Systems & Control Letters 92 (2016) 46–51

Contents lists available at ScienceDirect

Systems & Control Letters

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

A constrained optimization perspective on actor–critic algorithms and application to network routing

Prashanth L.A. [a,*], Prasad H.L. [b], Shalabh Bhatnagar [c], Prakash Chandra [d]

[a] Institute for Systems Research, University of Maryland, College Park, United States
[b] Astrome Technologies Pvt Ltd, Bangalore, India
[c] Department of Computer Science and Automation, Indian Institute of Science, Bangalore, India
[d] System Sciences and Automation, Indian Institute of Science, Bangalore, India

ARTICLE INFO

Article history:
Received 27 July 2015
Received in revised form
25 February 2016
Accepted 29 February 2016
Available online 5 May 2016

Keywords:
Actor–critic algorithm
Reinforcement learning
Constrained optimization

ABSTRACT

We propose a novel actor–critic algorithm with guaranteed convergence to an optimal policy for a discounted reward Markov decision process. The actor incorporates a descent direction that is motivated by the solution of a certain non-linear optimization problem. We also discuss an extension to incorporate function approximation and demonstrate the practicality of our algorithms on a network routing application.

1. Introduction

We consider a discounted MDP with state space $\mathcal{S}$, action space $\mathcal{A}$, both assumed to be finite. A randomized policy π specifies how actions are chosen, i.e., $\pi(s)$, for any $s \in \mathcal{S}$ is a distribution over the actions $\mathcal{A}$. The objective is to find the optimal policy π^* that is defined as follows:

$$\pi^*(s) = \operatorname*{argmax}_{\pi\in\Pi} \left\{ v^\pi(s) := E\left[\sum_n \beta^n \sum_{a\in\mathcal{A}(s_n)} r(s_n, a) \times \pi(s_n, a)|s_0 = s\right]\right\}, \quad (1)$$

where $r(s, a)$ is the instantaneous reward obtained in state s upon choosing action a, $\beta \in (0, 1)$ is the discount factor and Π is the set of all admissible policies. We shall use $v^*(=v^{\pi^*})$ to denote the optimal value function.

Actor–critic algorithms (cf. [1–3]) are popular stochastic approximation variants of the well-known policy iteration procedure for solving (1). The *critic* recursion provides estimates of the value function using the well-known temporal-difference (TD) algorithm, while the *actor* recursion performs a gradient search over the policy space. We propose an actor–critic algorithm with a novel descent direction for the actor recursion. The novelty of our approach is that we can motivate the actor-recursion in the following manner: the descent direction for the actor update is such that it (globally) minimizes the objective of a non-linear optimization problem, whose minima coincide with the optimal policy π^*. This descent direction is similar to that used in Algorithm 2 in [1], except that we use a different exponent for the policy and a similar interpretation can be used to explain Algorithm 2 (and also 5) of [1]. Using multi-timescale stochastic approximation, we provide global convergence guarantees for our algorithm.

While the proposed algorithm is for the case of full state representations, we also briefly discuss a function approximation variant of the same. Further, we conduct numerical experiments on a shortest-path network problem. From the results, we observe that our actor–critic algorithm performs on par with the well-known Q-learning algorithm on a smaller-sized network, while on a larger-sized network, the function approximation variant of our algorithm does better than the algorithm in [4].

* Corresponding author.
E-mail addresses: prashla@isr.umd.edu (Prashanth L.A.), prasad@astrome.co (Prasad H.L.), shalabh@csa.iisc.ernet.in (S. Bhatnagar), pchandra@ee.iisc.ernet.in (P. Chandra).

http://dx.doi.org/10.1016/j.sysconle.2016.02.020
0167-6911/

2. The non-linear optimization problem

With an objective of finding the optimal value and policy tuple, we formulate the following problem:

$$\left.\begin{array}{l}\min\limits_{v\in\mathbb{R}^{|\mathcal{S}|}}\min\limits_{\pi\in\Pi}\left(J(v,\pi):=\sum\limits_{s\in\mathcal{S}}\sum\limits_{a\in\mathcal{A}}\pi(s,a)\big[v(s)-Q(s,a)\big]\right)\\ \text{s.t.}\quad \forall s\in\mathcal{S},a\in\mathcal{A}\\ \text{(a)}\ \pi(s,a)\geq 0,\qquad \text{(b)}\ \sum\limits_{a\in\mathcal{A}}\pi(s,a)=1,\quad \text{and}\\ \text{(c)}\ g(s,a)\leq 0.\end{array}\right\}. \tag{2}$$

In the above, $g(s,a):=Q(s,a)-v(s)$, with $Q(s,a):=r(s,a)+\beta\sum_{s'}p(s'|s,a)v(s')$. Here $p(s'|s,a)$ denotes the probability of a transition from state s to s' upon choosing action a.

