Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Decomposable Markov Decision Processes: A Fluid Optimization Approach

Dimitris Bertsimas, Velibor V. Mišić

To cite this article:
Dimitris Bertsimas, Velibor V. Mišić (2016) Decomposable Markov Decision Processes: A Fluid Optimization Approach. Operations Research

Published online in Articles in Advance 13 Oct 2016

. http://dx.doi.org/10.1287/opre.2016.1531

OPERATIONS RESEARCH
Articles in Advance, pp. 1–19
ISSN 0030-364X (print) | ISSN 1526-5463 (online)

https://doi.org/10.1287/opre.2016.1531

Decomposable Markov Decision Processes: A Fluid Optimization Approach

Dimitris Bertsimas
Sloan School of Management and Operations Research Center, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, dbertsim@mit.edu

Velibor V. Mišić
Anderson School of Management, University of California, Los Angeles, Los Angeles, California 90095, velibor.misic@anderson.ucla.edu

Decomposable Markov decision processes (MDPs) are problems where the stochastic system can be decomposed into multiple individual components. Although such MDPs arise naturally in many practical applications, they are often difficult to solve exactly due to the enormous size of the state space of the complete system, which grows exponentially with the number of components. In this paper, we propose an approximate solution approach to decomposable MDPs that is based on re-solving a fluid linear optimization formulation of the problem at each decision epoch. This formulation tractably approximates the problem by modeling transition behavior at the level of the individual components rather than the complete system. We prove that our fluid formulation provides a tighter bound on the optimal value function than three state-of-the-art formulations: the approximate linear optimization formulation, the classical Lagrangian relaxation formulation, and a novel, alternate Lagrangian relaxation that is based on relaxing an action consistency constraint. We provide a numerical demonstration of the effectiveness of the approach in the area of multiarmed bandit problems, where we show that our approach provides near optimal performance and outperforms state-of-the-art algorithms.

Keywords: dynamic programming: optimal control; probability: Markov processes; programming: linear: applications.
Area of review: Stochastic Models.
History: Received February 2014; revisions received January 2015, September 2015, March 2016; accepted May 2016.
Published online in *Articles in Advance* October 13, 2016.

1. Introduction

Many real-world problems involving the control of a stochastic system can be modeled as Markov decision processes (MDPs). In a typical MDP, the system begins in a certain state $\mathbf{s}$ in some state space $\mathcal{S}$. The decision maker selects an action a from some action space $\mathcal{A}$. The system then transitions randomly to a new state $\mathbf{s}'$ with probability $p_a(\mathbf{s}, \mathbf{s}')$, and the decision maker garners some reward $g_a(\mathbf{s})$. Once the system is in this new state $\mathbf{s}'$, the decision maker once again selects a new action, leading to additional reward and causing the system to transition again. In the most basic form of the problem, the decision maker needs to make decisions over an infinite horizon, and the rewards accrued over this infinite horizon are discounted in time according to a discount factor $\beta \in (0, 1)$. The goal of the decision maker, then, is to find a policy π that prescribes an action $\pi(\mathbf{s})$ for each state $\mathbf{s}$ so as to maximize the expected total discounted reward

$$\mathbb{E}\left[\sum_{t=1}^{\infty} \beta^{t-1} g_{\pi(\mathbf{s}(t))}(\mathbf{s}(t))\right],$$

where $\mathbf{s}(t)$ is the random variable representing the state at time t, when the system is operated according to policy π. In other types of problems, the decision maker may only be making decisions over a finite time horizon; in those problems, the policy prescribing the action to take may depend not only on the state, but also on the time at which the decision is being made.

Although problems that are represented in this form can in principle be solved exactly with dynamic programming, this is often practically impossible. Exact methods based on dynamic programming require one to compute the optimal value function J^*, which maps states in the state space $\mathcal{S}$ to the optimal expected discounted reward when the system starts in that state. For many problems of practical interest, the state space $\mathcal{S}$ is so large that operating on, or even storing, the value function J^* becomes computationally infeasible. This is what is often referred to in the dynamic programming and MDP literature as the *curse of dimensionality* (Bellman 1961).

Where does the curse of dimensionality come from? That is, why is it that practical MDPs often have prohibitively large state spaces? For many practical problems, the system that is being modeled is often not a single, atomic system, but rather consists of a collection of smaller sub-systems or components. Mathematically, consider a system consisting of M components, where each component $m \in \{1, \ldots, M\}$ is endowed with a state s^m from an ambient state space $\mathcal{S}^m$. To represent the *complete* system, we must represent the state $\mathbf{s}$ as an M-tuple of the component states (i.e., $\mathbf{s} = (s^1, \ldots, s^M)$), and as a result, the state space of the complete system becomes the Cartesian product of the component

state spaces (i.e., $\mathcal{S} = \mathcal{S}^1 \times \cdots \times \mathcal{S}^M$). As the number of components M grows, the size of the state space of the complete system grows in an exponential fashion.

At the same time, the data of such systems are often not presented to us in terms of the complete system state. The probabilistic dynamics induced by each candidate action in $\mathcal{A}$ may be naturally expressed in terms of individual components or small combinations (e.g., pairs) of components. Similarly, the reward structure of the problem does not need to be specified in terms of the complete system state but can be specified in terms of the component states. In the remainder of the paper, we will refer to MDPs where the probabilistic dynamics and reward structure can be expressed in terms of the component states as *decomposable* MDPs.

Many practically relevant MDPs can be modeled as decomposable MDPs. One major class of MDPs that falls into the decomposable MDP framework is the class of multiarmed bandit problems. In the multiarmed bandit problem, the decision maker is presented with M machines ("bandits"), where each bandit m is initially in some state s^m from its state space $\mathcal{S}^m$. At each point in time, one of the bandits may be activated, in which case the chosen bandit changes state probabilistically and the decision maker earns some reward. The problem is then to decide, at each point in time, given the state of all of the bandits, which bandit to activate so as to maximize the total expected long term reward. In the basic form of the problem—the *regular* multiarmed bandit problem—when a bandit m is activated, the inactive bandits do not change state. In the *restless* bandit problem, the inactive bandits can also change state passively and the decision maker may earn a passive reward from bandits that are not activated. The multiarmed bandit, by its definition, is a decomposable MDP: the state space of the ensemble of bandits is the product of the state spaces of the individual bandits, and the probability transition structure is specified at the level of each bandit.

In this paper, we propose a new fluid optimization approach for approximately solving decomposable MDPs. The centerpiece of our approach is a linear optimization (LO) model in which, in its most basic form, the decision variables represent the marginal probabilities of each individual component being in each of its possible states and the action taken at a particular time, and the main constraints are conservation constraints that govern how these marginal probabilities "flow" from component states at time t to new states at $t+1$ under different action (hence the name *fluid*). The idea of the formulation is to approximate the behavior of the system when it is controlled optimally. The formulation achieves this in a tractable way by exploiting the decomposable nature of the problem: rather than modeling the precise transition behavior of the system at the level of tuples of component states, it models the macroscopic transitions of the system at the level of the individual components and their states. The optimal solution of the formulation, when it includes constraints that model the complete system starting in a certain state, can be used to derive an action for the state. In this way, the formulation leads naturally to a simple heuristic for solving the MDP.

Our contributions are as follows:

1. We propose a novel LO formulation for approximately modeling decomposable MDPs and an associated heuristic for solving the MDP. The formulation is tractable since the number of variables scales linearly with the number of individual components, as opposed to the exponential scaling that is characteristic of dynamic programming. We show that this formulation provides an upper bound on the optimal value of the MDP and provide idealized conditions under which our fluid formulation-based heuristic is optimal. We discuss how this basic, "first-order" formulation that models individual components can be extended to "higher-order" formulations that model combinations of components (e.g., a second-order formulation that models transitions of pairs of components). We also discuss how the basic formulation can be extended to address finite-horizon, time-dependent problems.

2. We theoretically compare our fluid formulation to three alternative proposals. In particular, we show that a finite version of our formulation that models the evolution of the system over a horizon of T periods provides provably tighter bounds on the optimal value function than three state-of-the-art formulations: the approximate linear optimization (ALO) formulation of de Farias and Van Roy (2003), the classical Lagrangian relaxation (CLR) formulation of Adelman and Mersereau (2008), and an alternate Lagrangian relaxation (ALR) that involves relaxing an action consistency constraint. The latter alternate Lagrangian relaxation is a novel formulation that is equivalent to the ALO and is of independent interest. Moreover, the fluid bound is nonincreasing with the time horizon T. Letting $J^*(\mathbf{s})$ denote the optimal value function at the state $\mathbf{s}$, $Z_T^*(\mathbf{s})$ denote the objective value of the fluid formulation with horizon T at $\mathbf{s}$, and $Z_{\text{ALO}}^*(\mathbf{s})$, $Z_{\text{ALR}}^*(\mathbf{s})$, and $Z_{\text{CLR}}^*(\mathbf{s})$ denote the objective values of the ALO, ALR and CLR formulations at $\mathbf{s}$, respectively, our results can be summarized in the following statement, which holds for any $T \in \{1, 2, \ldots\}$:

$$\begin{aligned} J^*(\mathbf{s}) \leqslant Z_T^*(\mathbf{s}) \leqslant \cdots \leqslant Z_2^*(\mathbf{s}) \\ \leqslant Z_1^*(\mathbf{s}) \leqslant Z_{\text{ALO}}^*(\mathbf{s}) = Z_{\text{ALR}}^*(\mathbf{s}) \leqslant Z_{\text{CLR}}^*(\mathbf{s}). \end{aligned}$$

In this way, our paper contributes to the overall understanding of the fluid approach and all previous proposals in a unified framework.

3. We demonstrate the effectiveness of our approach computationally on multiarmed bandit problems. We consider regular bandit problems (inactive bandits do not change state) and restless bandit problems (inactive bandits may change state). We show that bounds from our approach can be substantially tighter than those from state-of-the-art approaches, and that the performance of our fluid policy is near optimal and outperforms policies from state-of-the-art approaches.

The rest of this paper is organized as follows. In Section 2, we review the extant body of research related to this paper. In Section 3, we define the decomposable MDP and present our infinite LO fluid formulation. We prove a number of properties of this formulation, and motivated by the properties of this infinite LO formulation, we present a heuristic policy for generating actions at each decision epoch based on a finite version of this formulation. In Section 4, we compare the finite fluid formulation to the ALO formulation of de Farias and Van Roy (2003), the classical Lagrangian relaxation approach of Adelman and Mersereau (2008), and the alternate Lagrangian relaxation. In Section 5, we apply our framework to the multiarmed bandit problem and provide computational evidence for the strength of our approach in this class of problems. Finally, in Section 6, we state the conclusions of our study and offer some directions for future work.

