European Journal of Operational Research 000 (2017) 1–15

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Innovative Applications of O.R.

A comparison of Monte Carlo tree search and rolling horizon optimization for large-scale dynamic resource allocation problems

Dimitris Bertsimas[a], J. Daniel Griffith[b], Vishal Gupta[c], Mykel J. Kochenderfer[d], Velibor V. Mišić[e,*]

[a] Sloan School of Management and Operations Research Center, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA
[b] Lincoln Laboratory, Massachusetts Institute of Technology, 244 Wood Street, Lexington, MA 02420, USA
[c] Department of Data Sciences and Operations, Marshall School of Business, University of Southern California, 3670 Trousdale Parkway, Los Angeles, CA 90089, USA
[d] Department of Aeronautics and Astronautics, Stanford University, 496 Lomita Mall, Stanford, CA 94305, USA
[e] Anderson School of Management, University of California, Los Angeles, 110 Westwood Plaza, Los Angeles, CA 90024, USA

ARTICLE INFO

Article history:
Received 7 June 2016
Accepted 15 May 2017
Available online xxx

Keywords:
Dynamic resource allocation
Monte Carlo tree search
Rolling horizon optimization
Wildfire management
Queueing control

ABSTRACT

Dynamic resource allocation (DRA) problems constitute an important class of dynamic stochastic optimization problems that arise in many real-world applications. DRA problems are notoriously difficult to solve since they combine stochastic dynamics with intractably large state and action spaces. Although the artificial intelligence and operations research communities have independently proposed two successful frameworks for solving such problems—Monte Carlo tree search (MCTS) and rolling horizon optimization (RHO), respectively—the relative merits of these two approaches are not well understood. In this paper, we adapt MCTS and RHO to two problems – a problem inspired by tactical wildfire management and a classical problem involving the control of queueing networks – and undertake an extensive computational study comparing the two methods on large scale instances of both problems in terms of both the state and the action spaces. Both methods are able to greatly improve on a baseline, problem-specific heuristic. On smaller instances, the MCTS and RHO approaches perform comparably, but RHO outperforms MCTS as the size of the problem increases for a fixed computational budget.

1. Introduction

Dynamic resource allocation (DRA) problems are problems where one must assign resources to tasks over some finite time horizon. Many important real-world problems can be cast as DRA problems, including applications in air traffic control (Bertsimas & Stock Patterson, 1998), scheduling (Bertsimas, Gupta, & Lulli, 2014) and logistics, transportation and fulfillment (Acimovic & Graves, 2012). DRA problems are notoriously difficult to solve exactly since they typically exhibit stochasticity and extremely large state and action spaces. The artificial intelligence (AI) and operations research (OR) communities have sought more sophisticated techniques for addressing DRA and other dynamic stochastic optimization problems.

Within the AI community, one approach for dynamic stochastic optimization problems that has received increasing attention in the last 15 years is Monte Carlo tree search (MCTS) (Browne et al., 2012; Coulom, 2007). In any dynamic stochastic optimization problem, one can represent the possible trajectories of the system—the state and the action taken at each decision epoch—as a tree, where the root represents the initial state. In MCTS, one iteratively builds an approximation to this tree and uses it to inform the choice of action. MCTS's effectiveness stems from two key features: (1) bandit upper confidence bounds (see Auer, Cesa-Bianchi, & Fischer, 2002; Kocsis & Szepesvári, 2006) can be used to balance exploration and exploitation in learning, and (2) application-specific heuristics and knowledge can be used to customize the base algorithm (Browne et al., 2012). Moreover, MCTS can easily be tailored to a variety of problems. Indeed, the only prerequisite for implementing MCTS is a generative model that, given a state and an action at a given decision epoch, generates a new state for the next epoch. This flexibility makes MCTS particularly attractive as a general purpose methodology.

Most importantly, MCTS has been extremely successful in a number of applications, particularly in designing expert computer

* Corresponding author.
E-mail addresses: dbertsim@mit.edu (D. Bertsimas), dan.griffith@ll.mit.edu (J.D. Griffith), guptavis@usc.edu (V. Gupta), mykel@stanford.edu (M.J. Kochenderfer), velibor.misic@anderson.ucla.edu (V.V. Mišić).

http://dx.doi.org/10.1016/j.ejor.2017.05.032
0377-2217/

players for difficult games such as Go (Enzenberger, Muller, Arneson, Segal, 2010; Gelly & Silver, 2011), Hex (Arneson, Hayward, & Henderson, 2010), Kriegspiel (Ciancarini & Favini, 2010), and Poker (Rubin & Watson, 2011). Although MCTS is one of the top-performing algorithms for this class of games, games like Go and Hex are qualitatively different from DRAs: unlike typical DRA problems, the state of these games does not evolve stochastically, and the size of the feasible action space is often much smaller. For example, in the Go instances of Gelly and Silver (2011), the action branching factor is at most 81, whereas in one DRA instance we consider, a typical branching factor is approximately 230 million (cf. Eq. (12)). While MCTS has been applied to probabilistic problems (Eyerich, Keller, & Helmert, 2010) and problems with large action spaces (Couëtoux, Hoock, Sokolovska, Teytaud, & Bonnard, 2011), there is relatively little experience with MCTS in DRA-like problems.

