An Information-Theoretic Analysis of Return Maximization in Reinforcement Learning

Kazunori Iwata*

Graduate School of Information Sciences, Hiroshima City University, Hiroshima 731-3194, Japan

Abstract

We present a general analysis of return maximization in reinforcement learning. This analysis does not require assumptions of Markovianity, stationarity, and ergodicity for the stochastic sequential decision processes of reinforcement learning. Instead, our analysis assumes the asymptotic equipartition property fundamental to information theory, providing a substantially different view from that in the literature. As our main results, we show that return maximization is achieved by the overlap of typical and best sequence sets, and we present a class of stochastic sequential decision processes with the necessary condition for return maximization. We also describe several examples of best sequences in terms of return maximization in the class of stochastic sequential decision processes, which satisfy the necessary condition.

Keywords: reinforcement learning, stochastic sequential decision process, information theory, asymptotic equipartition property

*Corresponding author

1. Introduction

Reinforcement learning (RL) (Kaelbling et al., 1996; Mitchell, 1997; Sutton & Barto, 1998) refers to the framework of interactions between an agent and its environment. Stochastic sequential decision processes (SDPs) in RL consist of action-selection and state-transition probabilities determined by an agent's policy and the environment, respectively. During an SDP, the agent's policy for action-selection is improved by a learning algorithm, such as temporal difference learning (Watkins & Dayan, 1992; Dayan, 1992; Tsitsiklis, 1994). One of the outstanding features of RL is the definition of an optimal policy via return maximization (RM), and not through correct action-selections indicated by a supervisor. Basically, RM means that the state-action-reward sequences generated during an SDP tend to provide maximum returns as time progresses. Because this feature obviates the need to indicate correct action-selections individually for all state-actions, RL has been applied extensively to represent sequential decision-making processes arising in various applications (Tesauro, 1994; Likas, 1999; Doya, 2000; Morimoto & Doya, 2005; Pandana & Liu, 2005; Abbeel et al., 2007; Fujita & Ishii, 2007).

SDPs in RL are usually formulated as Markov decision processes (MDPs), in which the subsequent state is dependent only on the current state and the action taken by the agent. We often assume stationarity, ergodicity, or both of an MDP for its analysis. In brief, stationarity means that the probability measure that determines a stochastic process is not time-dependent, while ergodicity implies that the transition between any values drawn by a stochastic process occurs non-periodically with a positive probability. These assumptions ensure that the expectation of value function estimates are taken, which is required not only for simplicity of analysis of stochastic approximation, but also in the definition of a

stationary optimal policy. Nevertheless, most SDPs appearing in RL applications are, in fact, not necessarily Markovian, stationary, or ergodic. For example, if an environment varies with time, then the SDP is not a stationary process. This example leads us to question what conditions a non-Markovian, non-stationary, and/or non-ergodic SDP must satisfy to make RM possible. Answering this type of question requires a more general view of RL, and hence we take a substantially different view from the one found in the literature.

In this paper, we formulate RL as a much more general SDP than merely an MDP. In relation to this, we introduce several information-theoretic quantities to deal with general SDPs by building a novel bridge between RL and information theory. The aim of this paper is to shed further light on RM in RL, without the Markovianity, stationarity, or ergodicity assumptions for SDPs. Accordingly, we give a general definition of RM and present a class of SDPs with the necessary condition for RM. The definition of RM and the class of SDPs realize the first step in advancing RL theory in general SDPs. We also give several examples of best sequences in terms of RM within the class of SDPs.

The organization of this paper is as follows. Having reviewed MDPs in RL in Section 2, we consider more general SDPs in Section 3. The main results are discussed in Section 4. Finally, we give a summary of this paper in Section 5.

2. Markov Decision Processes

We concentrate on discrete-time SDPs with discrete states, actions, and rewards. We consider an episodic task throughout this paper. Let $\mathbb{N}$ denote the positive integers. Let $\mathbb{R}$ and $\mathbb{R}^+$ denote the real numbers and the positive real numbers, respectively. The sets of states, actions, and rewards of SDPs are de-

noted, respectively, as

$$\mathcal{S} \stackrel{\text{def.}}{=} \{ s_1, s_2, \ldots, s_I \}, \tag{1}$$

$$\mathcal{A} \stackrel{\text{def.}}{=} \{ a_1, a_2, \ldots, a_J \}, \tag{2}$$

$$\mathcal{R} \stackrel{\text{def.}}{=} \{ r_1, r_2, \ldots, r_K \}, \tag{3}$$

where $r_k \in \mathbb{R}$ and $|r_k| < \infty$ for all k. Throughout this paper, we assume that $\mathcal{S}$, $\mathcal{A}$, and $\mathcal{R}$ are non-empty finite sets with I, J, and K elements, respectively. Let $s(t)$, $a(t)$, and $r(t)$ be a state, action, and reward, respectively, at time $t \in \mathbb{N}$. At each time t, the agent senses the current state $s(t)$, and then according to its policy, selects an action $a(t)$, which is executed. According to the state-transition probabilities of the environment, the action changes the state into a subsequent state $s(t+1)$ and yields a scalar reward $r(t+1)$ as well. During each episode, these steps are repeated by incrementing time t by one. Training consists of a series of episodes. MDPs (Kaelbling et al., 1996; Mitchell, 1997; Sutton & Barto, 1998) are the most popular SDPs in RL for describing a framework of the interactions between an agent and its environment. For convenience of implementation, we make use of a six-tuple in defining an MDP, but we could equally well have used one of the other MDP definitions (Howard, 1960; Hopp et al., 1987; Sutton & Barto, 1998), which essentially mean the same. The six-tuple of an MDP is expressed as

$$\left(\mathcal{S}, \mathcal{A}, \mathcal{R}, \left\{ p_{ij}^{(t)} \right\}, \left\{ p_{iji'k}^{(t)} \right\}, \left\{ p_i^{(1)} \right\}\right),$$

where for all i, j, i', k,

$$p_{ij}^{(t)} \stackrel{\text{def.}}{=} \Pr(a(t) = a_j \mid s(t) = s_i), \tag{4}$$

$$p_{iji'k}^{(t)} \stackrel{\text{def.}}{=} \Pr(s(t+1) = s_{i'}, r(t+1) = r_k \mid s(t) = s_i, a(t) = a_j), \tag{5}$$

$$p_i^{(1)} \stackrel{\text{def.}}{=} \Pr(s(1) = s_i). \tag{6}$$

Equations (4) and (5) are called the action-selection probability and state-transition probability, respectively. The sequence of state, action, and reward is drawn from the respective Markov transitions. It is common in RL that the state-transition probabilities are unknown to the agent. This assumption creates a different problem-setting to that in dynamic programming (Bellman, 1957). If the state-transition probabilities of an MDP are not time-dependent, then its state-transition is said to be stationary. Furthermore, the policy of an agent whose action-selection probabilities are not time-dependent is said to be stationary. An MDP with a stationary state-transition and policy is called a stationary MDP. If the Markov chain of state, which is determined by a state-transition and a policy, is irreducible and non-periodic, then we say the MDP is ergodic on $\mathcal{S}$. Similarly, if the Markov chain of state-action is irreducible and non-periodic, then the MDP is said to be ergodic on $\mathcal{S} \times \mathcal{A}$. For simplicity, when an MDP is ergodic on $\mathcal{S}$ or $\mathcal{S} \times \mathcal{A}$, it is simply said to be ergodic, as long as we do not need to specify the ergodicity condition.

