Information Geometry and Information Theory in Machine Learning

Kazushi Ikeda[1] and Kazunori Iwata[2]

[1] Department of Systems Science, Kyoto University
Kyoto 606-8501 Japan
kazushi@i.kyoto-u.ac.jp
[2] Department of Intelligent Systems, Hiroshima City University
Hiroshima 731-3194 Japan
kiwata@hiroshima-cu.ac.jp

Abstract. Information geometry is a general framework of Riemannian manifolds with dual affine connections. Some manifolds (e.g. the manifold of an exponential family) have natural connections (e.g. e- and m-connections) with which the manifold is dually-flat. Conversely, a dually-flat structure can be introduced into a manifold from a potential function. This paper shows the case of quasi-additive algorithms as an example.

Information theory is another important tool in machine learning. Many of its applications consider information-theoretic quantities such as the entropy and the mutual information, but few fully recognize the underlying essence of them. The asymptotic equipartition property is one of the essence in information theory.

This paper gives an example of the property in a Markov decision process and shows how it is related to return maximization in reinforcement learning.

1 Introduction

Information geometry is a general framework of Riemannian manifolds with dual affine connections and was proposed by Amari [1] to give a clear view for the manifolds of statistical models. Since then, information geometry has widely been applied to other areas, such as statistical inference, information theory, neural networks, systems theory, mathematical programming, statistical physics, and stochastic reasoning [2], many of which are strongly related to machine learning community.

One example is that the Fisher information matrix appears as the Riemannian metric tensor of the statistical model in information geometry and another is that the Kullback-Leibler divergence and Hellinger distance are derived as the divergence defined for specific dual connections. Hence, if a study on machine learning considers the metric of a model or utilizes the mutual information, then it is based on information geometry in a sense, while there are a lot of more direct applications such as independent component analysis and semiparametric estimation. In this paper, we give another kind of applications of information geometry in Sec. 3.

M. Ishikawa et al. (Eds.): ICONIP 2007, Part II, LNCS 4985, pp. 295–304, 2008.

Another important tool in machine learning is information theory, which has much longer history than information geometry [3]. The asymptotic equipartition property (AEP) first stated by Shannon and developed through the method of types [3, Ch. 11] by Csiszár is based on a kind of the law of large numbers from the statistical viewpoint.

Although the AEP is an effective tool in analyzing learning algorithms, the importance of the AEP was not widely recognized in the machine learning community for a long time. However, some recent work utilizes the AEP for the analysis of learning algorithms such as genetic algorithms, since it holds in comprehensive stochastic processes related to machine learning. In this paper, we show that the AEP still holds in a Markov decision process (MDP) and discuss how it is related to return maximization in reinforcement learning (RL) in Sec. 5.

2 Preliminaries of Information Geometry

Information geometry discusses the properties of a manifold S, which is intuitively an n-dimensional differentiable subset of a Euclidean space with a coordinate system $\{\xi^i\}$ where ξ^i denotes the ith coordinate. Due to its smoothness, we can define the tangent space T_p at a point p in the manifold S as the space spanned by the tangent vectors $\{\partial_i \equiv \partial/\partial\xi^i\}$ of the coordinate curves, in other words, we locally linearize the manifold.

Since the tangent space T_p is a Euclidean space, an inner product can be defined as $g_{ij} \equiv \langle \partial_i, \partial_j \rangle$, where g_{ij} depends on the point $p \in S$ and it is called the Riemannian metric on S or simply the metric. Note that the metric is not naturally determined in general, the Fisher information matrix is a natural metric for the statistical manifold.

Since the tangent space T_p varies from point to point, we need to establish a linear mapping $\Pi_{p,p'} : T_p \to T_{p'}$ where p and p' are neighboring points and $\mathrm{d}\xi^i \equiv \xi^i(p') - \xi^i(p)$. Then, the difference between the vectors $\Pi_{p,p'}((\partial_j)_p)$ and $(\partial_j)_{p'}$ is a linear combination of $\{\mathrm{d}\xi^i\}$, that is,

$$\Pi_{p,p'}(\partial_j) = \partial'_j - \mathrm{d}\xi^i (\Gamma^k_{ij})_p \partial'_k, \tag{1}$$

where Γ^k_{ij} is the n^3 functions of p called the affine connection on S or simply the connection (Fig. 1). Using the connection of a manifold, any vector in T_p can be parallel-translated into another tangent space T_q along a curve connecting the two points p and q.

