Brain Inspired Cognitive Systems
August 29 – September 1, 2004
University of Stirling, Scotland, UK

Stochastic Complexity of Reinforcement Learning

Kazunori Iwata Kazushi Ikeda Hideaki Sakai

Department of Systems Science, Graduate School of Informatics, Kyoto University
Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501 Japan
{kiwata,kazushi,hsakai}@sys.i.kyoto-u.ac.jp

Abstract

Using the asymptotic equipartition property which holds on empirical sequences we elucidate the explicit performance of exploration, and the fact that the return maximization is characterized by two factors, the stochastic complexity and a quantity depending on the parameters of environment. We also examine the sensitivity of stochastic complexity, which is useful in appropriately tuning the parameters of the action selection strategy, and show the lower bound of the convergence speed of the divergence between the empirical sequence and the best empirical sequence which produces a maximal return.

Nomenclature reinforcement learning, Markov decision process, typical sequence, asymptotic equipartition property, stochastic complexity

1 Introduction

The weak law of large numbers in information theory is known as the asymptotic equipartition property (AEP) which was first stated in [1] and then developed by the type method in [2]. When a sequence of random variables is drawn independently according to an identical probability distribution for many times, 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. The AEP also holds on empirical sequences generated from a Markov decision process (MDP) in reinforcement learning [3]. It facilitates analysis of the learning process since most of our attention can be focused on the typical set of the empirical sequences. In this paper, with the AEP we elucidate the explicit performance of exploration, and the fact that the return maximization is characterized by two factors, the sum of conditional entropies and a quantity which depends on the parameters of environment. The sum of conditional entropies is referred to as stochastic complexity. We then examine the sensitivity of stochastic complexity, useful for appropriately tuning the parameters of the action selection strategy, and show that the lower bound of how fast the empirical sequence coincides with the best empirical sequence which yields a maximal return.

The organization of this paper is as follows. We introduce some notation and the AEP on empirical sequences in Section 2. Using the AEP we analyze the reinforcement learning process in Section 3. Finally, we give some conclusions in Section 4.

2 The AEP

We concentrate on the discrete-time MDP with discrete states and actions in this paper. Let $\mathcal{S} \triangleq \{s_1, s_2, \dots, s_I\}$ be the finite set of states of the environment, $\mathcal{A} \triangleq \{a_1, a_2, \dots, a_J\}$ be the finite set of actions, and $\Re_0 \triangleq \{r_1, r_2, \dots, r_K\} \subset \Re$ be the finite set of rewards which are discrete real numbers. Notice that $|\mathcal{S}| = I$, $|\mathcal{A}| = J$, and $|\Re_0| = K$. We assume that elements in these sets are recognized without error by the learner, hereinafter called the agent. We use t to denote a time step. The stochastic variables of state, action, and reward at time step t $(t = 1, 2, \dots)$ are written as $s(t)$, $a(t)$, and $r(t)$, respectively. 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 $\boldsymbol{x} \in (\mathcal{S} \times \mathcal{A} \times \Re_0)^n$ 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. Let $p^{(1)}(s_i) \triangleq \Pr(s(1) = s_i)$ be the initial probability distribution and $\mathbf{p}^{(1)} \triangleq (p^{(1)}(s_1), p^{(1)}(s_2), \dots, p^{(1)}(s_I))$. The agent learns the optimal policy which produces the maximal return by observing the empirical sequence. We use the term return to express the sum of rewards. The empirical sequence is drawn according to an ergodic MDP specified by the following three conditional probability distribution matrices. Henceforth, the conditional probability distribution matrix is simply called matrix. The policy matrix is an $I \times J$ matrix defined by

$$\mathbf{\Gamma}^{\pi} \triangleq \begin{pmatrix} \pi_{11} & \pi_{12} & \dots & \pi_{1J} \\ \pi_{21} & \pi_{22} & \dots & \pi_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ \pi_{I1} & \pi_{I2} & \dots & \pi_{IJ} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\pi}_{(1)} \\ \boldsymbol{\pi}_{(2)} \\ \vdots \\ \boldsymbol{\pi}_{(I)} \end{pmatrix}, \quad (1)$$

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

$$\mathbf{\Gamma}^{\mathrm{R}} \triangleq \begin{pmatrix} \mathrm{R}_{111} & \mathrm{R}_{112} & \dots & \mathrm{R}_{11K} \\ \mathrm{R}_{121} & \mathrm{R}_{122} & \dots & \mathrm{R}_{12K} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{R}_{1J1} & \mathrm{R}_{1J2} & \dots & \mathrm{R}_{1JK} \\ \mathrm{R}_{211} & \mathrm{R}_{212} & \dots & \mathrm{R}_{21K} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{R}_{IJ1} & \mathrm{R}_{IJ2} & \dots & \mathrm{R}_{IJK} \end{pmatrix} = \begin{pmatrix} \mathbf{R}_{(11)} \\ \mathbf{R}_{(12)} \\ \vdots \\ \mathbf{R}_{(1J)} \\ \mathbf{R}_{(21)} \\ \vdots \\ \mathbf{R}_{(IJ)} \end{pmatrix}, \quad (2)$$

where $\mathrm{R}_{ijk} \triangleq \Pr(r(t) = r_k | s(t) = s_i, a(t) = a_j)$. The state transition matrix is an $IJ \times I$ matrix defined by