The method for describing an agent's policy is called an action-selection strategy. A number of action-selection strategies have been proposed (Kaelbling, 1993; Dearden et al., 1998; Ishii et al., 2002; Iwata et al., 2004; Neumann & Peters, 2009). Here, we introduce one of the most popular methods, the softmax method (Kaelbling et al., 1996; Mitchell, 1997; Sutton & Barto, 1998). This

method depends on the estimates of action-value functions (Sutton & Barto, 1998), each of which indicates the (discounted) sum of rewards to be received in the future. Let Q_{ij} denote the action-value function estimate of state-action $(s_i, a_j) \in \mathcal{S} \times \mathcal{A}$. In this method, Equation (4) can be written as

$$p_{ij}^{(t)} = \frac{\exp\left(\beta Q_{ij}\right)}{\sum_{j' \in \mathcal{J}_i} \exp(\beta Q_{ij'})}, \tag{7}$$

where $\mathcal{J}_i$ denotes the set of action indices available in state s_i, and β is a parameter that determines how strongly the policy prefers actions with high estimates. Note that Equation (7) is not time-dependent only when all the estimates of Q_{ij} and parameter β are not time-dependent.

In fact, Equation (7) is time-dependent in most MDPs because an agent improves its policy by updating the estimates, and the parameter needs to be asymptotically increased to promote the acceptance of actions that may yield a higher return than others. Moreover, the state-transition of the environment is not stationary in some practical cases. However, we sometimes require the MDP to be stationary and/or ergodic for simplicity of analysis. At the least we frequently assume that the state-transition is stationary, the MDP is ergodic, and at each time, the future MDP is stochastically the same as the current one. These assumptions provide the assurance that the estimates will converge to their respective expectations by a stochastic approximation method and also that there exists a stationary optimal policy. Indeed, mathematical analyses of RL and learning algorithms for updating the estimates make use of several or all of these assumptions (Blackwell, 1962; Watkins & Dayan, 1992; Dayan, 1992; Tsitsiklis, 1994; Kushner & Yin, 1997; Borkar & Meyn, 2000; Singh et al., 2000; Gosavi, 2006; Iwata et al., 2006a,b).

3. Stochastic Sequential Decision Processes

We have explained that SDPs are usually formulated as MDPs with action-selection and state-transition probabilities, and also that these are supposed to be stationary, at least with respect to state-transition, ergodic, or both in the literature. Now, we dispense with these assumptions to give an example of more general SDPs. We consider that the estimates and parameters referred to in the softmax method are time-dependent and obey their respective stochastic processes. For all i, j, let $Q_{ij}(t)$ denote the action-value function estimate of state-action (s_i, a_j) at time t. The set of action-value function estimates on $\mathbb{R}^{IJ}$ is denoted by

$$Q(t) \stackrel{\text{def.}}{=} \left\{ Q_{ij}(t) \,|\, 1 \leq i \leq I, 1 \leq j \leq J \right\}. \tag{8}$$

Let $\beta(t)$ denote parameter β on $\mathbb{R}$ at time t. In this case, an action at each time t is selected according to the action-selection probability expressed as

$$p_{ij}^{(t)}(\beta, Q) \stackrel{\text{def.}}{=} \Pr(a(t) = a_j \,|\, \beta(t) = \beta, Q(t) = Q, s(t) = s_i). \tag{9}$$

In the same manner, we consider that the state-transition probability is parameterized with a vector of time-dependent parameters. For example, consider

$$p_{iji'k}^{(t)}(\zeta) \stackrel{\text{def.}}{=} \Pr(s(t+1) = s_{i'}, r(t+1) = r_k \,|\, \zeta(t) = \zeta, s(t) = s_i, a(t) = a_j), \tag{10}$$

where $\zeta(t)$ denotes a parameter vector on some space at time t. We can see that $p_{ij}^{(t)}$ and $p_{iji'k}^{(t)}$ are drawn according to probability measures that define $\beta(t)$, $Q(t)$, and $\zeta(t)$. Accordingly, in general, an SDP defined by $p_{ij}^{(t)}$ and $p_{iji'k}^{(t)}$ is no longer stationary, ergodic, or Markovian with respect to its action-selection and state-transition. This argument holds, even if we use another action-selection strategy.

Having noticed that the conventional form of MDPs is inadequate for dealing with the SDP defined by Equations (9) and (10), we formulate RL as a more

general SDP. For all $t \in \mathbb{N}$, we use $\theta(t)$ to denote all the factors that determine the action-selection and state-transition probabilities of the SDP. Θ denotes the sample space of all possible outcomes of $\theta(t)$. For all $n \in \mathbb{N}$, an outcome of $(\theta(1), \ldots, \theta(n))$ on Θ^n is simply denoted as θ. For example, if the action-selection and state-transition probabilities of an SDP are given by Equations (9) and (10), respectively, then $\theta(t)$ is expressed as

$$\theta(t) = (Q(t), \beta(t), \zeta(t)) \,. \tag{11}$$

Furthermore, if the SDP is a stationary MDP, then there exists a vector $\bar{\theta} \in \Theta$ such that $\theta(t) = \bar{\theta}$ for all $t \in \mathbb{N}$. For all $n \in \mathbb{N}$, let

$$\mathcal{X}^n \stackrel{\text{def.}}{=} \mathcal{S} \times (\mathcal{A} \times \mathcal{S} \times \mathcal{R})^n \,. \tag{12}$$

We denote the three-tuple of state, action, and reward at time t by

$$x(t) \stackrel{\text{def.}}{=} \begin{cases} s(1) & \text{for } t = 0, \\ (a(t), s(t+1), r(t+1)) & \text{for all } t \in \mathbb{N}. \end{cases} \tag{13}$$

Using the notation, the sequence of state, action, and reward,

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

is denoted in short by

$$x(0), x(1), \ldots, x(n).$$

We are now in a position to define the SDP.

