Automatica 58 (2015) 127–130

Contents lists available at ScienceDirect

Automatica

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

Technical communique

Random search for constrained Markov decision processes with multi-policy improvement☆

Hyeong Soo Chang [1]

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

ARTICLE INFO

Article history:
Received 7 August 2014
Received in revised form
19 April 2015
Accepted 5 May 2015
Available online 2 June 2015

Keywords:
Markov decision processes
Random search
Policy improvement
Constrained optimization

ABSTRACT

This communique first presents a novel multi-policy improvement method which generates a feasible policy at least as good as any policy in a given set of feasible policies in finite constrained Markov decision processes (CMDPs). A random search algorithm for finding an optimal feasible policy for a given CMDP is derived by properly adapting the improvement method. The algorithm alleviates the major drawback of solving unconstrained MDPs at iterations in the existing value-iteration and policy-iteration type exact algorithms. We establish that the sequence of feasible policies generated by the algorithm converges to an optimal feasible policy with probability one and has a probabilistic exponential convergence rate.

1. Introduction

We consider a finite constrained Markov decision process (CMDP; see, e.g., Altman (1998) for example problems) with a finite state set X, a finite action set A, a reward function R, and a transition function P. We denote the admissible action set at state x as $A(x)$ so that $A = \bigcup_{x\in X} A(x)$. The reward function R is given such that $R(x, a) \in \mathbb{R}$, $x \in X$, $a \in A(x)$ and P maps $\{(x, a)|x \in X, a \in A(x)\}$ to the set of probability distributions over X.

We define a (stationary non-randomized Markovian) policy π as a mapping from X to A with $\pi(x) \in A(x)$, $\forall x \in X$, and let Π be the set of all such policies. We define *the objective value of* $\pi \in \Pi$ with an initial state $x \in X$:

$$V^\pi(x) := E\left[\sum_{t=0}^{\infty} \gamma^t R(X_t, \pi(X_t)) \middle| X_0 = x\right],$$

where X_t is a random variable denoting state at time t by following π and $\gamma \in (0, 1)$ is a discounting factor.

The CMDP is also associated with a function κ defined over X and a constraint-cost function D such that $D(x, a) \in \mathbb{R}$, $x \in X$, $a \in A(x)$. A policy $\pi \in \Pi$ is called *feasible* if it satisfies the constraint inequality of $\sum_{x\in X} \delta(x)J^\pi(x) \le \sum_{x\in X} \delta(x)\kappa(x)$, where δ is an initial state distribution over X and the *constraint value of* π with an initial state $x \in X$ is defined such that

$$J^\pi(x) := E\left[\sum_{t=0}^{\infty} \beta^t D(X_t, \pi(X_t)) \middle| X_0 = x\right]$$

for a discounting factor $\beta \in (0, 1)$. Throughout the paper, we assume that δ *is fixed arbitrarily*.

We let $J^\pi_\delta = \sum_{x\in X} \delta(x)J^\pi(x)$, $V^\pi_\delta = \sum_{x\in X} \delta(x)V^\pi(x)$, and $\kappa_\delta = \sum_{x\in X} \delta(x)\kappa(x)$. The problem is then to obtain an optimal feasible policy π^* in the feasible policy set,

$$\Pi_f = \{\pi : \pi \in \Pi, J^\pi_\delta \le \kappa_\delta\},$$

which achieves $\max_{\pi\in\Pi_f} V^\pi_\delta$, if the problem is solvable, that is, $\Pi_f \neq \emptyset$. In the sequel, we assume that $\min_{\pi\in\Pi} J^\pi_\delta \le \kappa_\delta$ so that $\Pi_f \neq \emptyset$.