$$\mathbf{\Gamma}^{\mathrm{T}} \triangleq \begin{pmatrix} \mathrm{T}_{111} & \mathrm{T}_{112} & \dots & \mathrm{T}_{11I} \\ \mathrm{T}_{121} & \mathrm{T}_{122} & \dots & \mathrm{T}_{12I} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{T}_{1J1} & \mathrm{T}_{1J2} & \dots & \mathrm{T}_{1JI} \\ \mathrm{T}_{211} & \mathrm{T}_{212} & \dots & \mathrm{T}_{21I} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{T}_{IJ1} & \mathrm{T}_{IJ2} & \dots & \mathrm{T}_{IJI} \end{pmatrix} = \begin{pmatrix} \mathbf{T}_{(11)} \\ \mathbf{T}_{(12)} \\ \vdots \\ \mathbf{T}_{(1J)} \\ \mathbf{T}_{(21)} \\ \vdots \\ \mathbf{T}_{(IJ)} \end{pmatrix}, \quad (3)$$

where $\mathrm{T}_{iji'} \triangleq \Pr(s(t+1) = s_{i'} | s(t) = s_i, a(t) = a_j)$. The agent does not know $\mathbf{\Gamma}^{\mathrm{R}}$ and $\mathbf{\Gamma}^{\mathrm{T}}$ of the environment but the system is simulated and observed under any choice of actions. We assume that $\mathbf{\Gamma}^{\mathrm{R}}$ and $\mathbf{\Gamma}^{\mathrm{T}}$ are constant and that for simplicity of analysis $\mathbf{\Gamma}^{\pi}$ is fixed for n time steps where n is sufficiently large. For notational simplicity we define $\mathbf{\Gamma} \triangleq (\mathbf{\Gamma}^{\pi}, \mathbf{\Gamma}^{\mathrm{R}}, \mathbf{\Gamma}^{\mathrm{T}})$. Since MDPs are characterized by the finite sets, the initial probability distribution, and the matrices, we denote the MDP by $\mathrm{M}(\mathcal{S}, \mathcal{A}, \Re_0, \mathbf{p}^{(1)}, \mathbf{\Gamma})$.

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, $\boldsymbol{x} = (\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{r}) \in (\mathcal{S} \times \mathcal{A} \times \Re_0)^n$. In a similar manner, let n_{ij} $(n_{ij} \leq n_i)$ be the number of occurrences of t such that $(s(t), a(t)) = (s_i, a_j) \in \mathcal{S} \times \mathcal{A}$, and let n_{ijk} $(n_{ijk} \leq n_{ij})$ be the number of occurrences of t such that $(s(t), a(t), r(t)) = (s_i, a_j, r_k) \in \mathcal{S} \times \mathcal{A} \times \Re_0$ in the empirical sequence. With an additional "cyclic" convention that $s(n)$, $a(n)$, and $r(n)$ precede $s(1)$, $a(1)$, and $r(1)$, let $n_{iji'}$ $(n_{iji'} \leq n_{ij})$ denote the number of occurrences of t such that $(s(t), a(t), s(t+1)) = (s_i, a_j, s_{i'}) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ 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 nonnegative numbers n, n_i, n_{ij}, n_{ijk}, and $n_{iji'}$ is expressed as

$$n = \sum_{i=1}^{I} n_i = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} n_{ijk} = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{i'=1}^{I} n_{iji'}. \quad (4)$$

Now we define the type of $s_i \in \mathcal{S}$ by $f_i \triangleq n_i/n$. Also, the joint type of $(s_i, a_j) \in \mathcal{S} \times \mathcal{A}$ is defined as $f_{ij} \triangleq n_{ij}/n$. Let us denote all the types and the joint types by

$$\boldsymbol{F}(\boldsymbol{s}) \triangleq (f_1, f_2, \dots, f_I) \quad (5)$$

and

$$\boldsymbol{F}(\boldsymbol{s}, \boldsymbol{a}) \triangleq \begin{pmatrix} f_{11} & f_{12} & \dots & f_{1J} \\ f_{21} & f_{22} & \dots & f_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ f_{I1} & f_{I2} & \dots & f_{IJ} \end{pmatrix}, \quad (6)$$

respectively. In this case we say that $\boldsymbol{s}$ and $(\boldsymbol{s}, \boldsymbol{a})$ have the type $\boldsymbol{F}(\boldsymbol{s})$ and the joint type $\boldsymbol{F}(\boldsymbol{s}, \boldsymbol{a})$, respectively.

Conditional Type Relative to Policy If $n_i > 0$ for all i, then the conditional type of $(s_i, a_j) \in \mathcal{S} \times \mathcal{A}$ given a state sequence $\boldsymbol{s} \in \mathcal{S}^n$ is defined as $g_{ij} \triangleq n_{ij}/n_i$. However, if there exists i such that $n_i = 0$, then we can not uniquely determine the conditional type. To avoid such a case, we consider the set of action sequences given any state sequence having type $\boldsymbol{F}(\boldsymbol{s})$ and $I \times J$ matrix $\mathbf{\Phi}^{\pi} : \mathcal{S} \rightarrow \mathcal{A}$ expressed as

$$\mathbf{\Phi}^{\pi} \triangleq \begin{pmatrix} g_{11} & g_{12} & \cdots & g_{1J} \\ g_{21} & g_{22} & \cdots & g_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ g_{I1} & g_{I2} & \cdots & g_{IJ} \end{pmatrix} = \begin{pmatrix} \boldsymbol{G}^{\pi}_{(1)} \\ \boldsymbol{G}^{\pi}_{(2)} \\ \vdots \\ \boldsymbol{G}^{\pi}_{(I)} \end{pmatrix}. \quad (7)$$

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, is referred to as $\boldsymbol{\Phi}^{\pi}$-shell [2, p. 31] and denoted by $\mathcal{C}^n(\boldsymbol{\Phi}^{\pi})$. The entire set of possible matrices $\boldsymbol{\Phi}^{\pi}$ for any state sequence with the type $\boldsymbol{F}(\boldsymbol{s})$ is written as $\boldsymbol{\Lambda}_n^{\pi}$.