Definition 1 (Probability Measure). Given $\theta \in \Theta^n$, let P^n_θ be a probability measure on $\mathcal{X}^n$ expressed as

$$P^n_\theta(x) \stackrel{\text{def.}}{=} \Pr\left((x(0), x(1), \ldots, x(n)) = x \mid (\theta(1), \ldots, \theta(n)) = \theta\right), \tag{14}$$

for all $x \in \mathcal{X}^n$. We denote the stochastic variables drawn according to P^n_θ by X^n_θ. For all $x \in \mathcal{X}^n$, the mixed measure of P^n_θ is expressed as

$$P^n(x) \stackrel{\text{def.}}{=} \int_{\Theta^n} P^n_\theta(x) dy^n(\theta), \tag{15}$$

where dy^n is the probability measure on Θ^n. We denote the stochastic variables drawn according to P^n by X^n.

SDP X^n is a stochastic process defined as $(\mathcal{X}^n, P^n)$. Note that an MDP six-tuple can be rewritten as a two-tuple in this manner. In an episodic task discussed in this paper, the sequence of X^n is observed repeatedly.

The analysis of stochastic processes through their mixed process is well established in information theory and is known as information-spectrum analysis (Han, 2003). In this paper, we examine how the SDPs $\left\{X^n_\theta \,|\, \theta \in \Theta^n\right\}$ should be drawn for RM, by analyzing their mixed SDP X^n. We start by demonstrating in Example 1 that entropy is a useful quantity for verifying (asymptotically mean) stationary ergodic processes.

Example 1 (Stationary Ergodic SDPs). If P^n is stationary and ergodic, then it has the following two properties:

1. there exists a fixed constant $c \in \mathbb{R}^+$ such that

$$\lim_{n\to\infty} \frac{1}{n} H(P^n) = c, \tag{16}$$

where $H(P^n)$ denotes the entropy of P^n,

$$H(P^n) \stackrel{\text{def.}}{=} \sum_{x \in \mathcal{X}^n} P^n(x) \log \frac{1}{P^n(x)}, \tag{17}$$

and hence c is called the entropy rate of P^n.

2. the sequence of $(1/n)\log(1/P^n(X^n))$ converges to the entropy rate, that is,

$$\lim_{n\to\infty}\frac{1}{n}\log\frac{1}{P^n(X^n)} = \lim_{n\to\infty}\frac{1}{n}H(P^n) \qquad \text{a.e.,} \tag{18}$$

where a.e. means that the equation is true almost everywhere (Gray, 2010).

The proof is given in (Barron, 1985; Gray, 2010). Similar properties also hold when P^n is asymptotically mean stationary and ergodic (Gray & Kieffer, 1980; Gray, 2010).

Throughout this paper, all logarithms are expressed by the same arbitrary base. Note that an SDP, whose P^n has the two properties, is more general than a stationary ergodic process.

Example 2 (Non-Stationary and/or Non-Ergodic Processes). If P^n is neither stationary nor ergodic, then the sequence of $(1/n)\log(1/P^n(X^n))$ does not always converge to a constant fixed by P^n as $n \to \infty$. An example of this is found in (Han, 2003, Chapter 1).

We have explained in Examples 1 and 2 that the entropy is available for the analysis of stationary ergodic processes, but not always for more general processes than these. Accordingly, to analyze more general processes, we introduce a further information-theoretic quantity of SDPs, known as the spectral entropy rate (Han & Verdú, 1993; Han, 2003). The spectral superior entropy rate of P^n is defined as

$$\overline{H}(P^\infty) \stackrel{\text{def.}}{=} \inf\left\{b \in \mathbb{R} \;\middle|\; \lim_{n\to\infty}\Pr\left(\frac{1}{n}\log\frac{1}{P^n(X^n)} > b\right) = 0\right\}, \tag{19}$$

while the spectral inferior entropy rate of P^n is defined as

$$\underline{H}(P^\infty) \stackrel{\text{def.}}{=} \sup\left\{b \in \mathbb{R} \;\middle|\; \lim_{n\to\infty}\Pr\left(\frac{1}{n}\log\frac{1}{P^n(X^n)} < b\right) = 0\right\}. \tag{20}$$

The spectral superior and inferior entropy rates of P^n_θ are defined similarly.

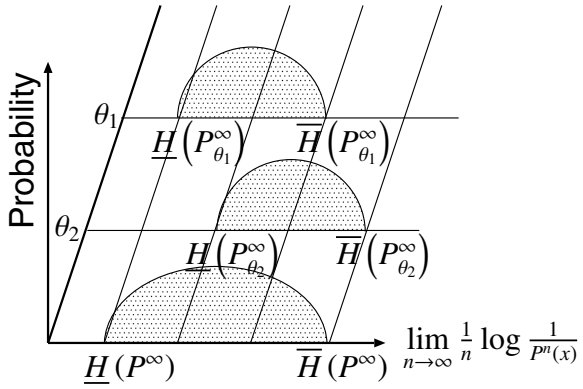

Figure 1: Entropy spectrum.

Example 3 (Entropy Spectrum). In SDPs, the quantity

$$\lim_{n\to\infty} \frac{1}{n} \log \frac{1}{P^n(X^n)}, \tag{21}$$

has a spectrum that ranges between $\overline{H}\left(P^\infty\right)$ and $\underline{H}\left(P^\infty\right)$. For example, consider

$$P^n(x) = P^n_{\theta_1}(x) dy^n\left(\theta_1\right) + P^n_{\theta_2}(x) dy^n\left(\theta_2\right), \tag{22}$$

where $dy^n\left(\theta_1\right) + dy^n\left(\theta_2\right) = 1$, for all $\theta_1, \theta_2 \in \Theta^n$. The probability density function of the quantity, illustrated in Figure 1, is referred to as the entropy spectrum (Han & Verdú, 1993; Han, 2003). In Figure 1, the horizontal and vertical axes show $\lim_{n\to\infty}(1/n)\log(1/P^n(X^n))$ and probability, respectively, and the oblique axis takes θ_1 and θ_2 only in this case. The figure illustrates that the entropy spectrum of P^n is the mixture of the entropy spectra of $P^n_{\theta_1}$ and $P^n_{\theta_2}$.

4. Main Results

The asymptotic equipartition property (AEP) (Cover & Thomas, 2006) is fundamental to information theory. In short, the AEP implies that there exists a typical set of sequences with probability nearly one, which are almost equi-probable.

