematical role of the parameters and the difference between the two strategies by studying their effect in RM.

4.2 Stochastic Complexity

We assume that the policy is improved sufficiently slowly such that the AEP holds. Figure 4 illustrates an RL process on the manifold spanned by $\mathbf{\Gamma}$. This manifold is called the information manifold (IM) (Amari & Han, 1989). Fuller explanation about the figure will be described in the following section. We use Q^*_{ij} to denote the expected value of Q_{ij} for all i, j, henceforth. Let $p^*_{ij} = \pi(\beta, Q^*_{ij})$ in the softmax method and $p^*_{ij} = \pi(\epsilon, Q^*_{ij})$ in the ϵ-greedy method. Let $\mathbf{\Gamma}^{\pi^*}$ be the policy matrix whose components are given by p^*_{ij}. We define $\mathbf{\Gamma}^* \stackrel{\text{def}}{=} (\mathbf{\Gamma}^{\pi^*}, \mathbf{\Gamma}^{\mathrm{T}})$ and write the set of $\mathbf{\Gamma}^*$ as $\Omega \stackrel{\text{def}}{=} \{\mathbf{\Gamma} | \mathbf{\Gamma}^{\pi} = \mathbf{\Gamma}^{\pi^*}\}$ for notational convenience. The set Ω is given by changing the parameter of AS strategy, such as β and ϵ. The optimal policy matrix is denoted by $\mathbf{\Gamma}^{\pi^\dagger}$ whose components are

$$p^\dagger_{ij} = \begin{cases} 1 & \text{if } j = \arg\max_{j' \in \mathcal{A}_i} Q^*_{ij'} \\ 0 & \text{if } j \neq \arg\max_{j' \in \mathcal{A}_i} Q^*_{ij'} \end{cases}. \tag{40}$$

For example, in the softmax method we can write it as $\mathbf{\Gamma}^{\pi^\dagger} = \{p^\dagger_{ij} = \pi(\infty, Q^*_{ij})\}$, and in the ϵ-greedy method we can also write it as $\mathbf{\Gamma}^{\pi^\dagger} = \{p^\dagger_{ij} = \pi(0, Q^*_{ij})\}$. Also, we define $\mathbf{\Gamma}^\dagger \stackrel{\text{def}}{=} (\mathbf{\Gamma}^{\pi^\dagger}, \mathbf{\Gamma}^{\mathrm{T}})$. Figure 5 shows Ω in each method. As shown in Figures 5(a) and 5(b) the set of matrices $\mathbf{\Phi}$ such that (21) holds, designated by the shaded circle, depends on n but not on β or ϵ. Note that the number of elements

in $\mathcal{C}^n_{\lambda_n}(\mathbf{\Gamma})$ depends on β or ϵ because the parameter affects the value of $\mathrm{H}(\mathbf{\Gamma}^\pi|\mathbf{V})$, while the set of such matrices does not depend. We assume that the neighborhood of the optimal matrix on the IM is smooth (differentiable) for the parameters of the AS strategy, such as β and ϵ.

[Figure 4 about here.]

[Figure 5 about here.]

According to Lemma A.1, the number of possible conditional type matrices on the IM is determined by n, I, J, and K. As n increases we can create empirical sequences with arbitrarily close conditional type matrices to $\mathbf{\Gamma}^*$. If the environment, or specifically, the state transition matrix $\mathbf{\Gamma}^{\mathrm{T}}$ is constant, $\mathbf{\Gamma}$ varies only with the changes of $\mathbf{\Gamma}^\pi$. Hence the area of possible $\mathbf{\Phi}$ on the IM is actually restricted. Now we define a stochastic complexity that will play an important role in the later discussion.

Definition 4.1 (Stochastic complexity) *The stochastic complexity (SC) is defined by*

$$\psi(\mathbf{\Gamma}) \stackrel{\mathrm{def}}{=} \mathrm{H}(\mathbf{\Gamma}^\pi|\mathbf{V}) + \mathrm{H}(\mathbf{\Gamma}^{\mathrm{T}}|\mathbf{W}). \tag{41}$$

This is referred to as complexity because the value of $\psi(\mathbf{\Gamma})$ is closely related to the algorithmic complexity as will be discussed in Section 4.4.

To understand the role of the SC in RL, it is worth to mention that the SC has a relationship to exploration (or exploitation) (Sutton & Barto, 1998, Chapter 2) in some cases. In short, the SC expresses the randomness of the agent's policy. Exploration is, in general, to search for policies better than the current one, instead of the simple randomness. One efficient way for such exploratory search is to give a randomness to the policy as is done in the softmax and ϵ-greedy methods. In this case, a policy for exploration is to enlarge the set of possible empirical sequences, that is, the $\mathbf{\Gamma}$-typical set in order to widely explore the environment. This is because the $\mathbf{\Gamma}$-typical set has probability almost one according to Theorem 3.1. On the other hand, using estimates of the action-value function the agent has to select the best action with the largest estimate of the action-value function to maximize the future return. Such a policy for exploitation is to make the $\mathbf{\Gamma}$-typical set smaller, so that only few empirical sequences which yield high return are allowed to be generated in practice. Thus, when the agent performs randomized exploration in AS strategy, if the value of $\psi(\mathbf{\Gamma})$ is large, then the policy is exploratory, and analogously if the value is small, then the policy is exploitative. Of course, since the SC does not assess the rewards of empirical sequence, we have to consider both the SC and the rewards when we argue the original sense of exploration in RL. In (Iwata, Ikeda, & Sakai, 2004) the estimated entropy (like the SC) of return with respect to each state-action was formulated and a novel criterion for AS strategy was proposed by combining the estimated entropy with the estimates of the action-value function.

4.3 Stochastic Complexity and Return Maximization

We will show the relationship between the SC and RM in RL. We use the term RM to maximize the probability that the best sequences appear, but does not mean a lucky case in which the best sequences appear unexpectedly.

Definition 4.2 (Return maximization) *We denote a proper subset of best sequences by*

$$\mathcal{X}^\dagger_n \stackrel{\mathrm{def}}{=} \{\boldsymbol{x} \in (\mathcal{S}\times\mathcal{A}\times\mathbb{R}_0)^n | \mathbf{\Phi}^\pi = \mathbf{\Gamma}^{\pi^\dagger} \ \textit{where}\ \mathbf{\Phi}^\pi \in \mathbf{\Lambda}^\pi_n\}. \tag{42}$$

Then, RM means that the subset of best sequences asymptotically has probability one, that is, $\mathcal{C}^n_{\lambda_n}(\mathbf{\Gamma}) = \mathcal{X}^\dagger_n$ *as* $n\to\infty$.

The following theorem states that RM can be performed under a proper AS strategy so that the estimates of the value function eventually converge to the expected values.

Theorem 4.1 (RM near the end of process) *If the agent's policy tends to be optimal, then* $\mathcal{X}_n^\dagger \subset \mathcal{C}_{\lambda_n}^n(\mathbf{\Gamma})$ *holds and the probability* $\Pr(\mathcal{X}_n^\dagger)$ *of RM satisfies*

$$\Pr(\mathcal{X}_n^\dagger) \to \frac{|\mathcal{X}_n^\dagger|}{|\mathcal{C}_{\lambda_n}^n(\mathbf{\Gamma})|}, \tag{43}$$

for sufficient large n*. Then, as* $n \to \infty$ *and* $\mathbf{\Gamma} \to \mathbf{\Gamma}^\dagger$, $\Pr(\mathcal{X}_n^\dagger) \to 1$.

The proof is given in Appendix B.7.

Hence, we consider that $\mathcal{X}_n^\dagger \subset \mathcal{C}_{\lambda_n}^n(\mathbf{\Gamma})$ and then reduce the $\mathbf{\Gamma}$-typical set such that $\mathcal{C}_{\lambda_n}^n(\mathbf{\Gamma}) \simeq \mathcal{X}_n^\dagger$. Here the key points are that

- by updating the estimates the agent has to improve the policy matrix $\mathbf{\Gamma}^\pi$ as quickly as possible such that the $\mathbf{\Gamma}$-typical set includes the empirical sequence having the conditional type matrix $\mathbf{\Gamma}^{\pi^*}$, that is,
$$\mathrm{D}(\mathbf{\Gamma}^{\pi^*} \| \mathbf{\Gamma}^\pi | \boldsymbol{F}_{\mathcal{S}}) \leq \lambda_n, \tag{44}$$
(see Figure 4(a)), and then
- the agent is required to shut out empirical sequences except the best sequences from the $\mathbf{\Gamma}$-typical set in order to assign high probability to the best sequences (see Figure 4(b)).

