A New Natural Policy Gradient by Stationary Distribution Metric

Tetsuro Morimura[1,2], Eiji Uchibe[1], Junichiro Yoshimoto[1,3], and Kenji Doya[1,3,4]

[1] Initial Research Project, Okinawa Institute of Science and Technology
[2] IBM Research, Tokyo Research Laboratory
[3] Graduate School of Information Science, Nara Institute of Science and Technology
[4] ATR Computational Neuroscience Laboratories
{morimura,uchibe,jun-y,doya}@oist.jp

Abstract. The parameter space of a statistical learning machine has a Riemannian metric structure in terms of its objective function. Amari [1] proposed the concept of "natural gradient" that takes the Riemannian metric of the parameter space into account. Kakade [2] applied it to policy gradient reinforcement learning, called a natural policy gradient (NPG). Although NPGs evidently depend on the underlying Riemannian metrics, careful attention was not paid to the alternative choice of the metric in previous studies. In this paper, we propose a Riemannian metric for the joint distribution of the state-action, which is directly linked with the average reward, and derive a new NPG named *"Natural State-action Gradient"* (NSG). Then, we prove that NSG can be computed by fitting a certain linear model into the immediate reward function. In numerical experiments, we verify that the NSG learning can handle MDPs with a large number of states, for which the performances of the existing (N)PG methods degrade.

Keywords: policy gradient reinforcement learning, natural gradient, Riemannian metric matrix, Markov decision process.

1 Introduction

Policy gradient reinforcement learning (PGRL) attempts to find a policy that maximizes the average (or time-discounted) reward, based on the gradient ascent in the policy parameter space [3,4,5]. As long as the policy is represented by a parametric statistical model that satisfies some mild conditions, PGRL can be instantly implemented in the Markov decision process (MDP). Moreover, since it is possible to treat the parameter controlling the randomness of the policy, PGRLs, rather than value-based RLs, can obtain the appropriate stochastic policy and be applied to the partially observable MDP (POMDP). Meanwhile, depending on the tasks, PGRL methods often take a huge number of learning steps. In this paper, we propose a new PGRL method that can improve the slow learning speed by focusing on the metric of the parameter space of the learning model.

W. Daelemans et al. (Eds.): ECML PKDD 2008, Part II, LNAI 5212, pp. 82–97, 2008.

It is easy to imagine that large-scale tasks suffer from a slow learning speed because the dimensionality of the policy parameters increases in conjunction with the task complexity. Besides the problem of dimensionality, the geometric structure of the parameter space also gives rise to slow learning. Ordinary PGRL methods omit the sensitivity of each element of the policy parameter and the correlation between the elements, in terms of the probability distributions of the MDP. However, most probability distributions expressed by the MDP have some manifold structures instead of Euclidean structures. Therefore, the updating direction of the policy parameter by the ordinary gradient method is different from the steepest direction on the manifold; thus, the optimization process occasionally falls into a stagnant state, commonly called a *plateau*. This is mainly due to the regions in which the geometric structure for the objective function with respect to the parameter coordinate system becomes fairly flat and its derivative becomes almost zero [6]. It was reported that a plateau was observed in a very simple MDP with only two states [2]. In order to solve such problem, Amari [1] proposed a "natural gradient" for the steepest gradient method in Riemannian space. Because the direction of the natural gradient is defined on a Riemannian metric, it is an important issue how to design the Riemannian metric. Nevertheless, the metric proposed by Kakade [2] has so far been the only metric in the application of the natural gradient for RL [7,8,9], commonly called *natural policy gradient* (NPG) reinforcement learning.

In this paper, we propose the use of the Fisher information matrix of the state-action joint distribution as the Riemannian metric for RL and derive a new robust NPG learning, *"natural state-action gradient"* (NSG) learning. It is shown that this metric considers the changes in the stationary state-action joint distribution, specifying the average reward as the objective function. In contrast, Kakade's metric takes into account only changes in the action distribution and omits changes in the state distribution, which also depends on the policy in general. A comparison with the Hessian matrix is also given in order to confirm the adequacy of the proposed metric. We also prove that the gradient direction as computed by NSG is equal to the adjustable parameter of the linear regression model with the basis function defined on the policy when it minimizes the mean square error for the rewards. Finally, we demonstrate that the proposed NSG learning improves the performance of conventional (N)PG-based learnings by means of numerical experiments with varying scales of MDP tasks.

2 Conventional Natural Policy Gradient Method

We briefly review PGRL in section 2.1 and the natural gradient [1] and the NPG in section 2.2. In section 2.3, we introduce the controversy of NPGs.

2.1 Policy Gradient Reinforcement Learning

PGRL is modeled on a discrete-time Markov decision process (MDP) [10,11]. It is defined by the quintuplet $(\mathcal{S}, \mathcal{A}, p, r, \pi_\theta)$, where $\mathcal{S} \ni s$ and $\mathcal{A} \ni a$ are finite sets of

states and actions, respectively. Further, $p : \mathcal{S}\times\mathcal{A}\times\mathcal{S} \to [0,1]$ is a state transition probability function of a state s_t, an action a_t, and the following state s_{t+1} at a time step t, i.e., $p(s_{t+1}|s_t,a_t) \equiv \Pr(s_{t+1}|s_t,a_t)$ [1]. $r : \mathcal{S}\times\mathcal{A}\times\mathcal{S} \to \mathcal{R}$ is a reward function of s_t, a_t, and s_{t+1}, and is bounded, which defines an immediate reward $r_{t+1} = r(s_t,a_t,s_{t+1})$ observed by a learning agent. $\pi : \mathcal{A}\times\mathcal{S}\times\mathcal{R}^d \to [0,1]$ is an action probability function of a_t, s_t, and a policy parameter $\boldsymbol{\theta} \in \mathcal{R}^d$, and is always differentiable with respect to $\boldsymbol{\theta}$ known as a policy, i.e., $\pi(a_t|s_t;\boldsymbol{\theta}) \equiv \Pr(a_t|s_t,\boldsymbol{\theta})$. It defines the decision-making rule of the learning agent and is adjustable by tuning $\boldsymbol{\theta}$. We make an assumption that the Markov chain $\mathrm{M}(\boldsymbol{\theta}) = \{\mathcal{S},\mathcal{A},p,\pi_\theta\}$ is ergodic for all $\boldsymbol{\theta}$. Then, there exists a unique stationary state distribution $d_{\boldsymbol{\theta}}(s) \equiv \Pr(s|\mathrm{M}(\boldsymbol{\theta}))$, which is equal to the limiting distribution and independent of the initial state, $d_{\boldsymbol{\theta}}(s') = \lim_{t\to\infty}\Pr(S_t = s'|S_0 = s,\mathrm{M}(\boldsymbol{\theta}))$, $^\forall s \in \mathcal{S}$. This distribution satisfies the balance equation:

$$d_{\boldsymbol{\theta}}(s') = \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} p(s'|s,a)\pi(a|s;\boldsymbol{\theta})d_{\boldsymbol{\theta}}(s). \tag{1}$$

The following equation instantly holds [10]:

$$d_{\boldsymbol{\theta}}(s') = \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\Pr(S_t = s'|S_0 = s,\mathrm{M}(\boldsymbol{\theta})),\ ^\forall s \in \mathcal{S}. \tag{2}$$