Since the AEP was first introduced by Shannon (Shannon, 1948) in a stationary ergodic process, there have been many studies on it in other stochastic processes (McMillan, 1953; Breiman, 1957, 1960; Moy, 1961; Kieffer, 1974; Gray & Kieffer, 1980; Barron, 1985; Verdú & Han, 1997). In particular, the concept of the AEP was ultimately generalized to arbitrary general processes in (Verdú & Han, 1997; Han, 2003). In the RL context, the AEP of a stationary ergodic MDP was first shown in (Iwata et al., 2006a,b) using the method of types (Csiszár & Körner, 1997; Csiszár, 1998). In this section, we introduce the AEP into RL, and explain the AEP-based analysis of RM.

Henceforth, whenever we refer to the AEP, we mean the generalized AEP defined as follows (Verdú & Han, 1997; Han, 2003).

Definition 2 (AEP). For all $n \in \mathbb{N}$, let M^n denote a probability measure on a sample space $\mathcal{Y}^n$. M^n is said to possess the AEP if it satisfies the following. For all $\delta \in \mathbb{R}^+$, as $n \to \infty$,

$$\begin{aligned} M^n\left(B_\delta\left(M^n\right)\right) &\to 0, \\ M^n\left(S_\delta\left(M^n\right)\right) &\to 0, \end{aligned} \tag{23}$$

where $B_\delta\left(M^n\right)$ denotes the set of sequences with atypically large probabilities,

$$B_\delta\left(M^n\right) \stackrel{\text{def.}}{=} \left\{ y \in \mathcal{Y}^n \,|\, M^n(y) \geq \exp\left(-(1-\delta)H\left(M^n\right)\right) \right\}, \tag{24}$$

and $S_\delta\left(M^n\right)$ denotes the set of sequences with atypically small probabilities,

$$S_\delta\left(M^n\right) \stackrel{\text{def.}}{=} \left\{ y \in \mathcal{Y}^n \,|\, M^n(y) \leq \exp\left(-(1+\delta)H\left(M^n\right)\right) \right\}. \tag{25}$$

The AEP implies that under some constraint on M^n, there exists a set of sequences in $\mathcal{Y}^n$ with probability of almost one. Definition 3 presents a new optimality criterion for SDPs, which is distinct from the conventional criteria for policies. This enables us to relate SDPs in RL to the AEP.

Definition 3 (Return Maximization). For all $n \in \mathbb{N}$, let $\mathcal{X}_\epsilon^{*n}$ be the set of best sequences in $\mathcal{X}^n$. RM means that for all $\epsilon \in \mathbb{R}^+$,

$$P^n(\mathcal{X}_\epsilon^{*n}) \to 1, \tag{26}$$

as $n \to \infty$.

In fact, the theorems in this section hold for any non-empty set of best sequences. In this sense, arbitrary sequences can be defined as the best sequences, but those in an RL context should be sequences that tend to yield a maximum return as time progresses. As an example, we specify the best sequences for a stationary ergodic MDP using the action-value functions.

Example 4 (Best Sequences for Stationary Ergodic MDPs). Assume that X^n is a stationary MDP, which is ergodic on $\mathcal{S}$ and $\mathcal{S} \times \mathcal{A}$. For all $t \in \mathbb{N}$, the stochastic variables of state, action, and reward at time t, drawn by X^n are denoted by $s_{X^n}(t)$, $a_{X^n}(t)$, and $r_{X^n}(t)$, respectively. Given $X^n = x \in \mathcal{X}^n$, the types thereof for $s_i \in \mathcal{S}$, $(s_i, a_j) \in \mathcal{S} \times \mathcal{A}$, and $(s_i, a_j, s_{i'}, r_k) \in \mathcal{X}$ are, respectively,

$$f_i(x) \overset{\text{def.}}{=} \frac{1}{n} \sharp\{\, t \in \{1, \ldots, n\} \mid s_{X^n}(t) = s_i \,\}, \tag{27}$$

$$f_{ij}(x) \overset{\text{def.}}{=} \frac{1}{n} \sharp\Big\{\, t \in \{1, \ldots, n\} \;\Big|\; (s_{X^n}(t), a_{X^n}(t)) = (s_i, a_j) \,\Big\}, \tag{28}$$

$$f_{iji'k}(x) \overset{\text{def.}}{=} \frac{1}{n} \sharp\Big\{\, t \in \{1, \ldots, n\} \;\Big|\; (s_{X^n}(t), a_{X^n}(t), s_{X^n}(t+1), r_{X^n}(t+1)) = (s_i, a_j, s_{i'}, r_k) \,\Big\}, \tag{29}$$

where $\sharp\{\cdot\}$ denotes the number of elements in a finite set. For all i, j, i', k, we define their conditional types $\hat{p}_{ij}(x)$ and $\hat{p}_{iji'k}(x)$ by

$$f_{ij}(x) = f_i(x)\hat{p}_{ij}(x), \tag{30}$$

$$f_{iji'k}(x) = f_{ij}(x)\hat{p}_{iji'k}(x). \tag{31}$$

Since in this case, the $p_{ij}^{(t)}$ and $p_{iji'k}^{(t)}$ in Equations (4) and (5) are invariant with respect to t, we denote these as p_{ij} and $p_{iji'k}$, respectively, dropping the t. Considering the AEP of the stationary ergodic MDPs (Iwata et al., 2006a,b), the best sequences based on the action-value functions are described as

$$\mathcal{X}_{\epsilon}^{*n} \stackrel{\text{def.}}{=} \{\, x \in \mathcal{X}^n \mid D\left(\hat{p}(x) \,\|\, p^* \mid f(x)\right) \leq \epsilon \,\}, \tag{32}$$

for all $\epsilon \in \mathbb{R}^+$, where

$$\begin{aligned} D\left(\hat{p}(x) \,\|\, p^* \mid f(x)\right) \stackrel{\text{def.}}{=} & \sum_{i=1}^{I} f_i(x) \sum_{j=1}^{J} \hat{p}_{ij}(x) \log \frac{\hat{p}_{ij}(x)}{p_{ij}^*} \\ & + \sum_{i=1}^{I} \sum_{j=1}^{J} f_{ij}(x) \sum_{i'=1}^{I} \sum_{k=1}^{K} \hat{p}_{iji'k}(x) \log \frac{\hat{p}_{iji'k}(x)}{p_{iji'k}}, \end{aligned} \tag{33}$$

and

$$p_{ij}^* \stackrel{\text{def.}}{=} \begin{cases} 1 & \text{if } j = \operatorname{argmax}_{j' \in \mathcal{J}_i} Q_{ij'}^*, \\ 0 & \text{if } j \neq \operatorname{argmax}_{j' \in \mathcal{J}_i} Q_{ij'}^*, \end{cases} \tag{34}$$

where Q_{ij}^* denotes the action-value function of (s_i, a_j). The first and second terms in Equation (33) denote the conditional information divergences of the action-selection and state-transition probabilities, respectively. These probabilities rely on the stationary distributions on state and state-action, which are derived from the ergodicity. Note that any action-value function becomes a fixed real number under the stationary ergodic MDPs.

