PAPER Special Section on Foundations of Computer Science

Policy Gradient Based Semi-Markov Decision Problems: Approximation and Estimation Errors

Ngo Anh VIEN[†a)], *Nonmember*, **SeungGwan LEE**[†*b)], *Member*, *and* **TaeChoong CHUNG**[†c)], *Nonmember*

SUMMARY In [1] and [2] we have presented a simulation-based algorithm for optimizing the average reward in a parameterized continuous-time, finite-state semi-Markov Decision Process (SMDP). We approximated the gradient of the average reward. Then, a simulation-based algorithm was proposed to estimate the approximate gradient of the average reward (called GSMDP), using only a single sample path of the underlying Markov chain. GSMDP was proved to converge with probability 1. In this paper, we give bounds on the approximation and estimation errors for GSMDP algorithm. The approximation error of that approximation is the size of the difference between the true gradient and the approximate gradient. The estimation error, the size of the difference between the output of the algorithm and its asymptotic output, arises because the algorithm sees only a finite data sequence.

key words: *Markov decision processes, dynamic programming, semi-Markov decision processes, policy gradient SMDP, approximation and estimation error bounds*

1. Introduction

A generalization of MDP is the SMDP in which the amount of time between one decision and the next is a random variable, either real- or integer-valued [3]–[6]. In the real-valued case, we have SMDPs model with continuous-time discrete-event systems. In a discrete-time SMDPs, decisions can be made only at (positive) integer multiples of an underlying time step. In this section, we restrict ourselves to continuous-time systems with a finite or a countable number of states.

We consider a semi-Markov decision process with finite state space $\mathcal{S} = \{1, \ldots, N\}$ and finite action space $\mathcal{U} = \{1, \ldots, M\}$. Transition probabilities in MDP are replaced by transition distributions $Q_{ij}(\tau, u)$, with a state $i \in \mathcal{S}$, $u \in \mathcal{U}(i)$, and $\tau \geq 0$ is time interval between the transition to state i and the transition to the next state. For a given pair (i, u), transition distributions are used to specify the joint distribution of the transition interval and the next state [4]:

$$Q_{ij}(\tau, u) = P\{t_{k+1} - t_k \leq \tau, x_{k+1} = j | x_k = i, u_k = u\} \quad (1)$$

where the state and control at any time t are denoted by $x(t)$ and $u(t)$ respectively. We will use the following notation for the whole paper:

t_k: The time of occurrence of the kth transition. We denote $t_0 = 0$.

$\tau_k = t_k - t_{k-1}$: The kth transition time interval.

$x_k = x(t_k)$: We have $x(t) = x_k$ for $t_k \leq t < t_{k+1}$.

$u_k = u(t_k)$: We have $u(t) = u_k$ for $t_k \leq t < t_{k+1}$.

We assume that for all states i and j, and actions $u \in \mathcal{U}(i)$, $Q_{ij}(\tau, a)$ is known that the average transition time is finite:

$$\int_0^\infty \tau Q_{ij}(d\tau, u) < \infty$$

Note that the transition distributions specify the ordinary transition probabilities via

$$p_{ij}(u) = P\{x_{k+1} = j | x_k = i, u_k = u\} = \lim_{\tau \to \infty} Q_{ij}(\tau, u)$$

For each pair (i, u), we denote $G(i, u)$ the single stage expected cost corresponding to state i and control u. We have

$$G(i, u) = g(i, u)\bar{\tau}(i, u) \quad (2)$$

where $g(i, u)$ is the immediate reward at each time step, i and u are the current state and action respectively. And $\bar{\tau}(i, u)$ is the expected value of the transition time corresponding to (i, u)

$$\bar{\tau}(i, u) = \sum_{j=1}^{N} \int_0^\infty \tau Q_{ij}(d\tau, u) \quad (3)$$

A randomized policy is defined as a mapping

$$\mu : \mathcal{S} \times \mathcal{U} \mapsto [0, 1]$$

such that:

$$\sum_{u \in \mathcal{U}(i)} \mu(i, u) = 1 \qquad \forall i \in \mathcal{S}$$

Under a policy μ, action u is chosen with probability $\mu(i, u)$ whenever the state is equal to i.

The approach we pursue here is to consider a class of policies parameterized by a parameter space $\{\theta \in \Re^K\}$, whose dimension K is tractable small, compute the gradient with respect to θ of the average reward (cost), and then improve the policy by adjusting the parameters in the gradient direction. Let $\theta = [\theta_1, \theta_2, \ldots, \theta_K]' \in \Re^K$ be the parameter

Manuscript received March 23, 2009.
Manuscript revised June 30, 2009.

[†]The authors are with Artificial Intelligence Laboratory, Department of Computer Engineering, Kyung Hee University, 1 Seocheon, Giheung, Yongin, Gyeonggi, 446–701, South Korea.
[*]The corresponding author.
a) E-mail: vienna@khu.ac.kr
b) E-mail: leesg@khu.ac.kr
c) E-mail: tcchung@khu.ac.kr
DOI: 10.1587/transinf.E93.D.271

that determines the control policy (we use the superscript $'$ to denote vector transposition). Because it is impossible to provide an arbitrary policy μ for problems having very large state spaces. In this paper, we are interested in parameterized policy methods that perform small incremental updates of the parameter θ. Hence, we choose to work with randomized policies that make policy have a smooth dependence on θ.

Let us consider a natural reward function for the continuous-time average reward problem:

$$J(\theta) = \lim_{T\to\infty} \frac{1}{T} E_\theta \left\{ \int_0^T g(x_t, u_t) dt \right\} \tag{4}$$

We parameterize a randomized policy $\mu(\theta)$, which at any given state i chooses an action u with probability $\mu(i, u, \theta)$. We assume that every $\mu(i, u, \theta)$ is nonnegative and that $\sum_{u\in U(i)} \mu(i, u, \theta) = 1$. Thus, the expected reward per stage is given by

$$g(i, \theta) = \sum_{u\in\mathcal{U}(i)} \mu(i, u, \theta) G(i, u) \tag{5}$$

The objective is to maximize the parameterized average reward function $J(\theta)$ under policy $\mu(\theta)$. Now we discuss various assumptions about the SMDP.

1.1 Assumptions

The following assumption assumes the bound of transition time:

Assumption 1: There is $\nu < \infty$ such that

$$0 < \bar{\tau}(i, u) < \nu \qquad i = 1, \ldots, N, \;\; u \in \mathcal{U}(i)$$

For every $\theta \in \Re^K$, let $P(\theta)$ is the stochastic matrix with entries $P_{ij}(\theta)$. We assume the transition probability and the reward matrices satisfying the following assumptions:

Assumption 2: For any $\theta \in \Re^K$, the embedded chain $\{x_k^\theta\}_{k\geq 0}$ with transition probability $P_{ij}(\theta) = P(x_{k+1}^\theta = j | x_k^\theta = i) = \sum_{u\in\mathcal{U}(i)} p_{ij}(u)\mu(i, u, \theta)$ is a unichain.