The algorithm for the former is simply TD learning. Figure 4(a) illustrates that $\mathbf{\Gamma}$ on the IM is refined by a TD learning such that the $\mathbf{\Gamma}$-typical set includes the empirical sequences having the matrix $\mathbf{\Gamma}^*$ of the conditional types, that is, (44) holds. It is known that the convergence order of TD learning is at most $1/\sqrt{n}$ (Kushner & Yin, 1997). After satisfying (44) the agent has to allot higher probability to the $\mathbf{\Gamma}$-typical set. The goal of the latter is to make the number of elements in the $\mathbf{\Gamma}$-typical set small while satisfying (44). This leads to the result that the subset of the best sequences occurs with high probability because according to Theorem 3.2 all the $\mathbf{\Gamma}$-typical sequences of n time steps have the same probability for sufficiently large n. From Theorem 3.3 we see that the number of elements in the $\mathbf{\Gamma}$-typical set is dependent on the SC $\psi(\mathbf{\Gamma})$ and the quantity λ_n, and that the smaller each value is, the smaller the number of elements. Recall that by tuning the parameters of the AS strategy the agent can control only the SC. This leads us to the question of how sensitive the parameters such as β and ϵ are for controlling the SC. The following theorems answer this question.

Theorem 4.2 (Relationship between β and SC) *The value of* $\psi(\mathbf{\Gamma})$ *decreases as* β *increases. The derivative of* $\psi(\mathbf{\Gamma})$ *with respect to* β *is*

$$\frac{d\psi(\mathbf{\Gamma})}{d\beta} = \sum_{i=1}^{I} v_i \left\{ \frac{-\beta}{2(\mathrm{Z}_i(\beta))^2} \sum_{j=1}^{J} \sum_{j'=1}^{J} (Q_{ij} - Q_{ij'})^2 \exp(\beta(Q_{ij} + Q_{ij'})) \right\}. \tag{45}$$

In particular, if $\beta \to \infty$*, then*

$$\psi(\mathbf{\Gamma}) \to \mathrm{H}(\mathbf{\Gamma}^{\mathrm{T}} | \mathbf{W}). \tag{46}$$

Theorem 4.3 (Relationship between ϵ and SC) *The value of* $\psi(\mathbf{\Gamma})$ *decreases as* $\epsilon \to 0$*. The derivative of* $\psi(\mathbf{\Gamma})$ *with respect to* ϵ *is*

$$-\frac{d\psi(\mathbf{\Gamma})}{d\epsilon} = \sum_{i=1}^{I} v_i \left\{ \left(1 - \frac{1}{J_i}\right) \left(\log \frac{\epsilon}{J_i} - \log \left(\frac{\epsilon}{J_i} + 1 - \epsilon \right) \right) \right\}. \tag{47}$$

In particular, if $\epsilon \to 0$*, then* $\psi(\mathbf{\Gamma})$ *coincides with* (46)*.*

Theorems 4.2 and 4.3 are proved in Appendices B.8 and B.9, respectively. The equations (45) and (47) denote the sensitivity for the randomness of the policy. The main difference between the two methods is that the estimates of the action-value function affect the derivative of the SC directly in the softmax method but not in the ϵ-greedy method. Theorems 4.2 and 4.3 draw an attention to the important dependence, often overlooked in tuning the parameter. In general, it is difficult to tune β and ϵ well and the tuning forms depend only on n in the literatures. For example, $\beta_n = c^n$ or $\epsilon_n = c/n$ is adopted and then c is optimized by trial-and-errors, although the result of the tuning strongly depends on the values of v_i, Q_{ij}, and J_i for every i, j, as explicitly shown in Theorems 4.2 and 4.3. In other words, one of the causes of the difficulty is that the sensitivity can not be considered on the above tuning. In fact, since all the values of Q_{ij} and J_i are available for the agent and all the value of v_i can be approximated by the values of the type, the agent can calculate the sensitivity asymptotically. Accordingly, the sensitivity may be a guide for tuning the parameters appropriately. The importance of knowing the sensitivity has been also pointed out in (Dearden, Friedman, & Russell, 1998), first.

Example 4.1 (A guide of RM) *The sensitivity is not something like a quantitative criterion to be directly used by itself in practical issues because of its generality. However, it can be used as a qualitative guide in choosing a tuning which depends on each case. Here, we calculate the sensitivity approximately to gain an insight into the RM speed. Let c be an arbitrary constant value. When we choose $\epsilon_n = c/n$, there exists a non-negative value c' such that*

$$-\frac{d\psi(\mathbf{\Gamma})}{d\epsilon} = -\sum_{i=1}^{I} v_i \left\{ \left(1 - \frac{1}{J_i}\right) \log \left(\frac{J_i}{c} n + 1 - J_i \right) \right\} \approx -c' \log n, \tag{48}$$

where n is sufficiently large. The value of c' can be computed by $\{f_i, J_i\}_{i=1}^{I}$. Due to $|\mathcal{C}_{\lambda_n}^n(\mathbf{\Gamma})| \approx \exp(n\psi(\mathbf{\Gamma}))$ for sufficient large n,

$$-\frac{d}{d\epsilon}\frac{1}{|\mathcal{C}_{\lambda_n}^n(\mathbf{\Gamma})|} = \frac{n}{|\mathcal{C}_{\lambda_n}^n(\mathbf{\Gamma})|}\frac{d\psi(\mathbf{\Gamma})}{d\epsilon} \approx c'' \frac{n \log n}{\exp(n\psi(\mathbf{\Gamma}))}, \tag{49}$$

where c'' is a non-negative value which depends on $\psi(\mathbf{\Gamma})$ and c'. This is a qualitative guide for checking the RM speed, $d\Pr(\mathcal{X}^\dagger)/d\epsilon$, near the optimal policy because from Theorem 4.1

$$\Pr(\mathcal{X}^\dagger) \to \frac{|\mathcal{X}^\dagger|}{|\mathcal{C}_{\lambda_n}^n(\mathbf{\Gamma})|}, \tag{50}$$

for sufficient large n. Thus, we can estimate the RM speed near the end of learning process and can select a tuning referring to it. If we think that the estimated speed is too fast for a given environment, choose more slower tuning such as $\epsilon_n = c/\log n$. Of course, this is a rough utility but have an interesting potential by combining it with other criteria. In the case of the softmax method, similarly, when $\beta = c^n$, there exists a non-negative value c' such that

$$\frac{d\psi(\mathbf{\Gamma})}{d\beta} \approx -c' c^n, \tag{51}$$

where n is sufficiently large. Then, we have

$$\frac{d}{d\beta}\frac{1}{|\mathcal{C}_{\lambda_n}^n(\mathbf{\Gamma})|} \approx c'' \frac{nc^n}{\exp(n\psi(\mathbf{\Gamma}))}, \tag{52}$$

for sufficient large n.

Next, we consider another important factor λ_n for making the number of elements in the $\mathbf{\Gamma}$-typical set smaller. Figure 4(b) shows the changes of λ_n with n where the lower bound of λ_n is given by (23). There may be a tighter bound in various situations such that MDPs have a deterministic rule because the bound was derived under the condition that $\mathbf{\Gamma}$ has no constraint. In other words, the bound means a sufficient condition for RM. Hence, the first order of the bound is tightest and valid only when the agent takes "randomized" AS strategies[2] such that $p_{ij} > 0$ for every i, j in the environments where $p_{iji'k} > 0$ for every i, j, i', k. In such cases, the bound suggests that the convergence rate of $\mathrm{D}(\mathbf{\Phi}_n \| \mathbf{\Gamma})$ going to zero is at most $(\log n)/n$ and its coefficient is $(IJ + I^2JK)$. The convergence rate indicates how fast the policy reflects on the structure of empirical sequence. The coefficient also implies that in applications a lot of time steps are required for agreement between the current matrix $\mathbf{\Gamma}$ and the matrix $\mathbf{\Phi}$ of the conditional types regarding the empirical sequence when the state, action, and reward sets are large.

4.4 Stochastic Complexity and Kolmogorov Complexity

In this section we show the relationship between the SC and the Kolmogorov complexity (KC) (Cover & Thomas, 1991; Li & Vitányi, 1997). The SC is also reasonable from the point of view of algorithmic complexity. Let $l(\boldsymbol{x})$ denote the length of the sequence $\boldsymbol{x}$. Let $\mathcal{U}(q)$ be the output of a universal computer $\mathcal{U}$ when presented with a program q. The KC of a sequence $\boldsymbol{x}$ is defined as the minimal description length of q (Cover & Thomas, 1991, pp. 147–148).

Definition 4.3 (KC and conditional KC) *The KC* $\mathrm{K}_{\mathcal{U}}(\boldsymbol{x})$ *of a sequence* $\boldsymbol{x}$ *with respect to a universal computer* $\mathcal{U}$ *is defined as*

$$\mathrm{K}_{\mathcal{U}}(\boldsymbol{x}) \stackrel{\text{def}}{=} \min_{q:\mathcal{U}(q)=\boldsymbol{x}} l(q), \tag{53}$$

the minimum length over all programs that print $\boldsymbol{x}$ *and halt. If we assume that the computer already knows the length of the sequence, then we can define the conditional KC knowing* $l(\boldsymbol{x})$ *as*

$$\mathrm{K}_{\mathcal{U}}(\boldsymbol{x}|l(\boldsymbol{x})) \stackrel{\text{def}}{=} \min_{q:\mathcal{U}(q,l(\boldsymbol{x}))=\boldsymbol{x}} l(q). \tag{54}$$

This is the shortest possible description length if the length of $\boldsymbol{x}$ *is made available to the computer* $\mathcal{U}$*.*

Since the length $l(\boldsymbol{x})$ of an empirical sequence $\boldsymbol{x} \in (\mathcal{S} \times \mathcal{A} \times \mathbb{R}_0)^n$ is $3n$, consider

$$\mathrm{K}_{\mathcal{U}}(\boldsymbol{x}|3n) = \min_{q:\mathcal{U}(q,3n)=\boldsymbol{x}} l(q). \tag{55}$$

Note that $\mathrm{K}_{\mathcal{U}}(\boldsymbol{x}|3n)$ denotes the algorithmic complexity to print $\boldsymbol{x}$ and halt. The following theorem shows that the expected value of $\mathrm{K}_{\mathcal{U}}(\boldsymbol{x}|3n)$ is asymptotically equal to the SC.

