European Journal of Operational Research 240 (2015) 697–705

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Stochastics and Statistics

Solving average cost Markov decision processes by means of a two-phase time aggregation algorithm

E.F. Arruda [a,*], M.D. Fragoso [b]

[a] Industrial Engineering Program, Alberto Luiz Coimbra Institute – Graduate School and Research in Engineering, Federal University of Rio de Janeiro, Caixa Postal 68507, Rio de Janeiro, RJ 21941-972, Brazil
[b] Center for Systems and Control, National Laboratory for Scientific Computation, Av. Getúlio Vargas, 333, Petrópolis, RJ 25651-075, Brazil

ARTICLE INFO

Article history:
Received 10 August 2012
Accepted 14 August 2014
Available online 26 August 2014

Keywords:
Dynamic programming
Markov decision processes
Embedding
Time aggregation
Stochastic optimal control

ABSTRACT

This paper introduces a two-phase approach to solve average cost Markov decision processes, which is based on state space embedding or time aggregation. In the first phase, time aggregation is applied for policy optimization in a prescribed subset of the state space, and a novel result is applied to expand the evaluation to the whole state space. This evaluation is then used in the second phase in a policy improvement step, and the two phases are then alternated until convergence is attained. Some numerical experiments illustrate the results.

1. Introduction

Recent developments in the theory and simulation techniques for Markov decision processes (MDP) (e.g., Busoniu, Ernst, Schutter, & Babuska, 2010; Cao, Ren, Bhatnagar, Fu, & Marcus, 2002; Chang, Fu, Hu, & Marcus, 2007; Leizarowitz & Shwartz, 2008; Powell, 2007) have led to a growing body of literature on MDP modeling for real world problems (e.g., Anderson, Boulanger, Powell, & Scott, 2011; Arruda & do Val, 2008; Pennesi & Paschalidis, 2010; Zhang & Archibald, 2011). Much of this increased interest is due, in part, to the development of powerful techniques to deal with MDPs of very large dimensions, encompassed in a framework known as approximate dynamic programming (ADP) (Bertsekas & Tsitsiklis, 1996; Powell, 2007; Sutton & Barto, 1998).

In the context of the ADP framework, there is a vast literature covering a variety of techniques, such as heuristic search (Hansen & Zilberstein, 2001) and real-time dynamic programming (Barto, Bradtke, & Singh, 1995; Bonet & Geffner, 2003), which make use of asynchronous updates and heuristic search to accelerate convergence, as well as topological value iteration (Dai & Goldsmith, 2007; Dai, Mausam, & Weld, 2009), that processes information related to the graphical features of MDPs to decide the optimal ordering of the value function updates. Asynchronous updates are also exploited in (Akramizadeh, Afshar, Menhaj, & Jafari, 2011; Moore & Atkeson, 1993), while a sequence of increasingly accurate approximate models is used in (Arruda, Ourique, LaCombe, & Almudevar, 2013).

Among the most popular ADP techniques one finds value function approximation (e.g., Arruda, Fragoso, & do Val, 2011; Boyan & Moore, 1995; Li & Littman, 2010), and simulation coupled with state space reduction (e.g., Arruda & do Val, 2008; Cao et al., 2002; Chang et al., 2007). There is a wide range of theory and applications within the ADP framework covering value function approximation techniques, especially for discounted cost MDP problems (see, e.g., Powell, 2012). In particular, the abstract representation of the value function in terms of algebraic decision diagrams (e.g., Hoey, St-aubin, Hu, & Boutilier, 1999; Joshi & Khardon, 2011; St-aubin, Hoey, & Boutilier, 2000) can be efficiently used to solve some large scale discounted MDPs. While much progress has been made and a few promising directions have been devised (e.g., Arruda et al., 2011; Ormoneit & Sen, 2002; Powell, 2007; Tsitsiklis & Van Roy, 1997), convergence results for general approximation architectures remain to be proved. Moreover, performance bounds for such techniques tend to be very specialized (e.g., Gordon, 1995; Tsitsiklis & Van Roy, 1997; Lin, Hui, Hua-Yong, & Lin-Cheng, 2009).