Assumption 2 ensures $J(\theta)$ is well defined quantity with a limit independent of the initial distribution for x_0. A finite state discrete-time homogeneous Markov chain with state space $\mathcal{S}$ and transition probability P_{ij} is said to be a unichain if the transition probability matrix $[P_{ij}]$ corresponding to every deterministic stationary policy consists of one single recurrent class plus a possibly empty set of transient states [4], [7]. Since any finite-sate Markov chain always ends up in a recurrent class, and it is the properties of this class that determine the long-term average reward, this assumption is mainly for convenience so that we do not have to include the recurrence class as a quantifier in our theorems.

We denote the Markov chain corresponding to $P(\theta)$ by $M(\theta)$. A stationary distribution of a Markov chain with transition probability matrix P is a probability distribution $\pi(\theta) = [\pi(\theta, 1), \ldots, \pi(\theta, N)]'$ over states. Assumption 2 implies that each $P(\theta)$ has a unique stationary distribution $\pi(\theta)$ satisfying the balance equations

$$\pi(\theta)' P(\theta) = \pi(\theta)' \tag{6}$$

where

$$\sum_{i=1}^N \pi(i, \theta) = 1 \tag{7}$$

Then, the average reward is equal to:

$$J(\theta) = \sum_{i=1}^N \pi(i, \theta) g(i, \theta) = \pi' g \tag{8}$$

To make a gradient method applicable, suitable derivatives must exist. The following assumption about parameterized stochastic policies suffices.

Assumption 3: The derivatives $\partial\mu(i, u, \theta)/\partial\theta_k$ exist for all $u \in \mathcal{U}$, $i \in \mathcal{S}$, $k = 1, \ldots, K$ and $\theta \in \Re^K$.

This assumption implies that the derivatives $\partial P_{ij}(\theta)/\partial\theta_k$ exist for all $i, j = 1, \ldots, N$, $k = 1, \ldots, K$ and $\theta \in \Re^K$.

Assumption 4: For every $i \in \mathcal{S}$, $u \in \mathcal{U}$ the magnitudes of the rewards, $|g(i, u)|$, are bounded by a $C < \infty$.

Assumption 5: There is a $B < \infty$ such that, for all $u \in \mathcal{U}$, $i \in \mathcal{S}$, $k = 1, \ldots, K$ and $\theta \in \Re^K$,

$$\frac{|\partial\mu(i, u, \theta)/\partial\theta_k|}{\mu(i, u, \theta)} \leq B$$

In the Assumption 5, we assume that $0/0 = 0$ in the case $\mu(i, u, \theta) = 0$.

2. Policy Gradient SMDP

In [1], [2], we pursued the approach that searches for a policy minimizing the average reward of SMDP problems directly. Our work was inspired by the works in [8] and [9]. In [8], the authors consider a class of stochastic policies parameterized by $\theta \in \Re^K$, and compute the gradient with respect to θ of the average reward, then improve the policy by adjusting the parameters in the gradient direction. We introduced GSMDP algorithm which extended the results in [8] and [9] with a continuous-time model. We can consider the average reward of SMDP problems as a function $J(\theta)$ of $\theta \in \Re^K$. Our main contribution in the papers [1], [2] is to develop an algorithm for computing an approximation $\nabla_\beta J(\theta)$, to the gradient $\nabla J(\theta)$ of the average reward, from a single path of the underlying Markov chain for a SMDP.

Theorem 1: For all $\theta \in \Re^K$,

$$\nabla J(\theta) = \lim_{\beta\to 0} \nabla_\beta J(\theta)$$

where

$$\nabla_\beta J(\theta) = \pi'(\nabla g) + \pi' \nabla P J_\beta(\theta) \tag{9}$$

Proof: See [1] and [2]. Where $J_\beta(\theta) = [J_\beta(1,\theta),\ldots, J_\beta(N,\theta)]'$ is the vector of the expected discounted rewards from each state i:

$$J_\beta(i,\theta) = E_\theta\left\{\int_0^\infty e^{-\beta t} g(x_t,u_t)\,\middle|\, x_0 = i\right\} \tag{10}$$

Then a gradient-based SMDP (GSMDP) algorithm was introduced to estimate the approximation $\nabla_\beta J(\theta)$ as described in Algorithm 1. This algorithm was proved in [1] and [2] to converge to the approximate gradient when the number of iterations $T \to \infty$: $\lim_{T\to\infty} \Delta_T(\theta) = \nabla_\beta J(\theta)$.

The accuracy of the approximation is controlled by a discount factor $\beta \in (0,\infty)$ of the algorithm. The approximation $\nabla_\beta J(\theta)$ has the property that $\nabla J(\theta) = \lim_{\beta\to 0} \nabla_\beta J(\theta)$. However, when β is close to 0, the variance of the algorithm's estimates increase as $\beta \to 0$. Inspired by the proofs in [8] and [9], we will prove the same results that the approximation error is small provided that the $1/(1-e^{-\beta\nu})$ (where ν is the bound of the magnitude of the transition time) is large compared to the mixing time of the derived Markov chain for a SMDP. And inherited by the proof in [9], we will give a bound on the estimation error of the GSMDP algorithm. The estimation error, which is the size of the difference between the output of the algorithm and its asymptotic output, arises because the algorithm sees only a finite data sequence.

Algorithm 1: Gradient-based Semi-Markov Decision Process (GSMDP) algorithm

1. Given :

- Parameter $\theta \in \Re^K$
- Parameterized class of randomized policies $\{\mu(.,.,\theta) : \theta \in \Re^K\}$ satisfy Assumption 3 and 5
- Discount factor $\beta \in (0,\infty)$
- State sequence $x_0, x_1, \ldots$ generated by the SMDP with controls $u_0, u_1, \ldots$ generated randomly according to the distribution $\{\mu(.,.,\theta)\}$.
- Reward sequence $g(x_0,u_0), g(x_1,u_1), \ldots$ satisfies Assumption 4, transition time sequence $\bar\tau(x_0,u_0), \bar\tau(x_1,u_1), \ldots$ satisfies Assumption 1

2: Set $z_0 = 0$ and $\Delta_0 = 0$ ($z_0, \Delta_0 \in \Re^K$)

3: For $t = 0$ to $T-1$ do

$$z_{t+1} = e^{-\beta} z_t + \nabla_t$$
$$\Delta_{t+1} = \Delta_t + \frac{1}{t+1}\left[\nabla_t g(x_t,u_t)\bar\tau(x_t,u_t) + g(x_{t+1},u_{t+1}) z_{t+1} - \Delta_t\right]$$

where

$$\nabla_t = \frac{\nabla\mu(x_t,u_t,\theta)}{\mu(x_t,u_t,\theta)}$$

4: End for

5: Return Δ_T

where we concatenated K gradients ∇_{θ_k} to one vector

$$\Delta = \left[\frac{\partial J(\theta)}{\partial\theta_1},\ldots,\frac{\partial J(\theta)}{\partial\theta_K}\right]'$$

and

$$\nabla_t = \left[\frac{\nabla_{\theta_1}\mu(\cdot,\cdot;\theta)}{\mu(\cdot,\cdot;\theta)},\ldots,\frac{\nabla_{\theta_K}\mu(\cdot,\cdot;\theta)}{\mu(\cdot,\cdot;\theta)}\right]'$$

Our estimation error bound is in terms of $1/(1-e^{-\beta\nu})^2$ and the mixing time of a certain stochastic process associated with the SMDP.

