European Journal of Operational Research 241 (2015) 85–98

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Decision Support

Approximate dynamic programming for stochastic linear control problems on compact state spaces

Stefan Woerner[a,*], Marco Laumanns[a], Rico Zenklusen[b], Apostolos Fertis[c]

[a] IBM Research, Saeumerstrasse 4, 8803 Rueschlikon, Switzerland
[b] ETH Zurich, Raemistrasse 101, 8092 Zurich, Switzerland
[c] SMA und Partner AG, Gubelstrasse 28, 8050 Zurich, Switzerland

ARTICLE INFO

Article history:
Received 28 November 2013
Accepted 2 August 2014
Available online 10 August 2014

Keywords:
Dynamic programming
Markov processes
Inventory
Dual sourcing
Multiple sourcing

ABSTRACT

This paper addresses Markov Decision Processes over compact state and action spaces. We investigate the special case of linear dynamics and piecewise-linear and convex immediate costs for the average cost criterion. This model is very general and covers many interesting examples, for instance in inventory management. Due to the curse of dimensionality, the problem is intractable and optimal policies usually cannot be computed, not even for instances of moderate size.

We show the existence of optimal policies and of convex and bounded relative value functions that solve the average cost optimality equation under reasonable and easy-to-check assumptions. Based on these insights, we propose an approximate relative value iteration algorithm based on piecewise-linear convex relative value function approximations. Besides computing good policies, the algorithm also provides lower bounds to the optimal average cost, which allow us to bound the optimality gap of any given policy for a given instance.

The algorithm is applied to the well-studied Multiple Sourcing Problem as known from inventory management. Multiple sourcing is known to be a hard problem and usually tackled by parametric heuristics. We analyze several MSP instances with two and more suppliers and compare our results to state-of-the-art heuristics. For the considered scenarios, our policies are always at least as good as the best known heuristic, and strictly better in most cases. Moreover, by using the computed lower bounds we show for all instances that the optimality gap has never exceeded 5%, and that it has been much smaller for most of them.

1. Introduction

In this paper we address Markov Decision Processes (MDPs) with compact state and action space, linear dynamics as well as piecewise-linear and convex immediate costs. We denote this class of MDPs as Stochastic Linear Control Problems (SLCPs). SLCPs are a broad class of problems that represent many practically relevant problems in various areas, for instance in inventory management.

We study existence and characterization policies that are optimal for the average cost criterion. In particular, we develop conditions that imply the existence of a *convex* solution to the Average Cost Optimality Equation (ACOE). Such solutions can be used to characterize optimal policies. In addition, we develop an Approximate Dynamic Programming (ADP) algorithm for SLCPs. The algorithm generates piecewise-linear and convex relative value function approximations, together with a non-decreasing sequence of lower bounds to the optimal average cost. The relative value functions approximation can then be used to derive a good policies. We show the effectiveness of this algorithm on the Multiple Sourcing Problem (MSP) known in inventory management and compare the results to the current best-known heuristics.

The paper is organized as follows. Section 2 surveys the related literature for average-cost MDPs, ADP, and the MSP. In Section 3, we formally introduce SLCPs. Section 4 discusses the related concepts and results of MDP theory. In Section 5, we review and extend results about the existence of optimal policies and relative value functions for general state space MDPs. In particular, we prove the existence of a convex solution to the ACOE under rather mild assumptions that are typically fulfilled in practice. This motivates our approximation algorithm, which is given in Section 6. In Section 7, we apply our algorithms to various MSP instances, compare the results with state-of-the-art heuristics, and discuss the differences.

[*] Corresponding author.
E-mail address: wor@zurich.ibm.com (S. Woerner).

http://dx.doi.org/10.1016/j.ejor.2014.08.003
0377-2217/© 2014 Elsevier B.V. All rights reserved.

2. Related work and our contributions

In this section we review the relevant literature about MDPs, ADP and MSP, and discuss our contributions in the context of existing results and open questions.

2.1. Average cost Markov decision processes

While average cost MDP with finite state and action spaces are well understood (Bertsekas, 2007), much less in known for MDPs with continuous state and action spaces. From the point of view of the present paper, the existing work on continuous state space MDPs can be classified into two groups, distinguished by continuity assumptions on the transition law.

Most authors assume the transition law to be strongly continuous (Hernández-Lerma & Lasserre, 1990; Montes-de Oca & Hernández-Lerma, 1996; Kurano, Nakagami, & Huang, 2000). This simplifies the analysis of the existence of optimal policies and relative value functions. However, in the more general model considered here, we only assume a weakly continuous transition law, so these results cannot be applied.

There are only very few papers that consider weakly continuous transition laws, most importantly Schäl (1993) as well as Hernández-Lerma and Lasserre (2002, chap. 12). Schäl (1993) follows the limiting discount factor approach to establish the existence of optimal policies. We apply these ideas to SLCPs and extend the results to the existence of a convex bounded solution to the ACOE.