The goal of PGRL is to find the policy parameter $\boldsymbol{\theta}^*$ that maximizes the average of the immediate rewards called the *average reward*:

$$R(\boldsymbol{\theta}) \equiv \lim_{T\to\infty}\frac{1}{T}\mathbb{E}\left\{\sum_{t=1}^{T} r_t \Big| s_0,\mathrm{M}(\boldsymbol{\theta})\right\}, \tag{3}$$

where $\mathbb{E}\{\cdot\}$ denotes expectation. It is noted that, under the assumption of ergodicity (eq.2), the average reward is independent of the initial state s_0 and can be shown to equal [10]:

$$\begin{aligned} R(\boldsymbol{\theta}) &= \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{s'\in\mathcal{S}} d_{\boldsymbol{\theta}}(s)\pi(a|s;\boldsymbol{\theta})p(s'|s,a)r(s,a,s') \\ &= \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d_{\boldsymbol{\theta}}(s)\pi(a|s;\boldsymbol{\theta})\bar{r}(s,a) \\ &\equiv \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \Pr(s,a|\mathrm{M}(\boldsymbol{\theta}))\bar{r}(s,a). \end{aligned} \tag{4}$$

where $\bar{r}(s,a) \equiv \sum_{s'\in\mathcal{S}} p(s'|s,a)r(s,a,s')$. The statistical model $\Pr(s,a|\mathrm{M}(\boldsymbol{\theta}))$ is called the stationary state-action (joint) distribution. Since $\bar{r}(s,a)$ is usually independent of the policy parameter, the derivative of the average reward with

[1] Although it should be $\Pr(S_{t+1} = s_{t+1}|S_t = s_t, A_t = a_t)$ for the random variables S_{t+1}, S_t, and A_t to be precise, we notate $\Pr(s_{t+1}|s_t,a_t)$ for simplicity. The same rule is applied to the other distributions.

respect to the policy parameter, $\nabla_{\boldsymbol{\theta}} R(\boldsymbol{\theta}) \equiv [\partial R(\boldsymbol{\theta})/\partial\theta_1, ..., \partial R(\boldsymbol{\theta})/\partial\theta_d]^\top$ is given by

$$\begin{aligned}\nabla_{\boldsymbol{\theta}} R(\boldsymbol{\theta}) &= \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \nabla_{\boldsymbol{\theta}} \left\{d_{\boldsymbol{\theta}}(s)\pi(a|s;\boldsymbol{\theta})\right\} \bar{r}(s,a) \\ &= \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d_{\boldsymbol{\theta}}(s)\pi(a|s;\boldsymbol{\theta})\bar{r}(s,a) \left\{\nabla_{\boldsymbol{\theta}} \ln \pi(a|s;\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}} \ln d_{\boldsymbol{\theta}}(s)\right\}, \end{aligned} \tag{5}$$

where $\top$ denotes transpose and $\nabla_{\boldsymbol{\theta}} a_{\boldsymbol{\theta}} b_{\boldsymbol{\theta}} \equiv (\nabla_{\boldsymbol{\theta}} a_{\boldsymbol{\theta}}) b_{\boldsymbol{\theta}}$. Therefore, the average reward $R(\boldsymbol{\theta})$ increases by updating the policy parameter as follows:

$$\boldsymbol{\theta} := \boldsymbol{\theta} + \alpha \nabla_{\boldsymbol{\theta}} R(\boldsymbol{\theta}),$$

where $:=$ denotes the right-to-left substitution and α is a sufficiently small learning rate. The above framework is called the PGRL method [5].

2.2 Natural Gradient [1]

Natural gradient learning is a gradient method on a Riemannian space. The parameter space being a Riemannian space implies that the parameter $\boldsymbol{\theta} \in \mathcal{R}^d$ is on the Riemannian manifold defined by the Riemannian metric matrix $\boldsymbol{G}(\boldsymbol{\theta}) \in \mathcal{R}^{d\times d}$ (positive definite matrix) and the squared length of a small incremental vector $\Delta\boldsymbol{\theta}$ connecting $\boldsymbol{\theta}$ to $\boldsymbol{\theta} + \Delta\boldsymbol{\theta}$ is given by

$$\|\Delta\boldsymbol{\theta}\|^2_{\boldsymbol{G}} = \Delta\boldsymbol{\theta}^\top \boldsymbol{G}(\boldsymbol{\theta}) \Delta\boldsymbol{\theta}.$$

Under the constraint $\|\Delta\boldsymbol{\theta}\|^2_{\boldsymbol{G}} = \varepsilon^2$ for a sufficiently small constant ε, the steepest ascent direction of a function $R(\boldsymbol{\theta})$ is given by

$$\widetilde{\nabla}_{\boldsymbol{G},\boldsymbol{\theta}} R(\boldsymbol{\theta}) = \boldsymbol{G}(\boldsymbol{\theta})^{-1} \nabla_{\boldsymbol{\theta}} R(\boldsymbol{\theta}). \tag{6}$$

It is called the natural gradient of $R(\boldsymbol{\theta})$ in the Riemannian space $\boldsymbol{G}(\boldsymbol{\theta})$. In RL, the parameter $\boldsymbol{\theta}$ is the policy parameter, the function $R(\boldsymbol{\theta})$ is the average reward, and the gradient $\widetilde{\nabla}_{\boldsymbol{G},\boldsymbol{\theta}} R(\boldsymbol{\theta})$ is called the natural policy gradient (NPG) [2]. Accordingly, in order to (locally) maximize $R(\boldsymbol{\theta})$, $\boldsymbol{\theta}$ is incrementally updated by

$$\boldsymbol{\theta} := \boldsymbol{\theta} + \alpha \widetilde{\nabla}_{\boldsymbol{G},\boldsymbol{\theta}} R(\boldsymbol{\theta}). \tag{7}$$

When we consider a statistical model of a variable x parameterized by $\boldsymbol{\theta}$, $\Pr(x|\boldsymbol{\theta})$, the Fisher information matrix (FIM) $\boldsymbol{F}_x(\boldsymbol{\theta})$ is often used as the Riemannian metric matrix: [12]

$$\begin{aligned}\boldsymbol{F}_x(\boldsymbol{\theta}) &\equiv \sum_{x\in\mathcal{X}} \Pr(x|\boldsymbol{\theta})\, \nabla_{\boldsymbol{\theta}} \ln \Pr(x|\boldsymbol{\theta})\, \nabla_{\boldsymbol{\theta}} \ln \Pr(x|\boldsymbol{\theta})^\top \\ &= -\sum_{x\in\mathcal{X}} \Pr(x|\boldsymbol{\theta})\, \nabla^2_{\boldsymbol{\theta}} \ln \Pr(x|\boldsymbol{\theta}), \end{aligned} \tag{8}$$

where $\mathcal{X}$ is a set of possible values taken by x. $\nabla^2_{\boldsymbol{\theta}} a_{\boldsymbol{\theta}}$ denotes $\nabla_{\boldsymbol{\theta}}(\nabla_{\boldsymbol{\theta}} a_{\boldsymbol{\theta}})$. The reason for using $\boldsymbol{F}(\boldsymbol{\theta})$ as $\boldsymbol{G}(\boldsymbol{\theta})$ comes from the fact that $\boldsymbol{F}(\boldsymbol{\theta})$ is a unique metric matrix of the second-order Taylor expansion of Kullback-Leibler (KL) divergence[2], which is known as a (pseudo) distance between two probability distributions. That is, the KL divergence of $\Pr(x|\boldsymbol{\theta}+\Delta\boldsymbol{\theta})$ from $\Pr(x|\boldsymbol{\theta})$ is represented by

$$D_{\mathrm{KL}}\{\Pr(x|\boldsymbol{\theta})|\Pr(x|\boldsymbol{\theta}+\Delta\boldsymbol{\theta})\} = \frac{1}{2}\Delta\boldsymbol{\theta}^\top \boldsymbol{F}_x(\boldsymbol{\theta})\,\Delta\boldsymbol{\theta} + O(\|\Delta\boldsymbol{\theta}\|^3),$$

where $\|\boldsymbol{a}\|$ denotes the Euclidean norm of a vector $\boldsymbol{a}$.

2.3 Controversy of Natural Policy Gradients

PGRL is regarded as an optimizing process of the policy parameter $\boldsymbol{\theta}$ on some statistical models relevant to both a stochastic policy $\pi(a|s;\boldsymbol{\theta})$ and a state transition probability $p(s'|s,a)$. If a Riemannian metric matrix $\boldsymbol{G}(\boldsymbol{\theta})$ can be designed on the basis of the FIM of an apposite statistical model, $\boldsymbol{F}^*(\boldsymbol{\theta})$, an efficient NPG $\widetilde{\nabla}_{\boldsymbol{F}^*,\boldsymbol{\theta}} R(\boldsymbol{\theta})$ is instantly derived by eq.6.