3. Policy Gradient in Partially Observable SMDP

In this section, we discuss an applicability of our proposed policy gradient algorithm (GSMDP) in Algorithm 1 for a Partially Observable SMDP (POSMDP). Specifically, we assume that there are Y observations $\mathcal{Y} = \{1,\ldots,Y\}$, and for each state $i \in \mathcal{S}$, there is a probability distribution $\nu(i)$ over observations in $\mathcal{Y}$. Denote the probability of observation $y \in \mathcal{Y}$ as $\nu(y,i)$. Thus, a randomized policy is defined as a function μ mapping observation $y \in \mathcal{Y}$ into the probability distribution over the controls $\mathcal{U}$. As a consequence, the Markov chain for the corresponding partially observable SMDP is that: state transitions are generated by choosing an observation y in state i according to the distribution $\nu(y,i)$, then choosing a control u according to the distribution $\mu(y,u)$, and then generating a transition to a next state j according to the probability $p_{ij}(u)$, the transition time τ is generated according to a certain distribution. If the policies are parameterized by a parameter $\theta \in \Re^K$, $\mu(y,u,\theta)$, then the Markov chain corresponding to θ has state transition matrix given by

$$p_{ij}(\theta) = E_{y\sim\nu(i)} E_{u\sim\mu(y,\theta)} p_{ij}(u)$$

which means

$$p_{ij}(\theta) = \sum_{u,y} \nu(y,i)\mu(y,u,\theta) p_{ij}(u)$$

then

$$\nabla p_{ij}(\theta) = \sum_{u,y} \nu(y,i)\nabla\mu(y,u,\theta) p_{ij}(u) = \sum_{u,y} \nu(y,i)\mu(y,u,\theta)\frac{\nabla\mu(y,u,\theta)}{\mu(y,u,\theta)} p_{ij}(u)$$

We introduce Algorithm 2 which is the policy gradient based POSMDP. The modification is in which the updates of z_t are now based on the parameterized policy function $\mu(y_t,u_t,\theta)$. The convergence proof of this algorithm can be derived straight-forward similarly to the proof of Algorithm 1 in [1], [2].

4. Approximation Error

The following results about approximation error bound are inspired by the works of Baxter [8], and Bartlett [9].

Theorem 1 showed that $\nabla_\beta J(\theta)$ is an accurate approximation to the gradient as β approaches 0. However, we will prove that β approaching 0 leads to large variance in the estimate of $\nabla_\beta J(\theta)$. The following theorem will give the approximation error bound depending on β. This theorem is defined under the assumption that the eigenvalues of the transition probability matrix $H = E \times P(\theta)$, where $E = diag(e^{-\beta\bar\tau(1)},\ldots,e^{-\beta\bar\tau(N)})$, are all distinct. The existence

Algorithm 2: Gradient-based Partially Observable Semi-Markov Decision Process (GPOSMDP) algorithm

1. Given :

• Parameter $\theta \in \Re^K$

• Parameterized class of randomized policies $\{\mu(.,.,\theta) : \theta \in \Re^K\}$ satisfy Assumption 3 and 5

• Discount factor $\beta \in (0, \infty)$

• Observation sequence $y_0, y_1, \ldots$ generated by the POSMDP with controls $u_0, u_1, \ldots$ generated randomly according to the distribution $\{\mu(y_t, ., \theta)\}$.

• Reward sequence $g(x_0, u_0), g(x_1, u_1), \ldots$ satisfies Assumption 4, transition time sequence $\bar{\tau}(x_0, u_0), \bar{\tau}(x_1, u_1), \ldots$. satisfies Assumption 1. Where $x_0, x_1, \ldots$ is the (hidden) sequence of states of the Markov decision process.

2: Set $z_0 = 0$ and $\Delta_0 = 0$ ($z_0, \Delta_0 \in \Re^K$)

3: For $t = 0$ to $T - 1$ do

$$z_{t+1} = e^{-\beta} z_t + \nabla_t$$

$$\Delta_{t+1} = \Delta_t + \frac{1}{t+1}\left[\nabla_t g(x_t, u_t)\bar{\tau}(x_t, u_t) + g(x_{t+1}, u_{t+1}) z_{t+1} - \Delta_t\right]$$

where

$$\nabla_t = \frac{\nabla \mu(y_t, u_t, \theta)}{\mu(y_t, u_t, \theta)}$$

4: End for

5: Return Δ_T

of a unique stationary distribution implies that the set of eigenvalues $\lambda_1 \geq |\lambda_2| \geq \cdots \geq |\lambda_N|$ have magnitudes less than 1 (some eigenvalues may be equal to each other) [10]. In the following theorem, $\kappa_2(X)$ is the spectral condition number of a nonsingular matrix X which is defined as

$$\kappa_2(X) = \|X\|_2 \left\|X^{-1}\right\|_2$$

where

$$\|X\|_2 = \max_{y:\|y\|=1} \|Xy\|$$

and $\|y\|$ is the Euclidean norm of a vector y.

Theorem 2: Let assumption 1–5 hold, with stationary distribution $\pi' = (\pi_1, \ldots, \pi_N)$ and denotes $\Pi = diag(\pi_1, \ldots, \pi_N)$, and $H = E \times P(\theta)$, where $E = diag(e^{-\beta\bar{\tau}_1}, \ldots, e^{-\beta\bar{\tau}_N})$, has distinct eigenvalues. Then the normalized inner product between $\nabla J(\theta)$ and $\nabla_\beta J(\theta)$ satisfies

$$1 - \frac{\nabla J e^{-\beta\nu} \nabla_\beta J}{\|\nabla J\|^2} \leq \frac{(1 - e^{-\beta\nu})BC\nu}{\|\nabla J\|} + \kappa_2(\Pi^{1/2} S) \frac{\left\|\nabla \sqrt{\pi'}\right\|}{\|\nabla J\|} \sqrt{g'\Pi g} \frac{1 - e^{-\beta\nu}}{1 - |\lambda_1|}$$

where $S = (x_1\ x_2\ \ldots\ x_N)$ is the matrix of the right eigenvectors of H corresponding, respectively, to the eigenvalues $\lambda_1 \geq |\lambda_2| \geq \cdots \geq |\lambda_N|$.

The proof of the Theorem 2 appears in Appendix A.

According to the result of Theorem 2, it turns out that if $|1 - e^{-\beta\nu}|$ is small compared to $|1 - \lambda_1|$, the gradient approximation is accurate, and large β gives large approximation error. By this theorem, we have trade-off method between the approximation error reduction and variance reduction of the algorithm. As seen in the proof of Theorem 1 in [1], if β is larger, the limit converges more quickly as $t \to \infty$ with the order of magnitude βt, that means, Algorithm 1 needs fewer iterations to make the approximate gradient $\nabla_\beta J$ converge to the true gradient $\nabla J(\theta)$. However, Theorem 2 shows that large β gives a larger approximation error. Inversely, small β makes the limit converges more slowly, that means Algorithm 1 needs many more iterations to make the approximate gradient $\nabla_\beta J$ converge to the true gradient $\nabla J(\theta)$. However, small β gives small approximation error.

Theorem 2 is similar to the work of Baxter [8]. It shows that the approximation is accurate, but the proof requires an assumption that the eigenvalues of the Markov chain $M(\theta)$ are all distinct. The following theorem has similar result, but without the assumption about the eigenvalues. The following result is inspired from the result of Bartlett [9]. The result is in terms of a different mixing time τ^*, based on χ^2 distance.