As well as the metric, the connection of a manifold can also be determined arbitrarily. However, if we require that the parallel translation of two vectors along a curve γ leaves their inner product unchanged, that is,

$$\langle \Pi_\gamma(D_1), \Pi_\gamma(D_2) \rangle_q = \langle D_1, D_2 \rangle_p \,, \tag{2}$$

then the connection is uniquely determined that satisfies

$$\partial_k g_{ij} = \Gamma_{ki,j} + \Gamma_{kj,i} \tag{3}$$

$$\Gamma_{ij,k} \equiv \Gamma^h_{ij} g_{hk}, \tag{4}$$

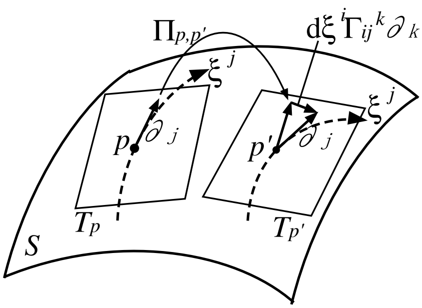

Fig. 1. Affine connection

which is called the Riemannian connection or the Levi-Civita connection with respect to g.

Information geometry introduces a pair of connections, called the dual connections, so that the inner product of two vectors is unchanged when one vector is parallel-translated with one connection and the other vector with the other connection, that is,

$$\langle D_1, D_2 \rangle_p = \left\langle \Pi_\gamma(D_1), \Pi^*_\gamma(D_2) \right\rangle_q . \tag{5}$$

It is known that the dual connections Γ and Γ^* satisfy

$$\partial_k g_{ij} = \Gamma_{ki,j} + \Gamma^*_{kj,i}. \tag{6}$$

This means that the Riemann connection is a special case where the connection is self-dual.

If a manifold has a coordinate system satisfying $\Gamma^k_{ij} = 0$, the manifold is called to be flat and the coordinate system is called affine. We denote an affine coordinate system by $\{\theta^i\}$ in this paper. It is known that if a manifold is flat for a connection Γ, it is also flat for its dual connection Γ^*. However, $\{\theta^i\}$ is never affine in general and we need to introduce another affine coordinate system $\{\eta_i\}$. These two coordinate systems called the dual coordinate systems have the relationship

$$\eta_i = \partial_i \psi(\boldsymbol{\theta}) \equiv \frac{\partial \psi(\boldsymbol{\theta})}{\partial \theta^i}, \tag{7}$$

$$\theta^i = \partial^i \phi(\boldsymbol{\eta}) \equiv \frac{\partial \phi(\boldsymbol{\eta})}{\partial \eta_i}, \tag{8}$$

$$\psi(\boldsymbol{\theta}) + \phi(\boldsymbol{\eta}) - \theta^i \eta_i = 0 \tag{9}$$

where $\psi(\boldsymbol{\theta})$ and $\phi(\boldsymbol{\eta})$ are respectively convex potential functions of $\boldsymbol{\theta} \equiv (\theta^1, \ldots, \theta^n)$ and $\eta \equiv (\eta_1, \ldots, \eta_n)$. In short, $\boldsymbol{\eta}$ is the Legendre transform of $\boldsymbol{\theta}$ and vice versa. The divergence which expresses a kind of the distance from p to q has a similar form to (9),

$$D(p\|q) \equiv \psi(\boldsymbol{\theta}(p)) + \phi(\boldsymbol{\eta}(q)) - \theta^i(p)\eta_i(q) \geq 0. \tag{10}$$

The divergence holds the generalized Pythagorean relation

$$D(p\|r) = D(p\|q) + D(q\|r) \tag{11}$$

when the Γ-geodesic between p and q and the Γ^8-geodesic between q and r are orthogonal at q. This relation is useful in optimization problems.