As Kakade [2] pointed out, the choice of the Riemannian metric matrix $\boldsymbol{G}(\boldsymbol{\theta})$ for PGRL is not unique and the question what metric is apposite to $\boldsymbol{G}(\boldsymbol{\theta})$ is still open. Nevertheless, all previous studies on NPG [13,14,8,7,9] did not seriously address the above problem and (naively) used the Riemannian metric matrix proposed by Kakade [2]. In the next section, we will discuss the statistical models and metric spaces for PGRL and propose a new Riemannian metric matrix.

3 Riemannian Metric Matrices for PGRL

In section 3.1, we propose a new Riemannian metric matrix for RL and derive its NPG named the NSG. In sections 3.2 and 3.3, we discuss the validity of this Riemannian metric by comparing it with the Riemannian metric proposed by Kakade [2] and the Hessian matrix of the average reward.

3.1 A Proposed Riemannian Metric Matrix and NPG Based on State-Action Probability

Since the only adjustable function in PGRL is the policy function $\pi(a|s;\boldsymbol{\theta})$, previous studies on NPG focused on the policy function $\pi(a|s;\boldsymbol{\theta})$, i.e., the statistical models $\Pr(a|s,\mathrm{M}(\boldsymbol{\theta}))$. However, the perturbations in the policy parameter $\boldsymbol{\theta}$ also give rise to the change in the probability of the state $\Pr(s|\mathrm{M}(\boldsymbol{\theta}))$. Because the average reward $R(\boldsymbol{\theta})$ as the objective function of PGRL is specified by the joint probability distribution of the state and the action $(s,a) \in \mathcal{S}\times\mathcal{A}$ (eq.4), it is natural and adequate to focus on the statistical model $\Pr(s,a|\mathrm{M}(\boldsymbol{\theta}))$. For this case, the FIM of $\Pr(s,a|\mathrm{M}(\boldsymbol{\theta}))$ can be used as the Riemannian metric $\boldsymbol{G}(\boldsymbol{\theta})$. Then, its NPG consists with the direction maximizing the average reward under

[2] It is same in the case of all f-divergences in general, except for scale [12].

the constraint that a measure of changes in the KL divergence of the stationary state-action distribution with respect to $\boldsymbol{\theta}$ is fixed by a sufficient small constant ε: $D_{\mathrm{KL}}\{\Pr(s,a|\mathrm{M}(\boldsymbol{\theta}))|\Pr(s,a|\mathrm{M}(\boldsymbol{\theta}+\Delta\boldsymbol{\theta}))\} = \varepsilon^2$. The FIM of this statistical model, $\boldsymbol{F}_{s,a}(\boldsymbol{\theta})$, is calculated with $\Pr(s,a|\mathrm{M}(\boldsymbol{\theta})) = d_{\boldsymbol{\theta}}(s)\pi(a|s;\boldsymbol{\theta})$ and eq.8 to be

$$\begin{aligned}\boldsymbol{F}_{s,a}(\boldsymbol{\theta}) &= \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \Pr(s,a|\mathrm{M}(\boldsymbol{\theta}))\nabla_{\boldsymbol{\theta}}\ln\Pr(s,a|\mathrm{M}(\boldsymbol{\theta}))\nabla_{\boldsymbol{\theta}}\ln\Pr(s,a|\mathrm{M}(\boldsymbol{\theta}))^{\top} \\ &= -\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d_{\boldsymbol{\theta}}(s)\pi(a|s;\boldsymbol{\theta})\nabla^2_{\boldsymbol{\theta}}\ln\left(d_{\boldsymbol{\theta}}(s)\pi(a|s;\boldsymbol{\theta})\right) \\ &= \boldsymbol{F}_s(\boldsymbol{\theta}) + \sum_{s\in\mathcal{S}} d_{\boldsymbol{\theta}}(s)\,\boldsymbol{F}_a(s,\boldsymbol{\theta}), \end{aligned} \tag{9}$$

where

$$\boldsymbol{F}_s(\boldsymbol{\theta}) = \sum_{s\in\mathcal{S}} d_{\boldsymbol{\theta}}(s)\nabla_{\boldsymbol{\theta}}\ln d_{\boldsymbol{\theta}}(s)\nabla_{\boldsymbol{\theta}}\ln d_{\boldsymbol{\theta}}(s)^{\top} \tag{10}$$

is the FIM defined from the statistical model comprising the state distribution, $\Pr(s|\mathrm{M}(\boldsymbol{\theta})) = d_{\boldsymbol{\theta}}(s)$, and

$$\boldsymbol{F}_a(s,\boldsymbol{\theta}) = \sum_{a\in\mathcal{A}} \pi(a|s;\boldsymbol{\theta})\nabla_{\boldsymbol{\theta}}\ln\pi(a|s;\boldsymbol{\theta})\nabla_{\boldsymbol{\theta}}\ln\pi(a|s;\boldsymbol{\theta})^{\top} \tag{11}$$

is the FIM of the policy comprising the action distribution given the state s, $\Pr(a|s,\mathrm{M}(\boldsymbol{\theta})) = \pi(a|s;\boldsymbol{\theta})$. Hence, the new NPG on the FIM of the stationary state-action distribution is

$$\widetilde{\nabla}_{\boldsymbol{F}_{s,a},\boldsymbol{\theta}} R(\boldsymbol{\theta}) = \boldsymbol{F}_{s,a}(\boldsymbol{\theta})^{-1}\,\nabla_{\boldsymbol{\theta}} R(\boldsymbol{\theta}).$$

We term it the "natural state-action gradient"(NSG).

3.2 Comparison with Kakade's Riemannian Metric Matrix

The only Riemannian metric matrix for RL that has been proposed so far is the following matrix, which was proposed by Kakade [2] and was the weighted sum of the FIMs of the policy by the stationary state distribution $d_{\boldsymbol{\theta}}(s)$,

$$\overline{\boldsymbol{F}}_a(\boldsymbol{\theta}) \equiv \sum_{s\in\mathcal{S}} d_{\boldsymbol{\theta}}(s)\,\boldsymbol{F}_a(s,\boldsymbol{\theta}). \tag{12}$$

This is equal to the second term in eq.9. If it is assumed that the stationary state distribution is not changed by a variation in the policy, i.e., if $\nabla_{\boldsymbol{\theta}} d_{\boldsymbol{\theta}}(s) = \mathbf{0}$ holds, then $\boldsymbol{F}_s(\boldsymbol{\theta}) = \mathbf{0}$ holds according to eq.10. While this assumption is not true in general, Kakade's metric $\overline{\boldsymbol{F}}_a(\boldsymbol{\theta})$ is equivalent to $\boldsymbol{F}_{s,a}(\boldsymbol{\theta})$ if it holds. These facts indicate that $\overline{\boldsymbol{F}}_a(\boldsymbol{\theta})$ is the Riemannian metric matrix ignoring the change in the stationary state distribution $d_{\boldsymbol{\theta}}(s)$ caused by the perturbation in the policy parameter $\boldsymbol{\theta}$ in terms of the statistical model of the stationary state-action distribution $\Pr(s,a|\mathrm{M}(\boldsymbol{\theta}))$.