Definition 1: Given two probability distributions p, π on $\{1, 2, \ldots, N\}$, with $\pi_i > 0$ for all i, the χ^2 distance between p and π is given by

$$d_{\chi^2}(p, \pi) = \left(\sum_{i=1}^{N} \frac{(p_i - \pi_i)}{\pi_i}\right)^{1/2}$$

We use the following lemma result from [9]

Lemma 1:

$$\left\|\Pi^{1/2}(P^{-t} - e'\pi)\Pi^{-1/2}\right\| \leq \sqrt{E_{x\sim\pi}\left\{d_{\chi^2}(p_x^t, \pi)^2\right\}}$$

See Lemma 22 in [9] for proof. And definition of p_x^t in the next theorem. Now we can obtain the bound of approximation error.

Theorem 3: Partition the state transition probability matrix P as

$$P^t = \begin{bmatrix} p_1^{t'} \\ \vdots \\ p_N^{t'} \end{bmatrix}$$

Suppose there are constants c, τ^* (the mixing time is defined based on χ^2 distance) that

$$\left(E_{x\sim\pi}\left\{d_{\chi^2}(p_x^t, \pi)^2\right\}\right)^{1/2} \leq c \exp\left(-\frac{t}{\tau^*}\right)$$

Then

$$\left\|\nabla J - e^{-\beta\nu}\nabla_\beta J\right\| \leq (1 - e^{-\beta\nu})BC\nu + \left\|\nabla\pi'\right\| c\tau^* \times (1 - e^{-\beta\nu}) \left\|\Pi^{-1/2}\right\| \left\|\Pi^{1/2} g\right\|$$

The proof of the Theorem 3 appears in Appendix B. The above theorem also justifies our discussion that if β is smaller, the right hand side reaches to 0, thus the approximation error is smaller; otherwise if β is larger, the right hand side becomes larger, thus the approximation error is larger. However, we previously argued that large β leads to small variance, and vice versa. Thus, we can derive the trade-off between bias/variance from either Theorem 2 or 3.

5. Estimation Error

In this section, we give bounds on the estimation error of the GSMDP algorithm. Our estimation error bound is in terms of the algorithm's discount factor and the mixing time of a certain stochastic process associated with the SMDP. We will generalize the work of Bartlett in [9] which finds estimation error bounds of GPOMDP algorithm associated with the Partially Observable MDP (POMDP) framework. In Algorithm 3, we also modify our algorithm similarly to GPOMDP. The new algorithm has three distinct phases, which extends for n_1, n_2, n_3 time steps. "The first phase involves waiting for the controlled SMDP to mix. The second involves gathering gradient information about actions that are taken. The third involves waiting for the long term outcomes of the actions for which the gradient information was gathered" (this modification is described in detail in [9]). Thus, we can rewrite Δ in Algorithm 1, the estimate produced by GSMDP algorithm in Algorithm 3:

$$\Delta = \frac{1}{n_2}\sum_{t=n_1}^{n_1+n_2-1}\nabla_t\Big[g(x_t,u_t)\bar{\tau}(x_t,u_t) + \sum_{s=0}^{n_1+n_2+n_3-1-t} e^{-\beta s} g(x_{t+1+s},u_{t+1+s})\Big] \tag{11}$$

Algorithm 3: GSMDP algorithm.

1. Given :

• Parameter $\theta \in \Re^K$, discount factor $\beta \in (0,\infty)$.

• Parameterized class of randomized policies $\{\mu(.,.,\theta) : \theta \in \Re^K\}$ satisfy Assumption 3 and 5.

• State sequence $x_0, x_1, \ldots$ generated by the SMDP with controls $u_0, u_1, \ldots$ generated randomly according to the distribution $\{\mu(.,.,\theta)\}$.

• Reward sequence $g(x_0,u_0), g(x_1,u_1), \ldots$ satisfies Assumption 4, transition time sequence $\bar{\tau}(x_0,u_0), \bar{\tau}(x_1,u_1), \ldots$ satisfies Assumption 1.

2: Set $z_0 = 0$ and $\Delta_0 = 0$ ($z_0, \Delta_0 \in \Re^K$)

3: For $t = 0$ to $n_1 - 1$ do

$$z_{t+1} = z_t$$
$$\Delta_{t+1} = \Delta_t$$

End for

4: For $t = n_1$ to $n_1 + n_2 - 1$ do

$$z_{t+1} = e^{-\beta} z_t + \nabla_t$$
$$\Delta_{t+1} = \Delta_t + \frac{1}{t-n_1+1}\Big[\nabla_t g(x_t,u_t)\bar{\tau}(x_t,u_t) + g(x_{t+1},u_{t+1})z_{t+1} - \Delta_t\Big]$$

End for

5: For $t = n_1 + n_2$ to $n_1 + n_2 + n_3 - 1$ do

$$z_{t+1} = e^{-\beta} z_t$$
$$\Delta_{t+1} = \Delta_t + \Big[\nabla_t g(x_t,u_t)\bar{\tau}(x_t,u_t) + g(x_{t+1},u_{t+1})z_{t+1}\Big]$$

End for

where

$$\nabla_t = \frac{\nabla\mu(x_t,u_t,\theta)}{\mu(x_t,u_t,\theta)}$$

6: Return Δ_T

where

$$\nabla_t = \frac{\nabla\mu(x_t,u_t,\theta)}{\mu(x_t,u_t,\theta)}$$

We also apply the k-blocked algorithm, which uses only k of the future reward values

$$\Delta^k = \frac{1}{n_2}\sum_{t=n_1}^{n_1+n_2-1}\nabla_t\Big[g(x_t,u_t)\bar{\tau}(x_t,u_t) + \sum_{s=0}^{k-1} e^{-\beta s} g(x_{t+1+s},u_{t+1+s})\Big] \tag{12}$$

Assume that $k \le n_3 + 1$. The estimate Δ^k is an average of n_2 terms, each of which is a function of a vector $S_t^k = (\nabla_t, g(t), \bar{\tau}(t), \ldots, g(t+k), \bar{\tau}(t+k))$. Define that

$$\Delta_t^k = \nabla_t\Big[g(x_t,u_t)\bar{\tau}(x_t,u_t) + \sum_{s=0}^{k-1} e^{-\beta s} g(x_{t+1+s},u_{t+1+s})\Big] \tag{13}$$

Then, use Assumptions 1, 4, 5 to obtain

$$\left\|\Delta_t^k\right\|_\infty \le BC\left(\nu + \frac{1-e^{-\beta k}}{1-e^{-\beta}}\right) \le BC\left(\nu + \frac{1}{1-e^{-\beta}}\right) \tag{14}$$

We now give the estimation error bound of the GSMDP algorithm

Theorem 4: If the process $S_t^k = (\nabla_t, g(t), \bar{\tau}(t), \ldots, g(t+k), \bar{\tau}(t+k))$ is $\tau^* - mixing$, $s \le n_2$, $k \le n_3 + 1$, then

$$Pr\left(\left\|\Delta - \nabla_\beta J\right\|_\infty \ge \epsilon + \frac{2BC}{1-e^{-\beta}}e^{-\beta k}\middle| X_{-\infty}^0\right) \le \frac{K}{2}\left\{se^{-\lfloor n_1/\tau^*\rfloor} + n_2 e^{-\lfloor s/\tau^*\rfloor} + 4se^{\left(\frac{-\epsilon^2 n_2(1-e^{-\beta})^2}{4B^2C^2 s}\right)}\right\}$$

where K is the parameter dimension, $\lfloor a/b\rfloor$ is integer division, and $X_{-\infty}^j$ is the infinite sequence $(\ldots, x_{j-1}, x_j)$.