The objective in (2) is to ensure that there is no Bellman error, i.e., the value estimates v are correct for the policy π. The constraints (2)(a)–(2)(b) ensure that π is a distribution, while the constraint (2)(c) is a proxy for the max in (1). Notice that the non-linear problem (2) has a quadratic objective and linear constraints.

From the definition of π^*, it is easy to infer the following claim:

Theorem 1. *Let $g^*(s,a):=Q^*(s,a)-v^*(s)$, with $Q^*(s,a):=r(s,a)+\beta\sum_{s'}p(s'|s,a)v^*(s')$, $\forall s\in\mathcal{S}$, $a\in\mathcal{A}$. Then,*

(i) *Any feasible (v^*,π^*) is optimal in the sense of (1) if and only if $J(v^*,\pi^*)=0$.*
(ii) *π^* is an optimal policy if and only if $\pi^*(s,a)g^*(s,a)=0$, $\forall a\in\mathcal{A}$, $s\in\mathcal{S}$.*

3. Descent direction

Proposition 1. *For the objective in (2), the direction $\sqrt{\pi(s,a)}g(s,a)$ is a non-ascent and in particular, a descent direction along $\pi(s,a)$ if $\sqrt{\pi(s,a)}g(s,a)\neq 0$, for all $s\in\mathcal{S}$, $a\in\mathcal{A}$.*

Proof. Consider any action $a\in\mathcal{A}$ for some $s\in\mathcal{S}$. We show that $\sqrt{\pi(s,a)}g(s,a)$ is a descent direction by the following Taylor series argument. Let

$$\hat{\pi}(s,a)=\pi(s,a)+\delta\sqrt{\pi(s,a)}g(s,a),$$

for a small $\delta>0$. We define $\hat{\pi}$ to be the same as π except with the probability of picking action a in state $s\in\mathcal{S}$ being changed to $\hat{\pi}(s,a)$ (and the rest staying the same). Then by Taylor's expansion of $J(\pi)$ up to the first order term, we have that

$$J(v,\hat{\pi})=J(v,\pi)+\delta\sqrt{\pi(s,a)}g(s,a)\frac{\partial J(v,\pi)}{\partial\pi(s,a)}.$$

Note that higher order terms are all zero since $J(v,\pi)$ is linear in π. It should be easy to see from definition of the objective that $\frac{\partial J(v,\pi)}{\partial\pi(s,a)}=-g(s,a)$. So,

$$J(v,\hat{\pi})=J(v,\pi)-\delta\sqrt{\pi(s,a)}(g(s,a))^2.$$

Thus, for $a\in\mathcal{A}$ and $s\in\mathcal{S}$ where $\pi(s,a)>0$ and $g(s,a)\neq 0$, $J(v,\hat{\pi})<J(v,\pi)$, while when $\sqrt{\pi(s,a)}g(s,a)=0$, $J(v,\hat{\pi})=J(v,\pi)$. □

The next section utilizes the descent direction to derive an actor–critic algorithm.

4. The actor–critic algorithm

Combining the descent procedure in π from the previous section, with a *TD*(0) [5] type update for the value function v on a faster time-scale, we have the following update scheme:

$$\begin{aligned}&\textbf{Q-Value}:\ Q_n(s,a)=r(s,a)+\beta v_n(s'),\\ &\textbf{TD Error}:\ g_n(s,a)=Q_n(s,a)-v_n(s),\\ &\textbf{Critic}:\ v_{n+1}(s)=v_n(s)+c(n)g_n(s,a),\\ &\textbf{Actor}:\ \pi_{n+1}(s,a)=\Gamma\left(\pi_n(s,a)+b(n)\sqrt{\pi_n(s,a)}g_n(s,a)\right).\end{aligned} \tag{3}$$

