Automatica 63 (2016) 182–186

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

Technical communique

Sleeping experts and bandits approach to constrained Markov decision processes[☆]

Hyeong Soo Chang[1]

Department of Computer Science and Engineering, Sogang University, Seoul, Republic of Korea

ARTICLE INFO

Article history:
Received 25 December 2014
Received in revised form
17 September 2015
Accepted 23 September 2015

Keywords:
Markov decision processes
Sleeping experts and bandits
Learning algorithm
Constrained optimization

ABSTRACT

This communique presents simple simulation-based algorithms for obtaining an approximately optimal policy in a given finite set in large finite constrained Markov decision processes. The algorithms are adapted from playing strategies for "sleeping experts and bandits" problem and their computational complexities are independent of state and action space sizes if the given policy set is relatively small. We establish convergence of their expected performances to the value of an optimal policy and convergence rates, and also almost-sure convergence to an optimal policy with an exponential rate for the algorithm adapted within the context of sleeping experts.

1. Introduction

Consider a discrete-time system with infinite horizon: $x_{t+1} = f(x_t, a_t, w_t)$ for $t = 0, 1, 2, \ldots,$ where x_t is the state at time t – ranging over a finite set X, a_t is the action at time t – to be chosen from a nonempty subset $A(x_t)$ of a given finite set of available actions A at time t, and w_t is a random disturbance uniformly and independently selected from $[0, 1]$ at time t, representing the uncertainty in the system, and f is a next-state function such that $f(x, a, w) \in X$ for $x \in X$, $a \in A(x)$, and $w \in [0, 1]$.

Define a (stationary non-randomized Markovian) policy $\pi : X \to A$ with $\pi(x) \in A(x)$ for all $x \in X$ and *value function* of π given by $V^{\pi}(x) = E_{w_0,\ldots,w_\infty}[\sum_{t=0}^{\infty} \gamma^t R(x_t, \pi(x_t), w_t)|x_0 = x]$, $x \in X$, with discount factor $\gamma \in (0, 1)$ and one-period reward function R such that $R(x, a, w) \in \mathcal{R}^+$ for $x \in X, a \in A(x)$, and $w \in [0, 1]$ and *constraint value function* of π given by $J^{\pi}(x) = E_{w_0,\ldots,w_\infty}[\sum_{t=0}^{\infty} \beta^t C(x_t, \pi(x_t), w_t)|x_0 = x]$, $x \in X$, with discount factor $\beta \in (0, 1)$ and one-period cost function C such that $C(x, a, w) \in \mathcal{R}^+$ for $x \in X, a \in A(x)$, and $w \in [0, 1]$. We also have a constraint constant $K > 0$ which determines the feasibility of π. A policy π is defined to be feasible at x if $J^{\pi}(x) \leq K$. The function f, together with X, A, R, C, and K comprise a constrained Markov decision process (CMDP) (Altman, 1998). For simplicity, we consider one constraint case. Extension to multiple case is straightforward. We let $R_{\max} = \sup_{x,a,w} R(x, a, w)$ and $C_{\max} = \sup_{x,a,w} C(x, a, w)$.

For a given $w = \{w_t\}$, we let $V^{\pi}(x, w) = \sum_{t=0}^{\infty} \gamma^t R(x_t, \pi(x_t), w_t)$ and $J^{\pi}(x, w) = \sum_{t=0}^{\infty} \beta^t C(x_t, \pi(x_t), w_t)$ with $x_0 = x$. We assume throughout that any sample of $V^{\pi}(x, w)$ and $J^{\pi}(x, w)$ is bounded. Without loss of generality, we take the bound to be 1, i.e., for any w, x, and π, $V^{\pi}(x, w) \in [0, 1]$ and $J^{\pi}(x, w) \in [0, 1]$. (The generalization to an arbitrary bound can be done by appropriate scaling. Or by defining a transformation of R into R' such that $R'(x, a, w) = R(x, a, w)(1-\gamma)/R_{\max}$ and C to C' similarly, we can construct an "equivalent" CMDP to the given CMDP which satisfies the assumption.) We also assume that an initial state x_0 is *fixed* by some $x \in X$ and a nonempty finite policy set Π is given. That is, Π, x_0, and K are problem specific parameters.

A policy $\pi \in \Pi$ is called ϵ*-feasible* (at x) if $J^{\pi}(x) \leq K + \epsilon$ for $\epsilon \geq 0$. We let ϵ-feasible policy set $\Pi_f^{\epsilon} = \{\pi : \pi \in \Pi, J^{\pi}(x) \leq K + \epsilon\}$. We then say that for $\epsilon \geq 0$, $\pi_{\epsilon}^* \in \Pi$ is an ϵ*-feasible optimal* policy if for some nonempty Δ such that $\Pi_f^{-\epsilon} \subseteq \Delta \subseteq \Pi_f^{\epsilon}$, $\pi_{\epsilon}^* \in \Delta$ and $\max_{\pi \in \Delta} V^{\pi}(x) = V^{\pi_{\epsilon}^*}(x)$. The problem we consider is obtaining or estimating a 0-feasible optimal policy in Π, if such a policy exists.