The proof of the theorem 4 appears in Appendix C. This result shows the relationship between the number of iterations in Algorithm 1 (n_2 or T) with the estimation error. We consider the right side in theorem 4 as a function of n_2. It is easy to derive that the function is an decreasing function if $n_2 > 0$ (by taking its derivative). Thus, if the number of iterations n_2 (or T) of GSMDP is infinite, the probability of the estimation error greater than a small number becomes zero.

6. Conclusion

In this paper, we give bounds on the approximation and estimation errors for our GSMDP algorithm in [1]. GSMDP algorithm is a simulation-based algorithm which was proposed to estimate the approximate gradient of the average reward, using only a single sample path of the underlying Markov chain. The approximate gradient of the average reward, with respect to the parameters in SMDP controlled by

parameterized stochastic policies, is computed as in Algorithm 1.

The approximation error of the above approximation is the size of the difference between the true gradient and the approximate gradient. The estimation error, the size of the difference between the output of the algorithm and its asymptotic output, arises because the algorithm sees only a finite data sequence. Through approximation error bounds, we show that the accuracy of the approximation depends on the relationship between the discount factor used by the approximation method and the mixing-time of the Markov chain for a SMDP. As a consequence, we derive a trade-off method between the approximation error reduction and variance reduction of the algorithm. The estimation error bound shows the relationship between the number of iterations of GSMDP with the estimation error. It is easy to derive that if the number of iterations of GSMDP is infinite, the probability of the estimation error greater than a small number becomes zero.

Acknowledgments

The authors would like to thank the anonymous reviewers. Their critical and detailed comments helped to substantially improve the paper.

References

[1] N.A. Vien and T. Chung, "Policy gradient semi-Markov decision process," 20th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2008), vol.2, pp.11–18, Dayton, Ohio, USA, Nov. 2008.
[2] N.A. Vien, N.H. Viet, S. Lee, and T. Chung, "Policy gradient SMDP for resource allocation and routing in integrated services networks," IEICE Trans. Commun., vol.E92-B, no.6, pp.2008–2022, June 2009.
[3] S.M. Ross, Applied Probability Models with Optimization Applications, Holden-Day, San Francisco, 1970.
[4] D.P. Bertsekas, Dynamic Programming and Optimal Control, Athena Scientific, Belmont, Mass, 2001.
[5] M.L. Puterman, Markov Decision Processes: Discrete Stochastic Dynamic Programming, John Wiley & Sons, New Jersey, 2005.
[6] E.A. Feinberg, "Continuous time discounted jump Markov decision processes: A discrete-event approach," Mathematics of Operations Research, vol.29, no.3, pp.492–524, 2004.
[7] H.C. Tijims, Stochastic Model: An Algorithm Approach, John Wiley & Sons, Chichester, 1994.
[8] J. Baxter and P.L. Bartlett, "Infinite-horizon policy-gradient estimation," J. Artifcial Intelligence Research (JAIR), vol.15, pp.319–350, 2001.
[9] P.L. Bartlett and J. Baxter, "Estimation and approximation bounds for gradient-based reinforcement learning," Thirteenth Annual Conference on Computational Learning Theory (COLT 2000), pp.133–141, 2000.
[10] A. Berman and R.J. Plemmons, Nonnegative Matrices in the Mathematical Sciences (Classics in Applied Mathematics), Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 1994.
[11] P. Lancaster and M. Tismenetsky, The Theory of Matrices, Academic Press, San Diego, CA, 1985.

Appendix A: Proof of Theorem 2

Lemma 2: Assume that $S = (x_1\ x_2\ \dots\ x_N)$ is the matrix of the right eigenvectors of $H = E \times P$ corresponding, where $E = diag(e^{-\beta\bar{\tau}(1)}, \dots, e^{-\beta\bar{\tau}(N)})$, respectively, to the eigenvalues $1 = \lambda_1 > |\lambda_2| \geq \cdots \geq |\lambda_N|$, and $S^{-1} = (y_1, \dots, y_N)'$.

Then y_i is the left eigenvector corresponding to eigenvalue λ_i, $i = 1, \dots, N$.

Proof: From Theorem 4.10.2, p153 in [11], the existence of N distinct eigenvalues implies that $H = S\Lambda S^{-1}$, where $\Lambda = diag(\lambda_1, \dots, \lambda_N)$, $S^{-1} = (y_1, \dots, y_N)'$.

$$
\begin{aligned}
&H = S\Lambda S^{-1} \Leftrightarrow S^{-1}H = \Lambda S^{-1} \\
&\Leftrightarrow (y_1, \dots, y_N)'H = (\lambda_1 y_1, \dots, \lambda_N y_N)' \\
&\Leftrightarrow (y'_1, \dots, y'_N)H = (\lambda_1 y'_1, \dots, \lambda_N y'_N) \\
&\Leftrightarrow \begin{pmatrix} y_{11} & \cdots & y_{1N} \\ \vdots & \ddots & \vdots \\ y_{N1} & \cdots & y_{NN} \end{pmatrix} \begin{pmatrix} H_{11} & \cdots & H_{1N} \\ \vdots & \ddots & \vdots \\ H_{N1} & \cdots & H_{NN} \end{pmatrix} \\
&= \begin{pmatrix} \lambda_1 y_{11} & \cdots & \lambda_1 y_{1N} \\ \vdots & \ddots & \vdots \\ \lambda_N y_{N1} & \cdots & \lambda_N y_{NN} \end{pmatrix} \\
&\Leftrightarrow \sum_{i=1}^{N} y_i H_{ij} = \lambda_j y_j \quad j = 1, \dots, N
\end{aligned}
$$

Then y_i is the left eigenvector corresponding to λ_i, $i = 1, \dots, N$. □

Lemma 3: For any polynomial function f, then $f(P) = S f(\Lambda) S^{-1}$.