2. Literature Review

MDPs have a long history, tracing back to the work of Bellman (1957) in the 1950s. The importance and significance of this area in the field of operations research is underscored both by the number of research papers that have been written in this area, as well as the numerous books written on the subject (examples include Howard 1971, Heyman and Sobel 1984, Puterman 1994, and Bertsekas 1995).

While MDPs can be solved exactly through methods such as value iteration, policy iteration, and the LO approach (see, e.g., Puterman 1994), these approaches become intractable in high-dimensional problems. As a result, much research has been conducted in the area of *approximate* dynamic programming (ADP). The interested reader is referred to Van Roy (2002) for a brief overview, and to Bertsekas and Tsitsiklis (1996) and Powell (2007) for more comprehensive treatments of the topic. The goal of ADP is to find an approximation to the true value function. By then applying the policy that is greedy with respect to this approximate value function, one hopes to achieve performance that is close to that of the true optimal policy.

Within the ADP literature, our work is most closely related to the approximate linear optimization (ALO) approach of de Farias and Van Roy (2003) and the Lagrangian relaxation approach of Adelman and Mersereau (2008). In the ALO approach to ADP, one approximates the value function as the weighted sum of a collection of basis functions and solves the LO formulation of the MDP with this approximate value function in place of the true value function. By doing so, the number of variables in the problem is significantly reduced, leading to a more tractable problem. In many applications, one can exploit the decomposable nature of the problem in selecting a basis function architecture: for example, in de Farias and Van Roy (2003) the approach is applied to a queueing control example, where the value function is approximated as a linear combination of all polynomials up to degree 2 of the individual queue lengths of the system. On the other hand, Hawkins (2003) and Adelman and Mersereau (2008) study MDPs where the problem can be viewed as a collection of subproblems, and the action that can be taken in each subproblem is constrained by a global linking constraint that couples the subproblems together. By dualizing the linking constraint, the complete problem decomposes along the subproblems, leading to an optimization problem that is significantly simpler than the exact LO model of the MDP. By solving this optimization problem, one obtains an upper bound on the optimal value at a given state, as well as a value function approximation.

Our fluid approach is closely related to the Lagrangian relaxation approach and builds on it in two important ways. First, we delineate two different types of Lagrangian relaxations: the "classical" Lagrangian relaxation, where the action space is implicitly defined by a coupling constraint, and a novel, "alternate" Lagrangian relaxation where the components are coupled by an "action consistency" constraint (the action taken in each component must be the same). Our formulation is not related to the former classical formulation, but to the latter alternate formulation. This alternate Lagrangian relaxation is significant because it extends the scope of the Lagrangian relaxation approach to problems that do not have a decomposable system action space. Furthermore, it turns out that this alternate Lagrangian relaxation is actually *equivalent* to the ALO: the two models lead to the same bound on the true optimal value function and the same value function approximation (Theorem 2). In contrast, for the classical Lagrangian relaxation, one can only show that the ALO bound is at least as tight (Adelman and Mersereau 2008), and one can find simple examples where the Lagrangian bound can be significantly worse than the ALO bound.

The second way in which our approach builds on the Lagrangian relaxation approach is through its view of time. Our formulation first models the state of each component separately at each decision epoch over a finite horizon before aggregating them over the remaining infinite horizon, whereas the formulation of Adelman and Mersereau (2008) aggregates them over the entire infinite horizon. While this may appear to be a superficial difference, it turns out to be rather significant because it allows us to prove that the fluid bound is at least as tight as the classical and alternate Lagrangian relaxation bounds (parts (a) and (c) of Theorem 3). As we will see in Section 5, the difference in the bounds and the associated performance can be considerable.

With regard to the ALO, the ALO and our fluid model differ in tractability. In particular, the size (number of variables and number of constraints) of our fluid model scales linearly in the number of components and actions. In contrast, in the ALO, while the number of variables may scale linearly in the number of components, the number of constraints still scales linearly with the number of *system* states, as in the exact LO model of the MDP. This necessitates the use of additional techniques to solve the problem, such as constraint sampling (de Farias and Van Roy 2004). Moreover, as

stated above, it turns out that when one uses a componentwise approximation architecture for the ALO, it is equivalent to the alternate Lagrangian relaxation formulation described above. Due to this equivalence, we are able to assert that our fluid model leads to better bounds than the ALO, and through our numerical results, that our fluid approach leads to better performance than the ALO approach.

Outside of ADP, many approximate approaches to stochastic control problems also exploit decomposability. One salient example of this is the performance region approach to stochastic scheduling. In this approach, one considers a vector of performance measures of the complete system (for an overview, see Bertsimas 1995). Using the probabilistic dynamics of the system, one can then derive *conservation laws* that constrain the values this vector of performance measures may take. The resulting set is the performance region of the system, over which one can solve an optimization problem to find the best vector of performance measures. It turns out that typically this vector of performance measures is achieved by simple policies or a randomization of simple policies. This approach was introduced by Coffman and Mitrani (1980) for multiclass scheduling in a single-server $M/M/1$ queue and later extended by Federgruen and Groenevelt (1988) and Shanthikumar and Yao (1992) to more general queueing systems. Bertsimas and Niño-Mora (1996) unified this framework and extended it beyond queueing control problems to such problems as the multiarmed bandit problem and branching bandits. Bertsimas and Niño-Mora (2000) later considered a performance region formulation for the restless bandit problem and used it to derive a high-quality heuristic for the restless bandit problem.

Our approach has some conceptual similarities to the performance region approach in the sense that one defines decision variables related to the proportion of time that components of the system are in certain states, imposes constraints that conserve these proportions with each transition, and optimizes an objective over the resulting feasible set. In spite of these commonalities, there are a number of key differences. Many existing performance region formulations, due to the nature of the stochastic system, possess attractive computational and theoretical properties. For example, for systems that satisfy generalized conservation laws, the performance region is an extended polymatroid or contra-polymatroid, and so a linear function can be optimized rapidly using a greedy algorithm, and the extreme points of the performance region correspond to deterministic priority rules (Bertsimas and Niño-Mora 1996). In contrast, our formulation does not appear to possess such special computational structure, and as we discuss in Section 3.3, optimal solutions of our fluid formulation may in general not be achieved by *any* policy, let alone a specific class of policies. At the same time, many extant performance region approaches are fragile, in that they exploit nontrivial properties of the underlying stochastic system and thus cannot be immediately extended to even simple generalizations. An example of this is the formulation of the multiarmed bandit problem in Bertsimas and Niño-Mora (1996), which exploits specific conservation properties of the regular multiarmed bandit problem and cannot be extended to restless bandits, necessitating the authors' exploration of an alternate approach in Bertsimas and Niño-Mora (2000). In contrast, our formulation is insensitive to these types of differences; it does not use any structure of the problem beyond the transition probabilities of individual components or groups of components.

Within the performance region literature, our fluid formulation of the restless bandit problem is similar to the performance region model of Bertsimas and Niño-Mora (2000). This model is actually equivalent to both the classical and the alternate Lagrangian relaxation formulations (Propositions 8 and 9). As a result, our comparison of the fluid formulation with the Lagrangian relaxation formulations allows us to assert that our approach leads to tighter state-wise bounds than the performance region formulation. Moreover, as we will see in Section 5.5, our fluid approach significantly outperforms the associated primal dual heuristic of Bertsimas and Niño-Mora (2000).

The first fluid formulation that we will propose in Section 3.2 is a countably infinite LO (CILO) problem. There exists a rich literature on this class of problems (see, e.g., Anderson and Nash 1987). Within this area, the works of Ghate and Smith (2013) and Lee et al. (2013) directly study MDPs; however, both of these papers specifically study *nonstationary* problems and do not additionally consider decomposability. Although we do not explore the application of the methods and theory from the CILO literature to our setting, we believe that it is an interesting direction for future research.

Lastly, the alternate Lagrangian relaxation we will develop in Section 4 arises by relaxing a certain type of action consistency constraint that requires that the action taken in any two components be equivalent. This bears some resemblance to the technique of "variable splitting" or "operator splitting" that is used in continuous optimization (see, for example, Boyd et al. 2011 and Goldfarb and Ma 2012). The exploration of the connections of the alternate Lagrangian relaxation to splitting-based formulations is an interesting direction for future research.

3. Methodology

We begin in Section 3.1 by defining a general decomposable, infinite horizon MDP. We then present an infinite LO formulation that is related to this MDP in Section 3.2. In Section 3.3, we prove some interesting properties of the formulation and, motivated by one of these properties, propose a solvable finite LO formulation and an associated heuristic in Section 3.4. The proofs of all theoretical results are provided in Section EC.1 of the electronic companion (available as supplemental material at https://doi.org/10.1287/opre.2016.1531).

3.1. Problem Definition

In this section, we define the decomposable MDP for which we will subsequently develop our fluid approach.

Let $\mathcal{S}$ be the state space of the complete system, and assume that the complete system state decomposes into M components, so that the complete system state space can be written as $\mathcal{S} = \mathcal{S}^1 \times \cdots \times \mathcal{S}^M$. Let $\mathcal{A}$ be a finite action space, and assume that any action in $\mathcal{A}$ can be taken at any state in $\mathcal{S}$. We make this assumption for simplicity; our approach can be extended to accommodate component-dependent constraints that restrict which actions can be action when specific components enter specific states (for example, an action a cannot be taken when component m is in state k). Let $\mathbf{s}(t)$ be the random variable that represents the state of the complete system at time t, and let $s^m(t)$ denote the state of component m of the system, so that $\mathbf{s}(t) = (s^1(t), \ldots, s^M(t))$. Let $\pi\colon \{1, 2, \ldots\} \times \mathcal{S} \to \mathcal{A}$ be the policy under which the system is operating, which maps a state $\mathbf{s}(t)$ at time t to an action $\pi(t, \mathbf{s}(t))$ in $\mathcal{A}$. Let $p_a(\mathbf{s}, \bar{\mathbf{s}})$ be the probability of the complete system transitioning from state $\mathbf{s}$ to state $\bar{\mathbf{s}}$ in one step when action $a \in \mathcal{A}$ is taken; i.e.,

$$p_a(\mathbf{s}, \bar{\mathbf{s}}) = \mathbb{P}(\mathbf{s}(t+1) = \bar{\mathbf{s}} \mid \mathbf{s}(t) = \mathbf{s}, \pi(t, \mathbf{s}(t)) = a)$$

for all $t \in \{1, 2, \ldots\}$. Let p^m_{kja} denote the probability of component m transitioning from state k to state j in one step when action $a \in \mathcal{A}$ is taken; i.e.,

$$p^m_{kja} = \mathbb{P}(s^m(t+1) = j \mid s^m(t) = k, \pi(t, \mathbf{s}(t)) = a),$$

for all $t \in \{1, 2, \ldots\}$. We assume that the components are independent, so that $p_a(\mathbf{s}, \bar{\mathbf{s}})$ can be written compactly as

$$p_a(\mathbf{s}, \bar{\mathbf{s}}) = \prod_{m=1}^{M} p^m_{s^m \bar{s}^m a}.$$

Let $g_a(\mathbf{s})$ be the reward associated with taking action a when the system is in state $\mathbf{s}$, and assume that it is additive in the components; that is, it can be written as

$$g_a(\mathbf{s}) = \sum_{m=1}^{M} g^m_{s^m a},$$

where g^m_{ka} is the reward associated with taking action a when component m is in state k. Assume that the system starts in state $\mathbf{s} = (s^1, \ldots, s^M)$. The problem is then to find a policy π that maximizes the expected total discounted reward:

$$\max_{\pi} \mathbb{E}\left[\sum_{t=1}^{\infty}\sum_{m=1}^{M} \beta^{t-1} g^m_{s^m(t),\, \pi(t, \mathbf{s}(t))} \,\middle|\, \mathbf{s}(1) = \mathbf{s}\right]. \tag{1}$$

