Systems & Control Letters 59 (2010) 760–766

Contents lists available at ScienceDirect

Systems & Control Letters

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

An actor–critic algorithm with function approximation for discounted cost constrained Markov decision processes

Shalabh Bhatnagar

Department of Computer Science and Automation, Indian Institute of Science, Bangalore 560 012, India

ARTICLE INFO

Article history:
Received 20 July 2010
Received in revised form
24 August 2010
Accepted 24 August 2010
Available online 8 October 2010

Keywords:
Constrained Markov decision processes
Infinite horizon discounted cost criterion
Function approximation
Actor–critic algorithm
Simultaneous perturbation stochastic approximation

ABSTRACT

We develop in this article the first actor–critic reinforcement learning algorithm with function approximation for a problem of control under multiple inequality constraints. We consider the infinite horizon discounted cost framework in which both the objective and the constraint functions are suitable expected policy-dependent discounted sums of certain sample path functions. We apply the Lagrange multiplier method to handle the inequality constraints. Our algorithm makes use of multi-timescale stochastic approximation and incorporates a temporal difference (TD) critic and an actor that makes a gradient search in the space of policy parameters using efficient simultaneous perturbation stochastic approximation (SPSA) gradient estimates. We prove the asymptotic almost sure convergence of our algorithm to a locally optimal policy.

1. Introduction

The problem that we are concerned with in this article is of finding an optimal policy for a constrained Markov decision process (C-MDP) when (a) the transition probabilities are not known and (b) a feature-based representation is used. The latter is particularly useful when the state and action spaces are large or unmanageable. A text book treatment of C-MDP can be found in [1]. Reinforcement learning (RL) [2,3] has proved to be a useful paradigm for such problems and has largely been studied in the case of regular (unconstrained) Markov decision processes (MDP) [4,5]. Algorithms based on both value and policy iteration techniques have been developed and studied in this scenario. Actor–critic algorithms [6,7] are a class of RL algorithms that are based on the policy iteration method. Whereas the critic addresses a problem of prediction by estimating the value function for a given policy update, the actor updates the policy itself and is concerned with the problem of control. Temporal difference (TD) learning [3,8] has been found to be one of the most effective methods for the problem of prediction. For the problem of control, policy gradient methods [9] have been successfully applied. The policy in these methods is represented via a parameterized class of functions that are differentiable in the parameter. In actor–critic algorithms based on policy gradients, the actor's parameter is updated along the direction of the performance gradient. The performance itself, for any given parameter update, is estimated by the critic.

In this paper we present an actor–critic algorithm with function approximation for a C-MDP and prove its convergence. Our aim is to find an optimal policy that minimizes an infinite horizon discounted cost function subject to prescribed bounds on additional cost functions under that policy. The (above) additional cost functions have a similar structure to the objective function itself i.e., they are also expected discounted sums of some other single-stage cost functions. Thus cost functions in both the objective and the constraints are policy dependent. It is also important to note that the value functions corresponding to the single-stage costs cannot in general be analytically determined as functions of the parameter. Hence the constraint set is *a priori* not properly defined as well. Thus any solution methodology cannot be based on directly projecting iterates after each update to the constraint set formed from the inequality constraints.

Constrained MDPs find immense applications in many domains. For instance, in communication networks, a problem of interest is to maximize the throughput (i.e., the rate at which packets are delivered to the destination) subject to constraints on the delay as well as packet loss in transmission, see [10,11]. One can find similar important problems in many other domains.

Our development is based on forming the Lagrangian by incorporating the (multiple) inequality constraints in the objective. The primal is a new actor–critic algorithm that combines ideas from temporal difference learning, policy gradient methods and simultaneous perturbation stochastic approximation (SPSA) [12,13]

E-mail address: shalabh@csa.iisc.ernet.in.