Proof: First, we prove that $P^k = S\Lambda^k S^{-1}$, k is integer. We have $S^{-1}P^k = S^{-1}PP^{k-1} = \Lambda S^{-1}P^{k-1} = \cdots = \Lambda^k S^{-1}$. Then $P^k = S\Lambda^k S^{-1}$. Assume that $f = a_0 + a_1 x + \cdots + a_n x_n$, then

$$
\begin{aligned}
f(P) &= \sum_{t=0}^{n} a_t P^t = \sum_{t=0}^{n} a_t S \Lambda^t S^{-1} \\
&= S \left(\sum_{t=0}^{n} a_t \Lambda^t \right) S^{-1} = S f(\Lambda) S^{-1}
\end{aligned}
$$

□

The following proof is for Theorem 2. We have the formula of the true gradient and the approximate gradient of the average reward Eq. (9)

$$\nabla J(\theta) = \pi'(\nabla g) + (\nabla \pi)' g$$

and

$$\nabla_\beta J(\theta) = \pi'(\nabla g) + \pi' \nabla P J_\beta$$

On the other hand, in [3]

$$J_\beta(i) = g(i) + e^{-\beta\bar{\tau}(i)} \sum_{j=0}^{\infty} P_{ij} J_\beta(j)$$

or

$$J_\beta(\theta) = g + EPJ_\beta(\theta) \tag{A·1}$$

where $E = diag(e^{-\beta\bar{\tau}(1)}, \dots, e^{-\beta\bar{\tau}(N)})$

Following from Assumption 1, the magnitude of the transition time satisfies $\bar{\tau}(i) \le \nu \Rightarrow e^{-\beta\bar{\tau}(i)} \ge e^{-\beta\nu}$. Thus, we have

$$\begin{aligned}(\nabla\pi)'g &= (\nabla\pi)'(J_\beta - EPJ_\beta)\\ &\le (\nabla\pi)'J_\beta - e^{-\beta\nu}(\nabla\pi)'PJ_\beta\end{aligned}$$

(by (A· 1)). Hence, from the definition of $\nabla J(\theta)$ and the above inequality, we have

$$\nabla J \le \pi'(\nabla g) + (\nabla\pi)'J_\beta - e^{-\beta\nu}(\nabla\pi)'PJ_\beta$$

According to the balance equation that $(\nabla\pi)'P = \nabla\pi' - \pi'\nabla P$, then

$$\nabla J \le \pi'(\nabla g) + (\nabla\pi)'J_\beta - e^{-\beta\nu}(\nabla\pi' - \pi'\nabla P)J_\beta$$

Subtract both sides of the above inequality by both sides of $\nabla_\beta J(\theta)$ formula multiplied with $e^{-\beta\nu}$ to obtain

$$\nabla J - e^{-\beta\nu}\nabla_\beta J \le (1 - e^{-\beta\nu})\pi'(\nabla g) + (1 - e^{-\beta\nu})(\nabla\pi)'J_\beta \tag{A· 2}$$

First, we will find the bound of the first part of the right-hand side, then the reduced form of the second of the above inequality. We have

$$\begin{aligned}\pi'(\nabla g) &= \sum_i \sum_u \pi(i)\mu(i,u)\frac{\nabla\mu(i,u)}{\mu(i,u)}g(i,u)\bar{\tau}(i,u)\\ &= \sum_i \sum_u E(Y_t)\end{aligned}$$

where $E(Y_t)$ is the expectation of the stationary and ergodic process $\{Y_t\}$ is defined by:

$$Y_t = \chi_i(x_t)\chi_u(u_t)\frac{\nabla\mu(i,u)}{\mu(i,u)}g(i,u)\bar{\tau}(i,u)$$

where $\chi_i(\cdot)$ denotes the indicator function:

$$\chi_i(x_t) = \begin{cases} 1 & \text{if } x_t = i\\ 0 & \text{otherwise}\end{cases}$$

then

$$\begin{aligned}\pi'\nabla g &= \sum_{i,u}\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\chi_i(x_t)\chi_u(u_t)\frac{\partial\mu(i,u)}{\mu(i,u)}g(i,u)\bar{\tau}(i,u)\\ &= \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\frac{\nabla\mu(x_t,u_t)}{\mu(x_t,u_t)}g(x_t,u_t)\bar{\tau}(x_t,u_t)\\ &\le \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}BC\nu = BC\nu\end{aligned} \tag{A· 3}$$

Use results from (A· 1), Lemmas 2 and 3, we obtain

$$\begin{aligned}(1 - e^{-\beta\nu})J_\beta &= (1 - e^{-\beta\nu})\sum_{t=0}^{\infty}(EP)^t g\\ &= (1 - e^{-\beta\nu})S\left(\sum_{t=0}^{\infty}\Lambda^t\right)S^{-1}g\\ &= (1 - e^{-\beta\nu})\sum_{j=1}^{n}x_j y_j'\left(\sum_{t=0}^{\infty}(\lambda_j)^t\right)g\\ &= SMS^{-1}g\end{aligned} \tag{A· 4}$$

where

$$M = diag\left(\frac{1 - e^{-\beta\nu}}{1 - \lambda_1}, \frac{1 - e^{-\beta\nu}}{1 - \lambda_2}, \ldots, \frac{1 - e^{-\beta\nu}}{1 - \lambda_N}\right)$$

Now we prove the main result of this theorem. From the inequality (A· 2), we have

$$\nabla J - (1 - e^{-\beta\nu})\pi'\nabla g - (1 - e^{-\beta\nu})(\nabla\pi)'J_\beta \le e^{-\beta\nu}\nabla_\beta J$$

Thus,

$$\begin{aligned}&1 - \frac{\nabla J e^{-\beta\nu}\nabla_\beta J}{\|\nabla J\|^2}\\ &\le 1 - \frac{\nabla J\left[\nabla J - (1 - e^{-\beta\nu})\pi'\nabla g - (1 - e^{-\beta\nu})(\nabla\pi)'J_\beta\right]}{\|\nabla J\|^2}\\ &= \frac{\nabla J\left[(1 - e^{-\beta\nu})\pi'\nabla g + (1 - e^{-\beta\nu})(\nabla\pi)'J_\beta\right]}{\|\nabla J\|^2}\\ &= \frac{\nabla J\left[(1 - e^{-\beta\nu})\pi'\nabla g + \nabla\pi' SMS^{-1}g\right]}{\|\nabla J\|^2}\end{aligned}$$

(Following to (A· 4))

$$\begin{aligned}&\le \frac{\nabla J\left[(1 - e^{-\beta\nu})BC\nu + \nabla\pi' SMS^{-1}g\right]}{\|\nabla J\|^2}\\ &\le \frac{(1 - e^{-\beta\nu})BC\nu}{\|\nabla J\|} + \frac{\left\|\nabla\pi' SMS^{-1}g\right\|}{\|\nabla J\|}\end{aligned}$$

(Following to (A· 3) and the Cauchy-Schwartz inequality).

From here we apply the result of Baxter [8] to reduce the above inequality of the normalized inner product between $\nabla J(\theta)$ and $\nabla_\beta J(\theta)$. Since $\nabla\pi' = (\nabla\sqrt{\pi'})\Pi^{1/2}$, then apply the Cauchy-Schwartz again to obtain

$$\begin{aligned}&1 - \frac{\nabla J e^{-\beta\nu}\nabla_\beta J}{\|\nabla J\|^2}\\ &\le \frac{(1 - e^{-\beta\nu})BC\nu}{\|\nabla J\|} + \frac{\left\|\nabla\sqrt{\pi'}\right\|\left\|\Pi^{1/2}SMS^{-1}g\right\|}{\|\nabla J\|}\end{aligned}$$

The second part of the right-hand side of the above inequality is

$$\begin{aligned}&\left\|\Pi^{1/2}SMS^{-1}g\right\| = \left\|\Pi^{1/2}SMS^{-1}\Pi^{-1/2}\Pi^{1/2}g\right\|\\ &\le \left\|\Pi^{1/2}S\right\|_2\left\|S^{-1}\Pi^{-1/2}\right\|_2\left\|\Pi^{1/2}g\right\|_2\|M\|_2\\ &\le \kappa_2(\Pi^{1/2}S)\sqrt{g'\Pi g}\frac{1 - e^{-\beta\nu}}{1 - \lambda_1}\end{aligned}$$