The most popular example of dual connections will be the ones for the manifold of an exponential family in statistics. The e-connection and m-connection are defined as

$$\Gamma^{(e)}_{ij,k} \equiv E[(\partial_i\partial_j l_\theta)(\partial_k l_\theta)] \tag{12}$$

$$\Gamma^{(m)}_{ij,k} = E[(\partial_i\partial_j l_\theta + \partial_i l_\theta \partial_j l_\theta)(\partial_k l_\theta)] \tag{13}$$

where $l_\theta \equiv \log p(x;\theta)$ and θ^i's and η_i's are the canonical and expectation parameters, respectively. The Kullback-Leibler divergence is derived from these connections.

3 Dually-Flat Structure of Learning Machines

In the above, the dual connections of a manifold lead to the dually-flat structure with two potential functions. Conversely, a dually-flat structure can be derived from a coordinate system with a convex potential function as below.

Let S be an n-dimensional manifold with a coordinate system $\boldsymbol{\theta}$ and $\psi(\boldsymbol{\theta})$ a smooth convex function on S. Then, the dual coordinate system $\boldsymbol{\eta}$ is defined as

$$\eta_i(\boldsymbol{\theta}) \equiv \partial_i \psi(\boldsymbol{\theta}), \tag{14}$$

and $\boldsymbol{\eta}(\boldsymbol{\theta}) = \partial\psi(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$, in short. From the convexity of $\psi(\boldsymbol{\theta})$, $\boldsymbol{\eta}$ is a one-to-one function of $\boldsymbol{\theta}$ and vice versa.

Let us define a function of $\boldsymbol{\eta}$ as

$$\phi(\boldsymbol{\eta}) \equiv \boldsymbol{\theta}(\boldsymbol{\eta}) \cdot \boldsymbol{\eta} - \psi(\boldsymbol{\theta}(\boldsymbol{\eta})), \tag{15}$$

where $\cdot$ is the canonical dot product and

$$\boldsymbol{\theta}(\boldsymbol{\eta}) \equiv \arg\max_{\theta^i} \left[\boldsymbol{\theta} \cdot \boldsymbol{\eta} - \psi(\boldsymbol{\theta})\right]. \tag{16}$$

It is easily shown $\partial^i \phi(\eta) = \theta^i$ and

$$\psi(\boldsymbol{\theta}) + \phi(\boldsymbol{\eta}) - \boldsymbol{\theta} \cdot \boldsymbol{\eta} = 0. \tag{17}$$

The divergence from P to Q is defined as

$$D(P\|Q) := \psi(\boldsymbol{\theta}_Q) + \phi(\boldsymbol{\eta}_P) - \boldsymbol{\theta}_Q \cdot \boldsymbol{\eta}_P, \tag{18}$$

which always takes a non-negative value and null if and only if $P = Q$, where $\boldsymbol{\theta}_P$ and $\boldsymbol{\theta}_Q$ respectively denote the $\boldsymbol{\theta}$-coordinates of two points $P \in S$ and $Q \in S$, and $\boldsymbol{\eta}_P$ and $\boldsymbol{\eta}_Q$ their $\boldsymbol{\eta}$-coordinates. Note that the divergence may be written as $D(\boldsymbol{\theta}_P, \boldsymbol{\theta}_Q)$ when we regard it as a function of $\boldsymbol{\theta}$-coordinates and $D(\boldsymbol{\eta}_P, \boldsymbol{\eta}_Q)$ when as a function of $\boldsymbol{\eta}$-coordinates.