Conditional Type Relative to Reward Similarly, we consider the set of reward sequences given any state and action sequences, hereinafter termed state-action sequences, having joint type $\boldsymbol{F}(\boldsymbol{s},\boldsymbol{a})$ and $IJ \times K$ matrix $\boldsymbol{\Phi}^{\mathrm{R}} : \mathcal{S} \times \mathcal{A} \to \Re_0$ denoted by

$$\boldsymbol{\Phi}^{\mathrm{R}} \triangleq \begin{pmatrix} g_{111} & g_{112} & \cdots & g_{11K} \\ g_{121} & g_{122} & \cdots & g_{12K} \\ \vdots & \vdots & \ddots & \vdots \\ g_{1J1} & g_{1J2} & \cdots & g_{1JK} \\ g_{211} & g_{212} & \cdots & g_{21K} \\ \vdots & \vdots & \ddots & \vdots \\ g_{IJ1} & g_{IJ2} & \cdots & g_{IJK} \end{pmatrix} = \begin{pmatrix} \boldsymbol{G}^{\mathrm{R}}_{(11)} \\ \boldsymbol{G}^{\mathrm{R}}_{(12)} \\ \vdots \\ \boldsymbol{G}^{\mathrm{R}}_{(1J)} \\ \boldsymbol{G}^{\mathrm{R}}_{(21)} \\ \vdots \\ \boldsymbol{G}^{\mathrm{R}}_{(IJ)} \end{pmatrix}. \qquad (8)$$

The set of reward sequences is termed $\boldsymbol{\Phi}^{\mathrm{R}}$-shell and denoted by $\mathcal{C}^n(\boldsymbol{\Phi}^{\mathrm{R}})$. The entire set of possible matrices $\boldsymbol{\Phi}^{\mathrm{R}}$ for any state-action sequences with the joint type $\boldsymbol{F}(\boldsymbol{s},\boldsymbol{a})$ is written as $\boldsymbol{\Lambda}_n^{\mathrm{R}}$.

Conditional Markov Type Relative to State Transition In a slightly different manner, we need to deal with the conditional Markov type. We consider the set of state sequences such that the joint type is $\boldsymbol{F}(\boldsymbol{s},\boldsymbol{a})$ given any action sequence and $IJ \times I$ matrix $\boldsymbol{\Phi}^{\mathrm{T}} : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ designated by

$$\boldsymbol{\Phi}^{\mathrm{T}} \triangleq \begin{pmatrix} g_{111} & g_{112} & \cdots & g_{11I} \\ g_{121} & g_{122} & \cdots & g_{12I} \\ \vdots & \vdots & \ddots & \vdots \\ g_{1J1} & g_{1J2} & \cdots & g_{1JI} \\ g_{211} & g_{212} & \cdots & g_{21I} \\ \vdots & \vdots & \ddots & \vdots \\ g_{IJ1} & g_{IJ2} & \cdots & g_{IJI} \end{pmatrix} = \begin{pmatrix} \boldsymbol{G}^{\mathrm{T}}_{(11)} \\ \boldsymbol{G}^{\mathrm{T}}_{(12)} \\ \vdots \\ \boldsymbol{G}^{\mathrm{T}}_{(1J)} \\ \boldsymbol{G}^{\mathrm{T}}_{(21)} \\ \vdots \\ \boldsymbol{G}^{\mathrm{T}}_{(IJ)} \end{pmatrix}. \qquad (9)$$

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

For simplicity, we define $\boldsymbol{\Phi} \triangleq (\boldsymbol{\Phi}^{\pi}, \boldsymbol{\Phi}^{\mathrm{R}}, \boldsymbol{\Phi}^{\mathrm{T}})$ and $\boldsymbol{\Lambda}_n \triangleq \boldsymbol{\Lambda}_n^{\pi} \times \boldsymbol{\Lambda}_n^{\mathrm{R}} \times \boldsymbol{\Lambda}_n^{\mathrm{T}}$. The set of empirical sequences that consists of the $\boldsymbol{\Phi}^{\pi}$-shell, $\boldsymbol{\Phi}^{\mathrm{R}}$-shell, and $\boldsymbol{\Phi}^{\mathrm{T}}$-shell is called $\boldsymbol{\Phi}$-shell and denoted by $\mathcal{C}^n(\boldsymbol{\Phi}) \triangleq \mathcal{C}^n(\boldsymbol{\Phi}^{\pi}) \times \mathcal{C}^n(\boldsymbol{\Phi}^{\mathrm{R}}) \times \mathcal{C}^n(\boldsymbol{\Phi}^{\mathrm{T}})$. In this case we write that the empirical sequence has the conditional type matrix $\boldsymbol{\Phi}$.