3.2. Fluid Linear Optimization Formulation

We now consider a fluid formulation of problem (1). We begin by defining, for each $t \in \{1, 2, 3 \ldots\}$, the decision variable $x^m_{ka}(t)$ to be the proportion of time that component $m \in \{1, \ldots, M\}$ is in state $k \in \mathcal{S}^m$ and action $a \in \mathcal{A}$ is taken at time t. For every $t \in \{1, 2, 3, \ldots\}$, we also define the decision variable $A_a(t)$ to be the proportion of time that action $a \in \mathcal{A}$ is taken at time t. As stated in Section 3.1, our data are the discount factor β, the reward g^m_{ka} (the reward accrued by the decision maker when component m is in state k and action a is taken) and the transition probability p^m_{kja} (the probability that component m transitions from state k to state j in one step when action a is taken).

Finally, we assume that the system starts deterministically in a state $\mathbf{s} \in \mathcal{S}$. For convenience, we will define $\alpha^m_k(\mathbf{s})$ as

$$\alpha^m_k(\mathbf{s}) = \begin{cases} 1, & \text{if } s^m = k, \\ 0, & \text{otherwise.} \end{cases}$$

The fluid problem for initial state $\mathbf{s}$ can now be formulated as follows:

$$\underset{\mathbf{x},\,\mathbf{A}}{\text{maximize}} \quad \sum_{t=1}^{\infty}\sum_{m=1}^{M}\sum_{k\in\mathcal{S}^m}\sum_{a\in\mathcal{A}} \beta^{t-1} \cdot g^m_{ka} \cdot x^m_{ka}(t) \tag{2a}$$

$$\text{subject to} \quad \sum_{a\in\mathcal{A}} x^m_{ja}(t) = \sum_{k\in\mathcal{S}^m}\sum_{\tilde{a}\in\mathcal{A}} p^m_{kj\tilde{a}} x^m_{k\tilde{a}}(t-1),$$
$$\forall m \in \{1, \ldots, M\},\ t \in \{2, 3, \ldots\},\ j \in \mathcal{S}^m, \tag{2b}$$

$$\sum_{k\in\mathcal{S}^m} x^m_{ka}(t) = A_a(t),$$
$$\forall m \in \{1, \ldots, M\},\ a \in \mathcal{A},\ t \in \{1, 2, \ldots\}, \tag{2c}$$

$$\sum_{a\in\mathcal{A}} x^m_{ka}(1) = \alpha^m_k(\mathbf{s}),$$
$$\forall m \in \{1, \ldots, M\},\ k \in \mathcal{S}^m, \tag{2d}$$

$$x^m_{ka}(t) \geqslant 0, \quad \forall m \in \{1, \ldots, M\},\ a \in \mathcal{A},$$
$$k \in \mathcal{S}^m,\ t \in \{1, 2, \ldots\}, \tag{2e}$$

$$A_a(t) \geqslant 0, \quad \forall a \in \mathcal{A},\ t \in \{1, 2, \ldots\}. \tag{2f}$$

Constraint (2b) ensures that probability is conserved from time $t-1$ to time t: the left-hand side represents the proportion of time that component m is in state j at time t in terms of the $x^m_{ja}(t)$ variables, while the right-hand side represents the same proportion, only in terms of the $x^m_{k\tilde{a}}(t-1)$ variables, which correspond to time $t-1$. Constraint (2c) ensures that, for each component, the proportion of time that action a is taken in terms of the $x^m_{ka}(t)$ variables is equal to $A_a(t)$ (which is precisely defined as the proportion of time that action a is taken at time t). Thus, the actions $a \in \mathcal{A}$ connect the variables corresponding to the different components. Constraint (2d) ensures that the initial frequency with which each component m is in a state k is exactly $\alpha^m_k(\mathbf{s})$. The remaining two constraints (2e) and (2f) ensure that all of the decision variables are nonnegative, as they represent proportions. Given the definition of $x^m_{ka}(t)$ as the proportion of time that component m is in state k at time t and action a is taken at time t, the objective can therefore be interpreted as the expected discounted long-term reward.

Note that constraints (2b) and (2d), together with the fact that $\sum_{j\in\mathcal{S}^m} p^m_{kja} = 1$ for any m, k, and a, imply that

$\sum_{k\in\mathcal{S}^m}\sum_{a\in\mathcal{A}} x^m_{ka}(t) = 1$ for each m and t. Together with constraint (2c), this also implies that $\sum_{a\in\mathcal{A}} A_a(t) = 1$ for each t.

The following result, which follows by standard arguments in infinite dimensional linear optimization (Romejin et al. 1992; see Section EC.1.1 in the electronic companion), establishes that the infinite horizon problem (2) is well defined.

PROPOSITION 1. *For each* $\mathbf{s}\in\mathcal{S}$, *problem* (2) *has an optimal solution.*

3.3. Properties of the Infinite Fluid LO

We will now develop some theoretical properties of the fluid LO model. Let $(\mathbf{x}(\mathbf{s}), \mathbf{A}(\mathbf{s}))$ and $Z^*(\mathbf{s})$ denote an optimal solution and the optimal objective value, respectively, to problem (2) corresponding to initial state $\mathbf{s}$. Denote by $J^*(\cdot)$ the optimal value function obtained using dynamic programming, that is,

$$J^*(\mathbf{s}) = \max_{\pi} \mathbb{E}\left[\sum_{t=1}^{\infty}\sum_{m=1}^{M} \beta^{t-1} g^m_{s^m(t),\,\pi(t,\mathbf{s}(t))} \,\middle|\, \mathbf{s}(1) = \mathbf{s}\right]$$

for every $\mathbf{s}\in\mathcal{S}$. We then have the following relationship between problem (2) and the optimal value function, whose proof appears as Section EC.1.2 in the electronic companion.

PROPOSITION 2. *For every* $\mathbf{s}\in\mathcal{S}$, $J^*(\mathbf{s}) \leqslant Z^*(\mathbf{s})$.

The idea behind the proof of Proposition 2 is that, by using an optimal policy π^*, it is possible to construct a feasible solution $(\mathbf{x}, \mathbf{A})$ to problem (2) whose objective value is the true optimal value $J^*(\mathbf{s})$. Unfortunately, the opposite inequality does not hold in general; see Section EC.2 of the electronic companion for a counterexample. Let us call the optimal solution $(\mathbf{x}(\mathbf{s}), \mathbf{A}(\mathbf{s}))$ *achievable* if there exists a (possibly nondeterministic and time-varying) policy π such that

$$x^m_{ka}(t,\mathbf{s}) = \mathbb{P}(s^m(t) = k, \pi(t,\mathbf{s}(t)) = a),$$
$$\forall\, m\in\{1,\dots,M\},\ k\in\mathcal{S}^m,\ a\in\mathcal{A},\ t\in\{1,2,\dots\}$$
$$A_a(t,\mathbf{s}) = \mathbb{P}(\pi(t,\mathbf{s}(t)) = a), \quad \forall\, a\in\mathcal{A},\ t\in\{1,2,\dots\},$$

where $\mathbf{s}(t)$ is the state of the complete system stochastic process at time t, operated according to π, starting from $\mathbf{s}$ (i.e., $\mathbf{s}(1) = \mathbf{s}$). (Note that we use $x^m_{ka}(t,\mathbf{s})$ and $A_a(t,\mathbf{s})$ to denote the optimal value of $x^m_{ka}(t)$ and $A_a(t)$ in the solution $(\mathbf{x}(\mathbf{s}), \mathbf{A}(\mathbf{s}))$ that corresponds to initial state $\mathbf{s}$.) Under the assumption of achievability, we have the following result.

PROPOSITION 3. *Let* $\mathbf{s}\in\mathcal{S}$. *If* $(\mathbf{x}(\mathbf{s}), \mathbf{A}(\mathbf{s}))$ *is achievable, then* $Z^*(\mathbf{s}) \leqslant J^*(\mathbf{s})$.

The proof is contained Section EC.1.3 of the electronic companion. The result follows since, under the assumption of achievability, $Z^*(\mathbf{s})$ is the total expected discounted reward of some policy, while $J^*(\mathbf{s})$ is the highest any such reward can be.

Under the assumptions of component independence and achievability, the fluid formulation allows us to construct an optimal policy.

THEOREM 1. *Suppose that for all* $\mathbf{s}\in\mathcal{S}$, $(\mathbf{x}(\mathbf{s}), \mathbf{A}(\mathbf{s}))$ *is achievable. Define the deterministic, stationary policy* $\pi\colon \mathcal{S}\to\mathcal{A}$ *as*

$$\pi(\mathbf{s}) = \arg\max_{a\in\mathcal{A}} A_a(1,\mathbf{s}).$$

Under these assumptions, the policy π *is an optimal policy*; *i.e.,* π *solves problem* (1).

The proof of the result (found in Section EC.1.4 of the electronic companion) follows by showing that any action a such that $A_a(1,\mathbf{s}) > 0$ is an action that is greedy with respect to the objective value $Z^*(\cdot)$ which, by combining Propositions 2 and 3, is equal to the optimal value function $J^*(\cdot)$.

3.4. Fluid-Based Heuristic

Theorem 1 tells us that, assuming that for every initial state $\mathbf{s}$ the optimal solution of problem (2) for initial state $\mathbf{s}$ is achievable, we immediately have an optimal policy by simply looking at the optimal values of the A variables at the first period ($t = 1$). Typically, however, the optimal solution of (2) will not be achievable. It nevertheless seems reasonable to expect that in many problems, the optimal solution $(\mathbf{x}(\mathbf{s}), \mathbf{A}(\mathbf{s}))$ may be close to being achievable for many states $\mathbf{s}$, because $(\mathbf{x}(\mathbf{s}), \mathbf{A}(\mathbf{s}))$ still respects the transition behavior of the system at the level of individual components. Consequently, the action $\arg\max_{a\in\mathcal{A}} A_a(1,\mathbf{s})$ should then be close to an optimal action for many states $\mathbf{s}$. It is therefore reasonable to expect that, by selecting the action a as $\arg\max_{a\in\mathcal{A}} A_a(1,\mathbf{s})$, one may often still be able to get good performance, even though the optimal solution of problem (2) may not be achievable.