Since the metric expresses the length of the infinitesimal segment, it is given by differentiating the divergence, that is,

$$G(\boldsymbol{\theta}) = [g_{ij}(\boldsymbol{\theta})] = \partial_i \partial_j \psi(\boldsymbol{\theta}) \tag{19}$$

$$H(\boldsymbol{\eta}) = [h^{ij}(\boldsymbol{\eta})] = \partial^i \partial^j \phi(\boldsymbol{\eta}) = G^{-1}(\boldsymbol{\theta}). \tag{20}$$

Since the dual connections and the geodesics for them are essentially equivalent, we determine the geodesics instead of explicitly defining the connections. Here, we assume that $\boldsymbol{\theta}$ is an affine coordinate system, that is, a geodesic for Γ is expressed as

$$\boldsymbol{\theta}(t) = \boldsymbol{c}t + \boldsymbol{b} \tag{21}$$

where $\boldsymbol{c}$ and $\boldsymbol{b}$ are constant vectors, and a geodesic for Γ^* is similarly expressed as

$$\boldsymbol{\eta}(t) = \boldsymbol{c}t + \boldsymbol{b}. \tag{22}$$

We apply the discussion above to the quasi-additive (QA) algorithms [4] according to [5]. The family of QA algorithms is a generalization of the perceptron learning for a linear dichotomy. It has two vectors, the parameter vector $\boldsymbol{\theta}$ to which a scaled input vector $\boldsymbol{x}$ is added and the weight vector $\boldsymbol{\eta}$ which is a nonlinear transform of $\boldsymbol{\theta}$ elementwise. More precisely,

$$\eta_i = f(\theta^i) \quad i = 1, \ldots, n, \tag{23}$$

where f is an monotonically increasing differentiable function. When f is an exponential function $\exp(\cdot)$, for instance, an addition to the parameter vector appears as a multiplication in the weight vector since

$$\eta_i^{(t)} = f({\theta^{(t)}}^i) = f({\theta^{(t-1)}}^i + {x^{(t)}}^i) = \eta_i^{(t-1)} \exp({x^{(t)}}^i). \tag{24}$$

The output of the linear dichotomy is the sign $y \in \{\pm 1\}$ of the dot product with the weight vector $\boldsymbol{\eta}$ for an input vector $\boldsymbol{x}$, that is, $y = \mathrm{sgn}[\boldsymbol{\eta} \cdot \boldsymbol{x}] \in \{\pm 1\}$. In total, QA algorithms have a general form of

$$\dot{\boldsymbol{\theta}} = C(\boldsymbol{\eta}, \boldsymbol{x}, y) y \boldsymbol{x}, \qquad \boldsymbol{\eta} = f(\boldsymbol{\theta}). \tag{25}$$

Suppose that f satisfies $f(0) = 0$ and define a potential function

$$\psi(\boldsymbol{\theta}) = \sum_{i=1}^{n} g(\theta^i), \qquad g(s) = \int_0^s f(\sigma) \mathrm{d}\sigma. \tag{26}$$

Then, we can introduce a dually-flat structure to QA algorithms from this potential function. In fact, the parameter vector $\boldsymbol{\theta}$ and the weight vector $\boldsymbol{\eta}$ of a QA algorithm are dual affine coordinate systems through the monotonically increasing function f as below:

$$\eta_i = \partial_i \psi(\boldsymbol{\theta}) = f(\theta^i), \quad g_{ij} = \partial_j \eta_i = f'(\theta^i)\delta_{ij} \tag{27}$$

$$\phi(\boldsymbol{\eta}) = \boldsymbol{\theta}^T \boldsymbol{\eta} - \psi(\boldsymbol{\theta}) = \sum_{i=1}^{n} \left[\theta^i f(\theta^i) - g(\theta^i)\right] = \sum_{i=1}^{n} h(f(\theta^i)) = \sum_{i=1}^{n} h(\eta_i), \tag{28}$$

$$\theta^i = \partial^i \phi(\boldsymbol{\eta}) = f^{-1}(\eta_i), \quad g^{ij} = \partial^j \theta^i = (f^{-1})'(\eta_i)\delta^{ij}, \tag{29}$$

where $'$ denotes the derivative and

$$h(s) = \int_0^s f^{-1}(\tau)\mathrm{d}\tau. \tag{30}$$

We can show that the QA algorithm is an approximate of the natural gradient descent method for the dually-flat structure derived from the potential (26). See [5] for details.