Theorem 4.4 (Relationship between KC and SC) *If* $\boldsymbol{s} \in \mathcal{C}^n_{\kappa_n}(\mathbf{V})$*,* $(\boldsymbol{s}, \boldsymbol{a}) \in \mathcal{C}^n_{\xi_n}(\mathbf{W})$*, and* $\boldsymbol{x} \in \mathcal{C}^n_{\lambda_n}(\mathbf{\Gamma})$*, then there exists a constant value* c *such that*

$$\frac{\log \nu}{n} - \rho_n \leq \frac{1}{n}\mathrm{E}_{\mathbf{\Gamma}}\left[\mathrm{K}_{\mathcal{U}}(\boldsymbol{x}|3n)\right] - \psi(\mathbf{\Gamma}) \leq \rho_n + \frac{\log I}{n} + (IJ + I^2JK)\frac{\log(n+1)}{n} + \frac{c}{n}, \tag{56}$$

for a computer $\mathcal{U}$ *and all* n*. In particular, if* $\mathbf{\Gamma} \to \mathbf{\Gamma}^*$ *as* $n \to \infty$*, then*

$$\lim_{n\to\infty} \frac{1}{n}\mathrm{E}_{\mathbf{\Gamma}^*}\left[\mathrm{K}_{\mathcal{U}}(\boldsymbol{x}|3n)\right] = \psi(\mathbf{\Gamma}^*). \tag{57}$$

The proof is given in Appendix B.10. The SC is so called because of this relationship.

[2]For example, the softmax and ϵ-greedy methods with the parameters $\beta < \infty$ and $\epsilon > 0$, respectively.

5 Conclusions

In this paper, we have formulated almost stationary ergodic MDPs by the type method and shown that the AEP holds on empirical sequences in such processes. Under a proper AS strategy which guarantees the convergence of the estimates, the RM is characterized by the SC $\psi(\mathbf{\Gamma})$ and the quantity λ_n. We examined the role of these factors on RM and then derived the sensitivity of the SC, which is a qualitative guide in tuning the parameters of AS strategy. Also, we showed the bound of the convergence speed of the empirical sequences tending to the best sequence in the worst cases. Using the results of (Merhav, 1991; Merhav & Neuhoff, 1992) the discussions in this paper can be readily extend to the more general case where the source of empirical sequences is a unifilar source (Han & Kobayashi, 2002, p. 77) in a similar manner.

Acknowledgments

We thank the anonymous referees for helpful comments particularly in the role of the SC and in the formulation of MDPs. This study was supported in part by a grant for scientific research (No. 14003714) from the Ministry of Education, Culture, Sports, Science, and Technology of Japan.

A Related Theorems

We will show a number of theorems related to the AEP. Obviously, from (Csiszár & Körner, 1997, Lemma 2.2) we obtain the following lemma which plays a major role in determining the AEP.

Lemma A.1 (Number of elements in the set of possible $\mathbf{\Phi}$) *The size of $\mathbf{\Lambda}_n^\pi$ is upper bounded by*

$$|\mathbf{\Lambda}_n^\pi| \leq (n+1)^{IJ}. \tag{58}$$

Analogously,

$$|\mathbf{\Lambda}_n^{\mathrm{T}}| \leq (n+1)^{I^2JK}. \tag{59}$$

Accordingly, the number of elements in the set of possible $\mathbf{\Phi}$ is upper bounded at most by a polynomial order of n, that is,

$$|\mathbf{\Lambda}_n| \leq (n+1)^{IJ+I^2JK}. \tag{60}$$

The following lemma states the fact that the discrepancy between the empirical entropy and the entropy asymptotically goes to zero.

Lemma A.2 *Let $\mathbf{\Phi} \in \mathbf{\Lambda}_n$ denote the matrix of the conditional types with respect to the empirical sequence which satisfies $\boldsymbol{s} \in \mathcal{C}_{\kappa_n}^n(\mathbf{V})$, $(\boldsymbol{s}, \boldsymbol{a}) \in \mathcal{C}_{\xi_n}^n(\mathbf{W})$, and $\boldsymbol{x} \in \mathcal{C}_{\lambda_n}^n(\mathbf{\Gamma})$. Then, if $\lambda_n \leq 1/8$, we obtain*

$$\left| \mathrm{H}(\mathbf{\Phi}^\pi|\boldsymbol{F}_{\mathcal{S}}) - \mathrm{H}(\mathbf{\Gamma}^\pi|\mathbf{V}) \right| \leq \sqrt{2\kappa_n}\log J - \sqrt{2\lambda_n}\log\frac{\sqrt{2\lambda_n}}{IJ}, \tag{61}$$

$$\left| \mathrm{H}(\mathbf{\Phi}^{\mathrm{T}}|\boldsymbol{F}_{\mathcal{SA}}) - \mathrm{H}(\mathbf{\Gamma}^{\mathrm{T}}|\mathbf{W}) \right| \leq \sqrt{2\xi_n}\log I + \sqrt{2\xi_n}\log K - \sqrt{2\lambda_n}\log\frac{\sqrt{2\lambda_n}}{I^2JK}. \tag{62}$$

For the proof of this lemma, see Appendix B.1.

Now we show that the number of sequences with the same conditional type matrix increases exponentially for n.

Theorem A.1 (Bound of $|\mathcal{C}^n(\mathbf{\Phi}, \boldsymbol{F}_{\mathcal{S}}, \boldsymbol{F}_{\mathcal{SA}})|$) *For every state sequence $\boldsymbol{s} \in \mathcal{S}^n$ with the type $\boldsymbol{F}_{\mathcal{S}}$ and matrix $\mathbf{\Phi}^\pi : \mathcal{S} \to \mathcal{A}$ such that $\mathcal{C}^n(\mathbf{\Phi}^\pi, \boldsymbol{s})$ is not empty,*

$$\frac{\exp\{n\mathrm{H}(\mathbf{\Phi}^\pi|\boldsymbol{F}_{\mathcal{S}})\}}{(n+1)^{IJ}} \leq |\mathcal{C}^n(\mathbf{\Phi}^\pi, \boldsymbol{s})| \leq \exp\{n\mathrm{H}(\mathbf{\Phi}^\pi|\boldsymbol{F}_{\mathcal{S}})\}. \tag{63}$$

Also, for every action sequence $\boldsymbol{a} \in \mathcal{A}^n$ *and matrix* $\boldsymbol{\Phi}^{\mathrm{T}} : \mathcal{S} \times \mathcal{A} \to \mathcal{S} \times \mathbb{R}_0$ *such that* $\mathcal{C}^n(\boldsymbol{\Phi}^{\mathrm{T}}, \boldsymbol{F}_{\mathcal{SA}})$ *is not empty and the joint-type is* $\boldsymbol{F}_{\mathcal{SA}}$*,*

$$\frac{\exp\left\{n\mathrm{H}(\boldsymbol{\Phi}^{\mathrm{T}}|\boldsymbol{F}_{\mathcal{SA}})\right\}}{n^{IJ}(n+1)^{I^2JK}} \leq |\mathcal{C}^n(\boldsymbol{\Phi}^{\mathrm{T}}, \boldsymbol{F}_{\mathcal{SA}})| \leq I\exp\left\{n\mathrm{H}(\boldsymbol{\Phi}^{\mathrm{T}}|\boldsymbol{F}_{\mathcal{SA}})\right\}. \tag{64}$$

Therefore, for every $\boldsymbol{x} \in (\mathcal{S} \times \mathcal{A} \times \mathbb{R}_0)^n$ *with the type* $\boldsymbol{F}_{\mathcal{S}}$ *and the joint-type* $\boldsymbol{F}_{\mathcal{SA}}$ *and for the matrix* $\boldsymbol{\Phi}$*,*

$$\frac{\exp\left[n\left\{\mathrm{H}(\boldsymbol{\Phi}^{\pi}|\boldsymbol{F}_{\mathcal{S}}) + \mathrm{H}(\boldsymbol{\Phi}^{\mathrm{T}}|\boldsymbol{F}_{\mathcal{SA}})\right\}\right]}{n^{IJ}(n+1)^{IJ+I^2JK}} \leq |\mathcal{C}^n(\boldsymbol{\Phi}, \boldsymbol{F}_{\mathcal{S}}, \boldsymbol{F}_{\mathcal{SA}})| \leq I\exp\left[n\left\{\mathrm{H}(\boldsymbol{\Phi}^{\pi}|\boldsymbol{F}_{\mathcal{S}}) + \mathrm{H}(\boldsymbol{\Phi}^{\mathrm{T}}|\boldsymbol{F}_{\mathcal{SA}})\right\}\right]. \tag{65}$$

The proof is given in Appendix B.2. There also exist the following bounds on the probability of the $\boldsymbol{\Phi}$-shell.