2.2 Main Theorems

To show the AEP we have to introduce the following sequences.

Definition 2.1 ($\mathbf{V}$- and $\mathbf{W}$-typical sequences [3]) *We assume the existence of the following two unique stationary probability distributions,*

$$\mathbf{V} \triangleq (v_1, v_2, \ldots, v_I) \qquad (10)$$

and

$$\mathbf{W} \triangleq \begin{pmatrix} w_{11} & w_{12} & \cdots & w_{1J} \\ w_{21} & w_{22} & \cdots & w_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ w_{I1} & w_{I2} & \cdots & w_{IJ} \end{pmatrix}, \qquad (11)$$

and that as $n \to \infty$ $\boldsymbol{F}(\boldsymbol{s})$ and $\boldsymbol{F}(\boldsymbol{s},\boldsymbol{a})$ tend to $\mathbf{V}$ and $\mathbf{W}$, respectively. The stationary probability distributions are uniquely determined by the MDP, $\mathrm{M}(\mathcal{S}, \mathcal{A}, \Re_0, \mathbf{p}^{(1)}, \boldsymbol{\Gamma})$. 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 $\boldsymbol{F}(\boldsymbol{s})$ of a state sequence $\boldsymbol{s} \in \mathcal{S}^n$ satisfies

$$\mathrm{D}(\boldsymbol{F}(\boldsymbol{s}) \| \mathbf{V}) = \sum_{i=1}^{I} f_i \log \frac{f_i}{v_i} \le \kappa_n, \qquad (12)$$

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

$$\mathrm{D}(\boldsymbol{F}(\boldsymbol{s},\boldsymbol{a}) \| \mathbf{W}) = \sum_{i=1}^{I} \sum_{j=1}^{J} f_{ij} \log \frac{f_{ij}}{w_{ij}} \le \xi_n \qquad (13)$$

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 $\mathcal{C}^n_{\xi_n}(\mathbf{W}) \triangleq \{(\boldsymbol{s},\boldsymbol{a}) \in (\mathcal{S} \times \mathcal{A})^n | \mathrm{D}(\boldsymbol{F}(\boldsymbol{s},\boldsymbol{a}) \| \mathbf{W}) \le \xi_n\}$.

Then, let us define

$$\mathrm{D}(\boldsymbol{\Phi}^{\pi} \| \boldsymbol{\Gamma}^{\pi} | \boldsymbol{F}(\boldsymbol{s})) \triangleq \sum_{i=1}^{I} \sum_{j=1}^{J} f_i g_{ij} \log \frac{g_{ij}}{\pi_{ij}}, \qquad (14)$$

$$\mathrm{D}(\boldsymbol{\Phi}^{\mathrm{R}} \| \boldsymbol{\Gamma}^{\mathrm{R}} | \boldsymbol{F}(\boldsymbol{s},\boldsymbol{a})) \triangleq \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} f_{ij} g_{ijk} \log \frac{g_{ijk}}{\mathrm{R}_{ijk}}, \qquad (15)$$

$$\mathrm{D}(\boldsymbol{\Phi}^{\mathrm{T}} \| \boldsymbol{\Gamma}^{\mathrm{T}} | \boldsymbol{F}(\boldsymbol{s},\boldsymbol{a})) \triangleq \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{i'=1}^{I} f_{ij} g_{iji'} \log \frac{g_{iji'}}{\mathrm{T}_{iji'}}. \qquad (16)$$

We give definitions of the typical sequence and the typical set of empirical sequences, which will lead us to show that the AEP holds for empirical sequences.

Definition 2.2 ($\mathbf{\Gamma}$-typical sequence and $\mathbf{\Gamma}$-typical set [3]) *If the matrix $\mathbf{\Phi} \in \mathbf{\Lambda}_n$ of the conditional types with respect to an empirical sequence $\boldsymbol{x} = (\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{r}) \in (\mathcal{S} \times \mathcal{A} \times \Re_0)^n$ satisfies*

$$\mathrm{D}(\mathbf{\Phi}^{\pi} \| \mathbf{\Gamma}^{\pi} | \boldsymbol{F}(\boldsymbol{s})) + \mathrm{D}(\mathbf{\Phi}^{\mathrm{R}} \| \mathbf{\Gamma}^{\mathrm{R}} | \boldsymbol{F}(\boldsymbol{s}, \boldsymbol{a})) + \mathrm{D}(\mathbf{\Phi}^{\mathrm{T}} \| \mathbf{\Gamma}^{\mathrm{T}} | \boldsymbol{F}(\boldsymbol{s}, \boldsymbol{a})) \le \lambda_n, \quad (17)$$

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

$$\mathcal{C}^n_{\lambda_n}(\mathbf{\Gamma}) \triangleq \bigcup_{\substack{\mathbf{\Phi} \in \mathbf{\Lambda}_n : \mathrm{D}(\mathbf{\Phi}^{\pi} \| \mathbf{\Gamma}^{\pi} | \boldsymbol{F}(\boldsymbol{s})) \\ + \mathrm{D}(\mathbf{\Phi}^{\mathrm{R}} \| \mathbf{\Gamma}^{\mathrm{R}} | \boldsymbol{F}(\boldsymbol{s}, \boldsymbol{a})) + \mathrm{D}(\mathbf{\Phi}^{\mathrm{T}} \| \mathbf{\Gamma}^{\mathrm{T}} | \boldsymbol{F}(\boldsymbol{s}, \boldsymbol{a})) \le \lambda_n}} \mathcal{C}^n(\mathbf{\Phi}). \quad (18)$$

We are in a position to show the following three theorems regarding the AEP on empirical sequences.