4 Preliminaries of Information Theory

Information theory gave answers to the two fundamental questions of the ultimate data compression and the ultimate data transmission in communication theory and has been applied to many other fields beyond the communication theory [3]. In this section, we introduce the so-called the asymptotic equipartition property (AEP) which is the analog of the low of large numbers.

The simplest version of the AEP is formalized in the following theorem.

Theorem 1 (AEP). *Let $p(x)$ be any probability density function defined over $\mathcal{X}$. If $X_1, X_2, \ldots$ are i.i.d. random variables drawn according to $p(x)$, then*

$$-\frac{1}{n}\log p(X_1, X_2, \ldots, X_n) \to H(p) \quad \textit{in probability}, \tag{31}$$

as $n \to \infty$, where $H(p)$ denotes the entropy of $p(x)$.

The AEP yields the typical set of sequences in this i.i.d. case.

Definition 1 (Typical Set). *The typical set $A_\epsilon^{(n)}$ with respect to $p(x)$ is defined as the set of sequences $(x_1, x_2, \ldots, x_n)$ such that for any $\epsilon > 0$,*

$$\exp[-n(H(p)+\epsilon)] \le p(x_1, x_2, \ldots, x_n) \le \exp[-n(H(p)-\epsilon)]. \tag{32}$$

Theorem 2 (Asymptotic Properties).

1. *If $(x_1, x_2, \ldots, x_n) \in A_\epsilon^{(n)}$, then $H(p) - \epsilon \le (-\log p(x_1, x_2, \ldots, x_n))/n \le H(p) + \epsilon$.*
2. *$\Pr(A_\epsilon^{(n)}) > 1 - \epsilon$ for n sufficiently large.*

3. $|A_\epsilon^{(n)}| \le \exp[n(H(p)+\epsilon)]$, *where* $|A|$ *is the number of elements in the set* A.
4. $|A_\epsilon^{(n)}| \ge (1-\epsilon)\exp[n(H(p)-\epsilon)]$ *for* n *sufficiently large.*

These properties state that there exists the typical set of sequences with probability nearly one, that all the elements of 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 density function. This means that the number of elements in the typical set is quite small compared to the number of possible sequences. Hence, we can focus most of our attention on the elements in the typical set since the others appear with probability nearly zero.

The AEP still holds in a number of stationary ergodic processes related to machine learning. In fact, it holds in a Markov chain model formulated by genetic algorithms and this has been applied in [6, 7] for the analysis of genetic algorithms. In the next section, we show that the AEP holds in a Markov decision process (MDP). According to [8], we also discuss how this is related to return maximization in reinforcement learning (RL).

5 The AEP in Reinforcement Learning

In general, RL is formulated as a discrete-time piecewise stationary ergodic MDP with discrete state-actions. The elements of the MDP are described as follows: the sets of states, actions and rewards are denoted as $\mathcal{S} \equiv \{s_1,\ldots,s_I\}$, $\mathcal{A} \equiv \{a_1,\ldots,a_J\}$ and $\mathcal{R} \equiv \{r_1,\ldots,r_K\}$, respectively. Let $s(t)$, $a(t)$ and $r(t)$ denote the random variables of state, action and reward at time-step $t \in \{1,2,\ldots\}$, defined over $\mathcal{S}$, $\mathcal{A}$ and $\mathcal{R}$, respectively. The policy matrix Γ^π of an agent and the state-transition matrix Γ^{T} of an environment are described as

$$\Gamma^\pi \equiv \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1J} \\ p_{21} & p_{22} & \cdots & p_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ p_{I1} & p_{I2} & \cdots & p_{IJ} \end{pmatrix}, \qquad \Gamma^{\mathrm{T}} \equiv \begin{pmatrix} p_{1111} & p_{1112} & \cdots & p_{11IK} \\ p_{1211} & p_{1212} & \cdots & p_{12IK} \\ \vdots & \vdots & \ddots & \vdots \\ p_{IJ11} & p_{IJ12} & \cdots & p_{IJIK} \end{pmatrix}, \tag{33}$$