Theorem A.2 (Bound on probability of $\mathcal{C}^n(\boldsymbol{\Phi}, \boldsymbol{F}_{\mathcal{S}}, \boldsymbol{F}_{\mathcal{SA}})$) *For every matrix* $\boldsymbol{\Phi}$ *such that* $\mathcal{C}^n(\boldsymbol{\Phi}, \boldsymbol{F}_{\mathcal{S}}, \boldsymbol{F}_{\mathcal{SA}})$ *is not empty,* $\Pr(\mathcal{C}^n(\boldsymbol{\Phi}, \boldsymbol{F}_{\mathcal{S}}, \boldsymbol{F}_{\mathcal{SA}}))$ *is bounded by*

$$\begin{aligned}\frac{\mu\exp\left[-n\left\{\mathrm{D}(\boldsymbol{\Phi}^{\pi}\|\boldsymbol{\Gamma}^{\pi}|\boldsymbol{F}_{\mathcal{S}}) + +\mathrm{D}(\boldsymbol{\Phi}^{\mathrm{T}}\|\boldsymbol{\Gamma}^{\mathrm{T}}|\boldsymbol{F}_{\mathcal{SA}})\right\}\right]}{n^{IJ}(n+1)^{IJ+I^2JK}} \\ \leq \Pr(\mathcal{C}^n(\boldsymbol{\Phi}, \boldsymbol{F}_{\mathcal{S}}, \boldsymbol{F}_{\mathcal{SA}})) \leq \\ \nu^{-1}I\exp\left[-n\left\{\mathrm{D}(\boldsymbol{\Phi}^{\pi}\|\boldsymbol{\Gamma}^{\pi}|\boldsymbol{F}_{\mathcal{S}}) + \mathrm{D}(\boldsymbol{\Phi}^{\mathrm{T}}\|\boldsymbol{\Gamma}^{\mathrm{T}}|\boldsymbol{F}_{\mathcal{SA}})\right\}\right],\end{aligned} \tag{66}$$

where ν *and* μ *are defined by* (24) *and* (27)*, respectively.*

The proof is given in Appendix B.3. This theorem implies that empirical sequences with conditional type matrix $\boldsymbol{\Phi}$ far from $\boldsymbol{\Gamma}$ are not likely to be generated in practice. The term "far" here means that the divergence between $\boldsymbol{\Phi}$ and $\boldsymbol{\Gamma}$ is large. We have mentioned the theorems to be used in the proofs of Theorem 3.1–3.3.

B Proofs

B.1 Proof of Lemma A.2

From (Kullback, 1967), if $\boldsymbol{s} \in \mathcal{C}^n_{\kappa_n}(\mathbf{V})$, then

$$\sum_{i=1}^{I} |f_i - v_i| \leq \sqrt{2\mathrm{D}(\boldsymbol{F}_{\mathcal{S}}\|\mathbf{V})}. \tag{67}$$

Hence, by $\mathrm{H}(\mathbf{P}_{(i)}) \leq \log J$ and (15),

$$\left|\ \mathrm{H}(\boldsymbol{\Gamma}^{\pi}|\boldsymbol{F}_{\mathcal{S}}) - \mathrm{H}(\boldsymbol{\Gamma}^{\pi}|\mathbf{V})\ \right| = \left|\ \sum_{i=1}^{I}(f_i - v_i)\mathrm{H}(\mathbf{P}_{(i)})\ \right| \leq \sqrt{2\kappa_n}\log J \tag{68}$$

is satisfied. In the same way as (Csiszár & Körner, 1997, Lemma 2.7), if $\lambda_n \leq 1/8$, then

$$\left|\ \mathrm{H}(\boldsymbol{\Phi}^{\pi}|\boldsymbol{F}_{\mathcal{S}}) - \mathrm{H}(\boldsymbol{\Gamma}^{\pi}|\boldsymbol{F}_{\mathcal{S}})\ \right| \leq \left|\ \mathrm{H}(\boldsymbol{\Phi}^{\pi}, \boldsymbol{F}_{\mathcal{S}}) - \mathrm{H}(\boldsymbol{\Gamma}^{\pi}, \boldsymbol{F}_{\mathcal{S}})\ \right| \leq -\sqrt{2\lambda_n}\log\frac{\sqrt{2\lambda_n}}{IJ}. \tag{69}$$

From

$$\left|\ \mathrm{H}(\boldsymbol{\Phi}^{\pi}|\boldsymbol{F}_{\mathcal{S}}) - \mathrm{H}(\boldsymbol{\Gamma}^{\pi}|\mathbf{V})\ \right| \leq \left|\ \mathrm{H}(\boldsymbol{\Phi}^{\pi}|\boldsymbol{F}_{\mathcal{S}}) - \mathrm{H}(\boldsymbol{\Gamma}^{\pi}|\boldsymbol{F}_{\mathcal{S}})\ \right| + \left|\ \mathrm{H}(\boldsymbol{\Gamma}^{\pi}|\boldsymbol{F}_{\mathcal{S}}) - \mathrm{H}(\boldsymbol{\Gamma}^{\pi}|\mathbf{V})\ \right|, \tag{70}$$

we have (61). The equation (62) can be derived similarly. ∎

B.2 Proof of Theorem A.1

First, we define

$$|\mathcal{C}^{n_i}(\boldsymbol{G}_{(i)})| \stackrel{\text{def}}{=} \frac{n_i!}{\prod_{j=1}^{J} n_{ij}!}. \tag{71}$$

Since the actions given by s_i have the type $\boldsymbol{G}_{(i)}$, from (Dueck & Körner, 1979) we have

$$\frac{\exp\left\{n_i \mathrm{H}(\boldsymbol{G}_{(i)})\right\}}{(n_i+1)^J} \le |\mathcal{C}^{n_i}(\boldsymbol{G}_{(i)})| \le \exp\left\{n_i \mathrm{H}(\boldsymbol{G}_{(i)})\right\}. \tag{72}$$

By $|\mathcal{C}^n(\boldsymbol{\Phi}^\pi, \boldsymbol{s})| = \prod_{i=1}^{I} |\mathcal{C}^{n_i}(\boldsymbol{G}_{(i)})|$, $|\mathcal{C}^n(\boldsymbol{\Phi}^\pi, \boldsymbol{s})|$ is bounded by

$$\frac{\exp\left\{n \mathrm{H}(\boldsymbol{\Phi}^\pi|\boldsymbol{F}_{\mathcal{S}})\right\}}{\prod_{i=1}^{I}(n_i+1)^J} \le |\mathcal{C}^n(\boldsymbol{\Phi}^\pi, \boldsymbol{s})| \le \exp\left\{n \mathrm{H}(\boldsymbol{\Phi}^\pi|\boldsymbol{F}_{\mathcal{S}})\right\}. \tag{73}$$

Therefore, by $\prod_{i=1}^{I}(n_i+1)^J \le (n+1)^{IJ}$ we obtain (63).

The proof of (64) is different from the above proof because $\mathcal{C}^n(\boldsymbol{\Phi}^{\mathrm{T}}, \boldsymbol{F}_{\mathcal{SA}})$ is not only dependent on $\boldsymbol{F}_{\mathcal{SA}}$ but also on the initial state $s(1)$ because of Markov property. So for any i, j we define

$$|\tilde{\mathcal{C}}^{n_{ij}}(\boldsymbol{G}_{(ij)})| \stackrel{\text{def}}{=} \frac{n_{ij}!}{\prod_{i'=1}^{I}\prod_{k=1}^{K} n_{iji'k}!}. \tag{74}$$

By following along the same lines as the proof of (63) above, we have

$$\frac{\exp\left\{n \mathrm{H}(\boldsymbol{\Phi}^{\mathrm{T}}|\boldsymbol{F}_{\mathcal{SA}})\right\}}{(n+1)^{I^2JK}} \le |\tilde{\mathcal{C}}^n(\boldsymbol{\Phi}^{\mathrm{T}}, \boldsymbol{F}_{\mathcal{SA}})| \le \exp\left\{n \mathrm{H}(\boldsymbol{\Phi}^{\mathrm{T}}|\boldsymbol{F}_{\mathcal{SA}})\right\}. \tag{75}$$