In Example 4, we have explained how to define best sequences using the action-value functions of Q_{ij}^*. We see that realizing RM with the best sequences determined by the action-value functions, as proposed in this paper, essentially corresponds to obtaining the optimal policy in existing RL studies under the stationary ergodic MDPs. The key to this example is that the best sequences can be

defined by p^*_{ij}, which represents a stationary optimal policy. However, for all i, j, Q^*_{ij} does not always become a real number fixed by P^n under more general SDPs than stationary ergodic MDPs. This implies that in general, an optimal policy is not stationary in such cases. Accordingly, the definition of best sequences in Example 4 is no longer applicable to such general SDPs.

Now, we consider the relationship between best sequences and the optimal policy in more general SDPs. For all $(x, \theta) \in \mathcal{X}^n \times \Theta^n$, let

$$\widetilde{P}^n(x, \theta) \stackrel{\text{def.}}{=} P^n_\theta(x) dy^n(\theta) \tag{35}$$

denote the joint probability measure of action-selection, state-transition, and their factors. Let $\widetilde{P^*}^n$ be the joint probability measure given by an optimal policy. Since the agent is able to control only action-selection probabilities (via some of the factors), for given state-transition probabilities, the area of $\widetilde{P^*}^n$ obtained by changing the action-selection probabilities is actually restricted over the domain of $\widetilde{P}^n$,

$$\left\{ \widetilde{P}^n \,|\, \widetilde{P}^n \text{ is a probability measure} \right\}.$$

For all $x \in \mathcal{X}^n$, we let

$$P^{*n}(x) \stackrel{\text{def.}}{=} \int_{\Theta^n} \widetilde{P^*}^n(x, \theta). \tag{36}$$

The possible P^{*n} is also restricted over its domain. We use $\mathcal{P}$ to denote the restricted domain of P^n. There are several ways of defining an optimal policy in more general SDPs. In Example 5, we show best sequences provided by an optimal policy in terms of average reward over time.

Example 5 (Best Sequences for More General SDPs). For all $n \in \mathbb{N}$, let

$$\overline{A}(P^n) \stackrel{\text{def.}}{=} \inf \left\{ b \in \mathbb{R} \;\middle|\; \Pr\left(\frac{1}{n} \sum_{t=1}^{n} r_{X^n}(t) > b \right) = 0 \right\}, \tag{37}$$

$$\underline{A}(P^n) \stackrel{\text{def.}}{=} \sup \left\{ b \in \mathbb{R} \;\middle|\; \Pr\left(\frac{1}{n} \sum_{t=1}^{n} r_{X^n}(t) < b \right) = 0 \right\}. \tag{38}$$

Since

$$-\infty < \min_{k:\ 1 \leq k \leq K} r_k \leq \frac{1}{n} \sum_{t=1}^{n} r_{X^n}(t) \leq \max_{k:\ 1 \leq k \leq K} r_k < \infty, \tag{39}$$

$\overline{A}(P^n)$ and $\underline{A}(P^n)$ are finite for all $n \in \mathbb{N}$. For all $n \in \mathbb{N}$, let

$$P^{*n} \stackrel{\text{def.}}{=} \underset{P^n \in \mathcal{P}}{\operatorname{argmax}} \int_{\underline{A}(P^n)}^{\overline{A}(P^n)} b \lambda_n(b) db, \tag{40}$$

where

$$\lambda_n(b) \stackrel{\text{def.}}{=} \Pr\left(\frac{1}{n} \sum_{t=1}^{n} r_{X^n}(t) = b \right). \tag{41}$$

For all $n \in \mathbb{N}$ and all $\epsilon \in \mathbb{R}^+$, the best sequences of n time steps in more general SDPs are described as

$$\mathcal{X}_\epsilon^{*n} \stackrel{\text{def.}}{=} \left\{ x \in \mathcal{X}^n \,|\, x \notin B_\epsilon\left(P^{*n}\right), x \notin S_\epsilon\left(P^{*n}\right) \right\}. \tag{42}$$

In Corollary 1, we explain why the set of best sequences should be defined by Equation (42). One of the main results is given by Theorem 1, which implies that RM is based on the AEP.

Theorem 1 (Role of the AEP in RM). *If*

1. *P^n has the AEP, and*
2. *for all δ, $\epsilon \in \mathbb{R}^+$,*

$$\lim_{n \to \infty} \left\{ P^n\left(\mathcal{X}_\epsilon^{*n} \cup C_\delta\left(P^n\right)\right) - P^n\left(\mathcal{X}_\epsilon^{*n} \cap C_\delta\left(P^n\right)\right) \right\} = 0, \tag{43}$$

where

$$C_\delta(P^n) \stackrel{def.}{=} \{\, x \in \mathcal{X}^n \mid x \notin B_\delta(P^n), x \notin S_\delta(P^n) \,\}, \tag{44}$$

then RM holds.

Proof. Since P^n has the AEP, for all $\delta \in \mathbb{R}^+$, Equation (23) gives

$$\lim_{n\to\infty} P^n(C_\delta(P^n)) = 1. \tag{45}$$

Since

$$P^n(C_\delta(P^n)) \leq P^n\left(\mathcal{X}_\epsilon^{*n} \cup C_\delta(P^n)\right), \tag{46}$$

$$P^n\left(\mathcal{X}_\epsilon^{*n} \cap C_\delta(P^n)\right) \leq P^n\left(\mathcal{X}_\epsilon^{*n}\right) \leq P^n\left(\mathcal{X}_\epsilon^{*n} \cup C_\delta(P^n)\right), \tag{47}$$

Equations (45) and (43) give

$$\lim_{n\to\infty} P^n\left(\mathcal{X}_\epsilon^{*n} \cup C_\delta(P^n)\right) = 1, \tag{48}$$

and

$$\lim_{n\to\infty} \{P^n\left(\mathcal{X}_\epsilon^{*n} \cup C_\delta(P^n)\right) - P^n\left(\mathcal{X}_\epsilon^{*n}\right)\} = 0, \tag{49}$$

respectively. Therefore, we reach the conclusion in Equation (26).