0167-6911/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.sysconle.2010.08.013

while the dual corresponds to a recursive search in the space of Lagrange multipliers. A Lagrange based approach for solving a C-MDP with a single constraint has been used in [14]. The algorithm there however works for the look-up table case and does not use function approximation i.e., it requires storage of all the states and actions, and performs updates in the space of all randomized policies. Further, it has been designed for the long-run average cost objective and not the discounted cost. We use feature based representations as a result of which our problem becomes one of optimizing parameters in a constrained optimization setting. We consider multiple inequality constraints and the infinite horizon discounted cost criterion. Ours is the first work that develops an actor–critic algorithm with function approximation for control with multiple inequality constraints and under the discounted cost objective.

The rest of the paper is organized as follows: in Section 2, we present the framework and problem formulation. We present in Section 3 our constrained actor–critic algorithm. In Section 4, we present the convergence proof. Finally, Section 5 contains the concluding remarks.

2. The framework and problem formulation

By a MDP we mean a stochastic process $\{X_n\}$ taking values in a set S (called the state space), that is governed by a control sequence $\{Z_n\}$ and satisfies the controlled Markov property (below). Let $A(i)$ be the set of feasible actions in state i and $A \stackrel{\triangle}{=} \cup_{i\in S} A(i)$ be the set of all actions or the action space. We assume that both S and A are finite sets. The controlled Markov property satisfied by $\{X_n\}$ is the following:

$$P(X_{n+1} = j \mid X_m, Z_m, m \leq n) = p(X_n, Z_n, j) \quad \text{a.s.},$$

where $p : S \times A \times S \to [0, 1]$ is a given function for which $\sum_{j\in S} p(i, a, j) = 1, \forall a \in A(i),\ i \in S$.

A policy is a decision rule for selecting actions. We call $\bar{\pi} \stackrel{\triangle}{=} \{\mu_0, \mu_1, \ldots\}$, where $\mu_n : S \to A$, an admissible policy when $\mu_n(i) \in A(i)\ \forall i \in S$. When $\mu_n \equiv \mu, \forall n \geq 0$, where μ is independent of n, we call $\bar{\pi}$ or many times μ itself a stationary deterministic policy (SDP). A stationary randomized policy (SRP) π is specified via a probability distribution $\pi(i, \cdot)$ over $A(i), \forall i \in S$. It is easy to see that under any given SDP or SRP, $\{X_n\}$ is a Markov chain. We make the following assumption:

Assumption 1. The Markov chain $\{X_n\}$ under any SRP π is irreducible.

It follows from Assumption 1 that $\{X_n\}$ is also positive recurrent under any SRP because S is a finite set. Let $c(n), g_1(n), \ldots, g_N(n), n \geq 0$ denote certain single-stage costs that we assume are non-negative, real-valued, uniformly bounded and mutually independent random variables. In addition, given the current state and action (X_n and Z_n), $c(n), g_k(n), k = 1, \ldots, N$ are conditionally independent of the previous states and actions (X_m, Z_m, $m < n$). The evolution in a C-MDP proceeds as follows: at instant n, the state X_n is observed and action Z_n is chosen. This results in the $(N + 1)$-length string of costs $c(n), g_1(n), \ldots, g_N(n)$ and the process moves to a new state X_{n+1} at instant $n + 1$. Let $c(i, a), g_k(i, a)$ be defined via $c(i, a) = E[c(n) \mid X_n = i, Z_n = a]$, $g_k(i, a) = E[g_k(n) \mid X_n = i, Z_n = a], \forall n \geq 0, k = 1, \ldots, N$, respectively. (Note the abuse of notation here.) It is shown in Theorem 3.1 of [1] that SRPs correspond to a complete class of policies for the problem that we consider, i.e., it is sufficient to find an optimal policy within the class of SRPs. Under an SRP π, let $d^\pi(i)$ denote the stationary probability of the Markov process being in state $i \in S$ and let $d^\pi \stackrel{\triangle}{=} (d^\pi(i), i \in S)^\top$.