The set $\tilde{\mathcal{C}}^n(\boldsymbol{\Phi}^{\mathrm{T}}, \boldsymbol{F}_{\mathcal{SA}})$ allows a unique reconstruction of empirical sequence with $\mathcal{C}^n(\boldsymbol{\Phi}^\pi, \boldsymbol{s})$ only when the initial state $s(1)$ is known. Then, from (Davisson et al., 1981), the upper bound of $\mathcal{C}^n(\boldsymbol{\Phi}^{\mathrm{T}}, \boldsymbol{F}_{\mathcal{SA}})$ is

$$|\mathcal{C}^n(\boldsymbol{\Phi}^{\mathrm{T}}, \boldsymbol{F}_{\mathcal{SA}})| \le I|\tilde{\mathcal{C}}^n(\boldsymbol{\Phi}^{\mathrm{T}}, \boldsymbol{F}_{\mathcal{SA}})| \le I \exp\left\{n \mathrm{H}(\boldsymbol{\Phi}^{\mathrm{T}}|\boldsymbol{F}_{\mathcal{SA}})\right\}, \tag{76}$$

because $s(1)$ is not specified by the I sequences. Next, in the same manner as (Davisson et al., 1981), we obtain the lower bound,

$$|\mathcal{C}^n(\boldsymbol{\Phi}^{\mathrm{T}}, \boldsymbol{F}_{\mathcal{SA}})| \ge \prod_{i=1}^{I}\prod_{j=1}^{J} \frac{(n_{ij}-1)!}{\prod_{i'=1}^{I}\prod_{k=1}^{K} n_{iji'k}!} \ge \frac{1}{n^{IJ}}|\tilde{\mathcal{C}}^n(\boldsymbol{\Phi}^{\mathrm{T}}, \boldsymbol{F}_{\mathcal{SA}})| \ge \frac{\exp\left\{n \mathrm{H}(\boldsymbol{\Phi}^{\mathrm{T}}|\boldsymbol{F}_{\mathcal{SA}})\right\}}{n^{IJ}(n+1)^{I^2JK}}. \tag{77}$$

Thus we have proved that (64) holds. Consequently, from (12) we obtain (65). ∎

B.3 Proof of Theorem A.2

The probability of $\boldsymbol{x} \in (\mathcal{S} \times \mathcal{A} \times \mathbb{R}_0)^n$ is

$$\Pr(\boldsymbol{x}) = \frac{\Pr(s(1))}{\Pr(s(1)|s(n), a(n))} \prod_{t=1}^{n} \left\{\Pr(a(t)|s(t)) \Pr(s(t+1), r(t+1)|s(t), a(t))\right\}, \tag{78}$$

$$= \frac{\Pr(s(1))}{\Pr(s(1)|s(n), a(n))} \left(\prod_{i=1}^{I}\prod_{j=1}^{J} {p_{ij}}^{n_{ij}}\right) \left(\prod_{i=1}^{I}\prod_{j=1}^{J}\prod_{i'=1}^{I}\prod_{k=1}^{K} {p_{iji'k}}^{n_{iji'k}}\right), \tag{79}$$

$$= \frac{\Pr(s(1))}{\Pr(s(1)|s(n), a(n))} \exp\big[-n\big\{\mathrm{D}(\boldsymbol{\Phi}^\pi\|\boldsymbol{\Gamma}^\pi|\boldsymbol{F}_{\mathcal{S}}) + \mathrm{D}(\boldsymbol{\Phi}^{\mathrm{T}}\|\boldsymbol{\Gamma}^{\mathrm{T}}|\boldsymbol{F}_{\mathcal{SA}}) + \mathrm{H}(\boldsymbol{\Phi}^\pi|\boldsymbol{F}_{\mathcal{S}}) + \mathrm{H}(\boldsymbol{\Phi}^{\mathrm{T}}|\boldsymbol{F}_{\mathcal{SA}})\big\}\big], \tag{80}$$

where $\Pr(s(n+1)|s(n),a(n)) = \Pr(s(1)|s(n),a(n))$ by the cyclic convention. With the definitions of (24) and (27), the probability of $\boldsymbol{x}$ is bounded by

$$\mu \exp\left[-n\left\{\mathrm{D}(\boldsymbol{\Phi}^{\pi}\|\boldsymbol{\Gamma}^{\pi}|\boldsymbol{F}_{\mathcal{S}}) + \mathrm{D}(\boldsymbol{\Phi}^{\mathrm{T}}\|\boldsymbol{\Gamma}^{\mathrm{T}}|\boldsymbol{F}_{\mathcal{SA}}) + \mathrm{H}(\boldsymbol{\Phi}^{\pi}|\boldsymbol{F}_{\mathcal{S}}) + \mathrm{H}(\boldsymbol{\Phi}^{\mathrm{T}}|\boldsymbol{F}_{\mathcal{SA}})\right\}\right] \leq \Pr(\boldsymbol{x}) \leq \nu^{-1}\exp\left[-n\left\{\mathrm{D}(\boldsymbol{\Phi}^{\pi}\|\boldsymbol{\Gamma}^{\pi}|\boldsymbol{F}_{\mathcal{S}}) + \mathrm{D}(\boldsymbol{\Phi}^{\mathrm{T}}\|\boldsymbol{\Gamma}^{\mathrm{T}}|\boldsymbol{F}_{\mathcal{SA}}) + \mathrm{H}(\boldsymbol{\Phi}^{\pi}|\boldsymbol{F}_{\mathcal{S}}) + \mathrm{H}(\boldsymbol{\Phi}^{\mathrm{T}}|\boldsymbol{F}_{\mathcal{SA}})\right\}\right]. \quad (81)$$

Using (65) and

$$\min_{\boldsymbol{x}\in\mathcal{C}^n(\boldsymbol{\Phi},\boldsymbol{F}_{\mathcal{S}},\boldsymbol{F}_{\mathcal{SA}})} |\mathcal{C}^n(\boldsymbol{\Phi},\boldsymbol{F}_{\mathcal{S}},\boldsymbol{F}_{\mathcal{SA}})|\Pr(\boldsymbol{x}) \leq \Pr(\mathcal{C}^n(\boldsymbol{\Phi},\boldsymbol{F}_{\mathcal{S}},\boldsymbol{F}_{\mathcal{SA}})) \leq \max_{\boldsymbol{x}\in\mathcal{C}^n(\boldsymbol{\Phi},\boldsymbol{F}_{\mathcal{S}},\boldsymbol{F}_{\mathcal{SA}})} |\mathcal{C}^n(\boldsymbol{\Phi},\boldsymbol{F}_{\mathcal{S}},\boldsymbol{F}_{\mathcal{SA}})|\Pr(\boldsymbol{x}), \quad (82)$$

we obtain (66). ∎

B.4 Proof of Theorem 3.1

Let us define the set of the matrix $\boldsymbol{\Phi}$ whose empirical sequence does not belong to the set of the $\boldsymbol{\Gamma}$-typical sequences as

$$\boldsymbol{\Lambda}'_n = \{\boldsymbol{\Phi}\in\boldsymbol{\Lambda}_n \mid \mathrm{D}(\boldsymbol{\Phi}^{\pi}\|\boldsymbol{\Gamma}^{\pi}|\boldsymbol{F}_{\mathcal{S}}) + \mathrm{D}(\boldsymbol{\Phi}^{\mathrm{T}}\|\boldsymbol{\Gamma}^{\mathrm{T}}|\boldsymbol{F}_{\mathcal{SA}}) > \lambda_n\}. \quad (83)$$

Then,

$$\Pr\left(\mathcal{C}^n_{\lambda_n}(\boldsymbol{\Gamma})\right) = 1 - \Pr\left(\bigcup_{\boldsymbol{\Phi}\in\boldsymbol{\Lambda}'_n} \mathcal{C}^n(\boldsymbol{\Phi},\boldsymbol{F}_{\mathcal{S}},\boldsymbol{F}_{\mathcal{SA}})\right). \quad (84)$$

Following along the same lines as (Csiszár & Körner, 1997, Theorem 2.15) with (66) we have

$$\Pr\left(\bigcup_{\boldsymbol{\Phi}\in\boldsymbol{\Lambda}'_n} \mathcal{C}^n(\boldsymbol{\Phi},\boldsymbol{F}_{\mathcal{S}},\boldsymbol{F}_{\mathcal{SA}})\right) \leq \nu^{-1}I(n+1)^{IJ+I^2JK}\exp\left[-n\min_{\boldsymbol{\Phi}\in\boldsymbol{\Lambda}'_n}\left\{\mathrm{D}(\boldsymbol{\Phi}^{\pi}\|\boldsymbol{\Gamma}^{\pi}|\boldsymbol{F}_{\mathcal{S}}) + \mathrm{D}(\boldsymbol{\Phi}^{\mathrm{T}}\|\boldsymbol{\Gamma}^{\mathrm{T}}|\boldsymbol{F}_{\mathcal{SA}})\right\}\right]. \quad (85)$$