respectively, where $p_{ij} \equiv \Pr(a(t)=a_j \,|\, s(t)=s_i)$ denotes the probability that the agent selects action $a_j \in \mathcal{A}$ in state $s_i \in \mathcal{S}$, and $p_{iji'k} \equiv \Pr(s(t+1) = s_{i'}, r(t+1) = r_k \,|\, s(t) = s_i, a(t) = a_j)$ denotes the probability that the agent receives scalar reward $r_k \in \mathcal{R}$ and observes subsequent state $s_{i'} \in \mathcal{S}$ of the environment when action $a_j \in \mathcal{A}$ is taken in state $s_i \in \mathcal{S}$. Let $\Gamma \equiv (\Gamma^\pi, \Gamma^{\mathrm{T}})$ for simplicity. Each of the initial state distribution in the environment is defined as $q_i \equiv \Pr(s(1)=s_i) > 0$ for any $s_i \in \mathcal{S}$. Note that the agent can determine the policy matrix Γ^π for action selection while it does not know the state-transition matrix Γ^{T}.

Suppose that the policy of the agent is improved sufficiently slowly such that the sequence of n time-steps, $\boldsymbol{x} \equiv \{s(1), a(1), r(2), s(2), a(2), \ldots, r(n), s(n), a(n), r(n+1)\}$, is drawn according to a stationary ergodic MDP described above. We let $r(n+1) = r(1)$ for notational convenience, and hence the sequence is simply written as $\boldsymbol{x} = \{s(t), a(t), r(t)\}_{t=1}^{n}$. As a result of actual trials by the agent, the empirical distributions $F_{\mathcal{S}}$, $F_{\mathcal{SA}}$, Φ^{π} and Φ^{T} are uniquely obtained according to the observed sequence of $\boldsymbol{x}$ in the trials, where $F_{\mathcal{S}} \equiv \{f_i\}$ and $F_{\mathcal{SA}} \equiv \{f_{ij}\}$ are the empirical state distribution and the empirical state-action distribution, respectively, where $f_i \equiv |\{t \in \{1, \ldots, n\} \mid s(t) = s_i \in \mathcal{S}\}|/n$ and $f_{ij} \equiv |\{t \in \{1, \ldots, n\} \mid s(t) = s_i \in \mathcal{S}, a(t) = a_j \in \mathcal{A}\}|/n$, and the empirical policy matrix Φ^{π} and the empirical state-transition matrix Φ^{T} are denoted as

$$\Phi^{\pi} \equiv \begin{pmatrix} g_{11} & g_{12} & \cdots & g_{1J} \\ g_{21} & g_{22} & \cdots & g_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ g_{I1} & g_{I2} & \cdots & g_{IJ} \end{pmatrix}, \qquad \Phi^{\mathrm{T}} \equiv \begin{pmatrix} g_{1111} & g_{1112} & \cdots & g_{11IK} \\ g_{1211} & g_{1212} & \cdots & g_{12IK} \\ \vdots & \vdots & \ddots & \vdots \\ g_{IJ11} & g_{IJ12} & \cdots & g_{IJIK} \end{pmatrix}, \tag{34}$$

respectively. We need to consider Φ^{π}-shell and Φ^{T}-shell for more strict discussion.

The following theorems are obtained from the AEP in the MDP.

Definition 2 (Typical Set). *The typical set $C_{\lambda_n}^{n}(\Gamma)$ in the MDP is defined as the set of sequences such that for any $\lambda_n > 0$, empirical distributions satisfy*

$$D(\Phi^{\pi} \| \Gamma^{\pi} \mid F_{\mathcal{S}}) + D(\Phi^{\mathrm{T}} \| \Gamma^{\mathrm{T}} \mid F_{\mathcal{SA}}) \leq \lambda_n, \tag{35}$$

where $D(\Phi^{\pi} \| \Gamma^{\pi} \mid F_{\mathcal{S}})$ denotes the conditional divergence between the elements in Φ^{π} and Γ^{π} given $F_{\mathcal{S}}$.