Notwithstanding the question of achievability, applying this intuition in practice is not immediately possible. The reason for this is that problem (2) is an LO problem with a countably infinite number of variables and constraints, and so cannot be solved using standard solvers. Toward the goal of developing a practical heuristic policy for problem (1), we now consider an alternate, finite problem that can be viewed as an approximation to problem (2). This new formulation, presented below as problem (3), requires the decision maker to specify a time horizon T over which the evolution of the system will be modeled. For $t\in\{1,\dots,T\}$, the variables $x^m_{ka}(t)$ and $A_a(t)$ have the same meaning as in problem (3). To model the evolution of the system beyond $t = T$, we use the variable $x^m_{ka}(T+1)$ to represent the expected discounted long-run frequency with which component m is in state k and action a is taken from $t = T+1$ on. Similarly, we use $A_a(T+1)$ to represent the expected discounted frequency with which action a is taken from $t = T+1$ on.

With these definitions, the formulation corresponding to initial state $\mathbf{s}$ is presented below.

$$\underset{\mathbf{x},\mathbf{A}}{\text{maximize}} \quad \sum_{t=1}^{T+1}\sum_{m=1}^{M}\sum_{k\in\mathcal{S}^m}\sum_{a\in\mathcal{A}} \beta^{t-1}\cdot g^m_{ka}\cdot x^m_{ka}(t) \tag{3a}$$

$$\text{subject to} \quad \sum_{a\in\mathcal{A}} x^m_{ja}(t) = \sum_{k\in\mathcal{S}^m}\sum_{\tilde{a}\in\mathcal{A}} p^m_{kj\tilde{a}}\cdot x^m_{k\tilde{a}}(t-1),$$

$$\forall m\in\{1,\dots,M\},\ t\in\{2,\dots,T\},\ j\in\mathcal{S}^m, \quad (3b)$$

$$\sum_{a\in\mathcal{A}} x^m_{ja}(T+1) = \sum_{k\in\mathcal{S}^m}\sum_{a\in\mathcal{A}} p^m_{kja}\cdot x^m_{ka}(T) + \beta\cdot\sum_{k\in\mathcal{S}^m}\sum_{a\in\mathcal{A}} p^m_{kja}\cdot x^m_{ka}(T+1),$$

$$\forall m\in\{1,\dots,M\},\ j\in\mathcal{S}^m, \quad (3c)$$

$$\sum_{k\in\mathcal{S}^m} x^m_{ka}(t) = A_a(t),$$

$$\forall m\in\{1,\dots,M\},\ a\in\mathcal{A},\ t\in\{1,\dots,T+1\} \quad (3d)$$

$$\sum_{a\in\mathcal{A}} x^m_{ka}(1) = \alpha^m_k(\mathbf{s}),$$

$$\forall m\in\{1,\dots,M\},\ k\in\mathcal{S}^m, \quad (3e)$$

$$x^m_{ka}(t)\geqslant 0, \quad \forall m\in\{1,\dots,M\},\ a\in\mathcal{A},\ k\in\mathcal{S}^m,\ t\in\{1,\dots,T+1\}, \quad (3f)$$

$$A_a(t)\geqslant 0, \quad \forall a\in\mathcal{A},\ t\in\{1,\dots,T+1\}. \quad (3g)$$

With regard to constraints, we retain the same conservation constraints that relate the x^m_{ka} variables at $t-1$ to t, the initial state constraint and the consistency constraints that relate the x^m_{ka} and the A_a variables at a time t, for $t\in\{1,\dots,T\}$. Beyond $t=T$, constraint (3c) models the long-run transition behavior of the system. This constraint can be interpreted as a conservation relation: the left-hand side represents the expected discounted number of times from $T+1$ on that we take an action out of component m being in state j, while the right-hand side represents the expected discounted number of times that we enter state j from $T+1$ on. More specifically, the first right-hand side term represents the expected number of times that we enter state j at time $T+1$ (which is not discounted, since $T+1$ is the first period of the horizon $\{T+1,T+2,T+3,\dots\}$) and the second term represents the expected discounted number of times that we enter state j from $T+2$ on. Note also that constraint (3d), which is the analog of constraint (2c), extends from $t=1$ to $t=T+1$, ensuring that the $x^m_{ka}(T+1)$ and the $A_a(T+1)$ variables are also consistent with each other. With regard to the objective, observe that rather than being an infinite sum from $t=1$, the objective of problem (3) is a finite sum that extends from $t=1$ to $t=T+1$.

Let $Z^*_T(\mathbf{s})$ denote the optimal value of problem (3). Problem (3), like problem (2), provides an upper bound on the optimal value function at $J^*(\mathbf{s})$, and this bound improves with T, as indicated by the following result.

Proposition 4. *For each* $\mathbf{s}\in\mathcal{S}$ *and all* $T\in\{1,2,\dots\}$:
(a) $Z^*_T(\mathbf{s})\geqslant J^*(\mathbf{s})$; *and*
(b) $Z^*_T(\mathbf{s})\geqslant Z^*_{T+1}(\mathbf{s})$.

The proof of part (a) of Proposition 4 follows along similar lines to Proposition 2, while the proof of part (b) follows by showing that a solution to problem (3) with $T+1$ can be used to construct a feasible solution for problem (3) with T that achieves an objective value of $Z^*_{T+1}(\mathbf{s})$. The proof of this proposition can be found in Section EC.1.5 of the electronic companion. Part (a) of the proposition is useful because in passing from the infinite to the finite formulation, we have not lost the useful property that the objective value provides an upper bound on the optimal value function. Part (b) is important because it suggests a tradeoff in bound quality and computation: by increasing T, the quality of the bound improves, but the size of the formulation (the number of variables and constraints) increases. We will see later in Sections 5.4 and 5.5 that typically T does not need to be very large to ensure strong bounds and performance.

With this formulation, our heuristic policy is then defined as Algorithm 1.

Algorithm 1 (Fluid LO heuristic for infinite horizon problem with known stationary probabilities)
Require: Parameter T; data $\mathbf{p}$, $\mathbf{g}$, β; current state $\mathbf{s}\in\mathcal{S}$.
Solve problem (3) corresponding to initial state $\mathbf{s}$, horizon T, and data $\mathbf{p}$, $\mathbf{g}$, β to obtain an optimal solution $(\mathbf{x}(\mathbf{s}),\mathbf{A}(\mathbf{s}))$.
Take action $\tilde{a}$, where $\tilde{a}=\arg\max_{a\in\mathcal{A}} A_a(1,\mathbf{s})$.

Before continuing, we comment on two important ways in which problem (3) can be extended and one limitation of formulation (3). First of all, in problem (3), we formulated the decomposable MDP problem by defining decision variables that correspond to first-order information: in particular, $x^m_{ka}(t)$ represents the frequency with which a *single* component (component m) is in state k and action a is taken at time t. As shown in Section 3.3, the resulting formulation provides an upper bound on the optimal expected discounted reward. We can improve on this by considering higher-order fluid formulations, where rather than defining our decision variables to correspond to one component being in a state, we can define decision variables corresponding to combinations of components being in combinations of states, while a certain action is taken at a certain time. For example, a second-order formulation would correspond to using decision variables that model how frequently *pairs* of components are in different *pairs* of states while an action is taken at each time. As the order of the formulation increases, the objective value becomes an increasingly tighter bound on the optimal value, and it may be reasonable to expect better performance from using Algorithm 1; however, the size of the formulation increases rapidly.

Second, problem (3) models an infinite horizon problem and Algorithm 1 is a heuristic for this problem. For finite horizon problems, we can apply our approach as follows. Problem (3) can be modified by setting T to the horizon of the actual problem and removing the terminal $T+1$ decision variables that model the long-run evolution of the system. Then, if we are at state $\mathbf{s}$ at period t', we restrict the fluid problem to $\{t',t'+1,\dots,T\}$ and use constraint (3e) to set the initial state at t' to $\mathbf{s}$. We then solve the problem to obtain the optimal solution $(\mathbf{x}(\mathbf{s}),\mathbf{A}(\mathbf{s}))$ and we take the action a

that maximizes $A_a(t, \mathbf{s})$. Note that if the transition probabilities change over time (i.e., rather than p^m_{kja} we have $p^m_{kja}(t)$ for $t \in \{1, \ldots, T-1\}$), we may also modify constraint (3b) and replace p^m_{kja} with $p^m_{kja}(t)$, without changing the size or the nature of the resulting formulation.

Finally, we comment on one limitation to the fluid formulation (3). Problem (3) is formulated in terms of the *system* action space $\mathscr{A}$; the actions that index the $x^m_{ka}(t)$ and $A_a(t)$ variables are elements of the system action space $\mathscr{A}$. For certain problems, the system action space $\mathscr{A}$ may be small and problem (3) may be easy to solve. For example, in a multiarmed bandit problem where exactly one bandit must be activated, $|\mathscr{A}| = M$ (one of the M bandits); similarly, in an optimal stopping problem, $|\mathscr{A}| = 2$ (stop or continue). For other problems, the action space of the problem may grow exponentially (e.g., a bandit problem where one may activate up to K of M bandits). For such problems, the fluid formulation (3) will be harder to solve; we do not consider this regime in the present paper. The development of a scalable solution method for large-scale versions of problem (3) constitutes an interesting direction for future research.

4. Comparisons to Other Approaches

In this section, we compare our finite fluid formulation (3) against three state-of-the-art formulations that can be used to solve decomposable MDPs. We begin by stating these formulations: in Sections 4.1–4.3, we present the ALO, classical Lagrangian relaxation and the alternate Lagrangian relaxation formulations, respectively. Then, in Section 4.4 we state a key theoretical result that asserts that the finite fluid formulation (3) provides a provably tighter bound than all three formulations. In Section 4.5 we discuss the sizes of the formulations, and in Section 4.6, we discuss how to extend the key idea of the fluid problem to the other formulations.