In the above, Γ is a projection operator that ensures that the updates to π stay within the simplex $\mathcal{D}=\{(x_1,\ldots,x_q)\mid x_i\geq 0,\forall i=1,\ldots,q,\sum_{j=1}^q x_j\leq 1\}$, where $q=|\mathcal{A}|$. Further, the step-sizes $b(n)$ and $c(n)$ satisfy

$$\sum_{n=1}^{\infty}c(n)=\sum_{n=1}^{\infty}b(n)=\infty,\qquad \sum_{n=1}^{\infty}\left(c^2(n)+b^2(n)\right)<\infty\quad\text{and}$$
$$b(n)=o(c(n)).$$

Remark 1 (*Connection to Algorithm 2 of [1]*)**.** From Proposition 1, we have that $\sqrt{\pi(s,a)}g(s,a)$ is a descent direction for $\pi(s,a)$. This implies $\pi(s,a)^\alpha\times\sqrt{\pi(s,a)}g(s,a)$ for any $\alpha\geq 0$, is also a descent direction. Hence,

$$\begin{aligned}&\text{a generic update rule for } \pi \text{ is}:\ \pi_{n+1}(s,a)\\ &=\Gamma\left(\pi_n(s,a)+b(n)(\pi_n(s,a))^{\alpha'}g_n(s,a)\right),\quad \text{for any } \alpha'\geq\frac{1}{2}.\end{aligned}$$

The special case of $\alpha'=1$ coincides with the π-recursion in Algorithm 2 of [1].

5. Convergence analysis

For the purpose of analysis, we assume that the underlying Markov chain for any policy $\pi\in\Pi$ is irreducible.

Main result. Let $v^\pi=[I-\beta P_\pi]^{-1}R_\pi$, where $R_\pi=<r(s,\pi),s\in\mathcal{S}>^T$ is the column vector of rewards and $P_\pi=[p(y|s,\pi),s\in\mathcal{S},y\in\mathcal{S}]$ is the transition probability matrix, both for a given π. Consider the ODE:

$$\frac{d\pi(s,a)}{dt}=\bar{\Gamma}\left(\sqrt{\pi(s,a)}g^\pi(s,a)\right),\quad \forall a\in\mathcal{A},s\in\mathcal{S},\text{ where} \tag{4}$$
$$g^\pi(s,a):=r(s,a)+\beta\sum_{y\in U(s)}p(y|s,a)v^\pi(y)-v^\pi(s). \tag{5}$$

In the above, $\bar{\Gamma}$ is a projection operator defined by $\bar{\Gamma}(\epsilon(\pi)):=\lim_{\alpha\downarrow 0}\frac{\Gamma(\pi+\alpha\epsilon(\pi))-\pi}{\alpha}$, for any continuous $\epsilon(\cdot)$.

Theorem 2. *Let K denote the set of all equilibria of the ODE (4), G the set of all feasible points of the problem (2) and $\hat{K}:=K\cap G$. Then, the iterates (v_n,π_n), $n\geq 0$ governed by (3) satisfy*

$$(v_n,\pi_n)\to K^*\ \text{a.s. as}\ n\to\infty,\quad \text{where } K^*=\{(v^*,\pi^*)\mid\pi^*\in\hat{K}\}.$$

The algorithm (3) comprises of updates to v on the faster time-scale and to π on the slower time-scale. Using the theory of two time-scale stochastic approximation [6, Chapter 6], we sketch the convergence of these recursions as well as prove global optimality in the following steps (the reader is referred to the appendix (see Appendix A) for proof details):

Step 1: Critic convergence. We assume π to be time-invariant owing to time-scale separation. Consider the ODE:

$$\frac{dv(s)}{dt}=r(s,\pi)+\beta\sum_{s'\in\mathcal{S}}p(s'|s,\pi)v(y)-v(s),\quad \forall s\in\mathcal{S}, \tag{6}$$

where $r(s,\pi)=\sum_{a\in\mathcal{A}}\pi(s,a)r(s,a)$ and $p(s'|s,\pi)=\sum_{a\in\mathcal{A}}\pi(s,a)p(s'|s,a)$. It is well-known (cf. [7]) that the above ODE has