The problem of obtaining a 0-feasible optimal policy is known to be NP-hard if Π contains all possible policies (in which case $|\Pi| = |A|^{|X|}$) and the problem size is characterized by the maximum of $|X|$ and $\max_{x \in X} |A(x)|$ and the number of constraints (Feinberg, 2000). It seems that there exist only two exact iterative algorithms for this problem that exploit structural properties of

[☆] The material in this paper was not presented at any conference. This paper was recommended for publication in revised form by Associate Editor Carlo Fischione under the direction of Editor André L. Tits.

E-mail address: hschang@sogang.ac.kr.

[1] Tel.: +82 2 705 8925; fax: +82 2 704 8273.

http://dx.doi.org/10.1016/j.automatica.2015.10.015
0005-1098/© 2015 Elsevier Ltd. All rights reserved.

CMDPs. Chen and Feinberg (2007) provided a value-iteration type algorithm based on certain dynamic programming equations and Chang (2014) presented a policy-iteration type algorithm based on a feasible-policy space characterization. Unfortunately, both require solving certain finite or infinite horizon MDP problems so that *computational complexities depend on state and action space sizes.* Note that linear programming used for finding a best *randomized* policy cannot be applied here due to non-linearity and non-convexity of this problem (cf., P1 in Feinberg, 2000, Theorem 3.1).

Even if there exists a body of works on simulation-based algorithms for solving unconstrained MDPs in order to break the curse of dimensionality (see, e.g., Chang, Fu, Hu, & Marcus, 2007, Powell, 2011 and the references therein), it seems that there has been no notable approach to CMDPs via simulation. This paper is probably the first step towards developing such algorithms. Because the algorithms proposed in this paper work with simulated sample-paths, computational complexities are independent of $|X|$ and $|A|$ as long as $|\Pi|$ is relatively small.

Our approach is simple and natural. We generate a sequence of $\{\Pi_{f,n,H}, n = 1, \ldots, N\}$ where $\Pi_{f,n,H}$ is an estimate of Π_f^0, similar to the sample average approximation (SAA) method (Kleywegt, Shapiro, & Homom-De-Mello, 2001), by using simulation over a finite horizon H. For each $\pi \in \Pi$, $J^{\pi}(x)$ is estimated with a sample mean and if the sample mean is less than or equal to K, π is included in $\Pi_{f,n,H}$. We then generate a sequence of policies $\{\pi(n), n = 1, \ldots, N\}$ from $\Pi_{f,n,H}$ at iteration n, where $\pi(n)$ is an estimate of a 0-feasible optimal policy. The selection of $\pi(n)$ from $\Pi_{f,n,H}$ follows the structure of the two playing strategies, called "follow-the-awake-leader" (FTAL) and "awake-upper-estimated-reward" (AUER), for "sleeping experts and bandits" problems (Kleinberg, Niculescu-Mizil, & Sharma, 2010) in on-line decision making. A major difference between FTAL and AUER is that for FTAL, we simulate each policy in $\Pi_{f,n,H}$ to update the sample mean of each policy but for AUER, we simulate only selected policy $\pi(n)$ to update the sample mean of $\pi(n)$.

Sleeping experts and bandits model (see, Kleinberg et al., 2010 for a formal description) formulates problems within the context of on-line sequential decision making process. We wish to develop a playing strategy which maximizes the sum of the sample rewards (over a finite horizon) obtained by choosing an action from currently "awaken" actions in each round. The set of available actions changes from one round to the next and when played, each arm provides a random reward from an unknown distribution specific to that arm. In the so-called expert setting, once an action is chosen and played, the sample rewards of all available actions are eventually revealed to the strategy at the end of each round. On the other hand, in the bandit setting, only the reward of the chosen action is revealed. The performance of the strategy is measured by "expected regret", roughly, the difference between the sum of the expected rewards of the strategy's chosen actions and that of the highest ranked actions in each round in terms of the expected reward, i.e., the best sequence of action choices in hindsight. We view $\Pi_{f,n,H}$ as the set of currently awaken or non-sleeping experts/bandits in Π and the sample value of the accumulated reward sum over the horizon H as the sample reward of playing the expert/bandit π. By proper adaptation of the results of the expected regret defined over the sleeping experts and bandits model then, we can establish convergence of the expected performance of our approach without the assumption that a 0-feasible optimal policy is unique. We show that when $\Pi_f^0 \neq \emptyset$, the expected performance $1/N \sum_{n=1}^{N} E[V_H^{\pi(n)}(x)]$ approaches the value of a 0-feasible optimal policy $\max_{\pi \in \Pi_f^0} V^{\pi}(x)$ as $N \to \infty$ and $H \to \infty$ with a rate of $O(1/N)$ (for $N \geq (\min_{\pi,\pi' \in \Pi}\{V^{\pi}(x) - V^{\pi'}(x) : V^{\pi}(x) - V^{\pi'}(x) > 0\})^{-1}$) in the FTAL case and of $O(\ln N/N)$ in the AUER case for such N. Here $V_H^{\pi}(x) = E_{w_0,\ldots,w_{H-1}}[\sum_{t=0}^{H-1} \gamma^t R(x_t, \pi(x_t), w_t)|x_0 = x]$ for $H < \infty$. For the FTAL case, we further provide almost-sure convergence of $\pi(N)$ to a 0-feasible optimal policy as N and H go to infinity with an exponential convergence rate at the expense of the assumption that value functions are all different among policies.