4.1. Approximate Linear Optimization

For the ALO formulation of de Farias and Van Roy (2003), we approximate the value function using the same functional form as in Adelman and Mersereau (2008):

$$J_{\text{ALO}}(\mathbf{s}) = \sum_{m=1}^{M} J^m_{s^m}; \tag{4}$$

i.e., we assume that each state of each component contributes an additive effect. For a given initial state $\mathbf{s} \in \mathscr{S}$, the corresponding ALO formulation is then

$$\underset{\mathbf{J}}{\text{minimize}} \quad \sum_{m=1}^{M} \sum_{k \in \mathscr{S}^m} \alpha^m_k(\mathbf{s}) \cdot J^m_k \tag{5a}$$

$$\text{subject to} \quad \sum_{m=1}^{M} J^m_{\bar{s}^m} \geqslant \sum_{m=1}^{M} g^m_{\bar{s}^m a} + \beta \sum_{m=1}^{M} \sum_{j \in \mathscr{S}^m} p^m_{\bar{s}^m ja} J^m_j, \quad \forall\, \bar{\mathbf{s}} \in \mathscr{S},\ a \in \mathscr{A}. \tag{5b}$$

To derive a policy from $\mathbf{J}$, we take the action $\tilde{a}$ that is greedy with respect to J_{ALO}; this action is defined as

$$\tilde{a} = \arg\max_{a \in \mathscr{A}} \left\{ \sum_{m=1}^{M} g^m_{s^m a} + \beta \cdot \sum_{m=1}^{M} \sum_{j \in \mathscr{S}^m} p^m_{s^m ja} J^m_j \right\}. \tag{6}$$

Let $Z^*_{\text{ALO}}(\mathbf{s})$ denote the objective value of problem (5) with initial state $\mathbf{s}$. The following result, due to Adelman and Mersereau (2008), establishes that $Z^*_{\text{ALO}}(\mathbf{s})$ upper bounds the optimal value function at $\mathbf{s}$. The proof can be found in Adelman and Mersereau (2008) and is thus omitted.

Proposition 5 (Proposition 4 of Adelman and Mersereau 2008). *For all* $\mathbf{s} \in \mathscr{S}$, $Z^*_{ALO}(\mathbf{s}) \geqslant J^*(\mathbf{s})$.

4.2. Classical Lagrangian Relaxation

We now present the classical Lagrangian relaxation (CLR) approach. To apply this approach to our decomposable MDP defined in Section 3.1, we require three additional assumptions.

Assumption 1. *In addition to the system state space being decomposable along components, the action space also decomposes along the components. More precisely, each component m is endowed with both a state space $\mathscr{S}^m$ and an action space $\mathscr{A}^m$. Thus, an action a in the system action space can be represented as a tuple of component actions, $a = (a^1, \ldots, a^m) \in \mathscr{A} \subseteq \mathscr{A}^1 \times \cdots \times \mathscr{A}^M$.*

Assumption 2. *The rewards and transition probabilities decompose with respect to the new action spaces $\mathscr{A}^m$. Let $R^m_{ka^m}$ denote the reward from component m when action a^m is taken in state k and let $\bar{p}^m_{ka^m}$ denote the transition probability of component m when action a^m is taken. We require that $p^m_{kja} = \bar{p}^m_{ka^m}$ and $g^m_{ka} = R^m_{ka^m}$ whenever the mth component of a is a^m.*

Assumption 3. *The system action state space $\mathscr{A}$ is defined implicitly through a linking constraint on the component actions*:

$$\mathscr{A} = \left\{ a = (a^1, \ldots, a^M) \in \mathscr{A}^1 \times \cdots \times \mathscr{A}^M \;\middle|\; \sum_{m=1}^{M} \mathbf{D}^m(a^m) \leqslant \mathbf{b} \right\}, \tag{7}$$

where $\mathbf{D}^m \colon \mathscr{A}^m \to \mathbb{R}^q$ is a function for each m and $\mathbf{b} \in \mathbb{R}^q$ for some finite q.

When these three assumptions hold, the Lagrangian approach involves dualizing the linking constraint $\sum_{m=1}^{M} \mathbf{D}^m(a^m) \leqslant \mathbf{b}$ by introducing a Lagrange multiplier vector $\boldsymbol{\lambda} \in \mathbb{R}^p$ for this linking constraint. The CLR formulation of the problem can be written as follows:

$$\underset{\boldsymbol{\lambda}, \mathbf{V}}{\text{minimize}} \quad \left\{ \frac{\boldsymbol{\lambda}^T \mathbf{b}}{1-\beta} + \sum_{m=1}^{M} \sum_{k \in \mathscr{S}^m} \alpha^m_k(\mathbf{s}) \cdot V^m_k \right\} \tag{8a}$$

$$\text{subject to} \quad V^m_k \geqslant R^m_{ka^m} - \boldsymbol{\lambda}^T \mathbf{D}^m(a^m) + \beta \cdot \sum_{j \in \mathscr{S}^m} \bar{p}^m_{kja^m} \cdot V^m_j, \quad \forall\, m \in \{1, \ldots, M\},\ k \in \mathscr{S}^m,\ a^m \in \mathscr{A}^m \tag{8b}$$

$$\boldsymbol{\lambda} \geqslant 0. \tag{8c}$$

The optimal variable $\mathbf{V}$ can be interpreted as a componentwise approximation to the value function. One can form a value function approximation that is analogous to the ALO approximation in Equation (4) and take the greedy action analogously to Equation (6).

The dual of problem (8) is

$$\underset{\mathbf{z}}{\text{maximize}} \quad \sum_{m=1}^{M} \sum_{k \in \mathcal{S}^m} \sum_{a^m \in \mathcal{A}^m} R^m_{ka^m} \cdot z^m_{ka^m} \tag{9a}$$

$$\text{subject to} \quad \sum_{a^m \in \mathcal{A}^m} z^m_{ja^m} = \alpha^m_j(\mathbf{s}) + \beta \cdot \sum_{k \in \mathcal{S}^m} \sum_{a^m \in \mathcal{A}^m} \bar{p}^m_{kja^m} \cdot z^m_{ka^m}, \quad \forall\, m \in \{1, \ldots, M\},\ j \in \mathcal{S}^m, \tag{9b}$$

$$\sum_{m=1}^{M} \sum_{k \in \mathcal{S}^m} \sum_{a^m \in \mathcal{A}^m} \mathbf{D}^m(a^m) z^m_{ka^m} \leqslant \frac{\mathbf{b}}{1-\beta}, \tag{9c}$$

$$z^m_{ka^m} \geqslant 0, \quad \forall\, m \in \{1, \ldots, M\},\ k \in \mathcal{S}^m,\ a^m \in \mathcal{A}^m. \tag{9d}$$

The variable $z^m_{ka^m}$ can be interpreted as the expected discounted frequency with which component m is in state k and action a^m is being taken over the entire infinite horizon. Constraint (9b) models the transition dynamics of component m in an expected discounted sense, while constraint (9c) can be interpreted as the expected discounted version of the linking constraint that defines the action space $\mathcal{A}$ in Equation (7).

Let $Z^*_{\text{CLR}}(\mathbf{s})$ be the optimal objective value of problem (8) corresponding to initial state $\mathbf{s}$. The following two results, due to Adelman and Mersereau (2008), establish that the CLR provides an upper bound on the optimal value function and that the ALO provides a tighter bound than the CLR. The proofs can be found in Adelman and Mersereau (2008) and are omitted.

Proposition 6 (Proposition 2 of Adelman and Mersereau 2008). *When Assumptions 1–3 hold, for all* $\mathbf{s} \in \mathcal{S}$, $Z^*_{\text{CLR}}(\mathbf{s}) \geqslant J^*(\mathbf{s})$.

Proposition 7 (Corollary 1 of Adelman and Mersereau 2008). *When Assumptions 1–3 hold, for all* $\mathbf{s} \in \mathcal{S}$, $Z^*_{\text{ALO}}(\mathbf{s}) \leqslant Z^*_{\text{CLR}}(\mathbf{s})$.

Furthermore, Adelman and Mersereau (2008) provide a simple parameterized problem (Section 3.3 of that paper) where the difference between $Z^*_{\text{ALO}}(\mathbf{s})$ and $Z^*_{\text{CLR}}(\mathbf{s})$ can be made arbitrarily large.

4.3. Alternate Lagrangian Relaxation

The CLR formulation requires Assumptions 1–3 to hold. When these assumptions hold, it is possible to exploit the definition of the system action space in Equation (7) to arrive at formulation (8). However, the decomposable MDP that we have defined in Section 3.1 may not be consistent with these assumptions; more precisely, the system action space may not naturally decompose along the components. Consider, for example, an optimal stopping problem where the system is actually M independent components. In this example, the action space (which consists of two actions, stop or continue) does not decompose along each component, and it does not make sense to think of the system action space as being the feasible set of a coupling constraint on the action spaces of M small MDPs.

Surprisingly, it turns out that there is a transformation by which one can convert any decomposable MDP with a general system action space $\mathcal{A}$ into a weakly coupled MDP and thus apply the Lagrangian relaxation approach, even when the action space $\mathcal{A}$ does not have a representation of the form in Equation (7). The steps of this transformation are as follows.

1. Construct M small MDPs, where the mth small MDP corresponds to component m of the decomposable MDP.
2. Set the state space of small MDP m to be $\mathcal{S}^m$, the state space of component m.
3. Set the action space of small MDP m to be $\mathcal{A}$, the action space of the complete system. (Thus, each small MDP involves controlling how component m evolves across its own state space, where we may choose any action from the *system* action space $\mathcal{A}$.)
4. Enforce the following coupling constraint:

$$\mathbb{I}\{a^m = a\} - \mathbb{I}\{a^{m+1} = a\} = 0, \quad \forall\, m \in \{1, \ldots, M-1\},\ a \in \mathcal{A}, \tag{10}$$

or equivalently, that

$$a^m = a^{m+1}, \quad \forall\, m \in \{1, \ldots, M-1\}.$$

The above constraint is simple: it requires that the actions taken in small MDP m and small MDP $m+1$ must be the same, or equivalently, the actions $a^m, a^{m'}$ taken in any pair of small MDPs $m, m' \in \{1, \ldots, M\}$ must be the same.

It is easy to see that this weakly coupled MDP is exactly the same as the decomposable MDP of Section 3.1. We now construct the Lagrangian relaxation of this weakly coupled MDP. Introducing the Lagrange multiplier λ^m_a for the (m, a) constraint in the family of constraints (10) and using $\boldsymbol{\lambda}$ to denote the vector of multipliers, the corresponding Lagrangian relaxation formulation of this weakly coupled MDP for initial state $\mathbf{s}$, can be shown to be

$$\underset{\boldsymbol{\lambda}, \mathbf{V}}{\text{minimize}} \quad \sum_{m=1}^{M} \sum_{k \in \mathcal{S}^m} \alpha^m_k(\mathbf{s}) V^m_k \tag{11a}$$

$$\text{subject to} \quad V^m_k \geqslant g^m_{ka} - \mathbb{I}\{m < M\} \cdot \lambda^m_a + \mathbb{I}\{m > 1\} \cdot \lambda^{m-1}_a + \beta \cdot \sum_{j \in \mathcal{S}^m} p^m_{kja} V^m_j, \quad \forall\, m \in \{1, \ldots, M\},\ k \in \mathcal{S}^m,\ a \in \mathcal{A}. \tag{11b}$$

We refer to this relaxation as the *alternate Lagrangian relaxation* (ALR). A more detailed derivation of the ALR can be found in Section EC.3 of the electronic companion. As with the CLR and the ALO, one can form a value function approximation of the form in Equation (4) using the optimal $\mathbf{V}$ values and take the greedy action analogously to Equation (6).