State space reduction techniques, known as *embedding* or *time aggregation*, can be traced back, in the context of control theory, at least to (Zhang & Yu-Chi, 1991), but we shall be particularly interested in the works of (Cao et al., 2002; Chang et al., 2007;

* Corresponding author. Tel.: +55 21 2562 8255; fax: +55 21 2270 9702.
E-mail addresses: efarruda@po.coppe.ufrj.br (E.F. Arruda), frag@lncc.br (M.D. Fragoso).

http://dx.doi.org/10.1016/j.ejor.2014.08.023
0377-2217/

Leizarowitz & Shwartz, 2008). It is well known that a great advantage of time aggregation is that, unlike state aggregation (e.g. Bertsekas, 2012), it preserves the Markov property. As a result, it can be used to produce an equivalent formulation with reduced state space. In that context, Fainberg (1986) studied the construction of embedded MDP models for the total cost criterion, whereas Leizarowitz and Shwartz (2008) investigated embedding techniques for average cost MDPs. An earlier work, (Cao et al., 2002), investigated embedding in a scenario where the control policy within a certain region of the state space is fixed and focused on reducing the computational burden of the solution procedure. This approach was later extended to deal with a continuous time stochastic control problem (Xu & Cao, 2011), and also inspired further work on algorithms for embedded (*time aggregated*) MDPs e.g., Ren and Krogh (2005), Sun, Zhao, and Luh (2007), Arruda and Fragoso (2011). Similar concepts were applied in the context of discount MDPs (Hauskrecht, Meuleau, Kaelbling, Dean, & Boutilier, 1998), where hierarchical models were employed to decompose the process. A thorough discussion of compact representations for MDPs can be found in (Boutilier, Dean, & Hanks, 1999).

The time aggregation approach, which transforms an MDP into another equivalent MDP with reduced state space, can be of great assistance when one wishes to find approximate solutions in reduced computational time. To accomplish such reduction, one can trade speed for accuracy and specify a priori an *outer policy* that prescribes a pure control action to each state outside of a prescribed region of interest F of the state space S. An appropriately optimized *inner policy* is then obtained and both outer and inner policies are composed to result in a sub-optimal policy over S. This policy minimizes the long term average cost over all control policies that adopt the prescribed *outer policy*. Note that optimality cannot be guaranteed unless the outer policy is optimal, i.e., it is comprised of optimal control actions for every state in $F^c = S \setminus F$. In particular, optimality can be attained for large scale MDPs with a large number of *uncontrollable* states, i.e. states for which only a single control action is available (see Cao et al., 2002).

A distinguishing feature of this paper is that, unlike the current literature in time aggregation, it addresses also the problem of iteratively refining the outer policy. The rationale is simple: to apply time aggregation iteratively, but refining the outer policy at each iteration, until the outer policy converges to an optimal outer policy. At this point, the time aggregation approach is able to retrieve an optimal policy for the original MDP, over the entire state space S. The proposed outer policy refinement routine can be seen as a policy improvement step of the classical policy iteration algorithm e.g., Bertsekas (2012), which makes use of the value function of the latest policy obtained by time aggregation. A novel contribution of this paper is the way we derive this value function, making use of some new results that are introduced in this paper. Firstly, we prove that the value function obtained by the time aggregation algorithm for each state in the subset F is numerically equal to the value function obtained by a classical policy evaluation algorithm for this same state. We then make use of this result to derive the value function for each state in F^c as the value of a classical stochastic shortest path problem starting from this state to reach any state in the target region F.

To sum up, we propose a two-phase time aggregation algorithm to solve MDPs to optimality. The two phases of the algorithm, which are applied successively up to convergence, work as follows: in the first phase, time aggregation is applied for some prescribed outer policy; then in the second phase, a policy improvement step is applied that refines the outer policy. We prove that the proposed algorithm converges monotonically to the optimal policy under general conditions on the structure of the MDP. It is worth pointing out that the proposed approach can be seen as a variation of the classical policy iteration algorithm with a policy search in the subset F at each iteration. The policy search is performed by the time aggregation algorithm, which finds the best possible policy in F given that the policy in F^c is fixed.