Meanwhile, Bagnell et al. [13] and Peters et al. [14] independently, showed the relationship between the Kakade's metric and the system trajectories $\xi_T = (s_0, a_0, s_1, ..., a_{T-1}, s_T) \in \Xi_T$. When the FIM of the statistical model for the system trajectory ξ_T,

$$\Pr(\xi_T|\mathrm{M}(\boldsymbol{\theta})) = \Pr(s_0) \prod_{t=0}^{T-1} \pi(a_t|s_t;\boldsymbol{\theta}) p(s_{t+1}|s_t, a_t),$$

is normalized by the time steps T with the limit $T \to \infty$, it is equivalent to the Kakade's Riemannian metric,

$$\begin{aligned}
\lim_{T\to\infty} \frac{1}{T} \boldsymbol{F}_{\xi_T}(\boldsymbol{\theta}) &= -\lim_{T\to\infty} \frac{1}{T} \sum_{\xi_T \in \Xi_T} \Pr(\xi_T|\mathrm{M}(\boldsymbol{\theta})) \nabla^2_{\boldsymbol{\theta}} \left\{ \sum_{t=0}^{T-1} \ln \pi(a_t|s_t;\boldsymbol{\theta}) \right\} \\
&= -\sum_{s\in\mathcal{S}} d_{\boldsymbol{\theta}}(s) \sum_{a\in\mathcal{A}} \pi(a|s;\boldsymbol{\theta}) \nabla^2_{\boldsymbol{\theta}} \ln \pi(a|s;\boldsymbol{\theta}) \\
&= \overline{\boldsymbol{F}}_a(\boldsymbol{\theta})
\end{aligned}$$

Since the PGRL objective, i.e., the maximization of the average reward, is reduced to the optimization of the system trajectory by eq.3 [13,14] suggested that the Kakade's metric $\overline{\boldsymbol{F}}_a(\boldsymbol{\theta})$ could be a good metric. However, being equal to $\overline{\boldsymbol{F}}_a(\boldsymbol{\theta})$, the normalized FIM for the infinite-horizon system trajectory obviously differs with $\boldsymbol{F}_{s,a}(\boldsymbol{\theta})$ and is the metric that ignores the information $\boldsymbol{F}_s(\boldsymbol{\theta})$ about the stationary state distribution $\Pr(s|\mathrm{M}(\boldsymbol{\theta}))$. This is due to the fact that the statistical model of the system trajectory considers not only the state-action joint distribution but also the progress for the (infinite) time steps, as follows.

Here, s_{+t} and a_{+t} are the state and the action, respectively, progressed in t time steps after converging to the stationary distribution. Since the distribution of the system trajectory for T time steps from the stationary distribution, $\xi_{+T} \equiv (s, a_{+0}, s_{+1}, ..., a_{+T-1}, s_{+T}) \in \Xi_T$, is

$$\Pr(\xi_{+T}|\mathrm{M}(\boldsymbol{\theta})) = d_{\boldsymbol{\theta}}(s) \prod_{t=0}^{T-1} \pi(a_{+t}|s_{+t};\boldsymbol{\theta}) p(s_{+t+1}|s_{+t}, a_{+t}),$$

its FIM is given by

$$\boldsymbol{F}_{\boldsymbol{\xi}_{+T}}(\boldsymbol{\theta}) = \boldsymbol{F}_s(\boldsymbol{\theta}) + T\overline{\boldsymbol{F}}_a(\boldsymbol{\theta}). \tag{13}$$

The derivation of which is shown in appendix A. Because of $\lim_{T\to\infty} \boldsymbol{F}_{\xi_{+T}}/T = \overline{\boldsymbol{F}}_a(\boldsymbol{\theta})$, the Kakade's metric $\overline{\boldsymbol{F}}_a(\boldsymbol{\theta})$ is regarded as the limit $T \to \infty$ of the system trajectory distribution for T time steps from the stationary state distribution. Consequently, $\overline{\boldsymbol{F}}_a(\boldsymbol{\theta})$ omits the FIM of the state distribution, $\boldsymbol{F}_s(\boldsymbol{\theta})$. On the other hand, the FIM of the system trajectory distribution for one time step is obviously equivalent to the FIM of the state-action joint distribution, i.e., $\boldsymbol{F}_{\xi_{+1}}(\boldsymbol{\theta}) = \boldsymbol{F}_{s,a}(\boldsymbol{\theta})$.

Now, we discuss which FIM is adequate for the average reward maximization. As discussed in section 3.1, the average reward in eq.4 is the expectation of $\bar{r}(s, a)$

over the distribution of the state-action (i.e. the +1-time-step system trajectory) and does not depend on the system trajectories after +2 time steps. It indicates that the Kakade's metric $\overline{\boldsymbol{F}}_a(\boldsymbol{\theta})$ supposed a redundant statistical model and the proposed metric for state-action distribution, $\boldsymbol{F}_{s,a}(\boldsymbol{\theta})$, would be more natural and adequate for PGRL. We give comparisons among various metrics such as $\boldsymbol{F}_{s,a}(\boldsymbol{\theta})$, $\overline{\boldsymbol{F}}_a(\boldsymbol{\theta})$, and a unit matrix $\boldsymbol{I}$ through the numerical experiments in section 5.

Similarly, when the reward function is extended a function of T time steps, $r(s_t, a_t, ..., a_{t+T-1}, s_{t+T})$, instead of one time step, $r(s_t, a_t, s_{t+1})$, the FIM of the T-time-step system trajectory distribution, $\boldsymbol{F}_{\xi_{+T}}(\boldsymbol{\theta})$, would be a natural metric because the average reward becomes $R(\boldsymbol{\theta}) = \sum_{\xi_{+T} \in \Xi_T} \Pr(\xi_{+T}|\mathrm{M}(\boldsymbol{\theta}))r(\xi_{+T})$.

3.3 Analogy with Hessian Matrix

We discuss the analogies between the Fisher information matrices $\boldsymbol{F}_{s,a}(\boldsymbol{\theta})$ and $\overline{\boldsymbol{F}}_a(\boldsymbol{\theta})$ and the Hessian matrix $\boldsymbol{H}(\boldsymbol{\theta})$, which is the second derivative of the average reward with respect to the policy parameter $\boldsymbol{\theta}$,

$$\begin{aligned}
\boldsymbol{H}(\boldsymbol{\theta}) &\equiv \nabla^2_{\boldsymbol{\theta}} R(\boldsymbol{\theta}) \\
&= \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \bar{r}(s,a) d_{\boldsymbol{\theta}}(s) \pi(a|s;\boldsymbol{\theta}) \\
&\quad \Big\{ \nabla^2_{\boldsymbol{\theta}} \ln\big(d_{\boldsymbol{\theta}}(s)\pi(a|s;\boldsymbol{\theta})\big) + \nabla_{\boldsymbol{\theta}} \ln\big(d_{\boldsymbol{\theta}}(s)\pi(a|s;\boldsymbol{\theta})\big) \nabla_{\boldsymbol{\theta}} \ln\big(d_{\boldsymbol{\theta}}(s)\pi(a|s;\boldsymbol{\theta})\big)^\top \Big\} \qquad (14) \\
&= \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \bar{r}(s,a) d_{\boldsymbol{\theta}}(s) \pi(a|s;\boldsymbol{\theta}) \\
&\quad \Big\{ \nabla^2_{\boldsymbol{\theta}} \ln \pi(a|s;\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}} \ln \pi(a|s;\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} \ln \pi(a|s;\boldsymbol{\theta})^\top + \nabla^2_{\boldsymbol{\theta}} \ln d_{\boldsymbol{\theta}}(s) \\
&\qquad + \nabla_{\boldsymbol{\theta}} \ln d_{\boldsymbol{\theta}}(s) \nabla_{\boldsymbol{\theta}} \ln d_{\boldsymbol{\theta}}(s)^\top + \nabla_{\boldsymbol{\theta}} \ln d_{\boldsymbol{\theta}}(s) \nabla_{\boldsymbol{\theta}} \ln \pi(a|s;\boldsymbol{\theta})^\top \\
&\qquad + \nabla_{\boldsymbol{\theta}} \ln \pi(a|s;\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} \ln d_{\boldsymbol{\theta}}(s)^\top \Big\}. \qquad (15)
\end{aligned}$$

Comparing eq.12 of the Kakade's metric matrix $\overline{\boldsymbol{F}}_a(\boldsymbol{\theta})$ with eq.15 of the Hessian matrix $\boldsymbol{H}(\boldsymbol{\theta})$, the Kakade's metric does not have any information about the last two terms in braces $\{\cdot\}$ of eq.15, as Kakade [2] pointed out[3]. This is because $\overline{\boldsymbol{F}}_a(\boldsymbol{\theta})$ is derived under $\nabla_{\boldsymbol{\theta}} d_{\boldsymbol{\theta}}(s) = \boldsymbol{0}$. By eq.9 and eq.14, meanwhile, the proposed metric $\boldsymbol{F}_{s,a}(\boldsymbol{\theta})$ obviously has some information about all the terms of $\boldsymbol{H}(\boldsymbol{\theta})$. This comparison with the Hessian matrix suggests that $\boldsymbol{F}_{s,a}(\boldsymbol{\theta})$ should be an appropriate metric for PGRL. Additionally, $\boldsymbol{F}_{s,a}(\boldsymbol{\theta})$ becomes equivalent to the Hessian matrix in the cases using an atypical reward function that depends on $\boldsymbol{\theta}$ (see Appendix B).