Theorem 2.1 (Probability of $\mathbf{\Gamma}$-typical set [3]) *If $\lambda_n \to 0$ as $n \to \infty$ and λ_n satisfies*

$$\lambda_n - \frac{(IJ + IJK + I^2 J) \log(n+1) + \log I - \log \nu}{n} > 0, \quad (19)$$

where

$$\nu \triangleq \min_{1 \le i, i' \le I, 1 \le j \le J} \mathrm{T}_{iji'}, \quad (20)$$

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

$$\Pr\left(\mathcal{C}^n_{\lambda_n}(\mathbf{\Gamma})\right) = 1 - \epsilon_n(I, J, K, \lambda_n). \quad (21)$$

Note that $n\lambda_n \to \infty$ because of (19). This theorem implies that the probability of the $\mathbf{\Gamma}$-typical set asymptotically goes to one independently of the underlying probabilistic structures, $\mathbf{\Gamma}^{\pi}$, $\mathbf{\Gamma}^{\mathrm{R}}$, 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 2.2 (Equi-probability of $\mathbf{\Gamma}$-typical sequence [3]) *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,*

$$\frac{\log \nu}{n} - \rho_n \le -\frac{1}{n} \log \Pr(\boldsymbol{x}) - \left\{\mathrm{H}(\mathbf{\Gamma}^{\pi} | \mathbf{V}) + \mathrm{H}(\mathbf{\Gamma}^{\mathrm{R}} | \mathbf{W}) + \mathrm{H}(\mathbf{\Gamma}^{\mathrm{T}} | \mathbf{W})\right\} \le -\frac{\log \mu}{n} + \lambda_n + \rho_n, \quad (22)$$

where

$$\mu \triangleq \min_{1 \le i \le I} p^{(1)}(s_i), \quad (23)$$

$$\mathrm{H}(\mathbf{\Gamma}^{\pi} | \mathbf{V}) \triangleq -\sum_{i=1}^{I} \sum_{j=1}^{J} v_i \pi_{ij} \log \pi_{ij}, \quad (24)$$

$$\mathrm{H}(\mathbf{\Gamma}^{\mathrm{R}} | \mathbf{W}) \triangleq -\sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} w_{ij} \mathrm{R}_{ijk} \log \mathrm{R}_{ijk}, \quad (25)$$

$$\mathrm{H}(\mathbf{\Gamma}^{\mathrm{T}} | \mathbf{W}) \triangleq -\sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{i'=1}^{I} w_{ij} \mathrm{T}_{iji'} \log \mathrm{T}_{iji'}. \quad (26)$$

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 2.3 (Bounds of number of $\mathbf{\Gamma}$-typical sequences [3]) *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 $\zeta_n(I, J, K, \kappa_n, \xi_n, \lambda_n) \to 0$ and $\eta_n(I, J, K, \kappa_n, \xi_n, \lambda_n) \to 0$, respectively. Then, the number of elements in the $\mathbf{\Gamma}$-typical set is bounded by*

$$\exp\left[n\left\{\mathrm{H}(\mathbf{\Gamma}^{\pi} | \mathbf{V}) + \mathrm{H}(\mathbf{\Gamma}^{\mathrm{R}} | \mathbf{W}) + \mathrm{H}(\mathbf{\Gamma}^{\mathrm{T}} | \mathbf{W}) - \zeta_n\right\}\right] \le |\mathcal{C}^n_{\lambda_n}(\mathbf{\Gamma})| \le \exp\left[n\left\{\mathrm{H}(\mathbf{\Gamma}^{\pi} | \mathbf{V}) + \mathrm{H}(\mathbf{\Gamma}^{\mathrm{R}} | \mathbf{W}) + \mathrm{H}(\mathbf{\Gamma}^{\mathrm{T}} | \mathbf{W}) + \eta_n\right\}\right]. \quad (27)$$

The ratio of the number of $\mathbf{\Gamma}$-typical sequences to that of all empirical sequences $\boldsymbol{x} \in (\mathcal{S} \times \mathcal{A} \times \Re_0)^n$ of n time steps is

$$\frac{|\mathcal{C}^n_{\lambda_n}(\mathbf{\Gamma})|}{(IJK)^n} \le \exp\Big[n\Big\{\mathrm{H}(\mathbf{\Gamma}^{\pi} | \mathbf{V}) + \mathrm{H}(\mathbf{\Gamma}^{\mathrm{R}} | \mathbf{W}) + \mathrm{H}(\mathbf{\Gamma}^{\mathrm{T}} | \mathbf{W}) + \eta_n - \log I - \log J - \log K\Big\}\Big] \to 0, \quad (28)$$

as $n \to \infty$. 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 that the total probability is almost one.

3 Analysis of Reinforcement Learning

The process of return maximization (RM) in reinforcement learning is analyzed using the AEP in this section. We first give a review of temporal difference (TD) learning and typical action selection (AS) strategies.