This paper is organized as follows. Section 2 presents the studied problem. Section 3 features the time aggregation approach and derives a novel result on the correspondence between the value functions of the embedded MDP and the original MDP, for a fixed control policy. This result is then applied in Section 4 to derive a two phase algorithm for the studied problem. The convergence of the proposed algorithm to the optimal solution is then proved in Section 4.1. Numerical experiments are presented in Section 5 to illustrate the approach, and Section 6 concludes the paper.

2. Preliminaries and the studied problem

Consider a time homogeneous discrete time Markov decision process (MDP) with a finite, possibly very large, state space S. Let $A(i) \in \mathbb{N}$ denote the set of feasible control actions at state i and define $A := \{A(i), i \in S\}$, and suppose that a function $f : S \times A \to \mathbb{R}_+$ represents the one-period cost of the process, where $\mathbb{R}_+$ denotes the set of nonnegative real numbers.

Let $\mathcal{L} : S \to A$ represent a stationary control policy over the state space S, and let $\mathbb{L}$ be the set of all feasible stationary control policies. Under policy $\mathcal{L}$, one selects control action $a = \mathcal{L}(i)$ at each time the controlled process visits state $i \in S$. Following a visit to state $i \in S$, and the application of a control action $a \in A(i)$, the process moves to state $j \in S$ with probability p_{ij}^a. Hence, the evolution of the controlled processes under a control policy $\mathcal{L} \in \mathbb{L}$ is governed by a Markov chain $\{X_t, t \geqslant 0\}$, and the one-period transitions are determined by the transition matrix $P^{\mathcal{L}} = \{p_{ij}^{\mathcal{L}(i)}\}, i, j \in S$. Following Cao et al. (2002), we assume that the controlled process is ergodic under all policies. Assume that the one-period cost function f is a measurable positive real-valued function and let

$$\eta^{\mathcal{L}} = \lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} f(X_k, \mathcal{L}(X_k)) \tag{1}$$

be the long term average cost of the controlled chain. Because the controlled chain is ergodic, this cost is independent of the initial state. The objective of the decision maker is to find the optimal policy $\mathcal{L}^* \in \mathbb{L}$, which satisfies

$$\eta^{\mathcal{L}^*} \leqslant \eta^{\mathcal{L}}, \ \forall \mathcal{L} \in \mathbb{L}. \tag{2}$$

3. Fixing an outer policy: the time aggregation approach

Now let us select a subset $F \subset S$, and define an *outer policy* $\mathcal{L}_{\text{out}} : F^c \to A$ as the set of control actions prescribed by policy $\mathcal{L}$ for all states outside of F, i.e. $\mathcal{L}_{\text{out}} = \{\mathcal{L}(i), i \in F^c\}$, where $F^c \triangleq S \setminus F$ is the complement of F. Similarly, an *inner policy* $\mathcal{L}_{\text{in}} : F \to A, \mathcal{L}_{\text{in}} = \{\mathcal{L}(i), i \in F\}$ denotes the control strategy prescribed by policy $\mathcal{L}$ for the subset F. Clearly, we have $\mathcal{L} = \mathcal{L}_{\text{in}} \cup \mathcal{L}_{\text{out}}$. We let $\mathbb{L}_{\text{in}}$ and $\mathbb{L}_{\text{out}}$ denote the sets of all feasible inner and outer policies, respectively. Fig. 1 illustrates the concepts of inner and outer policies.

Typically the set F is a relatively small subset of S. It may be comprised, for example, of the states which are more important from some control standpoint, or the states that are expected to be visited more frequently under some particular class of control policies. For example, in a storage control problem, one would expect the set F to be comprised of the states that are within some desirable vicinity of the zero-stock state.

Now let us select an outer policy $d : F^c \to A, d \in \mathbb{L}_{\text{out}}$, a priori and define the following problem