Because

$$\begin{aligned}\|M\|_2 &= \left\|diag\left(\frac{1 - e^{-\beta\nu}}{1 - \lambda_1}, \frac{1 - e^{-\beta\nu}}{1 - \lambda_2}, \ldots, \frac{1 - e^{-\beta\nu}}{1 - \lambda_N}\right)\right\|_2\\ &= \max_i\left|\frac{1 - e^{-\beta\nu}}{1 - \lambda_i}\right| = \frac{1 - e^{-\beta\nu}}{1 - \lambda_1}\end{aligned}$$

Then

$$\begin{aligned}1 - \frac{\nabla J e^{-\beta\nu}\nabla_\beta J}{\|\nabla J\|^2} &\le \frac{(1 - e^{-\beta\nu})BC\nu}{\|\nabla J\|}\\ &+ \kappa_2(^{1/2}S)\frac{\left\|\nabla\sqrt{\pi'}\right\|}{\|\nabla J\|}\sqrt{g'\Pi g}\frac{1 - e^{-\beta\nu}}{1 - \lambda_1}\end{aligned}$$

□

Appendix B: Proof of Theorem 3

Proof: The inequality (A· 2) shows that

$$\left\|\nabla J - e^{-\beta\nu}\nabla_\beta J\right\| \le (1-e^{-\beta\nu})\pi'(\nabla g) + (1-e^{-\beta\nu})(\nabla\pi)' J_\beta$$

And applying the inequality (A· 3) to obtain

$$\left\|\nabla J - e^{-\beta\nu}\nabla_\beta J\right\| \le (1-e^{-\beta\nu})BC\nu + (1-e^{-\beta\nu})(\nabla\pi)' J_\beta \quad \text{(A·5)}$$

According to (A· 1), we have

$$(1-e^{-\beta\nu})(\nabla\pi)' J_\beta = (1-e^{-\beta\nu})\nabla\pi' \sum_{t=0}^{\infty} E^t P^t g$$

$$= (1-e^{-\beta\nu})\nabla\pi' \sum_{t=0}^{\infty} E^t (P^t - e\pi')g$$

(because $\nabla\pi' e = \nabla(\pi' e) = 0$)

$$= \nabla\pi'(1-e^{-\beta\nu}) \sum_{t=0}^{\infty} E^t \Pi^{-1/2} \times \left(\Pi^{1/2}(P^t - e\pi')\Pi^{-1/2}\right)\Pi^{1/2} g$$

$$\le \left\|\nabla\pi'\right\|(1-e^{-\beta\nu}) \sum_{t=0}^{\infty} \left\|E^t\Pi^{-1/2}\right\| \times \sqrt{E_{x\sim\pi}\left\{d_{\chi^2}(p_x^t,\pi)^2\right\}}\left\|\Pi^{1/2}g\right\|$$

(by Lemma 1)

$$\le \left\|\nabla\pi'\right\|(1-e^{-\beta\nu}) \sum_{t=0}^{\infty} ce^{-t/\tau^*}\left\|E^t\Pi^{-1/2}\right\|\left\|\Pi^{1/2}g\right\|$$

$$\le \left\|\nabla\pi'\right\|(1-e^{-\beta\nu}) \sum_{t=0}^{\infty} ce^{-t/\tau^* - \beta t\bar{\tau}_{min}}\left\|\Pi^{-1/2}\right\|\left\|\Pi^{1/2}g\right\|$$

(because $\left\|E^t\right\| = \max\{e^{-\beta\bar{\tau}(1)},\ldots,e^{-\beta\bar{\tau}(N)}\} = e^{-\beta\bar{\tau}_{min}}$, where $\bar{\tau}_{min} = \min\{\bar{\tau}(1),\ldots,\bar{\tau}(N)\}$)

$$= \left\|\nabla\pi'\right\| \frac{c(1-e^{-\beta\nu})}{(1-e^{-1/\tau^* - \beta\bar{\tau}_{min}})}\left\|\Pi^{-1/2}\right\|\left\|\Pi^{1/2}g\right\|$$

Because

$$\frac{1}{1-e^{-\beta\bar{\tau}_{min}}e^{-1/\tau^*}} \le \frac{1}{1-e^{-1/\tau^*}} \le \tau^*$$

then

$$(1-e^{-\beta\nu})(\nabla\pi)' J_\beta \le \left\|\nabla\pi'\right\| c\tau^*(1-e^{-\beta\nu})\left\|\Pi^{-1/2}\right\|\left\|\Pi^{1/2}g\right\| \quad \text{(A·6)}$$

From (A· 5) and (A· 6), we have

$$\left\|\nabla J - e^{-\beta\nu}\nabla_\beta J\right\| \le (1-e^{-\beta\nu})BC\nu + \left\|\nabla\pi'\right\| c\tau^* \times (1-e^{-\beta\nu})\left\|\Pi^{-1/2}\right\|\left\|\Pi^{1/2}g\right\|$$

□

Appendix C: Proof of Theorem 4

Lemma 4: Assume that Assumptions 1, 2, 3, 4, and 5 hold, then

$$\left\|\Delta^k - \Delta\right\| \le BC\frac{e^{-\beta k}}{1-e^{-\beta}}$$

Proof: We have

$$\left\|\Delta^k - \Delta\right\| = \frac{1}{n_2} \times \left\| \sum_{t=n_1}^{n_1+n_2-1} \nabla_t \left\{ \begin{array}{l} \sum_{s=0}^{k-1} e^{-\beta s} g(x_{t+1+s}, u_{t+1+s}) \\ - \sum_{s=0}^{n_1+n_2+n_3-1-t} e^{-\beta s} g(x_{t+1+s}, u_{t+1+s}) \end{array} \right\} \right\|$$

$$\le \frac{1}{n_2} \sum_{t=n_1}^{n_1+n_2-1} \|\nabla_t\| \times \left| \begin{array}{l} \sum_{s=0}^{k-1} e^{-\beta s} g(x_{t+1+s}, u_{t+1+s}) \\ - \sum_{s=0}^{n_1+n_2+n_3-1-t} e^{-\beta s} g(x_{t+1+s}, u_{t+1+s}) \end{array} \right|$$

$$\le \frac{1}{n_2} \sum_{t=n_1}^{n_1+n_2-1} \|\nabla_t\| \sum_{s=k}^{n_1+n_2+n_3-1-t} e^{-\beta s} \times |g(x_{t+1+s}, u_{t+1+s})|$$

$$\le \frac{1}{n_2} BC \left\{ \frac{n_2 e^{-\beta k}}{1-e^{-\beta}} - \sum_{t=n_1}^{n_1+n_2-1} \frac{e^{-\beta(n_1+n_2+n_3)}e^{-\beta t}}{(1-e^{-\beta})} \right\}$$

$$= BC\frac{e^{-\beta k}}{1-e^{-\beta}} - \frac{BC}{n_2} \sum_{t=n_1}^{n_1+n_2-1} \frac{e^{-\beta(n_1+n_2+n_3)}e^{-\beta t}}{(1-e^{-\beta})}$$

$$\le BC\frac{e^{-\beta k}}{1-e^{-\beta}}$$

The equality is when $n_2 = \infty$. □

Lemma 5:

$$\left\|E_\pi \Delta_t^k - \nabla_\beta J\right\| \le BC\frac{e^{-\beta k}}{1-e^{-\beta}}$$