3.1 TD Learning and AS Strategy

Let Q_{ij} denote the estimate of an action value called the Q-function [4, 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 $s_i \in \mathcal{S}$ and let $|\mathcal{A}_i| = J_i$. We use α_n to denote the learning rate at time step n and γ to denote the discount factor that controls the relative importance of an immediate reward and delayed rewards. For any i, j and i' the (one-step) TD learning [4, chapter 6] has the update form,

$$Q_{ij} \leftarrow Q_{ij} + \alpha_n \delta Q_{iji'}, \tag{29}$$

where $\delta Q_{iji'}$ is written as

$$\delta Q_{iji'} = r + \gamma \max_{j' \in \mathcal{A}_{i'}} Q_{i'j'} - Q_{ij}, \tag{30}$$

where r denotes an immediate reward, in Q-learning [5], for example. The update is done after every transition from state-action (s_i, a_j) to subsequent state $s_{i'}$. By sufficient iterations of (29) for every i, j, all the estimates of the Q-function converge to the expected values. We denote the expected value by $Q^*_{ij} \triangleq \mathrm{E}[Q_{ij}]$ henceforth. In Q-learning the convergence is guaranteed under certain conditions such as sufficient iterations [5].

Now we review the following two AS strategies which have been employed in many cases. The softmax method [4, chapter 2] is the most popular strategy and is also termed the Boltzman method when the exponential function is used. Recall that π_{ij} denotes the probability that the agent chooses an action a_j in a state s_i. The policy probability is defined as

$$\pi_{ij} \triangleq \pi\left(\beta, Q_{ij}\right) = \frac{\exp(\beta Q_{ij})}{\mathrm{Z}_i(\beta)}, \tag{31}$$

where the partition function is $\mathrm{Z}_i(\beta) \triangleq \sum_{j' \in \mathcal{A}_i} \exp(\beta Q_{ij'})$. 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.

In the ε-greedy Method [4, chapter 2], with probability ε, the agent randomly chooses an action. On the other hand, the agent chooses the best action with the largest estimated value with probability $1 - \varepsilon$. That is, π_{ij} is given by

$$\pi_{ij} \triangleq \pi\left(\varepsilon, Q_{ij}\right) = \frac{\varepsilon}{J_i} + (1 - \varepsilon)\theta_{ij}, \tag{32}$$

where

$$\theta_{ij} \triangleq \begin{cases} 1 & \text{if } j = \arg\max_{j' \in \mathcal{A}_i} Q_{ij'} \\ 0 & \text{if } j \neq \arg\max_{j' \in \mathcal{A}_i} Q_{ij'} \end{cases}. \tag{33}$$

The parameter ε is gradually decreased such that $\varepsilon \to 0$ 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 [4, 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, or more concretely, their effect on RM and the sensitivity of exploratory performance with respect to the parameters.

3.2 Performance of Exploration

We assume that the policy is improved sufficiently slowly such that the AEP holds. Figures 1 and 2 illustrate a reinforcement learning process on the manifold spanned by $\mathbf{\Gamma}$. This manifold is called the information manifold (IM) [6]. Recall that for any i, j the expected value of Q_{ij} is denoted by Q^*_{ij}. Let $\pi^*_{ij} = \pi(\beta, Q^*_{ij})$ for any β in the softmax method and $\pi^*_{ij} = \pi(\varepsilon, Q^*_{ij})$ for any ε in the ε-greedy method. Let $\mathbf{\Gamma}^{\pi^*}$ be the policy matrix whose components are given by π^*_{ij}. We define $\mathbf{\Gamma}^* \triangleq (\mathbf{\Gamma}^{\pi^*}, \mathbf{\Gamma}^{\mathrm{R}}, \mathbf{\Gamma}^{\mathrm{T}})$ and write the set of $\mathbf{\Gamma}^*$ as $\mathbf{\Omega} \triangleq \{\mathbf{\Gamma} | \mathbf{\Gamma}^{\pi} = \mathbf{\Gamma}^{\pi^*}\}$ for notational convenience. The set $\mathbf{\Omega}$ is given by changing the parameter of AS strategy, such as β and ε. Let us denote the set of best empirical sequences that yield maximal return by $\{\boldsymbol{x}^\dagger\}$. The optimal policy matrix that has the largest probability of the set $\{\boldsymbol{x}^\dagger\}$ appearing is denoted by $\mathbf{\Gamma}^{\pi^\dagger}$ where components are

$$\pi^\dagger_{ij} = \begin{cases} 1 & \text{if } j = \arg\max_{j' \in \mathcal{A}_i} Q^*_{ij'} \\ 0 & \text{if } j \neq \arg\max_{j' \in \mathcal{A}_i} Q^*_{ij'} \end{cases}. \tag{34}$$

For example, in the softmax method we can write it as $\mathbf{\Gamma}^{\pi^\dagger} = \{\pi^\dagger_{ij} = \pi(\infty, Q^*_{ij})\}_{ij}$, and in the ε-greedy method we can also write it as $\mathbf{\Gamma}^{\pi^\dagger} = \{\pi^\dagger_{ij} = \pi(0, Q^*_{ij})\}_{ij}$. Also, we define $\mathbf{\Gamma}^\dagger \triangleq (\mathbf{\Gamma}^{\pi^\dagger}, \mathbf{\Gamma}^{\mathrm{R}}, \mathbf{\Gamma}^{\mathrm{T}})$. We assume that the neighborhood of the optimal matrix on the IM is smooth for the parameters of the AS strategy, such as β and ε. If the environment, or specifically, the reward matrix $\mathbf{\Gamma}^{\mathrm{R}}$ and the state transition matrix $\mathbf{\Gamma}^{\mathrm{T}}$ are 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 (SC) which will play an important role in the later discussion.