Hernández-Lerma and Lasserre (2002, chap. 12) follow the infinite dimensional Linear Programming (LP) approach. They provide conditions that imply existence of optimal policies and relative value functions. However, their results only hold almost everywhere with respect to an invariant measure. The conditions we propose in Section 5 imply that all their results are applicable as well.

Arapostathis, Borkar, Fernández-Gaucherand, Ghosh, and Marcus (1993) provide a general survey about MDPs with the average cost criterion ranging from finite to Borel state and action spaces.

Finally, we also note that convexity of value functions of MDPs has been an important object of study for a long time. However, most of these structural results have been established for finite horizon or discounted cost problems only (Dynkin, 1972; Hinderer, 1984; Hernández-Lerma & Runggaldier, 1994; Hernández-Lerma, Piovesan, & Runggaldier, 1995).

2.2. Approximate dynamic programming

ADP is an active research area with a variety of literature. Standard textbooks include Bertsekas and Tsitsiklis (1996) as well as Powell (2007).

The typical approach in ADP is to approximate an optimal relative value function with a combination of a fixed set of basis functions. Such approaches were shown to work well for some applications (Schweitzer & Seidmann, 1985; De Farias & Van Roy, 2003, 2004; Farias & Van Roy, 2006, chap. 6). However, determining a good set of basis functions is often very difficult and problem-specific.

As we will show in Section 5, there exists a convex solution to the ACOE for SLCPs, therefore we propose to use a piecewise-linear convex approximation of the relative value function instead. Thus, instead of fixing a set of basis functions, we take the maximum of a set of hyperplanes, which we generate in a particular way during the run of the algorithm.

Piecewise-linear convex value function approximations were also considered by Lincoln and Rantzer (2006) and Shapiro (2011). Lincoln and Rantzer (2006) also focus on SLCPs, but instead of the average cost criterion, they consider discounted cost, for which the existence of optimal policies and relative value functions is well established. Lincoln and Rantzer introduce a ‘‘Relaxed Value Iteration’’ and exploit the fact that the Bellman operator can be expressed as a Multi-Parametric LP (MPLP) and thus preserves piecewise-linearity and convexity. In contrast to our approach, Lincoln and Rantzer (2006) have to solve an MPLP in every iteration to obtain the hyperplane representation of the next relative value function, which is a very complex operation (exponential in the worst case and practically solvable only in low dimensions). In order to reduce the complexity of the relative value function, some of the hyperplanes are dropped as long as given deviation bounds are respected. Satisfying the bounds implies performance guarantees for the resulting policies and allows to control the trade-off between computational effort and performance of the policy. However, to get meaningful performance guarantees, the bounds have to be chosen quite tight, which implies that most hyperplanes have to be kept. Unfortunately, solving an MPLP with many hyperplanes is impractical even in smaller dimensions. This drawback might be the reason why this approach has never been applied to problems of realistic size. For more details on MPLPs in control, including finite horizon approximation techniques, we refer to Jones, Baric, and Morari (2007) as well as Jones and Morari (2009).

In this paper, we develop a different approach to cope with the inherent complexity of the exact value iteration step. Instead of solving an MPLP to find all hyperplanes in every step and then dropping some of them, we compute only a subset in the first place. We choose the subsets in such a way that the associated lower bounds are non-decreasing.

Piecewise-linear convex value functions have also been used by Shapiro (2011) to compute a lower approximation of the optimal value function in the Stochastic Dual Dynamic Programming (SDDP) method, which originated in Pereira and Pinto (1991). SDDP is typically applied to the Sample Average Approximation problem. It is designed for finite horizon problems, whereas the technique proposed here targets infinite horizon problems.

2.3. Multiple sourcing

Inventory management with a single supplier is a well studied problem. Scarf (1960) and Iglehart (1963) proved that the optimal policy of an infinite horizon discounted cost problem with deterministic lead times and stationary stochastic demand is an (s, S) policy. Veinott and Wagner (1965) concluded that this holds also for infinite horizon average cost per stage problems.

While single sourcing is well understood, there exist only few results about the structure of optimal policies if multiple suppliers are available. Fukuda (1964) proved that for two suppliers, a dual base stock policy is optimal when no fixed order costs are accounted, lead times are deterministic, and the difference of the lead times of the two suppliers is exactly equal to one. However, an optimal policy for a general MSP is highly state-dependent (Whittemore & Saunders, 1977). Therefore, parametric heuristics and the computation of their optimal parameters is an active field of research.

Minner (2003) provided a detailed review of MSPs and discussed different heuristics and different setups. Scheller-Wolf, Veeraraghavan, and Van Houtum (2006) showed how the optimal parameters of the Single Index Policy (SIP) can be computed for a general dual sourcing problem with deterministic lead times. In Veeraraghavan and Scheller-Wolf (2008), the more complex Dual Index Policy (DIP) was analyzed and a simulation-based approach to determine the optimal parameters was presented, while Arts,