Our aim is to find a SRP that for a given initial distribution β (over states), minimizes (over all SRPs π) the discounted cost

$$J^\beta(\pi) = \sum_{i\in S} \beta(i) U^\pi(i), \tag{1}$$

subject to the constraints

$$S_k^\beta(\pi) = \sum_{i\in S} \beta(i) W_k^\pi(i) \leq \alpha_k, \tag{2}$$

$k = 1, \ldots, N$. Here

$$U^\pi(i) \stackrel{\triangle}{=} E\left[\sum_{m=0}^{\infty} \gamma^m c(m) \mid X_0 = i, \pi\right],$$

$$W_k^\pi(i) \stackrel{\triangle}{=} E\left[\sum_{m=0}^{\infty} \gamma^m g_k(m) \mid X_0 = i, \pi\right],$$

$k = 1, \ldots, N$. Here $0 < \gamma < 1$ is a given discount factor. Also, in (2), $\alpha_1, \ldots, \alpha_N$ are given positive constants. We assume that there exists at least one SRP π for which all the inequality constraints (2) are satisfied. Under this requirement, it is shown in Theorem 3.8 of [1] that an optimal SRP π^* that uses at most N randomizations exists. The constraints (2) can be alternatively written as

$$G_k^\beta(\pi) \stackrel{\triangle}{=} S_k^\beta(\pi) - \alpha_k \leq 0, \tag{3}$$

$k = 1, \ldots, N$.

Let $\bar{\lambda} = (\lambda_1, \ldots, \lambda_N)^\top$ denote the vector of Lagrange multipliers $\lambda_1, \ldots, \lambda_N \in \mathcal{R}^+ \cup \{0\}$ and let $L^\beta(\pi, \bar{\lambda})$ denote the Lagrangian

$$L^\beta(\pi, \bar{\lambda}) = J^\beta(\pi) + \sum_{k=1}^{N} \lambda_k G_k^\beta(\pi). \tag{4}$$

For a relaxed MDP problem for which the single-stage cost is $c(i, a) + \sum_{k=1}^{N} \lambda_k (g_k(i, a) - \alpha_k)$ in state i when the action chosen is a, the Bellman equation for optimality corresponds to

$$V^{*,\bar{\lambda}}(i) = \min_a \left(c(i, a) + \sum_{k=1}^{N} \lambda_k (g_k(i, a) - \alpha_k) + \gamma \sum_{j\in S} p(i, j, a) V^{*,\bar{\lambda}}(j) \right), \tag{5}$$

for all $i \in S$, where $V^{*,\bar{\lambda}}(\cdot)$ denotes the value function for a given vector $\bar{\lambda}$ of Lagrange parameters. Further, let $V^{\pi,\bar{\lambda}}(\cdot)$ denote the value function under a given SRP π and Lagrange parameters $\bar{\lambda}$. The Poisson equation in this case corresponds to

$$V^{\pi,\bar{\lambda}}(i) = \sum_{a\in A(i)} \pi(i, a) \left(c(i, a) + \sum_{k=1}^{N} \lambda_k (g_k(i, a) - \alpha_k) + \gamma \sum_{j\in S} p(i, j, a) V^{\pi,\bar{\lambda}}(j) \right), \tag{6}$$

$\forall i \in S$. The solutions $V^{*,\bar{\lambda}}(\cdot)$ and $V^{\pi,\bar{\lambda}}(\cdot)$ to (5) and (6) respectively can both be seen to be unique [4,5].

In what follows, we shall restrict our attention to SRPs π that depend on a parameter $\theta \stackrel{\triangle}{=} (\theta_1, \ldots, \theta_d)^\top$ and consider an analogous problem of finding the optimum $(\theta, \bar{\lambda})$-tuple. Let $\{\pi^\theta(i, a),\ i \in S,\ a \in A(i),\ \theta \in C \subset \mathcal{R}^d\}$ denote the parameterized class of SRP. Here, the set C in which θ takes values is assumed to be a compact and convex subset of $\mathcal{R}^d$. From now on, π itself will represent the parameterized SRP π^θ. The following is a standard requirement in policy gradient methods.