Feinberg (2000) showed that if the size of the above problem is characterized by the maximum of $|X|$ and $\max_{x\in X} |A(x)|$ and the number of constraints, then it is NP-hard. In particular, he provided a mathematical program (MP) formulation for this problem (cf., P1 in Feinberg (2000, Theorem 3.1)) such that the MP is feasible if and only if $\Pi_f \neq \emptyset$. Due to its non-linearity and non-convexity properties, linear programming (LP), which can be used for finding a best *randomized* policy, cannot be applied to the MP. A policy-iteration (PI) type algorithm, called "exact policy search (EPS)", has been recently presented by Chang (2014) based on a characterization of the entire feasible policy space and it is shown that EPS converges to an optimal feasible policy in a finite number of iterations, $|\Pi|$ iterations in the worst case. But EPS requires solving unconstrained

☆ The material in this paper was not presented at any conference. This paper was recommended for publication in revised form by Associate Editor Henrik Sandberg under the direction of Editor André 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.2015.05.016
0005-1098/

infinite-horizon MDPs *at each iteration* involved with feasible policies. The computational complexity of exactly solving an infinite-horizon MDP is well-known to be high in general. In particular, even if the time-complexity of the value iteration (VI) algorithm for convergence in terms of the number of iterations is polynomial in $|X|$, $|A|$, $1/(1-\gamma)$, and the size of representing the inputs R and P, the dependence on $1/(1-\gamma)$ is a major drawback (Blondel & Tsitsiklis, 2000). On the other hand, PI's time-complexity for convergence is known to be exponential in general (Hollanders, Delvenne, & Jungers, 2012) even if it is strongly polynomial when γ is fixed (Ye, 2011). Note that the per-iteration computational complexity of VI is $O(|A||X|^2)$ and that of PI is $O(|X|^3+|A||X|^2)$. Scherrer (2013) improves the result of the upper bound on the number of iterations of PI by Ye (2011) by providing the upper bound of $O(\frac{|X|^2|A|}{1-\gamma}\log(\frac{1}{1-\gamma}))$ but again the dependence on $1/(1-\gamma)$ is a major drawback. He provided some upper bounds that are independent of γ under some structural assumptions on MDPs but we do not impose any such assumptions here. Even though in theory MDP can be solved in polynomial time by LP, it is known that the existing polynomial-time LP algorithms run very slowly in practice. Simplex methods seem to perform well in practice but can take an exponential number of iterations on some problems (Littman, Dean, & Kaelbling, 1995).

Based on certain dynamic programming equations in Chen and Blankenship (2004), Chen and Feinberg (2007) provided a VI-type algorithm addressing CMDPs with non-randomized, but possibly non-stationary policies, while the present communique deals with only stationary policies. Therefore, it is not directly applicable to the problem here. In fact, Chen and Feinberg's approach is only rather theoretically interesting because a sequence of finite-horizon MDPs needs to be solved with increasing the horizon size but the MDPs cannot be solved exactly in practice due to the arbitrarily increasing horizon size.

In this communique, we first provide a novel *multi-policy improvement* method which generates a feasible policy at least as good as any feasible policy in a given nonempty subset of Π_f. We stress that the improving policy is not necessarily in the subset. We then properly adapt the improvement method to develop a convergent random search algorithm for obtaining an optimal feasible policy. In particular, we follow the spirit of "policy set iteration (PSI)" in Chang (2013) for solving *unconstrained* MDPs. However, because we need to deal with both feasible and infeasible policies unlike the unconstrained case, we need (1) to test the feasibility of policies in a set and (2) to generate a feasible policy that improves only the feasible policies of the set while ignoring the infeasible ones. At each iteration, we sample independently $N>0$ policies from Π by a given fixed distribution d and generate a feasible policy that improves all feasible policies sampled at the current iteration and all feasible policies sampled at the previous iterations. This produces a sequence of monotonically improving feasible policies $\{\pi_k^*\}$ where π_k^* improves all feasible policies sampled at iterations $0,1,\ldots,k$, i.e., all feasible policies among $N(k+1)$ sampled policies. The sequence $\{\pi_k^*\}$ approaches an optimal feasible policy with probability one as k goes to infinity.

Furthermore, we establish that such monotonically improving sequence has a probabilistic exponential convergence rate; for any newly sampled feasible-policy π from Π by d, the probability that V_δ^π is bigger than $V_\delta^{\pi_k^*}$ converges to zero with $O(N^{-k})$ rate.