The set of sequences given by Equation (44) is referred to in information theory as the typical set. If the first condition of Theorem 1 holds, then the second condition is equivalent to

$$\lim_{n\to\infty} C_\delta(P^n) = \lim_{n\to\infty} \mathcal{X}_\epsilon^{*n}, \tag{50}$$

because Equation (43) can be rewritten as

$$\lim_{n\to\infty} P^n\left(\overline{\mathcal{X}_\epsilon^{*n}} \cap C_\delta(P^n)\right) = 0, \tag{51}$$

$$\lim_{n\to\infty} P^n\left(\mathcal{X}_\epsilon^{*n} \cap \overline{C_\delta(P^n)}\right) = 0, \tag{52}$$

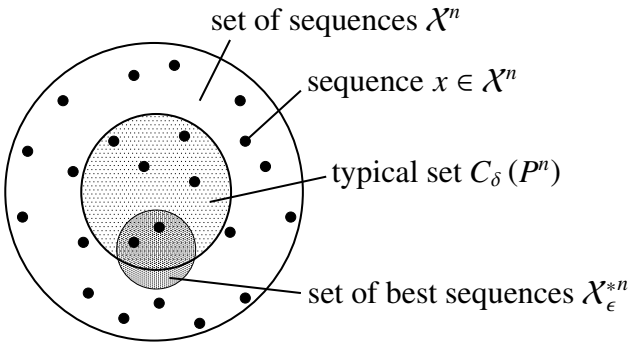

Figure 2: Best sequences and the typical set.

where $\bar{\cdot}$ denotes the complement of a set, and these equations yield Equation (50) immediately. Figure 2 shows the set of best sequences and the typical set. The dots, light shaded circle, and dark shaded circle represent the sequences, typical set, and set of best sequences in $\mathcal{X}^n$, respectively. Theorem 1 suggests that if $\mathcal{X}_\epsilon^{*n}$ and $C_\delta(P^n)$ intersect by policy improvement, as illustrated in Figure 2, then $P^n(\mathcal{X}_\epsilon^{*n})$ tends toward a positive probability. Intuitively, this means that it is possible for the agent to achieve RM. Thus, RM based on the AEP does not rely on Markovianity, stationarity, or ergodicity to be satisfied.

Corollary 1 gives an example of RM with the set of best sequences.

Corollary 1. *If the set of best sequences is given by*

$$\mathcal{X}_\epsilon^{*n} \stackrel{def.}{=} C_\epsilon(P^{*n}), \tag{53}$$

and P^n satisfies the following properties:

1. *P^n has the AEP, and*
2. *there exist $\kappa \in \mathbb{R}^+$, $\sigma \in \mathbb{R}^+$, and $n_0 \in \mathbb{N}$ such that for all $n_0 \leq n$,*

$$\max_{x \in \mathcal{X}^n} \left| P^n(x) - P^{*n}(x) \right| \leq \kappa \exp(-\sigma n), \tag{54}$$

where

$$\sigma > (1+\epsilon)\left(\log \sharp\{\mathcal{S}\} + \log \sharp\{\mathcal{A}\} + \log \sharp\{\mathcal{R}\}\right), \tag{55}$$

then RM holds.

Proof. The probability of the typical set of P^n is bounded by

$$\sharp\{C_\epsilon(P^n)\} \exp\left(-(1+\epsilon)H(P^n)\right) < P^n\left(C_\epsilon(P^n)\right) \leq 1, \tag{56}$$

where $\sharp\{C_\epsilon(P^n)\}$ denotes the number of elements in the typical set. Meanwhile, from (Cover & Thomas, 2006, Theorem 2.6.4), the entropy of P^n is bounded by

$$H(P^n) \leq \log \sharp\{\mathcal{S}\} + nl, \tag{57}$$

where l is a constant given by

$$l \stackrel{\text{def.}}{=} \log \sharp\{\mathcal{S}\} + \log \sharp\{\mathcal{A}\} + \log \sharp\{\mathcal{R}\}. \tag{58}$$

For all $x \in \mathcal{X}^n$ and a sufficiently large n, we have

$$P^n\left(C_\epsilon(P^n)\right) \leq P^{*n}\left(C_\epsilon(P^n)\right) + \kappa \sharp\{C_\epsilon(P^n)\} \exp(-\sigma n), \tag{59}$$

$$\leq P^{*n}\left(C_\epsilon(P^n)\right) + \exp\left(n\left((1+\epsilon)l - \sigma\right) + (1+\epsilon)\log \sharp\{\mathcal{S}\} + \log \kappa\right), \tag{60}$$

where the first inequality is derived from the second property of P^n, and the second one is given by the bounds (56) and (57). Since

$$\lim_{n\to\infty} P^n\left(C_\epsilon(P^n)\right) = 1, \tag{61}$$

and

$$\lim_{n\to\infty} \exp\left(n\left((1+\epsilon)l - \sigma\right) + (1+\epsilon)\log \sharp\{\mathcal{S}\} + \log \kappa\right) = 0, \tag{62}$$

we obtain

$$\lim_{n\to\infty} P^{*n}\left(C_\epsilon\left(P^n\right)\right) = 1. \tag{63}$$

Since $P^{*n}\left(C_\epsilon\left(P^n\right)\right) \leq P^{*n}\left(\mathcal{X}_\epsilon^{*n}\right)$ holds, we have

$$\lim_{n\to\infty} P^{*n}\left(\mathcal{X}_\epsilon^{*n}\right) = 1. \tag{64}$$

These equations yield

$$\lim_{n\to\infty} C_\epsilon\left(P^n\right) = \lim_{n\to\infty} \mathcal{X}_\epsilon^{*n}, \tag{65}$$

and thus we reach the conclusion.

This suggests that if the conditions in Corollary 1 are satisfied by a learning algorithm, then RM holds in terms of the set of best sequences.

According to Theorem 1, the AEP is a necessary condition for RM. Hence, in the rest of this section, we describe it in more detail. We readily obtain Corollary 2 by following (Verdú & Han, 1997).

Corollary 2. *P^n satisfies*

$$\underline{H}\left(P^\infty\right) = \overline{H}\left(P^\infty\right), \tag{66}$$

if and only if P^n has the AEP.

Equation (66) is called the strong converse property (Han, 2003). It is well known in information theory that probability measures with the strong converse property are more general than stationary ergodic probability measures; that is, stationary ergodic probability measures have the property. Interestingly, this implies that there are more general SDPs than stationary ergodic MDPs, such that RM holds.

Recall that in an episodic task, the sequence of X^n is sampled repeatedly. As Corollary 2 is stated in terms of P^n, we develop it further in Theorem 2 so that it

is described in terms of P_θ^n for all $\theta \in \Theta^n$. This is convenient to obtain available knowledge on what form SDPs $\left\{X_\theta^n \,|\, \theta \in \Theta^n\right\}$ should take for RM, since the agent can observe sequences drawn according to P_θ^n by giving an outcome $\theta \in \Theta^n$, rather than sequences drawn directly according to P^n.