Theorem 3 (Probability of Typical Set). *If $\lambda_n \to 0$ as $n \to \infty$ such that*

$$\lambda_n > \frac{(IJ + I^2JK)\log(n+1) + \log I - \min p_{iji'k}}{n}, \tag{36}$$

there exists a sequence $\{\epsilon(\lambda_n)\}$ such that $\epsilon(\lambda_n) \to 0$, and $\Pr(C_{\lambda_n}^{n}(\Gamma)) = 1 - \epsilon(\lambda_n)$.

Theorem 4 (Equi-Probability of Elements). *If $\boldsymbol{x} \in C_{\lambda_n}^{n}(\Gamma)$, then there exists a sequence $\{\rho_n\}$ such that $\rho_n \to 0$ as $n \to \infty$, and*

$$\frac{\min p_{iji'k}}{n} - \rho_n \leq -\frac{1}{n}\log \Pr(\boldsymbol{x}) - \phi(\Gamma) \leq -\frac{\min q_i}{n} + \lambda_n + \rho_n, \tag{37}$$

where $\phi(\Gamma)$ is the stochastic complexity of the MDP, defined as

$$\phi(\Gamma) \equiv H(\Gamma^{\pi}|V) + H(\Gamma^{\mathrm{T}}|W), \tag{38}$$

where V and W are the limits of $F_{\mathcal{S}}$ and $F_{\mathcal{SA}}$ with respect to n.

Theorem 5 (Typical Set Size). *There exist two sequences $\{\zeta_n\}$ and $\{\eta_n\}$ such that $\zeta_n \to 0$ and $\eta_n \to 0$ as $n \to \infty$, and*

$$\exp[n\{\phi(\Gamma) - \zeta_n\}] \leq |C_{\lambda_n}^{n}(\Gamma)| \leq \exp[n\{\phi(\Gamma) + \eta_n\}]. \tag{39}$$

Now let us consider how we can maximize the return in RL. In this paper, return maximization means maximizing the probability that the best sequences appear in trials. Since only the sequences in the typical set appear with probability nearly one, the typical set must be large enough such that it includes the best sequences. On the other hand, from the equi-probability of elements in the typical set, the size of the typical set should be minimized to increase the ratio of the best sequences to the elements in the typical set. This tradeoff is essentially identical to the exploration-exploitation dilemma widely recognized in RL. Because the size of the typical set is characterized by the stochastic complexity, and it is an important guide to solve the dilemma. For example, we can derive the dependency of the stochastic complexity on the learning parameter such as β in the softmax method and ϵ in the ϵ-greedy method, which gives some insight into an appropriate control of the parameter when the learning proceeds.

Information theory can also be applied to the multi-agent problem [9] which is the analog of the multi-terminal information theory. Let the sequence $\boldsymbol{x}_m$ of the m-th of M agents be $\{s_m(1), a_m(1), r_m(2), s_m(2), a_m(2), \ldots, r_m(n), s_m(n), a_m(n), r_m(n+1)\}$. The AEP still holds in an MDP in the multi-agent case, where p_{ij} and $p_{iji'k}$ in the elements of the matrices in (33) are extended to

$$p_{i_1\cdots i_M, j_1\cdots j_M} \equiv \Pr(\boldsymbol{a}(t) = \boldsymbol{a}_{j_1\cdots j_M} \mid \boldsymbol{s}(t) = \boldsymbol{s}_{i_1\cdots i_M}), \tag{40}$$

$$\begin{aligned} p_{i_1\cdots i_M, j_1\cdots j_M, i'_1\cdots i'_M, k_1\cdots k_M} \equiv \Pr(&\boldsymbol{s}(t+1) = \boldsymbol{s}_{i'_1\cdots i'_M}, \boldsymbol{r}(t+1) = \boldsymbol{r}_{k_1\cdots k_M} \\ &\mid \boldsymbol{s}(t) = \boldsymbol{s}_{i_1\cdots i_M}, \boldsymbol{a}(t) = \boldsymbol{a}_{j_1\cdots j_M}), \end{aligned} \tag{41}$$