We note that PSI in Chang (2013) is not a random search algorithm but a randomized variant of PI which generalizes PI with a finite-time convergence–$|\Pi|$ iterations in the worst case while guaranteeing no slower convergence speed than PI in terms of the number of iterations. The multi-policy improvement with randomly sampled policies is mainly used to expedite the convergence rate of PI. On the other hand, the algorithm proposed in this paper is a random search. Multiple i.i.d. samples are drawn by a given (sampling) distribution, instead of a single sample used in the well-known pure random search (see, e.g., Floudas and Pardalos (2009) and Spall (2003)), and bookkeeping of the "maximum" among all solutions generated so far is done. The multi-policy improvement is basically used for the maximum bookkeeping. Unlike PSI, even if the consecutive policies are the same in the monotonically improving sequence $\{\pi_k^*\}$, it does not necessarily mean that an optimal policy has been found. We cannot guarantee a finite-time convergence but the probability one convergence.

The main motivation for studying such a random approach is to alleviate the computational complexities of the two exact algorithms in Chang (2014) and Chen and Feinberg (2007) having a viable approach to solving CMDPs. The random search algorithm *does not solve any unconstrained MDP* (except possibly once for setting an initial feasible policy) and only has the γ-independent per-iteration complexity of $O(N(|X|^3+|A||X|^2))$. That is, the complexity amounts to the per-iteration complexity of PI multiplied by a factor $O(N)$.

Because the CMDP problem under consideration is just an NP-hard combinatorial optimization problem, any meta-heuristic algorithm (Floudas & Pardalos, 2009), e.g., genetic algorithm (Hirayama & Kawai, 2000), simulated annealing (Wah, Chen, & Wang, 2007), etc., designed for constrained problems can be also adapted into our setting with incorporating the multi-policy improvement method. We here focus on a simplest form of meta-heuristic, i.e., pure random search.

2. Multi-policy improvement

Let $B(X)$ be the set of all real-valued functions on X. We denote the probability of making a transition to state $y\in X$ when taking an action $a\in A(x)$ at state $x\in X$ by P_{xy}^a. Given a value function $w\in B(X)$, we let $w_\delta=\sum_{x\in X}\delta(x)w(x)$ and define *w-inducing feasible action set*

$$A_w(x) := \Big\{a\in A(x) : D(x,a)+\beta\sum_{y\in X}P_{xy}^a w(y) \le w(x)+(1-\beta)(\kappa_\delta-w_\delta)\Big\},\quad x\in X.$$

Roughly, in order for action a in $A(x)$ to be feasible at state x when the total constraint-cost at each state is measured by w, the excess of the total constraint-cost made by taking a at x at the first time-step needs to be within the "slackness" induced by w (cf., Lemma 1 in Chang (2014)).

Given a nonempty set Λ of feasible policies in Π_f, define ψ such that

$$\psi\in\operatorname*{arg\,max}_{\phi\in\{\phi_\pi:\pi\in\Lambda\}} V_\delta^\phi, \tag{1}$$

where for $\pi\in\Lambda$, ϕ_π is given such that

$$\phi_\pi(x)\in\operatorname*{arg\,max}_{a\in A_{J^\pi}(x)}\Big(R(x,a)+\gamma\sum_{y\in X}P_{xy}^a V^\pi(y)\Big),\quad x\in X.$$

Note that ψ as defined above is not necessarily in Λ and ϕ_π, $\pi\in\Lambda$, is the policy obtained by applying the original policy improvement method in MDPs (by using a single policy as a base-policy for improvement) to π with respect to the J^π-inducing feasible action set. In other words, the policy improvement method in PI is applied to π with the actions that ensure the feasibility of the improving policy. We now show that ψ improves all policies in Λ.

Theorem 1. *Given a nonempty finite set Λ of feasible policies in Π_f, consider ψ given in (1). Then ψ is feasible and $V_\delta^\psi\ge\max_{\pi\in\Lambda}V_\delta^\pi$ for any δ.*