Since $\mathrm{D}(\boldsymbol{\Phi}^{\pi}\|\boldsymbol{\Gamma}^{\pi}|\boldsymbol{F}_{\mathcal{S}}) + \mathrm{D}(\boldsymbol{\Phi}^{\mathrm{T}}\|\boldsymbol{\Gamma}^{\mathrm{T}}|\boldsymbol{F}_{\mathcal{SA}}) > \lambda_n$ when $\boldsymbol{\Phi}\in\boldsymbol{\Lambda}'_n$, substituting λ_n for the minimum value we obtain

$$\Pr\left(\bigcup_{\boldsymbol{\Phi}\in\boldsymbol{\Lambda}'} \mathcal{C}^n(\boldsymbol{\Phi},\boldsymbol{F}_{\mathcal{S}},\boldsymbol{F}_{\mathcal{SA}})\right) \leq \nu^{-1}I(n+1)^{IJ+I^2JK}\exp(-n\lambda_n), \quad (86)$$

$$= \exp\left[-n\left\{\lambda_n - \frac{(IJ+I^2JK)\log(n+1)+\log I-\log\nu}{n}\right\}\right]. \quad (87)$$

We define

$$\varepsilon_n(I,J,K,\lambda_n) \stackrel{\mathrm{def}}{=} \exp\left[-n\left\{\lambda_n - \frac{(IJ+I^2JK)\log(n+1)+\log I-\log\nu}{n}\right\}\right], \quad (88)$$

and hence $\varepsilon_n \to 0$ as $n\to\infty$ if (23) is satisfied. Also, by (84) Theorem 3.1 holds. ∎

B.5 Proof of Theorem 3.2

First, let us derive the lower bound. We define

$$\rho_n(I,J,K,\kappa_n,\xi_n,\lambda_n) \stackrel{\mathrm{def}}{=} \sqrt{2\kappa_n}\log J + \sqrt{2\xi_n}\log K + \sqrt{2\xi_n}\log I - \sqrt{2\lambda_n}\log\frac{\sqrt{2\lambda_n}}{IJ} - \sqrt{2\lambda_n}\log\frac{\sqrt{2\lambda_n}}{I^2JK}. \quad (89)$$

By (81) we have

$$
\begin{aligned}
-\log \Pr(\boldsymbol{x}) &\geq \log \nu + n \left\{ \mathrm{D}(\boldsymbol{\Phi}^{\pi} \| \boldsymbol{\Gamma}^{\pi} | \boldsymbol{F}_{\mathcal{S}}) + \mathrm{D}(\boldsymbol{\Phi}^{\mathrm{T}} \| \boldsymbol{\Gamma}^{\mathrm{T}} | \boldsymbol{F}_{\mathcal{SA}}) + \mathrm{H}(\boldsymbol{\Phi}^{\pi} | \boldsymbol{F}_{\mathcal{S}}) + \mathrm{H}(\boldsymbol{\Phi}^{\mathrm{T}} | \boldsymbol{F}_{\mathcal{SA}}) \right\}, && (90) \\
&\geq \log \nu + n \left\{ \mathrm{H}(\boldsymbol{\Phi}^{\pi} | \boldsymbol{F}_{\mathcal{S}}) + \mathrm{H}(\boldsymbol{\Phi}^{\mathrm{T}} | \boldsymbol{F}_{\mathcal{SA}}) \right\}, && (91) \\
&\geq \log \nu + n \left\{ \mathrm{H}(\boldsymbol{\Gamma}^{\pi} | \mathbf{V}) + \mathrm{H}(\boldsymbol{\Gamma}^{\mathrm{T}} | \mathbf{W}) - \rho_n \right\}, && (92)
\end{aligned}
$$

where (91) is obtained by the non-negativity of the divergence and (92) follows from Lemma A.2.

Analogously, the upper bound is obtained as follows:

$$
\begin{aligned}
-\log \Pr(\boldsymbol{x}) &\leq -\log \mu + n \left\{ \mathrm{D}(\boldsymbol{\Phi}^{\pi} \| \boldsymbol{\Gamma}^{\pi} | \boldsymbol{F}_{\mathcal{S}}) + \mathrm{D}(\boldsymbol{\Phi}^{\mathrm{T}} \| \boldsymbol{\Gamma}^{\mathrm{T}} | \boldsymbol{F}_{\mathcal{SA}}) + \mathrm{H}(\boldsymbol{\Phi}^{\pi} | \boldsymbol{F}_{\mathcal{S}}) + \mathrm{H}(\boldsymbol{\Phi}^{\mathrm{T}} | \boldsymbol{F}_{\mathcal{SA}}) \right\}, && (93) \\
&\leq -\log \mu + n \left\{ \mathrm{H}(\boldsymbol{\Phi}^{\pi} | \boldsymbol{F}_{\mathcal{S}}) + \mathrm{H}(\boldsymbol{\Phi}^{\mathrm{T}} | \boldsymbol{F}_{\mathcal{SA}}) + \lambda_n \right\}, && (94) \\
&\leq -\log \mu + n \left\{ \mathrm{H}(\boldsymbol{\Gamma}^{\pi} | \mathbf{V}) + \mathrm{H}(\boldsymbol{\Gamma}^{\mathrm{T}} | \mathbf{W}) + \lambda_n + \rho_n \right\}. && (95)
\end{aligned}
$$

Thus dividing (92) and (95) by n we have (26). ∎

B.6 Proof of Theorem 3.3

We first prove the lower bound. Using the fact that $\mathcal{C}^n_{\lambda_n}(\boldsymbol{\Gamma}) \supseteq \mathcal{C}^n(\boldsymbol{\Phi}, \boldsymbol{F}_{\mathcal{S}}, \boldsymbol{F}_{\mathcal{SA}})$ and (65), we get

$$
\begin{aligned}
|\mathcal{C}^n_{\lambda_n}(\boldsymbol{\Gamma})| &\geq |\mathcal{C}^n(\boldsymbol{\Phi}, \boldsymbol{F}_{\mathcal{S}}, \boldsymbol{F}_{\mathcal{SA}})|, && (96) \\
&\geq \frac{\exp\left[n \left\{ \mathrm{H}(\boldsymbol{\Phi}^{\pi} | \boldsymbol{F}_{\mathcal{S}}) + \mathrm{H}(\boldsymbol{\Phi}^{\mathrm{T}} | \boldsymbol{F}_{\mathcal{SA}}) \right\} \right]}{n^{IJ} (n+1)^{IJ + I^2 JK}}, && (97) \\
&= \exp\left[n \left\{ \mathrm{H}(\boldsymbol{\Phi}^{\pi} | \boldsymbol{F}_{\mathcal{S}}) + \mathrm{H}(\boldsymbol{\Phi}^{\mathrm{T}} | \boldsymbol{F}_{\mathcal{SA}}) - \frac{(IJ + I^2 JK) \log(n+1) + IJ \log n}{n} \right\} \right], && (98) \\
&\geq \exp\left[n \left\{ \mathrm{H}(\boldsymbol{\Gamma}^{\pi} | \mathbf{V}) + \mathrm{H}(\boldsymbol{\Gamma}^{\mathrm{T}} | \mathbf{W}) - \zeta_n \right\} \right], && (99)
\end{aligned}
$$

where (99) is derived from Lemma A.2 and

$$
\zeta_n(I, J, K, \kappa_n, \xi_n, \lambda_n) \stackrel{\mathrm{def}}{=} \rho_n(I, J, K, \kappa_n, \xi_n, \lambda_n) + \frac{(IJ + I^2 JK) \log(n+1) + IJ \log n}{n}. \tag{100}
$$

Next, we consider the upper bound. By (65) and (22), we have

$$
\begin{aligned}
|\mathcal{C}^n_{\lambda_n}(\boldsymbol{\Gamma})| &\leq \bigcup_{\substack{\boldsymbol{\Phi} \in \boldsymbol{\Lambda}_n : \\ \mathrm{D}(\boldsymbol{\Phi}^{\pi} \| \boldsymbol{\Gamma}^{\pi} | \boldsymbol{F}_{\mathcal{S}}) + \mathrm{D}(\boldsymbol{\Phi}^{\mathrm{T}} \| \boldsymbol{\Gamma}^{\mathrm{T}} | \boldsymbol{F}_{\mathcal{SA}}) \leq \lambda_n}} |\mathcal{C}^n(\boldsymbol{\Phi}, \boldsymbol{F}_{\mathcal{S}}, \boldsymbol{F}_{\mathcal{SA}})|, && (101) \\
&\leq I (n+1)^{IJ + I^2 JK} \exp\left[n \left\{ \mathrm{H}(\boldsymbol{\Phi}^{\pi} | \boldsymbol{F}_{\mathcal{S}}) + \mathrm{H}(\boldsymbol{\Phi}^{\mathrm{T}} | \boldsymbol{F}_{\mathcal{SA}}) \right\} \right], && (102) \\
&\leq \exp\left[n \left\{ \mathrm{H}(\boldsymbol{\Gamma}^{\pi} | \mathbf{V}) + \mathrm{H}(\boldsymbol{\Gamma}^{\mathrm{T}} | \mathbf{W}) + \eta_n \right\} \right], && (103)
\end{aligned}
$$

where (103) is derived from Lemma A.2 and

$$
\eta_n(I, J, K, \kappa_n, \xi_n, \lambda_n) \stackrel{\mathrm{def}}{=} \rho_n(I, J, K, \kappa_n, \xi_n, \lambda_n) + \frac{(IJ + I^2 JK) \log(n+1) + \log I}{n}. \tag{104}
$$