It is noted that the average reward would not be a quadratic form with respect to the policy parameter $\boldsymbol{\theta}$ in general. Especially when $\boldsymbol{\theta}$ is far from the optimal parameter $\boldsymbol{\theta}^*$, the Hessian matrix $\boldsymbol{H}(\boldsymbol{\theta})$ occasionally gets into an indefinite matrix. Meanwhile, FIM $\boldsymbol{F}(\boldsymbol{\theta})$ is always positive (semi-)definite, assured

[3] Strictly speaking, $\boldsymbol{H}(\boldsymbol{\theta})$ is sligthtly different from the Hessian matrix used in [2]. However, the essence of argument is the same as in [2].

by its definition in eq.8. Accordingly, the natural gradient method using FIM might be a more versatile covariant gradient ascent for PGRL than the Newton-Raphson method [15], in which the gradient direction is given by $\widetilde{\nabla}_{\boldsymbol{H},\boldsymbol{\theta}} R(\boldsymbol{\theta})$. Comparison experiments are presented in section 5.

4 Computation of Natural State-Action Gradient

In this section, we view the estimation of the NSG. It will be shown that this estimation can be reduced to the regression problem of the immediate rewards.

Consider the following linear regression model

$$f_\theta(s,a;\boldsymbol{\omega}) \equiv \boldsymbol{\phi}_\theta(s,a)^\top \boldsymbol{\omega}, \tag{16}$$

where $\boldsymbol{\omega}$ is the adjustable parameter and $\boldsymbol{\phi}_\theta(s,a)$ is the basis function of the state and action, also depending on the policy parameter $\boldsymbol{\theta}$,

$$\begin{aligned}\boldsymbol{\phi}_\theta(s,a) &\equiv \nabla_{\boldsymbol{\theta}} \ln \left(d_{\boldsymbol{\theta}}(s)\pi(a|s;\boldsymbol{\theta})\right) \\ &= \nabla_{\boldsymbol{\theta}} \ln d_{\boldsymbol{\theta}}(s) + \nabla_{\boldsymbol{\theta}} \ln \pi(a|s;\boldsymbol{\theta}).\end{aligned} \tag{17}$$

Then, the following theorem holds:

Theorem 1. *Let the Markov chain* $\mathrm{M}(\boldsymbol{\theta})$ *have the fixed policy parameter* $\boldsymbol{\theta}$*, if the objective is to minimize the mean square error* $\epsilon(\boldsymbol{\omega})$ *of the linear regression model* $f_\theta(s_t,a_t;\boldsymbol{\omega})$ *in eq.16 for the rewards* r_{t+1}*,*

$$\epsilon(\boldsymbol{\omega}) = \lim_{T\to\infty} \frac{1}{2T} \sum_{t=0}^{T-1} \left\{r_{t+1} - f_\theta(s_t,a_t;\boldsymbol{\omega})\right\}^2, \tag{18}$$

then the optimal adjustable parameter $\boldsymbol{\omega}^*$ *is equal to NSG as the natural policy gradient on* $\boldsymbol{F}_{s,a}(\boldsymbol{\theta})$*:*

$$\widetilde{\nabla}_{\boldsymbol{F}_{s,a},\boldsymbol{\theta}} R(\boldsymbol{\theta}) = \boldsymbol{\omega}^*.$$

Proof: By the ergodic property of $\mathrm{M}(\boldsymbol{\theta})$, eq.18 is written as

$$\epsilon(\boldsymbol{\omega}) = \frac{1}{2} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} d_{\boldsymbol{\theta}}(s)\pi(a|s;\boldsymbol{\theta}) \left(\bar{r}(s,a) - f_\theta(s,a;\boldsymbol{\omega})\right)^2.$$

Since $\boldsymbol{\omega}^*$ satisfies $\nabla_{\boldsymbol{\omega}}\epsilon(\boldsymbol{\omega})|_{\boldsymbol{\omega}=\boldsymbol{\omega}^*} = \mathbf{0}$, we have

$$\sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} d_{\boldsymbol{\theta}}(s)\pi(a|s;\boldsymbol{\theta})\boldsymbol{\phi}_\theta(s,a)\boldsymbol{\phi}_\theta(s,a)^\top \boldsymbol{\omega}^* = \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} d_{\boldsymbol{\theta}}(s)\pi(a|s;\boldsymbol{\theta})\boldsymbol{\phi}_\theta(s,a)\bar{r}(s,a).$$

By the definition of the basis function (eq.17), the following equations hold,

$$\begin{aligned}\sum_{s,a} d_{\boldsymbol{\theta}}(s)\pi(a|s;\boldsymbol{\theta})\boldsymbol{\phi}_\theta(s,a)\boldsymbol{\phi}_\theta(s,a)^\top &= \boldsymbol{F}_{s,a}(\boldsymbol{\theta}), \\ \sum_{s,a} d_{\boldsymbol{\theta}}(s)\pi(a|s;\boldsymbol{\theta})\boldsymbol{\phi}_\theta(s,a)\bar{r}(s,a) &= \nabla_{\boldsymbol{\theta}} R(\boldsymbol{\theta}).\end{aligned}$$

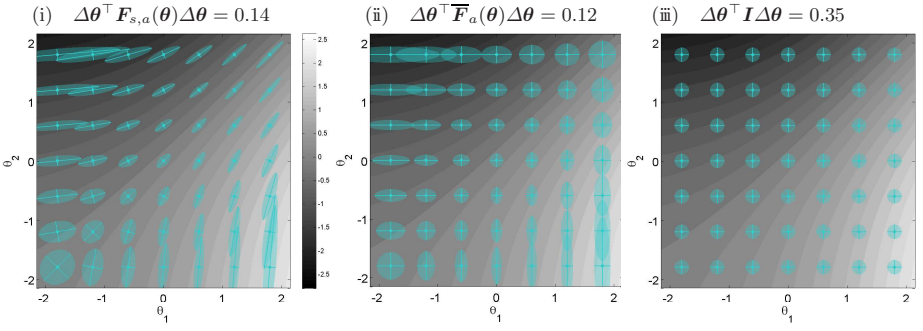

Fig. 1. Phase planes of a policy parameter in a two-state MDP: The gray level denotes $\ln d_{\boldsymbol{\theta}}(1)/d_{\boldsymbol{\theta}}(2)$. Each ellipsoid denotes the fixed distance spaces by each metric $G(\boldsymbol{\theta}) :=$ (i) $\boldsymbol{F}_{s,a}(\boldsymbol{\theta})$, (ii) $\overline{\boldsymbol{F}}_a(\boldsymbol{\theta})$, or (iii) $\boldsymbol{I}$.