It should be clear that the ALR problem (11) is *not* the same as the CLR problem (8). Problem (11) only

decomposes the state and does not decompose the system action space; it accomplishes this by endowing each component with the system action space and enforcing the action consistency constraint (10). Problem (8), on the other hand, decomposes the state *and* the action space by using the structure of the action space given in Equation (7). The resulting formulations thus differ in their sizes; typically, the ALR will be larger than the CLR because the dimensions of the ALR problem (numbers of variables and constraints) scale with the number of *system* actions. Note that problem (8) can be formulated only when Assumptions 1–3 hold. On the other hand, problem (11) can *always* be formulated, regardless of the structure of the action space. To the best of our knowledge, this type of alternate Lagrangian relaxation has not been proposed before.

To understand how the ALR relates to the fluid formulation, it is helpful to formulate the dual of problem (11):

$$\underset{\mathbf{z}}{\text{maximize}} \quad \sum_{m=1}^{M} \sum_{k\in\mathcal{S}^m} \sum_{a\in\mathcal{A}} g^m_{ka} z^m_{ka} \tag{12a}$$

$$\text{subject to} \quad \sum_{a'\in\mathcal{A}} z^m_{ja'} - \beta \sum_{k\in\mathcal{S}^m} \sum_{a\in\mathcal{A}} p^m_{kja} \cdot z^m_{ka} = \alpha^m_j(\mathbf{s}), \quad \forall\, m\in\{1,\dots,M\},\ j\in\mathcal{S}^m, \tag{12b}$$

$$\sum_{k\in\mathcal{S}^m} z^m_{ka} = \sum_{k\in\mathcal{S}^{m+1}} z^{m+1}_{ka}, \quad \forall\, m\in\{1,\dots,M-1\},\ a\in\mathcal{A}, \tag{12c}$$

$$z^m_{ka} \geqslant 0, \quad \forall\, m\in\{1,\dots,M\},\ k\in\mathcal{S}^m,\ a\in\mathcal{A}. \tag{12d}$$

The dual variable z^m_{ka} can be interpreted as the expected discounted number of times that the component m is in state k and action a is taken over the entire horizon. Constraint (12b) models the long-term transition behavior of small MDP m, while constraint (12c) can be interpreted as the expected discounted version of the linking constraint (10); the expected discounted number of times that we take action a in small MDP m must be the same as the expected discounted number of times that we take action a in small MDP $m+1$.

Having formed the dual problem (12), we can see that the ALR dual (12) and the finite fluid formulation (3) bear some resemblance, in terms of accounting for how frequently components are in specific states while a specific action is taken, accounting for the transition behavior and accounting for the fact that component state-action frequencies (the $x^m_{ka}(t)$ variables in problem (3) and the z^m_{ka} variables in problem (12)) are linked across components through the action. However, the key difference lies in the fact that in problem (12), time is fully aggregated: the z^m_{ka} variables represent the *long-run* expected discounted frequency with which component m is in state k and action a is taken from $t=1$ on. In contrast, in problem (3), time is partially disaggregated: for $t=1$ to $t=T$, the transition behavior of the system is modeled separately for each t, and for $t=T+1$ and beyond, the transition behavior is modeled in the same aggregate sense (compare constraints (12b) and (3c)). One can thus interpret the fluid formulation as a partially disaggregated version of the ALR dual (12). If one imagines the finite fluid formulation (3) with $T=0$—i.e., the formulation does not account for the transition behavior separately for any periods and only accounts for transition behavior in a long-term discounted sense, from period 1 ($=T+1$) on—then one can see that it would be equivalent to the ALR dual problem (12).

Let $Z^*_{ALR}(\mathbf{s})$ denote the optimal value of the ALR formulation (11) corresponding to initial state $\mathbf{s}$ and let $Z^*_{ALO}(\mathbf{s})$ denote the optimal value of the ALO formulation (5) corresponding to initial state $\mathbf{s}$. The following result establishes that problems (11) and (5) are equivalent.

THEOREM 2. *For each* $\mathbf{s}\in\mathcal{S}$:

(a) $Z^*_{ALO}(\mathbf{s}) = Z^*_{ALR}(\mathbf{s})$; *and*

(b) *Let* $\mathbf{V}\in\mathbb{R}^{\sum_{m=1}^{M}|\mathcal{S}^m|}$. *There exists* $\boldsymbol{\lambda}$ *such that* $(\mathbf{V},\boldsymbol{\lambda})$ *is an optimal solution for the Lagrangian relaxation formulation* (11) *corresponding to state* $\mathbf{s}$ *if and only if* $\mathbf{V}$ *is an optimal solution for the ALO formulation* (5) *corresponding to state* $\mathbf{s}$.

The proof of this result, found in Section EC.1.6 of the electronic companion, follows by essentially showing that the optimal solution of one problem leads to a feasible solution for the other problem with the same optimal value.

We offer two remarks on Theorem 2. First, we believe Theorem 2 to be valuable because the alternate Lagrangian relaxation problem (11) is considerably more tractable than the ALO problem (5). Specifically, the former has a number of variables and constraints that is linear in the problem dimensions, while the latter has a number of constraints that is in general exponential in the number of components. One of the challenges of applying the ALO approach is that, although applying a basis function approximation as in Equation (4) allows one to reduce the number of *variables*, one is still left with a large number of *constraints* (one for each pair $(\mathbf{s},a)\in\mathcal{S}\times\mathcal{A}$). To cope with the large number of constraints, one might use constraint sampling (de Farias and Van Roy 2004) or column generation techniques (see, e.g., Adelman 2007 and Adelman and Mersereau 2008). Theorem 2 implies that in cases where $\mathcal{A}$ is not too large, one can avoid resorting to these techniques by directly solving problem (11): by part (a) of the theorem, the resulting bound will be the same as that of the ALO formulation, and by part (b), the resulting value function approximation is also a valid ALO value function approximation (since the optimal $\mathbf{V}$ for the ALR problem (11) is also a valid optimal solution for the ALO problem (5)).

Second, it is valuable to contrast Theorem 2 to Proposition 7, which pertains to the relationship between the CLR and the ALO formulations. Theorem 2 asserts that the ALR and the ALO are equivalent, whereas Proposition 7 asserts

that the CLR is no tighter than the ALO (moreover, as discussed in Section 4.2 there exist simple examples where the difference between $Z^*_{\text{ALO}}(\mathbf{s})$ and $Z^*_{\text{CLR}}(\mathbf{s})$ can be extremely large). Thus, problem (11) is a tighter formulation of the MDP than problem (8), as summarized in the following corollary.

COROLLARY 1. *When Assumptions 1–3 hold, for all* $\mathbf{s} \in \mathcal{S}$, $Z^*_{\text{ALR}}(\mathbf{s}) \leqslant Z^*_{\text{CLR}}(\mathbf{s})$.

4.4. Comparison of Bounds

With Theorem 2 in hand, we are ready to state our key theoretical result, whose proof appears in Section EC.1.7 of the electronic companion.

THEOREM 3. *For each* $\mathbf{s} \in \mathcal{S}$ *and all* $T \in \{1, 2, \ldots\}$:
(a) $Z^*_T(\mathbf{s}) \leqslant Z^*_{\text{ALR}}(\mathbf{s})$;
(b) $Z^*_T(\mathbf{s}) \leqslant Z^*_{\text{ALO}}(\mathbf{s})$; *and*
(c) $Z^*_T(\mathbf{s}) \leqslant Z^*_{\text{CLR}}(\mathbf{s})$ (*when Assumptions 1–3 hold*).

Part (a) follows by using a solution of problem (3) with T to construct a feasible solution with objective value $Z^*_T(\mathbf{s})$ for problem (12); part (b) follows by using Theorem 2 and part (a); and part (c) follows by combining part (b) with Proposition 7. In Section 5, where we apply our fluid approach to multiarmed bandits, we provide numerical examples that show these three inequalities can be strict and the bounds can be significantly different. The above result, Theorem 2, and Propositions 4 and 7 can together be summarized in the following corollary.

COROLLARY 2. *When Assumptions 1–3 hold, for all* $\mathbf{s} \in \mathcal{S}$ *and all* $T \in \{1, 2, \ldots\}$:

$$\begin{aligned} J^*(\mathbf{s}) &\leqslant Z^*_T(\mathbf{s}) \leqslant \cdots \leqslant Z^*_2(\mathbf{s}) \\ &\leqslant Z^*_1(\mathbf{s}) \leqslant Z^*_{\text{ALO}}(\mathbf{s}) = Z^*_{\text{ALR}}(\mathbf{s}) \leqslant Z^*_{\text{CLR}}(\mathbf{s}). \end{aligned}$$

Theorem 3 essentially asserts that the fluid formulation provides a provably tighter bound than all three alternate approaches: the classical Lagrangian relaxation, the alternate Lagrangian relaxation and the ALO. The classical Lagrangian relaxation, and the ALO formulations have been widely applied to solve practical problems; it is fair to say that these approaches constitute the state-of-the-art in solving large-scale MDPs of practical interest. Our result is therefore significant because we have shown that the finite fluid formulation (3) leads to bounds that are at least as good as those of the Lagrangian relaxation and ALO formulations. In Section 5.4, we show that these differences can in fact be significant. More significantly, although Theorem 3 pertains to bounds, it is reasonable to expect that a formulation that produces a tighter bound will also produce better policies. Indeed, we will later show numerically that the heuristic policy given as Algorithm 1 based on the finite fluid problem (3) can significantly outperform the Lagrangian relaxation and ALO approaches.

4.5. Comparison of Formulation Sizes

As a complement to Theorem 3, where we compare the formulation bounds, we now compare the formulations in terms of their sizes. Table 1 summarizes the sizes of the four types of formulations in terms of the number of variables and the number of constraints. (Recall that q is the number of constraints that define the action space in Equation (7) for the CLR approach.)

Although the exact numbers of variables and constraints will depend on the specific values of $|\mathcal{S}^m|$, $|\mathcal{A}|$, $|\mathcal{A}^m|$, T, and q, we can derive some general qualitative insights:

- When Assumptions 1–3 hold and the action space can be described by a small number q of linking constraints as in Equation (7), then the CLR problem (8) will in general be the smallest formulation, as its dimensions are not dependent on $|\mathcal{A}|$ and $|\mathcal{S}|$.
- The largest formulation will in general be the ALO problem (5), as the number of constraints in the ALO scales with the size of the system state space $|\mathcal{S}|$ and the size of the system action space $|\mathcal{A}|$.
- The ALR problem (11) and the finite fluid formulation (3) will be somewhere in between the CLR problem (8) and the ALO problem (5), as the numbers of variables and constraints depend on the size of the system action space $|\mathcal{A}|$. Between the two, the fluid formulation will be larger than the ALR formulation due to the dependence on T.