Thus we have proved the upper and lower bounds in Theorem 3.3. ∎

B.7 Proof of Theorem 4.1

When the agent's return is maximized, obviously the subset of best sequences has to be included within the $\mathbf{\Gamma}$-typical set. Hence, $\mathcal{X}_n^\dagger \subset \mathcal{C}_{\lambda_n}^n(\mathbf{\Gamma})$ holds and then

$$\Pr(\mathcal{X}_n^\dagger) = \Pr(\mathcal{C}_{\lambda_n}^n(\mathbf{\Gamma}))\Pr(\mathcal{X}_n^\dagger|\mathcal{C}_{\lambda_n}^n(\mathbf{\Gamma})). \tag{105}$$

From Theorem 3.1, for sufficient large n,

$$\Pr(\mathcal{C}_{\lambda_n}^n(\mathbf{\Gamma})) \approx 1. \tag{106}$$

Since from Theorem 3.2 every element in $\mathcal{C}_{\lambda_n}^n(\mathbf{\Gamma})$ has the same probability for sufficient large n, we have

$$\Pr(\mathcal{X}_n^\dagger|\mathcal{C}_{\lambda_n}^n(\mathbf{\Gamma})) \to \frac{|\mathcal{X}_n^\dagger|}{|\mathcal{C}_{\lambda_n}^n(\mathbf{\Gamma})|}. \tag{107}$$

Therefore, (43) holds. Then, from the definition of $\mathcal{X}_n^\dagger$, as $n \to \infty$ and $\mathbf{\Gamma} \to \mathbf{\Gamma}^\dagger$ clearly we obtain $\Pr(\mathcal{X}_n^\dagger) \to 1$. ∎

B.8 Proof of Theorem 4.2

Differentiating (36) with respect to β, we have

$$\frac{d\pi(\beta, Q_{ij})}{d\beta} = \frac{\exp(\beta Q_{ij})\left(Q_{ij}\mathrm{Z}_i(\beta) - \sum_{j=1}^J \{Q_{ij}\exp(\beta Q_{ij})\}\right)}{(\mathrm{Z}_i(\beta))^2}. \tag{108}$$

Using (108) we get

$$\begin{aligned}
\frac{d}{d\beta}\mathrm{H}(\boldsymbol{G}_{(i)}) &= -\sum_{j=1}^J \frac{d}{d\beta}\left(\pi(\beta, Q_{ij})\log\pi(\beta, Q_{ij})\right), && (109)\\
&= -\sum_{j=1}^J (\log\pi(\beta, Q_{ij}) + 1)\frac{d\pi(\beta, Q_{ij})}{d\beta}, && (110)\\
&= \frac{\beta}{(\mathrm{Z}_i(\beta))^2}\left\{\left(\sum_{j=1}^J Q_{ij}\exp(\beta Q_{ij})\right)^2 - \mathrm{Z}_i(\beta)\left(\sum_{j=1}^J Q_{ij}^2\exp(\beta Q_{ij})\right)\right\}, && (111)\\
&= -\frac{\beta}{2(\mathrm{Z}_i(\beta))^2}\sum_{j=1}^J\sum_{j'=1}^J (Q_{ij} - Q_{ij'})^2\exp(\beta(Q_{ij} + Q_{ij'})), && (112)\\
&\le 0. && (113)
\end{aligned}$$

Therefore, by $d\mathrm{H}(\mathbf{\Gamma}^\pi|\mathbf{V})/d\beta = \sum_{i=1}^I v_i(d\mathrm{H}(\boldsymbol{G}_{(i)})/d\beta)$ we obtain (45). Also, on the limit $\beta \to \infty$, $\mathrm{H}(\mathbf{P}_{(i)}) = 0$ holds for all i. Hence we obtain (46). ∎

B.9 Proof of Theorem 4.3

Differentiating (38) with respect to ϵ, we have

$$\frac{d\pi(\epsilon, Q_{ij})}{d\epsilon} = \frac{1}{J_i} - \theta_{ij}. \tag{114}$$

By (114) we have

$$-\frac{d}{d\epsilon}\mathrm{H}(\boldsymbol{G}_{(i)}) = \sum_{j=1}^{J} \frac{d}{d\epsilon}\left(\pi(\epsilon, Q_{ij}) \log \pi(\epsilon, Q_{ij})\right), \tag{115}$$

$$= \sum_{j=1}^{J} (\log \pi(\epsilon, Q_{ij}) + 1) \frac{d\pi(\epsilon, Q_{ij})}{d\epsilon}, \tag{116}$$

$$= \frac{1}{J_i} \sum_{j=1}^{J} \log\left(\frac{\epsilon}{J_i} + (1-\epsilon)\theta_{ij}\right) - \sum_{j=1}^{J} \theta_{ij} \log\left(\frac{\epsilon}{J_i} + (1-\epsilon)\theta_{ij}\right) - \sum_{j=1}^{J} \theta_{ij} + 1, \tag{117}$$

$$= \left(1 - \frac{1}{J_i}\right)\left(\log \frac{\epsilon}{J_i} - \log\left(\frac{\epsilon}{J_i} + 1 - \epsilon\right)\right), \tag{118}$$

$$\leq 0. \tag{119}$$

Accordingly, using $d\mathrm{H}(\boldsymbol{\Gamma}^{\pi}|\mathbf{V})/d\epsilon = \sum_{i=1}^{I} v_i (d\mathrm{H}(\boldsymbol{G}_{(i)})/d\epsilon)$ we obtain (47). Also, on the limit $\epsilon \to \infty$, $\mathrm{H}(\mathbf{P}_{(i)}) = 0$ holds for all i. Therefore we have (46). ∎

B.10 Proof of Theorem 4.4

First, let us consider the lower bound. Since the length $l(q)$ of the program q satisfies the Kraft inequality (Cover & Thomas, 1991, p. 154), we have

$$\sum_{q:\mathcal{U}(q)\mathrm{halts}} \exp(-l(q)) \leq 1. \tag{120}$$

We assign to each $\boldsymbol{x}$ the length of the shortest program q such that $\mathcal{U}(q, 3n) = \boldsymbol{x} \in \mathcal{C}^n_{\lambda_n}(\boldsymbol{\Gamma})$. These shortest programs also satisfy the Kraft inequality. Since the expected codeword length must be greater than the entropy, we obtain the following lower bound,

$$\mathrm{E}_{\boldsymbol{\Gamma}}\left[\mathrm{K}_{\mathcal{U}}(\boldsymbol{x}|3n)\right] \geq \mathrm{E}_{\boldsymbol{\Gamma}}\left[-\log \Pr(\boldsymbol{x})\right], \tag{121}$$

$$\geq \log \nu + n\left\{\psi(\boldsymbol{\Gamma}) - \rho_n\right\}, \tag{122}$$

by (92).

Next we consider the upper bound. Let c denote a constant value. We describe the matrix $\boldsymbol{\Phi} \in \boldsymbol{\Lambda}_n$ with respect to the empirical sequence using $\log|\boldsymbol{\Lambda}_n|$ bits. Also, to describe the index of the empirical sequence within the set of all sequences having the same matrix of conditional types, $\log|\mathcal{C}^n(\boldsymbol{\Phi}, \boldsymbol{F}_{\mathcal{S}}, \boldsymbol{F}_{\mathcal{SA}})|$ bits are required because the set has less than $|\mathcal{C}^n(\boldsymbol{\Phi}, \boldsymbol{F}_{\mathcal{S}}, \boldsymbol{F}_{\mathcal{SA}})|$ elements. Hence,

$$\mathrm{K}_{\mathcal{U}}(\boldsymbol{x}|3n) \leq \log|\mathcal{C}^n(\boldsymbol{\Phi}, \boldsymbol{F}_{\mathcal{S}}, \boldsymbol{F}_{\mathcal{SA}})| + \log|\boldsymbol{\Lambda}_n| + c, \tag{123}$$

$$\leq n\left\{\mathrm{H}(\boldsymbol{\Phi}^{\pi}|\boldsymbol{F}_{\mathcal{S}}) + \mathrm{H}(\boldsymbol{\Phi}^{\mathrm{T}}|\boldsymbol{F}_{\mathcal{SA}})\right\} + \log I + (IJ + I^2JK)\log(n+1) + c, \tag{124}$$

$$\leq n\left\{\psi(\boldsymbol{\Gamma}) + \rho_n\right\} + \log I + (IJ + I^2JK)\log(n+1) + c, \tag{125}$$

where (124) follows from Lemma A.1 and (65), and (125) follows from Lemma A.2. Again, taking the expectation and dividing (122) and (125) by n yields the upper bound of (56). ∎

References

Amari, S., & Han, T. S. (1989, March). Statistical inference under multiterminal rate restrictions: a differential geometric approach. *IEEE Transactions on Information Theory*, *35*(2), 217–227.

Chaitin, G. J. (1977). Algorithmic information theory. *IBM journal of research and development*, *21*(4), 350–359.

Chaitin, G. J. (1987). *Algorithmic information theory* (Vol. 1). Cambridge, UK: Cambridge University Press. (reprinted with revisions in 1988)

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory* (1st ed.). New York: John Wiley & Sons, Inc.