Therefore, the following equation holds:

$$\boldsymbol{\omega}^* = \boldsymbol{F}_{s,a}(\boldsymbol{\theta})^{-1}\nabla_{\boldsymbol{\theta}} R(\boldsymbol{\theta}) = \widetilde{\nabla}_{\boldsymbol{F}_{s,a},\boldsymbol{\theta}} R(\boldsymbol{\theta}). \qquad \square$$

It is confirmed by theorem 1 that if the least-square regression to the immediate reward r_{t+1} by the linear function approximator $f_\theta(s_t, a_t; \boldsymbol{\omega})$ with the basis function $\boldsymbol{\phi}_\theta(s,a) \equiv \nabla_{\boldsymbol{\theta}} \ln\left(d_{\boldsymbol{\theta}}(s)\pi(a|s;\boldsymbol{\theta})\right)$ is performed, the adjustable parameter $\boldsymbol{\omega}$ becomes the unbiased estimate of NSG $\widetilde{\nabla}_{\boldsymbol{F}_{s,a},\boldsymbol{\theta}} R(\boldsymbol{\theta})$. Therefore, since the NSG estimation problem is reduced to the regression problem of the reward function, NSG would be simply estimated by the least-square technique or by such a gradient descent technique as the method with the eligibility traces proposed by Morimura et al. [7], where the matrix inversion is not required.

It should be noted that, in order to implement this estimation, the computation of both the derivatives, $\nabla_{\boldsymbol{\theta}} \ln \pi(a|s;\boldsymbol{\theta})$ and $\nabla_{\boldsymbol{\theta}} \ln d_{\boldsymbol{\theta}}(s)$, is required for the basis function $\boldsymbol{\phi}_\theta(s,a)$. While $\nabla_{\boldsymbol{\theta}} \ln \pi(a|s;\boldsymbol{\theta})$ can be instantly calculated, $\nabla_{\boldsymbol{\theta}} \ln d_{\boldsymbol{\theta}}(s)$ cannot be solved analytically because the state transition probabilities are generally unknown in RL. However, an efficient online estimation manner for $\nabla_{\boldsymbol{\theta}} \ln d_{\boldsymbol{\theta}}(s)$, which is similar to the method of estimating the value function, has been established by Morimura et al. [16]. However, we have not discussed the concrete implementations in this paper.

5 Numerical Experiments

5.1 Comparison of Metrics

We first looked into the differences among the Riemannian metric matrices $\boldsymbol{G}(\boldsymbol{\theta})$—the proposed metric $\boldsymbol{F}_{s,a}(\boldsymbol{\theta})$, Kakade's metric $\overline{\boldsymbol{F}}_a(\boldsymbol{\theta})$, and unit matrix $\boldsymbol{I}$—in a simple two-state MDP [2], where each state $s \in \{1,2\}$ has self- and cross-transition actions $\mathcal{A} = \{l, m\}$ and each state transition is deterministic.

The policy with $\boldsymbol{\theta} \in \mathcal{R}^2$ is represented by the sigmoidal function: $\pi(l|s;\boldsymbol{\theta}) = 1/(1+\exp(-\boldsymbol{\theta}^\top \boldsymbol{\psi}(s)))$, where $\boldsymbol{\psi}(1) = [1,0]^\top$ and $\boldsymbol{\psi}(2) = [0,1]^\top$. Figure 1 shows the phase planes of the policy parameter $\boldsymbol{\theta}$. The gray level denotes the log ratio of the stationary state distribution, and each ellipsoid corresponds to the set of $\Delta\boldsymbol{\theta}$ satisfying a constant distance $\Delta\boldsymbol{\theta}^\top \boldsymbol{G}(\boldsymbol{\theta})\Delta\boldsymbol{\theta} = \varepsilon^2$, in which NPG looks for the steepest direction maximizing the average reward. It is confirmed that the ellipsoids by the proposed metric $\boldsymbol{F}_{s,a}(\boldsymbol{\theta})$ coped with the changes in the state distribution by the perturbation in $\boldsymbol{\theta}$ because the alignment of the minor axis of the ellipsoid on $\boldsymbol{F}_{s,a}(\boldsymbol{\theta})$ complied with the direction significantly changing the $d_{\boldsymbol{\theta}}(s)$. This indicates that the policy update with NSG does not drastically change $d_{\boldsymbol{\theta}}(s)$. As we see theoretically. the other metrics could not grasp the changes even though $\overline{\boldsymbol{F}}_a(\boldsymbol{\theta})$ is the expectation of $\boldsymbol{F}_a(\boldsymbol{\theta})$ over $d_{\boldsymbol{\theta}}(s)$.

5.2 Comparison of Learnings

We compared NSG with Kakade's NPG, the ordinary PG, and the (modified) Newton PG learnings in terms of the optimizing performances for $\boldsymbol{\theta}$ through randomly synthesized MDPs with a varying number of states, $|\mathcal{S}| \in \{3, 10, 20, 35, 50, 65, 80, 100\}$. Note that the only difference among these gradients is the definition of the matrix $\boldsymbol{G}(\boldsymbol{\theta})$ in eq.6. The Newton PG uses a modified Hessian matrix $\boldsymbol{H}^\star(\boldsymbol{\theta})$ to assure the negative definiteness:

$$\boldsymbol{H}^\star(\boldsymbol{\theta}) = \boldsymbol{H}(\boldsymbol{\theta}) - \max(0, \lambda_{\max} - \lambda'_{\max})\boldsymbol{I},$$

where $\lambda_{\max}$ and $\lambda'_{\max}$ are the maximum and the largest-negative eigenvalues of $\boldsymbol{H}(\boldsymbol{\theta})$, respectively[4].

It is noted that each gradient was computed analytically because we focussed on the direction of the gradients rather than the sampling issue in this paper.

Experimental Setup. We initialized the $|\mathcal{S}|$-state MDP in each episode as follows. The set of the actions was always $|\mathcal{A}| = \{l, m\}$. The state transition probability function was set by using the Dirichlet distribution $\mathtt{Dir}(\boldsymbol{\alpha} \in \mathcal{R}^2)$ and the uniform distribution $\mathtt{U}(|\mathcal{S}|; b)$ generating an integer from 1 to $|\mathcal{S}|$ other than b: we first initialized it such that $p(s'|s,a) := 0$, $^\forall(s', s, a)$ and then, with $\boldsymbol{q}(s,a) \sim \mathtt{Dir}(\boldsymbol{\alpha}{=}[.3, .3])$ and $x_{\setminus b} \sim \mathtt{U}(|\mathcal{S}|; b)$,

$$\begin{cases} p(s+1|s,l) := q_1(s,l) \\ p(x_{\setminus s+1}|s,l) := q_2(s,l) \end{cases} \qquad \begin{cases} p(s|s,m) := q_1(s,m) \\ p(x_{\setminus s}|s,m) := q_2(s,m) \end{cases}$$

where $s'{=}1$ and $s'{=}|\mathcal{S}|{+}1$ are the identical states. The reward function $r(s,a,s')$ was temporarily set for each argument by Gaussian distribution $\mathrm{N}(\mu{=}0, \sigma^2{=}1)$ and was normalized such that $\max_{\boldsymbol{\theta}} R(\boldsymbol{\theta}) = 1$ and $\min_{\boldsymbol{\theta}} R(\boldsymbol{\theta}) = 0$;

$$r(s,a,s') := \frac{r(s,a,s') - \min_{\boldsymbol{\theta}} R(\boldsymbol{\theta})}{\max_{\boldsymbol{\theta}} R(\boldsymbol{\theta}) - \min_{\boldsymbol{\theta}} R(\boldsymbol{\theta})}.$$

[4] We examined various Hessian modifications [15]. The modification adopted here worked best in this task.