Thus, while the fluid formulation provides a provably tighter bound than the other three formulations, it will in general not be the smallest formulation. In situations where the system action space $\mathcal{A}$ is not too large, the improved quality of the bound may justify the additional computational effort required for the fluid formulation.

4.6. Disaggregating the ALO and the ALR

The key idea in the fluid problem (3) is to partially disaggregate time in the first T periods, and then aggregate time in a discounted way from period $T+1$ on. This disaggregation is what allows us to prove that the fluid problem is tighter than the ALR (part (a) of Theorem 3). One might then wonder if this type of disaggregation can be applied in the ALR and the ALO. To understand how this disaggregation applies in the ALR and ALO formulations, let us define two new partially disaggregated formulations: the ALR(T) formulation and the ALO(T) formulation.

The ALR(T) formulation is

$$\underset{\boldsymbol{\lambda}, \mathbf{V}}{\text{minimize}} \quad \sum_{m=1}^{M} \sum_{k \in \mathcal{S}^m} \alpha^m_k(\mathbf{s}) V^m_k(1) \tag{13a}$$

$$\begin{aligned} \text{subject to} \quad & V^m_k(t) \geqslant g^m_{ka} - \mathbb{I}\{m < M\} \cdot \lambda^m_a(t) \\ & + \mathbb{I}\{m > 1\} \cdot \lambda^{m-1}_a(t) + \beta \cdot \sum_{j \in \mathcal{S}^m} p^m_{kja} V^m_j(t+1), \\ & \forall\, m \in \{1, \ldots, M\},\ k \in \mathcal{S}^m, \\ & a \in \mathcal{A},\ t \in \{1, \ldots, T\} \end{aligned} \tag{13b}$$

Table 1. Comparison of sizes of formulations.

Formulation	Number of variables	Number of constraints
ALO problem (5)	$\sum_{m=1}^{M} \lvert\mathcal{S}^m\rvert$	$\lvert\mathcal{S}\rvert \cdot \lvert\mathcal{A}\rvert$
CLR problem (8)	$\sum_{m=1}^{M} \lvert\mathcal{S}^m\rvert + q$	$\sum_{m=1}^{M} \lvert\mathcal{S}^m\rvert \cdot \lvert\mathcal{A}^m\rvert$
ALR problem (11)	$\sum_{m=1}^{M} \lvert\mathcal{S}^m\rvert + (M-1)\cdot\lvert\mathcal{A}\rvert$	$\sum_{m=1}^{M} \lvert\mathcal{S}^m\rvert \cdot \lvert\mathcal{A}\rvert$
Fluid problem (3)	$(T+1)\left(\sum_{m=1}^{M} \lvert\mathcal{S}^m\rvert \cdot \lvert\mathcal{A}\rvert + \lvert\mathcal{A}\rvert\right)$	$(T+1)\left(\sum_{m=1}^{M} \lvert\mathcal{S}^m\rvert + M\lvert\mathcal{A}\rvert\right)$

Note. The number of constraints quoted for each formulation does not count any nonnegativity constraints.

$$V_k^m(T+1) \geqslant g_{ka}^m - \mathbb{I}\{m<M\}\cdot\lambda_a^m(T+1) + \mathbb{I}\{m>1\}\cdot\lambda_a^{m-1}(T+1) + \beta\cdot\sum_{j\in\mathcal{S}^m} p_{kja}^m V_j^m(T+1), \quad \forall m\in\{1,\ldots,M\},\ k\in\mathcal{S}^m,\ a\in\mathcal{A}. \tag{13c}$$

Let $Z^*_{\mathrm{ALR}(T)}(\mathbf{s})$ denote the optimal value of problem (13). The ALO(T) formulation is

$$\underset{\mathbf{J}}{\text{minimize}} \quad \sum_{m=1}^{M}\sum_{k\in\mathcal{S}^m} \alpha_k^m(\mathbf{s})\cdot J_k^m(1) \tag{14a}$$

$$\text{subject to} \quad \sum_{m=1}^{M} J_{\bar{s}^m}^m(t) \geqslant \sum_{m=1}^{M} g_{\bar{s}^m a}^m + \beta\sum_{m=1}^{M}\sum_{j\in\mathcal{S}^m} p_{\bar{s}^m ja}^m J_j^m(t+1), \quad \forall \bar{\mathbf{s}}\in\mathcal{S},\ a\in\mathcal{A},\ t\in\{1,\ldots,T\}, \tag{14b}$$

$$\sum_{m=1}^{M} J_{\bar{s}^m}^m(T+1) \geqslant \sum_{m=1}^{M} g_{\bar{s}^m a}^m + \beta\sum_{m=1}^{M}\sum_{j\in\mathcal{S}^m} p_{\bar{s}^m ja}^m J_j^m(T+1), \quad \forall \bar{\mathbf{s}}\in\mathcal{S},\ a\in\mathcal{A}. \tag{14c}$$

Let $Z^*_{\mathrm{ALO}(T)}(\mathbf{s})$ denote the optimal value of problem (14). We then have the following theoretical result.

THEOREM 4. *For all* $\mathbf{s}\in\mathcal{S}$, $T\in\{1,2,\ldots\}$, $Z_T^*(\mathbf{s}) = Z^*_{\mathrm{ALR}(T)}(\mathbf{s}) = Z^*_{\mathrm{ALO}(T)}(\mathbf{s})$.

The first part of the equality restates more rigorously the earlier observation from Section 4.3, which is that the fluid problem can be viewed as the ALR problem with time disaggregated over a horizon of T periods. The second equality asserts that the fluid problem is equivalent to a time-disaggregated version of the ALO, analogously to Theorem 2.

5. Application to Multiarmed Bandit Problems

5.1. Problem Definition

In the multiarmed bandit problem, the decision maker is presented with a set of bandits/arms, and each arm is endowed with some state space. At each point in time, the decision maker needs to select one of the arms to activate so as to maximize his long-term (over an infinite horizon) expected discounted reward. We consider the regular multiarmed bandit problem, where only the activated arm changes state and generates reward, and the restless multiarmed bandit problem, where inactive arms may also change state state (i.e., passive transitions are allowed) and generate reward (i.e., there are passive rewards).

5.2. Fluid Model

The multiarmed problem can be readily formulated in our fluid framework. The components M of the stochastic system correspond to the individual bandits. The action space $\mathcal{A}$ is defined here as $\mathcal{A}=\{1,\ldots,M\}$. The reward g_{ka}^m, for $m\in\{1,\ldots,M\}$, $k\in\mathcal{S}^m$, and $a\in\mathcal{A}$, is the reward that is earned when arm m is in state k and arm a is activated. Similarly, p_{kja}^m is the probability that bandit m transitions from state $k\in\mathcal{S}^m$ to state $j\in\mathcal{S}^m$ when arm a is activated. In the case of the regular bandit problem, we need to ensure that whenever $m\neq a$, $g_{ka}^m=0$ for every $k\in\mathcal{S}^m$ and $p_{kja}^m=\mathbb{I}\{k=j\}$ for every pair of states $k,\ j\in\mathcal{S}^m$. In the case of the restless bandit problem, we only need to ensure that whenever $a, a', m\in\{1,\ldots,M\}$, with $m\neq a$ and $m\neq a'$, that $g_{ka}^m = g_{ka'}^m$ for every $k\in\mathcal{S}^m$ and $p_{kja}^m = p_{kja'}^m$ for every $k,\ j\in\mathcal{S}^m$.

5.3. Relation to Bertsimas and Niño-Mora (2000)

One interesting property of the fluid formulation is how it relates to the performance measure formulation developed in Bertsimas and Niño-Mora (2000). Let w_{j0}^m be defined for every bandit $m\in\{1,\ldots,M\}$, state $j\in\mathcal{S}^m$ as the expected discounted number of times that bandit m is in state j and it is not activated. Similarly, let w_{j1}^m be defined as the expected discounted number of times that bandit m is in state j and *is*

activated. Let $\bar{p}^m_{ij1}$ and $\bar{p}^m_{ij0}$ be the active and passive transition probabilities of bandit m from state i to state j, respectively, and let R^m_{k0} and R^m_{k1} be the passive and active rewards from activating bandit m when it is in state k, respectively. The w^m_{ja} variables are referred to as *performance measures*. Finally, suppose that the system starts in state $\mathbf{s} \in \mathcal{S}$ at time $t = 1$ and as assumed in Section 5.1, we must activate exactly one arm at any period.

For a given collection of performance measures, the corresponding reward is the sum $\sum_{m=1}^{M} \sum_{k \in \mathcal{S}^m} \sum_{a \in \{0,1\}} R^m_{ka} w^m_{ka}$, which forms the objective of the problem. The performance measures, by their definition, satisfy certain conservation laws, and the feasible set of performance measures resulting from those laws is referred to as the *performance region*. The formulation developed in Bertsimas and Niño-Mora (2000) is to maximize this reward over the performance region:

$$\underset{\mathbf{w}}{\text{maximize}} \quad \sum_{m=1}^{M} \sum_{k \in \mathcal{S}^m} \sum_{a \in \{0,1\}} R^m_{ka} w^m_{ka} \tag{15a}$$

$$\text{subject to} \quad \sum_{m=1}^{M} \sum_{k \in \mathcal{S}^m} w^m_{k1} = \frac{1}{1-\beta}, \tag{15b}$$

$$w^m_{j0} + w^m_{j1} = \alpha^m_j(\mathbf{s}) + \beta \sum_{i \in \mathcal{S}^m} \sum_{a \in \{0,1\}} \bar{p}^m_{ija} w^m_{ia}, \quad \forall m \in \{1, \ldots, M\},\ j \in \mathcal{S}^m, \tag{15c}$$

$$w^m_{j0},\ w^m_{j1} \geqslant 0, \quad \forall m \in \{1, \ldots, M\},\ j \in \mathcal{S}^m. \tag{15d}$$

In words, the formulation finds the vector of performance measures $\mathbf{w}$ that satisfies the transition constraints at the level of the components and maximizes the total expected discounted reward, which is just the sum of the performance measures weighted by their corresponding rewards. Note that in terms of the data defining the fluid formulation, the data in problem (15) and in the finite fluid problem (3) identify as follows. For every bandit m, states i and j, we have $\bar{p}^m_{ij1} = p^m_{ijm}$, while $\bar{p}^m_{ij0} = p^m_{ija}$ for every $a \neq m$. Similarly, for the rewards, we have $R^m_{k1} = g^m_{km}$ for every bandit m and state k, while $R^m_{k0} = g^m_{ka}$ for every $a \neq m$. To make decisions, Bertsimas and Niño-Mora (2000) propose a primal dual heuristic where one solve problem (15) at each new state $\mathbf{s}$. We describe how the heuristic operates for the case when exactly one arm must be activated at each period; for more details, the interested reader is referred to Bertsimas and Niño-Mora (2000). Using the solution of the problem, the heuristic considers the optimal variables $w^m_{s^m 1}$ and $w^m_{s^m 0}$ for each $m \in \{1, \ldots, M\}$ and proceeds as follows:

1. If exactly one of $w^m_{s^m 1}$ for $m \in \{1, \ldots, M\}$ is positive, say bandit m', then activate bandit m'. (Intuitively, $w^m_{s^m 1}$ represents the expected discounted amount of time that bandit m is in its initial state s^m and it is activated; if there is only one bandit for which this value is positive, the solution suggests that we should activate this bandit.)