Csiszár, I. (1998, October). The method of types. *IEEE Transactions on Information Theory*, *44*(6), 2505–2523.

Csiszár, I., & Körner, J. (1997). *Information theory: coding theorems for discrete memoryless systems* (3rd ed.). Budapest, Hungary: Akadémiai Kiadó. (1st impression 1981, 2nd impression 1986)

Davisson, L. D., Longo, G., & Sgarro, A. (1981, July). The error exponent for the noiseless encoding of finite ergodic markov sources. *IEEE Transactions on Information Theory*, *27*(4), 431–438.

Dayan, P. (1992, May). The convergence of TD(λ) for general λ. *Machine Learning*, *8*(3), 341–362.

Dearden, R., Friedman, N., & Russell, S. (1998, July). Bayesian Q-learning. In *Proceedings of the 15th national conference on artificial intelligence* (pp. 761–768). Madison, Wisconsin: AAAI Press.

Dueck, G., & Körner, J. (1979, January). Reliability function of a discrete memoryless channel at rates above capacity. *IEEE Transactions on Information Theory*, *25*(1), 82–85.

Han, T. S., & Kobayashi, K. (2002). *Mathematics of information and coding* (Vol. 203). Providence, RI: American Mathematical Society.

Iwata, K., Ikeda, K., & Sakai, H. (2004, July). A new criterion using information gain for action selection strategy in reinforcement learning. *IEEE Transactions on Neural Networks*, *15*(4), 792–799.

Jaakkola, T., Jordan, M. I., & Singh, S. P. (1994). On the convergence of stochastic iterative dynamic programming algorithms. *Neural Computation*, *6*(6), 1185–1201.

Kullback, S. (1967, January). A lower bound for discrimination information in terms of variation. *IEEE Transactions on Information Theory*, *13*(1), 126–127.

Kushner, H. J., & Yin, G. G. (1997). *Stochastic approximation algorithms and applications* (Vol. 35). New York: Springer-Verlag.

Li, M., & Vitányi, P. (1997). *An introduction to kolmogorov complexity and its applications* (2nd ed.). New York: Springer-Verlag. (1st edition 1993)

Merhav, N. (1991, May). Universal coding with minimum probability of codeword length overflow. *IEEE Transactions on Information Theory*, *37*(3), 556–563.

Merhav, N., & Neuhoff, D. L. (1992, January). Variable-to-fixed length codes provide better large deviations performance than fixed-to-variable length codes. *IEEE Transactions on Information Theory*, *38*(1), 135–140.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423, 623–656.

Singh, S., Jaakkola, T., Littman, M. L., & Szepesvári, C. (2000). Convergence results for single-step on-policy reinforcement-learning algorithms. *Machine Learning*, *39*, 287–308.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

Tsitsiklis, J. N. (1994, September). Asynchronous stochastic approximation and Q-learning. *Machine Learning*, *16*(3), 185–202.

Watkins, C. J. C. H., & Dayan, P. (1992). Technical note: Q-learning. *Machine Learning*, *8*, 279–292.

Wolfowitz, J. (1978). *Coding theorems of information theory* (3rd ed.). Berlin: Springer-Verlag. (1st edition 1961, 2nd edition 1964)

List of Figures

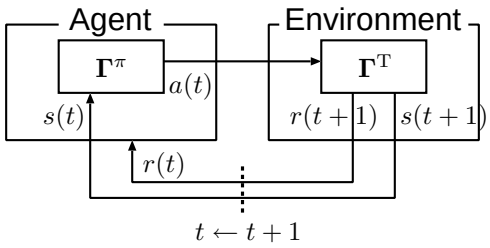


Figure 1: Interactions between the agent and the environment.

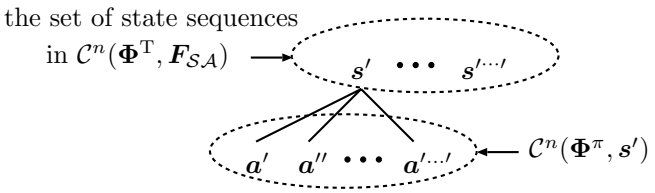


Figure 2: Structure of the $\boldsymbol{\Phi}$-shell.

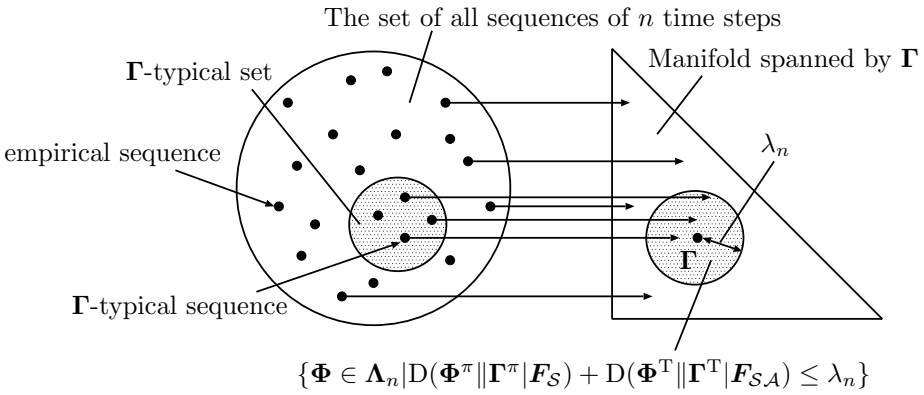
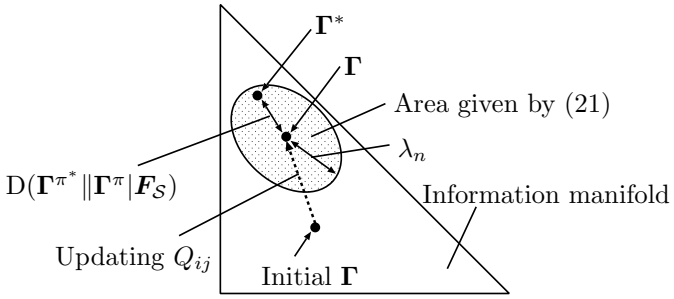
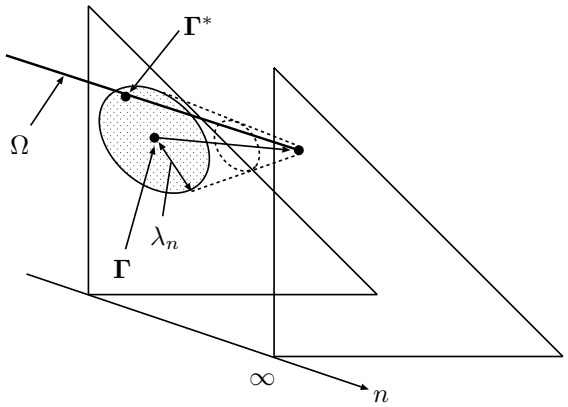


Figure 3: $\mathbf{\Gamma}$-typical set and $\mathbf{\Gamma}$-typical sequence.

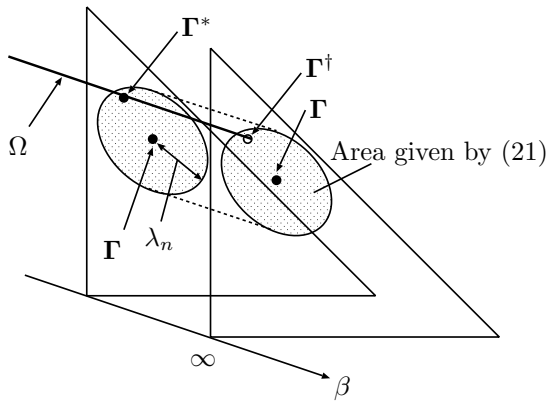


(a) Trajectory of $\mathbf{\Gamma}$ by updating Q_{ij}.

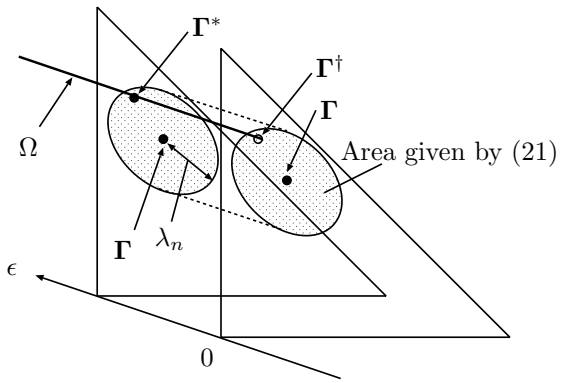


(b) Asymptotic decrease of λ_n.

Figure 4: Matrix $\mathbf{\Gamma}$ on the information manifold. Figure 4(a) illustrates the trajectory, drawn by updating the estimates of the action-value function using TD methods. Figure 4(b) shows the changes of λ_n with n.



(a) Ω in the softmax method.



(b) Ω in the ϵ-greedy method.

Figure 5: Set Ω in the softmax and ϵ-greedy methods. The matrix $\mathbf{\Gamma}^{\pi^*}$ varies with the changes of the parameter of the AS strategy, so that the set Ω is drawn as shown here.