The policy parameterization was the same as that for previous experiment. Accordingly, in this MDP setting, there is no local optimum except for the global optimum. Each element of $\boldsymbol{\theta}_0 \in \mathcal{R}^{|\mathcal{S}|}$ and $\boldsymbol{\psi}(s) \in \mathcal{R}^{|\mathcal{S}|}$ for any state s was drawn from $\mathrm{N}(0, .5)$ and $\mathrm{N}(0, 1)$, respectively. We set the total episode time step at $T = 300$ and the initial learning rate α_0 in eq.7 for each (N)PG before each episode at the inverse of RMS,

$$\alpha_0 = \sqrt{|\mathcal{S}|}/\left\|\widetilde{\nabla}_{\boldsymbol{G},\boldsymbol{\theta}} R(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right\|.$$

If the learning rate α is decent, $R(\boldsymbol{\theta})$ will always increase by the policy update of eq.7. Hence, when the policy update decreased $R(\boldsymbol{\theta})$, we tuned the learning rate "$\alpha := \alpha/2$" and reattempted the update in the same time step. This tuning was kept until $\Delta R(\theta) \geq 0$. On the other hand, when $\alpha_0 > \alpha$ held true at the following time step, we also tuned "$\alpha := 2\alpha$" to avoid standstills of the learning.

Results and Discussions. Figure 2 shows the learning curves for ten individual episodes in 100-state MDPs and reveals that NSG learning was able to succeed in optimizing the policy parameter uniformly and robustly though, compared with the other gradients, NSG was not infrequently slow in improving of performance at a moment. These are consistent with the results about the application of the natural gradient method to the learning of the multilayer perceptron [17].

Figure 3(A) shows the success rate of the learning by 300 episodes at each number of states. Since the maximum of the average reward was set to 1, we regarded the episodes satisfying $R(\boldsymbol{\theta}_T) \geq 0.95$ as "successful" episodes. This suggests that, in the case of the MDPs with a small number of states, NSG and Kakade's NPG methods could avoid falling into the severe plateau phenomena and robustly optimize the policy parameter $\boldsymbol{\theta}$, compared with the other methods. The reason why Kakade's NPG could work as well as NSG would be that the Riemannian metric used in Kakade's method has partial information about the statistical model $\Pr(s, a|\mathrm{M}(\boldsymbol{\theta}))$. Meanwhile, Kakade's method frequently failed to improve the average reward in the cases of the MDPs with a large number of states. This could be due to the fact that Kakade's metric omits the FIM about the state distribution, $\boldsymbol{F}_s(\boldsymbol{\theta})$ unlike the proposed metric, as discussed theoretically in section 3.2. It is also confirmed that Kakade's NPG was inferior to the modified Newton PG in the cases of many states. This could also be a result of whether the gradient has the information about the derivative of $d_{\boldsymbol{\theta}}(s)$ or not.

Finally, we analyzed how severe was the plateau in which these PG learnings were trapped. As this criterion, we utilized the smoothness of the learning curve (approximate curvature),

$$\Delta^2 R(\boldsymbol{\theta}_t) = \Delta R(\boldsymbol{\theta}_{t+1}) - \Delta R(\boldsymbol{\theta}_t),$$

where $\Delta R(\boldsymbol{\theta}_t) \equiv R(\boldsymbol{\theta}_t) - R(\boldsymbol{\theta}_{t-1})$. The criterion for the plateau measure of the episode was defined by

$$\mathtt{PM} = \sum_{t=1}^{T-1} \|\Delta^2 R(\boldsymbol{\theta}_t)\|.$$

Figure 3(B) represents the average of PM over all episodes for each PG and shows that NSG learning could learn very smoothly. This result indicates that the

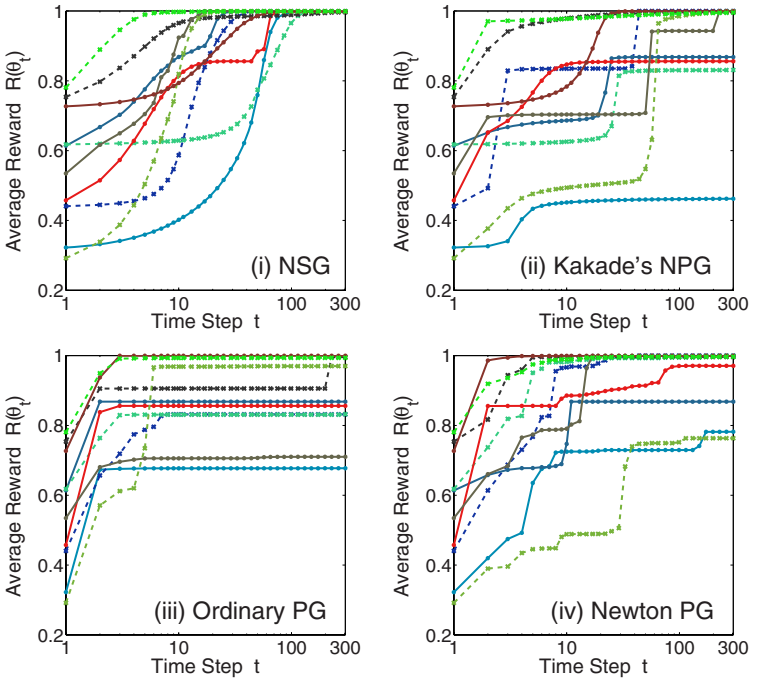

Fig. 2. Time courses of $R(\boldsymbol{\theta})$ for ten individual runs by (i) $\widetilde{\nabla}_{F_{s,a},\boldsymbol{\theta}} R(\boldsymbol{\theta})$, (ii) $\widetilde{\nabla}_{\overline{F}_a,\boldsymbol{\theta}} R(\boldsymbol{\theta})$, (iii) $\widetilde{\nabla}_{I,\boldsymbol{\theta}} R(\boldsymbol{\theta})$, (iv) $\widetilde{\nabla}_{-H^\star,\boldsymbol{\theta}} R(\boldsymbol{\theta})$

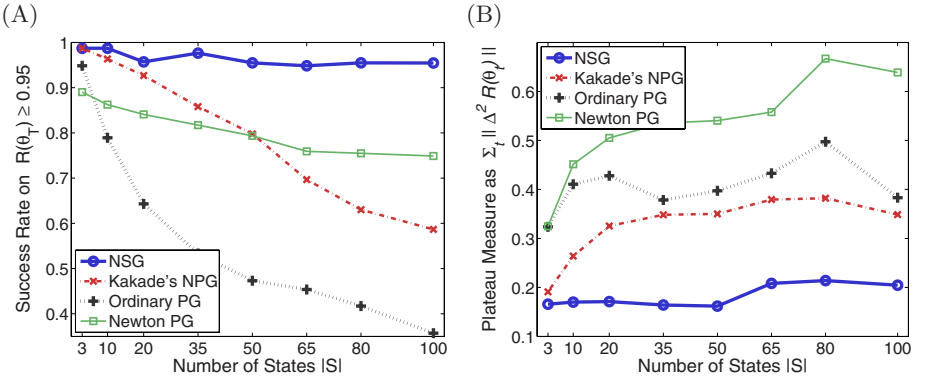

Fig. 3. (A) Learning success rates and (B) plateau measures for each number of states

learning by NSG could most successfully escape from a plateau; this is consistent with all other results.

Since NSG could avoid the plateau and robustly optimize $\boldsymbol{\theta}$ without any serious effect of the setting of the MDP and the initial policy parameter, we conclude that NSG could be a more robust and natural NPG than the NPG by Kakade [2].

6 Summary and Future Work

This paper proposed a new Riemannian metric matrix for the natural gradient of the average reward, which was the Fisher information matrix of the stationary state-action distribution. We clarified that Kakade's NPG [2], which has been widely used in RL, does not consider the changes in the stationary state distribution caused by the perturbation of the policy, while our proposed NSG does. The difference was confirmed in numerical experiments where NSG learning could dramatically improve the performance and rarely fell into the plateau. Additionally, we proved that, when the immediate rewards were fitted by using the linear regression model with the basis function defined on the policy, its adjustable parameter represented the unbiased NSG estimate.

More algorithmic and experimental studies are necessary to further emphasize the effectiveness of NSG. The significant ones would be to establish an efficient Monte-Carlo estimation way of NSG along with estimating the derivative of the stationary state distribution [16], and then to clarify whether or not the proposed NSG method can still be useful even when the gradient is computed from samples. We will investigate them in future work.