Theorem 2. *We identify the following three properties of P^n.*

1. *For all $\theta \in \Theta^\infty$ where $dy^\infty(\theta) > 0$,*

$$\underline{H}\left(P_\theta^\infty\right) = \overline{H}\left(P_\theta^\infty\right). \tag{67}$$

2. *For all $\theta_1, \theta_2 \in \Theta^\infty$ where $dy^\infty(\theta_1) > 0$ and $dy^\infty(\theta_2) > 0$,*

$$\begin{aligned} \overline{H}\left(P_{\theta_1}^\infty\right) &= \overline{H}\left(P_{\theta_2}^\infty\right), \\ \underline{H}\left(P_{\theta_1}^\infty\right) &= \underline{H}\left(P_{\theta_2}^\infty\right). \end{aligned} \tag{68}$$

3. *For all $\theta_1, \theta_2 \in \Theta^n$ where $dy^n(\theta_1) > 0$ and $dy^n(\theta_2) > 0$,*

$$\lim_{n\to\infty} \sum_{x\in\mathcal{X}^n} \left| P_{\theta_1}^n(x) - P_{\theta_2}^n(x) \right| = 0. \tag{69}$$

P^n has properties 1 and 2 if and only if it has the AEP. Furthermore, if P^n has properties 1 and 3, then it also has the AEP.

Proof. Assume first that P^n has properties 1 and 2. Then, we obtain

$$\overline{H}\left(P^\infty\right) = \sup\left\{ \overline{H}\left(P_\theta^\infty\right) \,|\, dy^\infty(\theta) > 0 \right\}, \tag{70}$$

$$\underline{H}\left(P^\infty\right) = \inf\left\{ \underline{H}\left(P_\theta^\infty\right) \,|\, dy^\infty(\theta) > 0 \right\}. \tag{71}$$

Hence, for all $\theta \in \Theta^\infty$ where $dy^\infty(\theta) > 0$,

$$\overline{H}\left(P^\infty\right) \geq \overline{H}\left(P_\theta^\infty\right) \geq \underline{H}\left(P_\theta^\infty\right) \geq \underline{H}\left(P^\infty\right). \tag{72}$$

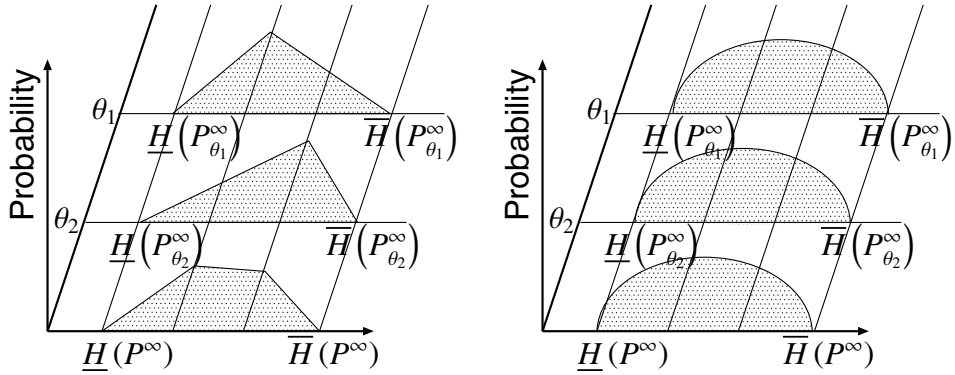

Figure 3: Difference in entropy spectrum. The left and right figures show entropy spectra for P^n with properties 2 and 3, respectively.

Equations (67) and (68) yield

$$\overline{H}\left(P^{\infty}\right) = \overline{H}\left(P^{\infty}_{\theta}\right) = \underline{H}\left(P^{\infty}_{\theta}\right) = \underline{H}\left(P^{\infty}\right), \tag{73}$$

and by Corollary 2, the AEP holds.

Conversely, assume that P^n has the AEP. From Corollary 2, we have Equation (66). From Equation (72), this can be rewritten as Equation (73), and properties 1 and 2 follow from Equation (73).

The second assertion follows immediately from (Han, 2003, Corollary 2.1.1).

Clearly, property 3 is a stronger condition on P^n than property 2. Figure 3 illustrates the difference in the entropy spectrum for properties 2 and 3. We see from Theorem 2 that P^n_{θ} is allowed to have a spectrum when $n < \infty$.

If an SDP is defined by P^n, which has a certain property, we say that the SDP has the property or refer to the SDP with the property. Theorem 1 states that RL applications must be designed in such a way that their SDPs have properties 1 and 2, otherwise RM does not hold. Although the following proposition is almost

obvious, its proof provides a specific example of SDPs with properties 1 and 3.

Proposition 1. *SDPs with properties 1 and 3 are more general than stationary MDPs that are ergodic on $\mathcal{S}$ and $\mathcal{S} \times \mathcal{A}$.*

Proof. It is obvious that the stationary ergodic MDPs on $\mathcal{S}$ and $\mathcal{S} \times \mathcal{A}$ have property 1. Accordingly, we show that they have property 3. As described in Section 3, $\theta(t)$ is not time-dependent in the stationary ergodic MDPs on $\mathcal{S}$ and $\mathcal{S} \times \mathcal{A}$, and hence there exists a unique vector $\bar{\theta} \in \Theta$, such that $\theta(t) = \bar{\theta}$ for all $t \in \mathbb{N}$. In this case, for all $n \in \mathbb{N}$,

$$dy^n(\theta) = \begin{cases} 1 & \text{if } \theta = (\bar{\theta}, \ldots, \bar{\theta}) \in \Theta^n, \\ 0 & \text{otherwise.} \end{cases} \tag{74}$$

Accordingly, letting $\theta_1 = \theta_2 = (\bar{\theta}, \ldots, \bar{\theta}) \in \Theta^n$ in Equation (69) ensures that property 3 holds.

Figure 4 summarizes the relationship among the stationary ergodic MDPs on $\mathcal{S}$ and $\mathcal{S} \times \mathcal{A}$, SDPs with properties 1 and 3, and SDPs with properties 1 and 2 (i.e., having the AEP). Again, we should note that Theorem 1 states that an SDP arising in RL must be one of the SDPs with the AEP for RM.

5. Discussion and Summary

Every state-action-reward sequence is classified, not only as a typical or untypical sequence, but also based on whether it is a best sequence. We have defined RM using the overlap of the typical and best sequence sets. This is substantially different from the existing RL literature. This view of classifying sequences is

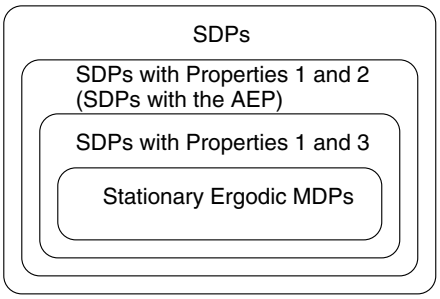

Figure 4: Classes of SDPs.

fairly normal in information theory, but not in RL. Thus, by building a bridge between RL and information theory, we have provided the first step for advancing RL theory in general SDPs.