Definition 3.1 (Stochastic complexity) *The SC is defined by*

$$\psi(\mathbf{\Gamma}) \triangleq \mathrm{H}(\mathbf{\Gamma}^{\pi} | \mathbf{V}) + \mathrm{H}(\mathbf{\Gamma}^{\mathrm{R}} | \mathbf{W}) + \mathrm{H}(\mathbf{\Gamma}^{\mathrm{T}} | \mathbf{W}). \tag{35}$$

This is referred to as complexity since the value of $\psi(\mathbf{\Gamma})$ is closely related to the algorithmic complexity [7].

We shall show that the SC sets the performance of exploration. Within the framework of reinforcement learning the

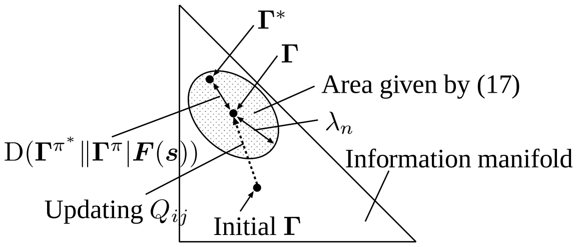
$\mathbf{\Gamma}^*$
$\mathbf{\Gamma}$
Area given by (17)
λ_n
$\mathrm{D}(\mathbf{\Gamma}^{\pi^*}\|\mathbf{\Gamma}^{\pi}|\boldsymbol{F}(\boldsymbol{s}))$
Information manifold
Updating Q_{ij}
Initial $\mathbf{\Gamma}$

Figure 1: Trajectory of updating Q_{ij}

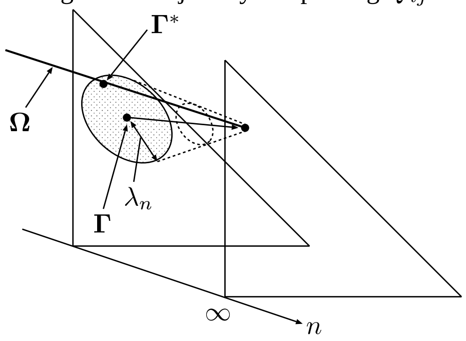

Figure 2: Asymptotic decrease of λ_n

agent learns a policy based only on observed rewards because the optimal selections are not directly instructed. The agent accordingly has to perform an explicit trial-and-error search for finding better actions, especially in the early stages of learning. In other words, the agent is required 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 occurs with probability almost one according to Theorem 2.1. Such a policy for exploratory search is termed exploration. On the other hand, using estimates of the Q-function the agent has to select the best action with the largest estimate of the Q-function to maximize the future return. This aim corresponds to making the $\mathbf{\Gamma}$-typical set smaller, so that only few empirical sequences which yield high return are allowed to be generated in practice. Such a policy for RM is termed exploitation. This tradeoff is well-known as the exploration-exploitation dilemma in reinforcement learning [4, chapter 2]. Theorem 2.3 states that the number of elements in the $\mathbf{\Gamma}$-typical set is characterized by the SC $\psi(\mathbf{\Gamma})$ and the quantity λ_n. Since λ_n depends on n the agent can control only the SC by changing the AS strategy, and the larger the value of $\psi(\mathbf{\Gamma})$, the greater the number of the $\mathbf{\Gamma}$-typical sequences. Thus, this naturally leads to the following definition.

Definition 3.2 (Performance of exploration) *Under the reward matrix $\mathbf{\Gamma}^{\mathrm{R}}$ and the state transition matrix $\mathbf{\Gamma}^{\mathrm{T}}$ of the environment the performance of the exploration of the policy matrix $\mathbf{\Gamma}^{\pi}$ is given by the size $|\mathcal{C}^n_{\lambda_n}(\mathbf{\Gamma})|$ of the $\mathbf{\Gamma}$-typical set, which is determined essentially by the SC $\psi(\mathbf{\Gamma})$.*

This implies that if the value of $\psi(\mathbf{\Gamma})$ is large, then the policy is exploratory and that if the value is small, then the policy is exploitative. Thus we can mathematically describe the term "exploration" from the aspect of information theory. Using the property of this definition, a neat AS strategy has been proposed in [8].

3.3 Return Maximization

We will show the relationship between the SC and RM in reinforcement learning. Maximizing return corresponds to the set of best empirical sequences having probability nearly one, that is, $\mathcal{C}^n_{\lambda_n}(\mathbf{\Gamma}) \simeq \{\boldsymbol{x}^{\dagger}\}$ under a proper AS strategy so that the estimates of the Q-function eventually converge to the expected values. Hence we consider that $\{\boldsymbol{x}^{\dagger}\} \subset \mathcal{C}^n_{\lambda_n}(\mathbf{\Gamma})$ and then reduce the $\mathbf{\Gamma}$-typical set such that $\mathcal{C}^n_{\lambda_n}(\mathbf{\Gamma}) \simeq \{\boldsymbol{x}^{\dagger}\}$. Here the key points are that

- by updating the estimates we have 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,

$$\mathrm{D}(\mathbf{\Gamma}^{\pi^*}\|\mathbf{\Gamma}^{\pi}|\boldsymbol{F}(\boldsymbol{s})) \leq \lambda_n, \tag{36}$$

  (see Figure 1), and then while keeping (36)

- we are required to shut out empirical sequences except the best empirical sequence from the $\mathbf{\Gamma}$-typical set in order to assign high probability to the best empirical sequence (see Figure 2).