References

1. Amari, S.: Natural gradient works efficiently in learning. Neural Computation 10(2), 251–276 (1998)
2. Kakade, S.: A natural policy gradient. In: Advances in Neural Information Processing Systems, vol. 14. MIT Press, Cambridge (2002)
3. Williams, R.J.: Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine Learning 8, 229–256 (1992)
4. Kimura, H., Miyazaki, K., Kobayashi, S.: Reinforcement learning in pomdps with function approximation. In: International Conference on Machine Learning, pp. 152–160 (1997)
5. Baxter, J., Bartlett, P.: Infinite-horizon policy-gradient estimation. Journal of Artificial Intelligence Research 15, 319–350 (2001)
6. Fukumizu, K., Amari, S.: Local minima and plateaus in hierarchical structures of multilayer perceptrons. Neural Networks 13(3), 317–327 (2000)
7. Morimura, T., Uchibe, E., Doya, K.: Utilizing natural gradient in temporal difference reinforcement learning with eligibility traces. In: International Symposium on Information Geometry and its Applications, pp. 256–263 (2005)
8. Peters, J., Vijayakumar, S., Schaal, S.: Natural actor-critic. In: European Conference on Machine Learning (2005)

9. Richter, S., Aberdeen, D., Yu, J.: Natural actor-critic for road traffic optimisation. In: Advances in Neural Information Processing Systems. MIT Press, Cambridge (2007)
10. Bertsekas, D.P.: Dynamic Programming and Optimal Control, vol. 1, 2. Athena Scientific (1995)
11. Sutton, R.S., Barto, A.G.: Reinforcement Learning. MIT Press, Cambridge (1998)
12. Amari, S., Nagaoka, H.: Method of Information Geometry. Oxford University Press, Oxford (2000)
13. Bagnell, D., Schneider, J.: Covariant policy search. In: Proceedings of the International Joint Conference on Artificial Intelligence (July 2003)
14. Peters, J., Vijayakumar, S., Schaal, S.: Reinforcement learning for humanoid robotics. In: IEEE-RAS International Conference on Humanoid Robots (2003)
15. Nocedal, J., Wright, S.J.: Numerical Optimization. Springer, Heidelberg (2006)
16. Morimura, T., Uchibe, E., Yoshimoto, J., Doya, K.: Reinforcement learning with log stationary distribution gradient. Technical report, Nara Institute of Science and Technology (2007)
17. Amari, S., Park, H., Fukumizu, K.: Adaptive method of realizing natural gradient learning for multilayer perceptrons. Neural Computation 12(6), 1399–1409 (2000)

Appendix

A Derivation of eq.13

For simplicity, we denote $\pi_{+t} \equiv \pi(a_{+t}|s_{+t};\boldsymbol{\theta})$ and $p_{+t} \equiv p(s_{+t}|s_{+t-1}, a_{+t-1})$. Since ξ_{+T} is the system trajectory for T time steps from $d_{\boldsymbol{\theta}}(s)$, $\boldsymbol{F}_{\boldsymbol{\xi}_{+T}}(\boldsymbol{\theta})$ is calculated to be

$$\begin{aligned}
\boldsymbol{F}_{\boldsymbol{\xi}_{+T}}(\boldsymbol{\theta}) &= -\sum_{\xi_{+T}\in\Xi_T} \Pr(\xi_{+T})\, \nabla^2_{\boldsymbol{\theta}} \Big\{ \ln d_{\boldsymbol{\theta}}(s) + \sum_{t=0}^{T-1} \ln \pi(a_{+t}|s_{+t};\boldsymbol{\theta}) \Big\} \\
&= -\sum_{s\in\mathcal{S}} d_{\boldsymbol{\theta}}(s) \Big(\nabla^2_{\boldsymbol{\theta}} \ln d_{\boldsymbol{\theta}}(s) + \sum_{a_{+0}\in\mathcal{A}} \pi_{+0} \Big(\nabla^2_{\boldsymbol{\theta}} \ln \pi_{+0} + \\
&\qquad \sum_{s_{+1}\in\mathcal{S}} p_{+1} \sum_{a_{+1}\in\mathcal{A}} \pi_{+1} \Big(\nabla^2_{\boldsymbol{\theta}} \ln \pi_{+1} + \cdots + \\
&\qquad\quad \sum_{s_{+T-1}\in\mathcal{S}} p_{+T-1} \sum_{a_{+T-1}\in\mathcal{A}} \pi_{+T-1}\, \nabla^2_{\boldsymbol{\theta}} \ln \pi_{+T-1} \Big) \cdots \Big) \Big).
\end{aligned}$$

By using the balance equation of the $d_{\boldsymbol{\theta}}(s)$ in eq.1,

$$\begin{aligned}
\boldsymbol{F}_{\boldsymbol{\xi}_{+T}}(\boldsymbol{\theta}) &= \boldsymbol{F}_s(\boldsymbol{\theta}) + \sum_{t=0}^{T-1} \Big(\sum_{s_{+t}\in\mathcal{S}} d_{\boldsymbol{\theta}}(s_{+t}) \boldsymbol{F}_a(\boldsymbol{\theta}|s_{+t}) \Big) \\
&= \boldsymbol{F}_s(\boldsymbol{\theta}) + T\overline{\boldsymbol{F}}_a(\boldsymbol{\theta}).
\end{aligned}$$

□

B Consistency of $\boldsymbol{F}_{\boldsymbol{s},\boldsymbol{a}}(\boldsymbol{\theta})$ and $\boldsymbol{H}(\boldsymbol{\theta})$

If the immediate reward is dependent on $\boldsymbol{\theta}$

$$r(s,a;\boldsymbol{\theta}) = \frac{\Pr(s,a|\mathrm{M}(\boldsymbol{\theta}^*))}{\Pr(s,a|\mathrm{M}(\boldsymbol{\theta}))} \ln \Pr(s,a|\mathrm{M}(\boldsymbol{\theta})), \tag{19}$$

then the average reward becomes the negative cross entropy,

$$R(\boldsymbol{\theta}) = \sum_{s,a} \Pr(s,a|\mathrm{M}(\boldsymbol{\theta}^*)) \ln \Pr(s,a|\mathrm{M}(\boldsymbol{\theta})).$$

Hence, $\Pr(s,a|\mathrm{M}(\boldsymbol{\theta}^*)) = \Pr(s,a|\mathrm{M}(\boldsymbol{\theta}))$ holds, if the average reward is maximized. The Hessian matrix becomes $\boldsymbol{H}(\boldsymbol{\theta}) = \sum_{s,a} \Pr(s,a|\mathrm{M}(\boldsymbol{\theta}^*)) \nabla^2_{\boldsymbol{\theta}} \ln \Pr(s,a|\mathrm{M}(\boldsymbol{\theta}))$. If the policy parameter is nearly optimal $\boldsymbol{\theta} \approx \boldsymbol{\theta}^*$, $\Pr(s,a|\mathrm{M}(\boldsymbol{\theta})) \approx \Pr(s,a|\mathrm{M}(\boldsymbol{\theta}^*))$ holds by the assumption of the smoothness of $\pi(a|s;\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. Therefore, at this time, the Hessian matrix approximately equates the negative, proposed FIM:

$$\begin{aligned} \boldsymbol{H}(\boldsymbol{\theta}) &\approx \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \Pr(s,a|\mathrm{M}(\boldsymbol{\theta})) \nabla^2_{\boldsymbol{\theta}} \ln \Pr(s,a|\mathrm{M}(\boldsymbol{\theta})) \\ &= -\boldsymbol{F}_{s,a}(\boldsymbol{\theta}). \end{aligned}$$

$\boldsymbol{H}(\boldsymbol{\theta}^*) = -\boldsymbol{F}_{s,a}(\boldsymbol{\theta}^*)$ obviously holds. Therefore, when the reward function is in eq.19 and the policy parameter is close to the optimal, NSG almost consists with the Newton direction and the NSG learning attains quadratic convergence.