In this paper, we elucidated an AEP-based analysis of RM in RL. Our analysis does not assume Markovianity, stationarity, or ergodicity of the SDPs. The AEP is an important property because it is a necessary condition for RM in general SDPs. Accordingly, it should be taken fully into account when considering applications of RL or their design thereof. We have also shown several examples of best sequences in terms of RM.

It is an important future task to find an algorithm that assures RM in an SDP with the AEP. Then, AEP-analysis might provide a reason why RL applications such as robot learning, sometimes work well even in more general SDPs than MDPs, as reported in the literature. Since the AEP-based view liberates RL applications from the Markovianity, stationarity, or ergodicity assumptions, adopting it for the description of application optimality should be fruitful. Finally, Definition 3 is a reasonable RM definition in the information-theoretic sense and implies that we can focus on those SDPs with the AEP. Accordingly, other definitions can be used for general SDPs, although analyses based thereon remain as open issues.

Acknowledgments

We wish to thank Prof. Hideaki Sakai and Prof. Kazushi Ikeda for their helpful comments on an earlier draft of this work. This work was supported partially by Grant-in-Aid 22700152 for scientific research from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

References

Abbeel, P., Coates, A., Quigley, M., & Ng, A. Y. (2007). An application of reinforcement learning to aerobatic helicopter flight. In B. Schölkopf, J. Platt, & T. Hoffman (Eds.), *Advances in Neural Information Processing Systems* (pp. 1–8). Cambridge, MA: MIT Press volume 19.

Barron, A. R. (1985). The strong ergodic theorem for densities: Generalized Shannon-Mcmillan-Breiman theorem. *Annals of Probability*, *13*, 1292–1303.

Bellman, R. E. (1957). *Dynamic Programming*. Princeton, NJ: Princeton University Press.

Blackwell, D. (1962). Discrete dynamic programming. *Annals of Mathematical Statistics*, *33*, 719–726.

Borkar, V. S., & Meyn, S. P. (2000). The O.D.E. method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization*, *38*, 447–469.

Breiman, L. (1957). The individual ergodic theorem of information theory. *Annals of Mathematical Statistics*, *28*, 809–811.

Breiman, L. (1960). Correction to the individual ergodic theorem of information theory. *Annals of Mathematical Statistics*, *31*, 809–810.

Cover, T. M., & Thomas, J. A. (2006). *Elements of Information Theory*. (2nd ed.). New York: John Wiley & Sons, Inc.

Csiszár, I. (1998). The method of types. *IEEE Transactions on Information Theory*, *44*, 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.

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

Dearden, R., Friedman, N., & Russell, S. (1998). Bayesian Q-learning. In *Proceedings of the 15th National Conference on Artificial Intelligence* (pp. 761–768). American Association for Artificial Intelligence Madison, Wisconsin: AAAI Press.

Doya, K. (2000). Reinforcement learning in continuous time and space. *Neural Computation*, *12*, 219–245.

Fujita, H., & Ishii, S. (2007). Model-based reinforcement learning for partially observable games with sampling-based state estimation. *Neural Computation*, *19*, 3051–3087.

Gosavi, A. (2006). Boundedness of iterates in Q-learning. *Systems & Control Letters*, *55*, 347–349.

Gray, R. M. (2010). *Entropy and Information Theory*. (2nd ed.). New York: Springer.

Gray, R. M., & Kieffer, J. C. (1980). Asymptotically mean stationary measures. *Annals of Probability*, *8*, 962–973.

Han, T. S. (2003). *Information-Spectrum Methods in Information Theory* volume 50 of *Applications of mathematics*. Berlin: Springer.

Han, T. S., & Verdú, S. (1993). Approximation theory of output statistics. *IEEE Transactions on Information Theory*, *39*, 752–772.

Hopp, W. J., Bean, J. C., & Smith, R. L. (1987). A new optimality criterion for nonhomogeneous Markov decision processes. *Operations Research*, *35*, 875–883.

Howard, R. A. (1960). *Dynamic Programming and Markov Processes*. Cambridge, MA: MIT Press.

Ishii, S., Yoshida, W., & Yoshimoto, J. (2002). Control of exploitation-exploration meta-parameter in reinforcement learning. *Neural Networks*, *15*, 665–687.

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

Iwata, K., Ikeda, K., & Sakai, H. (2006a). The asymptotic equipartition property in reinforcement learning and its relation to return maximization. *Neural Networks*, *19*, 62–75.

Iwata, K., Ikeda, K., & Sakai, H. (2006b). A statistical property of multiagent learning based on Markov decision process. *IEEE Transactions on Neural Networks*, *17*, 829–842.

Kaelbling, L. P. (1993). *Learning in Embedded Systems*. Cambridge, MA: MIT Press.

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, *4*, 237–285.

Kieffer, J. C. (1974). A simple proof of the Moy-Perez generalization of the Shannon-Mcmillan theorem. *Pacific Journal of Mathematics*, *51*, 203–206.

Kushner, H. J., & Yin, G. G. (1997). *Stochastic Approximation Algorithms and Applications* volume 35 of *Applications of Mathematics*. New York: Springer-Verlag.

Likas, A. (1999). A reinforcement learning approach to online clustering. *Neural Computation*, *11*, 1915–1932.

McMillan, B. (1953). The basic theorems of information theory. *Annals of Mathematical Statistics*, *24*, 196–219.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill computer science series. New York: McGraw-Hill.

Morimoto, J., & Doya, K. (2005). Robust reinforcement learning. *Neural Computation*, *17*, 335–359.

Moy, S. C. (1961). Generalizations of Shannon-Mcmillan theorem. *Pacific Journal of Mathematics*, *11*, 705–714.

Neumann, G., & Peters, J. (2009). Fitted Q-iteration by advantage weighted regression. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Advances in Neural Information Processing Systems* (pp. 1177–1184). Cambridge, MA: MIT Press volume 22.

Pandana, C., & Liu, K. J. R. (2005). Near-optimal reinforcement learning framework for energy-aware sensor communications. *IEEE Journal on Selected Areas in Communications*, *23*, 788–797.

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*, *38*, 287–308.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. Adaptive Computation and Machine Learning. Cambridge, MA: MIT Press.

Tesauro, G. (1994). TD-gammon, a self-teaching backgammon program, achieves master-level play. *Neural Computation*, *6*, 215–219.

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

Verdú, S., & Han, T. S. (1997). The role of the asymptotic equipartition property in noiseless source coding. *IEEE Transactions on Information Theory*, *43*, 847–857.

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