The algorithm for the former is simply TD learning. It is known that the convergence order of TD learning is at most $1/\sqrt{n}$ [9]. The goal of the latter is to make the number of elements in the $\mathbf{\Gamma}$-typical set small while satisfying (36). This leads to the result that the set of the best empirical sequences occurs with high probability because according to Theorem 2.2 all the $\mathbf{\Gamma}$-typical sequences of length n have the same probability for sufficiently large n. From Theorem 2.3 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 we 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 3.1 (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*

$$\frac{d\psi(\mathbf{\Gamma})}{d\beta} = \sum_{i=1}^{I} v_i \Bigg\{ \frac{-\beta}{2(\mathrm{Z}_i(\beta))^2} \sum_{j=1}^{J} \sum_{j'=1}^{J} \Bigg((Q_{ij} - Q_{ij'})^2 \exp(\beta(Q_{ij} + Q_{ij'})) \Bigg) \Bigg\}. \tag{37}$$

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

$$\psi(\mathbf{\Gamma}) \to \mathrm{H}(\mathbf{\Gamma}^{\mathrm{R}}|\mathbf{W}) + \mathrm{H}(\mathbf{\Gamma}^{\mathrm{T}}|\mathbf{W}). \tag{38}$$

Theorem 3.2 (Relationship between ε and SC) *The value of $\psi(\mathbf{\Gamma})$ decreases as $\varepsilon \to 0$. The derivative of $\psi(\mathbf{\Gamma})$ with respect to ε is*

$$\frac{d\psi(\mathbf{\Gamma})}{d\varepsilon} = \sum_{i=1}^{I} v_i \Bigg\{ \Bigg(\frac{1}{J_i} - 1 \Bigg) \log \Bigg(\frac{\varepsilon}{J_i} + 1 - \varepsilon \Bigg) + \Bigg(1 - \frac{1}{J_i} \Bigg) \log \frac{\varepsilon}{J_i} \Bigg\}. \tag{39}$$

In particular, if $\varepsilon \to 0$, then $\psi(\mathbf{\Gamma})$ coincides with (38).

We omit the proofs because of the limitation of paper length. Note that from Definition 3.2 (37) and (39) denote the sensitivity for the performance of exploration. The sensitivity may be an important guide for tuning the parameters appropriately. The main difference between the two methods is that estimates of the Q-function affect the derivative of the SC directly in the softmax method but not in the ε-greedy method. Next, we will consider another important factor λ_n for making the number of elements in the $\mathbf{\Gamma}$-typical set smaller. Figure 2 shows the changes of λ_n with n where the lower bound of λ_n is given by (19). The lower 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 $(IJ + IJK + I^2 J)$. This means that we can not accomplish the RM faster than this rate even if we know all the values of Q^*_{ij} in advance. 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 Conclusions

In this paper, using the AEP on empirical sequences we elucidated that the SC exhibits the performance of exploration in the sense of information theory, and the fact that the RM is characterized by the SC $\psi(\mathbf{\Gamma})$ and the quantity λ_n under a proper AS strategy. We can control only the SC by tuning the parameters of the AS strategy, such as β and ε. We then examined the relationship between the parameters and the SC, which is important for tuning, and showed the lower bound of the convergence speed of the empirical sequences tending to the best empirical sequence.

References

[1] C. E. Shannon, “A mathematical theory of communication,” *Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, 1948.

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

[3] K. Iwata, K. Ikeda, and H. Sakai, “Asymptotic equipartition property on empirical sequence in reinforcement learning,” in *Proceedings of the 2nd IASTED International Conference on Neural Networks and Computational Intelligence*, IASTED. Grindelwald, Switzerland: ACTA Press, Feb. 2004, in press.

[4] R. S. Sutton and A. G. Barto, *Reinforcement Learning : An Introduction*, ser. Adaptive Computation and Machine Learning. Cambridge, MA: MIT Press, Mar. 1998.

[5] C. J. C. H. Watkins and P. Dayan, “Technical note : Q-learning,” *Machine Learning*, vol. 8, pp. 279–292, 1992.

[6] S. Amari and T. S. Han, “Statistical inference under multiterminal rate restrictions: a differential geometric approach,” *IEEE Transactions on Information Theory*, vol. 35, no. 2, pp. 217–227, Mar. 1989.

[7] G. J. Chaitin, *Algorithmic information theory*, ser. Cambridge tracts in theoretical computer science. Cambridge, UK: Cambridge University Press, 1987, vol. 1, reprinted with revisions in 1988.

[8] K. Iwata, K. Ikeda, and H. Sakai, “A new criterion using information gain for action selection strategy in reinforcement learning,” *IEEE Transaction on Neural Networks*, May 2004, in press.

[9] H. J. Kushner and G. G. Yin, *Stochastic Approximation Algorithms and Applications*, ser. Applications of Mathematics. New York: Springer-Verlag, 1997, vol. 35.