Proof: From the proof of Theorem 2 in [1] and [2], we have

$$\nabla_\beta J(\theta) = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \nabla_t \Big[g(x_t,u_t)\bar{\tau}(x_t,u_t) + \sum_{s=0}^{\infty} e^{-\beta s} g(x_{s+t+1}, u_{s+t+1}) \Big]$$

On the other hand,

$$E_\pi \Delta_t^k = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \nabla_t \Big[g(x_t,u_t)\bar{\tau}(x_t,u_t) + \sum_{s=0}^{k-1} e^{-\beta s} g(x_{s+t+1}, u_{s+t+1}) \Big]$$

Thus,

$$\begin{aligned}
&\left\|E_\pi \Delta_t^k - \nabla_\beta J(\theta)\right\| \\
&= \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \nabla_t \sum_{s=k}^{\infty} e^{-\beta s} g(x_{s+t+1}, u_{s+t+1}) \\
&\le \lim_{T\to\infty} \frac{BC}{T} \sum_{t=0}^{T-1} \sum_{s=k}^{\infty} e^{-\beta s} \\
&= BC \frac{e^{-\beta k}}{1-e^{-\beta}}
\end{aligned}$$

□

To prove Theorem 4, we use the following theorem from [9] (see Theorem 15): if $\{X_t\}$ is $\tau^* - mixing$ and $f : X \to [a,b]^K$, and $s \le n_2$, then

$$\begin{aligned}
&\Pr\left(\left\|\frac{1}{n_2}\sum_{i=n_1}^{n_1+n_2-1} f(X_i) - E_\pi(f)\right\|_\infty \ge \epsilon \,\middle|\, X_{-\infty}^0\right) \\
&\le \frac{K}{2}\left(s e^{-\lfloor n_1/\tau^*\rfloor} + n_2 e^{-\lfloor s/\tau^*\rfloor} + 4s\exp\left(\frac{-\epsilon^2 n_2}{4(b-a)^2 s}\right)\right)
\end{aligned}$$

(where K is the parameter dimension and $\lfloor a/b\rfloor$ is integer division)

Theorem 4 is now easy to be proved by using results of Lemmas 4, 5 and the above theorem. We apply the above Theorem to the function Δ_t^k of the vector S_t^k to obtain

$$\begin{aligned}
&\Pr\left(\left\|\frac{1}{n_2}\sum_{i=n_1}^{n_1+n_2-1} f(X_i) - E_\pi(f)\right\|_\infty + \frac{2BC}{1-e^{-\beta}} e^{-\beta k} \right. \\
&\left. \ge \epsilon + \frac{2BC}{1-e^{-\beta}} e^{-\beta k} \,\middle|\, X_{-\infty}^0\right) \\
&= \Pr\left(\left\|\Delta^k - E_\pi \Delta_t^k\right\|_\infty + \frac{2BC}{1-e^{-\beta}} e^{-\beta k} \right. \\
&\qquad \left. \ge \epsilon + \frac{2BC}{1-e^{-\beta}} e^{-\beta k} \,\middle|\, X_{-\infty}^0\right)
\end{aligned}$$

where,

$$\Delta^k = \frac{1}{n_2}\sum_{t=n_1}^{n_1+n_2-1} \Delta_t^k$$

On the other hand,

$$\begin{aligned}
&\left\|\Delta^k - E_\pi \Delta_t^k\right\|_\infty + \frac{2BC}{1-e^{-\beta}} e^{-\beta k} \\
&\ge \left\|\Delta^k - E_\pi \Delta_t^k\right\|_\infty + \left\|\Delta - \Delta^k\right\|_\infty \\
&\qquad + \left\|E_\pi \Delta_t^k - \nabla_\beta J\right\|_\infty \\
&\ge \left\|\Delta - \nabla_\beta J\right\|_\infty
\end{aligned}$$

(Following to Lemmas 4, 5). Thus,

$$\begin{aligned}
&\Pr\left(\begin{array}{l} \left\|\Delta^k - E_\pi \Delta_t^k\right\|_\infty + \frac{2BC}{1-e^{-\beta}} e^{-\beta k} \\ \ge \varepsilon + \frac{2BC}{1-e^{-\beta}} e^{-\beta k} \end{array} \,\middle|\, X_{-\infty}^0\right) \\
&\ge \Pr\left(\left\|\Delta - \nabla_\beta J\right\|_\infty \ge \varepsilon + \frac{2BC}{1-e^{-\beta}} e^{-\beta k} \,\middle|\, X_{-\infty}^0\right)
\end{aligned}$$

According to (14) (or actually Theorem 15 in [9]), if we set

$$\begin{cases} a = 0 \\ b = BC\left(\nu + \frac{1}{1-e^{-\beta}}\right) \end{cases}$$

Then finally, we obtain

$$\begin{aligned}
&\Pr\left(\left\|\Delta - \nabla_\beta J\right\|_\infty \ge \epsilon + \frac{2BC}{1-e^{-\beta}} e^{-\beta k} \,\middle|\, X_{-\infty}^0\right) \\
&\le \frac{K}{2}\left\{\begin{array}{l} s e^{-\lfloor n_1/\tau^*\rfloor} + n_2 e^{-\lfloor s/\tau^*\rfloor} \\ +4s\exp\left(\frac{-\epsilon^2 n_2 (1-e^{-\beta})^2}{4B^2C^2(\nu - \nu e^{-\beta}+1)^2 s}\right) \end{array}\right\} \\
&\le \frac{K}{2}\left\{s e^{-\lfloor n_1/\tau^*\rfloor} + n_2 e^{-\lfloor s/\tau^*\rfloor} + 4s e^{\frac{-\epsilon^2 n_2 (1-e^{-\beta})^2}{4B^2C^2 s}}\right\}
\end{aligned}$$

□

Ngo Anh Vien received the B.S. degree in Computer Science from Hanoi University of Technology, Hanoi, Vietnam, in 2005. He is now working toward a Ph.D. degree at Artificial Intelligence Laboratory, Department of Computer Engineering, Kyung Hee University, South Korea. His current research interests include Reinforcement Learning, Approximate Dynamic Programming, Bayesian nonparametrics and Statistical Learning Theory.

SeungGwan Lee received Ph.D., M.S. and B.S. degrees in the Department of Computer Engineering at Kyung Hee University in 2004, 1999 and 1997. He has been a Professor in the College of Liberal Arts at Kyung Hee University since 2006.9. He was a Visiting Professor in the School of Computer Science and Information Engineering at Catholic University from 2004 to 2006.8. His research interests include artificial intelligence, meta-search algorithm, multi-agents, ubiquitous computing and robot soccer.

TaeChoong Chung received the B.S. degree in electronic engineering from Seoul National University, Korea, in 1980, and the M.S. and Ph.D. degrees in electronic engineering from KAIST, Korea, in 1982 and 1987, respectively. Since 1988, he has been with Department of Computer Engineering, Kyung Hee University, Suwon, Korea, where he is now a Professor. His research interests include Reinforcement Learning, Heuristic Search, Robot Soccer.