On the other hand, within the OR community, the study of DRAs has proceeded along different lines. A prominent stream of research is based upon mathematical optimization (MO). In contrast to MCTS which only requires access to a generative model, MO approaches model the dynamics of the system *explicitly* via a constrained optimization problem. The solution to this optimization problem then yields a control policy for the system. We consider a specific MO-based approach that is sometimes called *rolling horizon optimization* (RHO). Specifically, we replace uncertain parameters in a MO formulation with their expected values and periodically re-solve the formulation for an updated policy as the true system evolves. This paradigm goes by many other names such as fluid approximation, certainty equivalent control or model predictive control. It is known to have excellent practical performance in applications like queueing (Avram, Bertsimas, & Ricard, 1995) and network revenue management (Ciocan & Farias, 2012), and in some special cases, also enjoys strong theoretical guarantees (e.g., Ciocan & Farias, 2012; Gallego & van Ryzin, 1994).

The widespread use and success of RHO approaches for DRAs contrasts strongly with a lack of computational experience with MCTS for DRAs. Furthermore, the two methods differ philosophically. MCTS involves directly simulating the *true* system and efficiently searching through the tree of state-action trajectories. In contrast, RHO involves first constructing an *approximation* of the true system and then solving an optimization problem based on this approximation to determine a policy; this policy is generally not guaranteed to be optimal for the true system. MCTS and RHO also differ in their informational requirements. MCTS only requires a generative model for simulating transitions, and one can interact with this model in a "black-box" fashion, without being able to precisely and compactly describe its dynamics. On the other hand, RHO requires one to know something about the dynamics of the system in order to specify the underlying MO model.

In this paper, we aim to understand the relative merits of MCTS and RHO by applying them to two challenging DRA problems:

1. **Tactical wildfire management.** The decision maker controls the spread of a fire on a discrete grid (representing a wildland area) by deploying suppression resources to cells on this grid. This problem is computationally intractable: each cell on the grid may be burning or not burning, resulting in an exponentially large state space, while the allocation decision involves choosing a subset of the burning cells to extinguish, resulting in an exponentially large action space. This problem is also of practical importance: for example, in the US, increasing wildfire severity has resulted in increased government spending on wildfire management, amounting to $3.5 billion in 2013 (Bracmort, 2013).
2. **Queueing network control.** The decision maker controls a network of servers that serve jobs of different classes and at each decision epoch, must decide which job class each server should process so as to minimize the average long-run number of jobs in the system. The system state is encoded by the number of jobs of each class and is exponential in the number of job classes. The action is to decide which class each server should service, and is also exponential in the number of servers. The problem thus constitutes a challenging DRA. At the same time, queueing networks arise in many domains such as manufacturing (Buzacott & Shanthikumar, 1993), computer systems (Harchol-Balter, 2013) and neuroscience (Mišić, Sporns, & McIntosh, 2014).

We summarize our contributions as follows:

1. We develop an MCTS approach for the tactical wildfire management problem and the queueing network control problem. To the best of our knowledge, this represents the first application of MCTS to challenging DRA problems motivated by real-world applications. Towards this end, we combine a number of classical features of MCTS, such as bandit upper confidence bounds, with new features such as double progressive widening (Couëtoux, Hoock, Sokolovska, Teytaud, & Bonnard, 2011). For the wildfire problem, we also propose a novel action generation approach to cope with the size of the state and action spaces of the DRA.
2. We propose an RHO approach based on a mixed-integer optimization (MIO) model of the wildfire problem that approximates the original discrete and stochastic elements of the MDP by suitable continuous and deterministic counterparts. This particular formulation incorporates elements of a linear dynamical system which may be of independent interest in other DRA problems. For the queueing control problem, we apply an existing fluid optimization approach (Avram, Bertsimas, & Ricard, 1995).
3. Through extensive computational experiments in both problems, we show the following:
   (a) MCTS and RHO both produce high-quality solutions, generally performing as well or better than a baseline heuristic. MCTS and RHO perform comparably when the problem instance is small. With a fixed computational budget, however, the RHO approach begins to outperform the MCTS approach as the size of the problem instance grows, either in state space or action space. Indeed, in the wildfire problem, MCTS can begin to perform worse than our baseline heuristic when the action space grows very large; the RHO approach, by comparison, still performs quite well. Similarly, for queueing network control, MCTS with an informed rollout policy (the $c\mu$ rule) often performs worse than the same rollout policy on its own for larger queueing systems.
   (b) The choice of hyperparameters in MCTS—such as the exploration bonus and the progressive widening parameters—can significantly affect its overall performance. The interdependence between these parameters is complex and in general, they cannot be chosen independently. Some care must be taken to appropriately tune the algorithm to a specific DRA.

In tactical wildfire management, there have been a number of empirically validated, deterministic models for wildfire spread proposed (e.g., Tymstra, Bryce, Wotton, Taylor, & Armitage, 2010). However, there have been fewer works that incorporate the stochastic elements of fire spread (Boychuck, Braun, Kulperger, Krougly, & Stanford, 2008; Fried, Gilless, & Spero, 2006; Ntaimo, Arrubla, Stripling, Young, & Spencer, 2012). Most works focus on developing simulation models; few consider the associated problem of managing suppression resources. A notable exception is the research stream of Hu and Ntaimo (2009) and Ntaimo et al. (2013), which considers the problem of determining how many and what

Please cite this article as: D. Bertsimas et al., A comparison of Monte Carlo tree search and rolling horizon optimization for large-scale dynamic resource allocation problems, European Journal of Operational Research (2017), http://dx.doi.org/10.1016/j.ejor.2017.05.032