respectively, where $\boldsymbol{s}(t) \equiv (s_1(t), \ldots, s_M(t))$ and $\boldsymbol{s}_{i_1\cdots i_M} \equiv (s_{i_1}, \ldots, s_{i_M}) \in \mathcal{S}^M$. When the agents that exist in the same environment can communicate with each other, i.e., know their states and decide their actions together, the probability of their policy is expressed as (40). When each agent can know all of the other agents' states but cannot know how the others' actions are taken, $p_{i_1\cdots i_M, j_1\cdots j_M}$ in this case cannot take a general form but it is expanded as

$$p_{i_1\cdots i_M, j_1\cdots j_M} = \prod_{m=1}^{M} \Pr(a_m(t) = a_{j_m} \mid \boldsymbol{s}(t) = \boldsymbol{s}_{i_1\cdots i_M}). \tag{42}$$

This visible case is more limited in the communication among the agents. When no agent can recognize any of the other agents' states nor actions, it is also

$$p_{i_1\cdots i_M, j_1\cdots j_M} = \prod_{m=1}^{M} \Pr(a_m(t) = a_{j_m} \mid s_m(t) = s_{i_m}). \tag{43}$$

This blind case is much more limited than the visible case. The limitations in the communication increase the entropy $H(\Gamma^\pi|V)$ and make the performance of the agents worse. The multi-agent studies should take the limitations into account.

6 Conclusions

In this paper, we briefly introduced an essence of the information geometry, that is, the duality was shown to be one of the most important properties. When a manifold is dually-flat, the divergence is naturally derived. From a convex potential function, on the other hand, we can introduce a dually-flat structure to the space. One example on quasi-additive algorithms was given in Sec. 3.

Another important tool in machine learning is information theory. Although it has a wide diversity, we concentrate our attention on the asymptotic equipartition property (AEP), which is known as the law of large numbers in statistics. We showed that the AEP on the sequences generated from a Markov decision process using an example on the sequences in reinforcement learning (RL) in Sec. 5. This property should be taken into account in the analysis of algorithms since only the typical sequences appear with probability nearly one.

Information geometry and information theory are so powerful tools that there are a lot of fields to be applied in the future.

Acknowledgment

This study is supported in part by a Grant-in-Aid for Scientific Research (18300078, 18700157) from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

1. Amari, S.I.: Differential-Geometrical Methods in Statistics. Lecture Notes in Statistics, vol. 28. Springer, Heidelberg (1985)
2. Amari, S.I., Nagaoka, H.: Methods of Information Geometry. Translations of Mathematical Monographs, vol. 191. AMS and Oxford Univ. Press, Oxford (2000)
3. Cover, T.M., Thomas, J.A.: Elements of Information Theory, 2nd edn. John Wiley and Sons, Inc., Hoboken (2006)
4. Grove, A.J., Littlestone, N., Schuurmans, D.: General convergence results for linear discriminant updates. Machine Learning 43(3), 173–210 (2001)
5. Ikeda, K.: Geometric properties of quasi-additive learning algorithms. IEICE Trans. Fundamentals E89-A(10), 2812–2817 (2006)
6. Suzuki, J.: A markov chain analysis on simple genetic algorithms. IEEE Trans. on Systems, Man and Cybernetics 25(4), 655–659 (1995)
7. Suzuki, J.: A further result on the markov chain model of genetic algorithms and its application to a simulated annealing-like strategy. IEEE Trans. on Systems, Man and Cybernetics, Part B, Cybernetics 28(1), 95–102 (1998)
8. Iwata, K., Ikeda, K., Sakai, H.: The asymptotic equipartition property in reinforcement learning and its relation to return maximization. Neural Networks 19(1), 62–75 (2006)
9. Iwata, K., Ikeda, K., Sakai, H.: A statistical property of multi-agent learning based on Markov decision process. IEEE Trans. on Neural Networks 17(4), 829–842 (2006)