The works on the problem of finding the best solution from a finite set of solutions given stochastic objective and constraint functions by simulation are relatively sparse (see Pasupathy, Hunter, Pujowidianto, Lee, & Chen, 2015 and the related references therein). These works study allocating different (Monte-Carlo) simulation budgets to the solutions to (approximately) maximize the probability of selecting the best solution from sample-mean estimates but provide explicit forms of such allocation only in an asymptotic limit, i.e., when the total number of samples approaches infinity. This is also typically given under the assumption that the best solution is unique and the distribution of samples are normal and in terms of the unknown true means and variances. Even if heuristic iterative approximation procedures of such results are given, the convergences of those are not known. In our context, the best policy is not necessarily unique and the normality assumption is not necessarily valid. Although Pasupathy et al. (2015) consider general distribution case, the optimal allocation is only characterized by an optimization problem so that explicit forms of budget allocation are difficult to obtain even in an asymptotic limit except for some special cases. Without the uniqueness and the normality assumptions, Li, Sava, and Xie (2009) consider a sequence of penalty cost functions to combine objective and constraint functions with certain budget allocation strategy among the solutions but obtaining the sequence of the penalty cost functions is not straightforward and their algorithm converges to a locally optimal solution when some restrictive assumptions are satisfied.

Our setting also covers that in which explicit forms for f, R, and C are not available, but they can be simulated. In this setting, another approach to consider is to employ a stochastic-approximation based learning-algorithm as for unconstrained MDPs (see, e.g., Bhatnagar, Hemachandra, & Mishra, 2011 and Djonin & Krishnamurthy, 2007). But this works when Π is the set of all possible policies and the convergence speed is typically very slow and finite-time behaviours of such methods are not known. Moreover, it is not immediate how to adapt such approach when Π is a subset of the set of all possible policies.

2. Algorithm

We first provide the pseudocode of the FTAL algorithm below. It mainly consists of the **Feasible-Policy Set Estimation** step and the **Feasible Optimal Policy Estimation** step. The **Feasible-Policy Set Estimation** step obtains $\Pi_{f,n,H} = \{\pi : J_{n,H}^{\pi}(x) \leq K, \pi \in \Pi\}$ at iteration n. Here $J_{n,H}^{\pi}(x)$ is the sample mean obtained by n independent samples of $J_H^{\pi}(x, w^{\pi}) = \sum_{t=0}^{H-1} \beta^t C(x_t, \pi(x_t), w_t^{\pi})$ for $w^{\pi} = \{w_0^{\pi}, \ldots, w_{H-1}^{\pi}\}$ and $H < \infty$. We let $J_H^{\pi}(x) := E_{w^{\pi}}[J_H^{\pi}(x, w^{\pi})]$. The **Feasible Optimal Policy Estimation** step selects $\pi(n)$ that achieves $\max_{\pi \in \Pi_{f,n,H}} V_{\tau(\pi),H}^{\pi}(x)$ if $\Pi_{f,n,H} \neq \emptyset$ and $\tau(\pi) \neq 0$ for all $\pi \in \Pi_{f,n,H}$. (That is, we "follow the current best" among non-sleeping experts.) Similarly, $V_{n,H}^{\pi}(x)$ is the sample mean obtained by n independent samples of $V_H^{\pi}(x, w^{\pi}) = \sum_{t=0}^{H-1} \gamma^t R(x_t, \pi(x_t), w_t^{\pi})$, and recall that $V_H^{\pi}(x) = E_{w^{\pi}}[V_H^{\pi}(x, w^{\pi})]$. Note that w^{π} generated in $J_H^{\pi}(x, w^{\pi})$ for all $\pi \in \Pi$ is reused in $V_H^{\pi}(x, w^{\pi})$ for all $\pi \in \Pi_{f,n,H}$. The counter $\tau(\pi)$ keeps track of the number of times π has been simulated to obtain a sample of $V_H^{\pi}(x, w)$. Whenever π is included in $\Pi_{f,n,H}$ at some n, π is simulated. If there exists π in $\Pi_{f,n,H}$ such that $\tau(\pi) = 0$, $\pi(n)$ is set to be any such π. If $\Pi_{f,n,H} = \emptyset$, $\pi(n)$ is set to be any $\pi \in \Pi$.