2. If all $w^m_{s^m 1}$ are zero, then activate the bandit $m \in \{1, \ldots, M\}$ with the lowest reduced cost of the active performance measure $w^m_{s^m 1}$. (Intuitively, the reduced cost of $w^m_{s^m 1}$ represents the marginal decrease in the objective value per unit increase in $w^m_{s^m 1}$; by selecting the m with the lowest reduced cost of $w^m_{s^m 1}$, we select the bandit that will have the lowest detriment to the objective.)

3. If more than one $w^m_{s^m 1}$ for $m \in \{1, \ldots, M\}$ is positive, then activate the bandit m with the largest reduced cost of the passive performance measure $w^m_{s^m 0}$ among those m with $w^m_{s^m 1} > 0$. (Similarly to the previous case, the reduced cost of $w^m_{s^m 0}$ represents the marginal decrease in the objective for a unit increase in the passive performance measure $w^m_{s^m 0}$; by activating the bandit with the largest passive reduced cost we try to counteract this effect.)

Before continuing on to the results, it is important to establish how the performance region formulation relates to the formulations presented in Section 4 and to the fluid method. Let $Z^*_{BNM}(\mathbf{s})$ be the optimal objective value of problem (15) when the system starts in state $\mathbf{s} \in \mathcal{S}$. First, problem (15) and the CLR problem (8) are equivalent; this connection was originally observed by Hawkins (2003). To see this, for each m set $\mathcal{A}^m = \{0, 1\}$, where 1 indicates that bandit m is activated and 0 indicates that it is not activated, and plug in the following choices of $\mathbf{D}^m(\cdot)$ and $\mathbf{b}$ for the coupling constraint in Equation (7):

$$\mathbf{D}^m(a^m) = \begin{bmatrix} \mathbb{I}\{a^m = 1\} \\ -\mathbb{I}\{a^m = 1\} \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}. \tag{16}$$

This coupling constraint requires that exactly one bandit be activated. It is then easy to see that the dual CLR problem (9) exactly coincides with problem (15), leading to the following result.

Proposition 8. *For each* $\mathbf{s} \in \mathcal{S}$, $Z^*_{\text{CLR}}(\mathbf{s}) = Z^*_{\text{BNM}}(\mathbf{s})$.

It turns out that problem (15) and the ALR problem (11) for the problem as defined in Section 5.1 are in fact the same, in that they lead to the same objective value.

Proposition 9. *For each* $\mathbf{s} \in \mathcal{S}$, $Z^*_{\text{ALR}}(\mathbf{s}) = Z^*_{\text{BNM}}(\mathbf{s})$.

The proof (found in Section EC.1.9 in the electronic companion) consists of showing that the optimal solution of one can be used to construct a feasible solution for the other. An immediate corollary of this result and Theorem 3 is that the finite fluid formulation bound $Z^*_T(\mathbf{s})$ is at least as tight as the performance region bound $Z^*_{BNM}(\mathbf{s})$.

Corollary 3. *For each* $\mathbf{s} \in \mathcal{S}$ *and* $T \in \{1, 2, \ldots\}$, $Z^*_T(\mathbf{s}) \leqslant Z^*_{\text{BNM}}(\mathbf{s})$.

Combining Theorem 2 and Propositions 4, 8, and 9, we obtain the following corollary which summarizes the ordering of all bounds for this problem.

Corollary 4. *For the bandit problem defined in Section 5.1, for all* $\mathbf{s} \in \mathcal{S}$ *and* $T \in \{1, 2, \ldots\}$:

$$\begin{aligned} J^*(\mathbf{s}) &\leqslant Z^*_T(\mathbf{s}) \leqslant \cdots \leqslant Z^*_2(\mathbf{s}) \leqslant Z^*_1(\mathbf{s}) \\ &\leqslant Z^*_{\text{ALO}}(\mathbf{s}) = Z^*_{\text{ALR}}(\mathbf{s}) = Z^*_{\text{CLR}}(\mathbf{s}) = Z^*_{\text{BNM}}(\mathbf{s}). \end{aligned}$$

Table 2. Objective value results (in %) for infinite horizon experiment, $M = 5$, $n = 4$, for instance 1 of sets REG.SAR and RSTLS.SAR.

Set	Instance	β	Method (h)	$\mathcal{O}_{\text{mean},\,h}$	$\mathcal{O}_{95,\,h}$	$\mathcal{O}_{\max,\,h}$
REG.SAR	1	0.5	Fluid, $T = 1$	1.3511	2.7801	4.3253
			Fluid, $T = 5$	0.4161	0.8019	**1.2562**
			Fluid, $T = 10$	**0.4015**	**0.7976**	**1.2562**
			ALR	2.0973	5.2572	7.0816
		0.9	Fluid, $T = 1$	1.8020	3.9777	6.7438
			Fluid, $T = 5$	0.6383	1.2568	1.9959
			Fluid, $T = 10$	**0.3879**	**0.6368**	**0.6997**
			ALR	2.5094	5.4317	8.7091
		0.95	Fluid, $T = 1$	0.7311	1.7283	2.7162
			Fluid, $T = 5$	0.2679	0.6317	0.9857
			Fluid, $T = 10$	**0.1518**	**0.2938**	**0.3639**
			ALR	0.9774	2.2040	3.4264
		0.99	Fluid, $T = 1$	0.0334	0.0862	0.1391
			Fluid, $T = 5$	0.0132	0.0343	0.0556
			Fluid, $T = 10$	**0.0081**	**0.0170**	**0.0224**
			ALR	0.0436	0.1063	0.1708
RSTLS.SAR	1	0.5	Fluid, $T = 1$	2.9222	5.2561	6.8915
			Fluid, $T = 5$	**2.6406**	**4.1177**	**5.2122**
			Fluid, $T = 10$	**2.6406**	**4.1177**	**5.2122**
			ALR	4.5558	10.0487	52.7464
		0.9	Fluid, $T = 1$	4.7079	5.6050	6.5985
			Fluid, $T = 5$	**4.5995**	**5.1573**	**5.5452**
			Fluid, $T = 10$	**4.5995**	**5.1573**	**5.5452**
			ALR	5.4381	8.3128	13.6692
		0.95	Fluid, $T = 1$	4.8905	5.3635	5.8881
			Fluid, $T = 5$	**4.8336**	**5.1322**	**5.3391**
			Fluid, $T = 10$	**4.8336**	**5.1322**	**5.3391**
			ALR	5.2499	6.6973	9.2057
		0.99	Fluid, $T = 1$	5.0305	5.1297	5.2383
			Fluid, $T = 5$	**5.0187**	**5.0810**	**5.1239**
			Fluid, $T = 10$	**5.0187**	**5.0810**	**5.1239**
			ALR	5.1015	5.3931	5.8688

Note. In each instance, value of β and metric, the best value is indicated in bold.

In Sections 5.4, we will show that the inequality between $Z_T^*(\mathbf{s})$ and the four equivalent bounds—$Z_{\text{ALO}}^*(\mathbf{s})$, $Z_{\text{ALR}}^*(\mathbf{s})$, $Z_{\text{CLR}}^*(\mathbf{s})$, and $Z_{\text{BNM}}^*(\mathbf{s})$—can be strict.

Since the fluid formulation provides a bound that is at least as tight as the performance region formulation, it would seem reasonable to expect that the heuristic policy derived from the fluid formulation to give performance that is generally as good as, if not better than, that of the primal dual heuristic derived from the performance region formulation in Bertsimas and Niño-Mora (2000). In Section 5.5, we will show that this is indeed the case, and that in fact the fluid-based heuristic significantly outperforms the primal dual heuristic of Bertsimas and Niño-Mora (2000).

5.4. Bound Comparison

We begin the discussion of our numerical results by comparing the bound generated by our fluid optimization model to the bound generated by the ALR on medium-scale instances. For the fluid approach, we considered T values of 1, 5, and 10; for values of $T > 10$, the metric values changed negligibly relative to $T = 10$.

We set the number of bandits M to 5 and the number of states of each bandit n to 4, resulting in $5^4 = 1024$ system states. In each bandit state space, we number the states from 1 to n (i.e., $\mathcal{S}^m = \{1, 2, \ldots, n\}$). We generated four different sets of five instances, with the following structure:

- REG.SAR, consisting of regular multiarmed bandits, where the reward g_{km}^m was set as $g_{km}^m = (10/n) \cdot k$ for every bandit m and state k. Each active transition probability vector was drawn uniformly from the $(n-1)$-dimensional unit simplex.
- RSTLS.SAR, consisting of restless bandits, with the same reward structure as REG.SAR. Each active and passive transition probability vector was drawn uniformly from the $(n-1)$-dimensional unit simplex.
- RSTLS.SBR, consisting of restless bandits, where the active reward g_{km}^m was set as $g_{km}^m = (10/n) \cdot k$ for every m and k, and the passive reward g_{ka}^m for $a \neq m$ was set to ρ_k^m, where $\rho_k^m = (1/M) \cdot (10/n) \cdot k$ for each m and k. Each active and passive transition probability vector was drawn uniformly from the $(n-1)$-dimensional unit simplex.

Table 3. Objective value results (in %) for infinite horizon experiment, $M = 5$, $n = 4$, for instance 1 of sets RSTLS.SBR and RSTLS.DET.SBR.

Set	Instance	β	Method (h)	$\mathcal{O}_{\text{mean},h}$	$\mathcal{O}_{95,h}$	$\mathcal{O}_{\max,h}$
RSTLS.SBR	1	0.5	Fluid, $T = 1$	1.1795	2.0612	3.2705
			Fluid, $T = 5$	**1.1426**	**1.8816**	**2.4596**
			Fluid, $T = 10$	**1.1426**	**1.8816**	**2.4596**
			ALR	2.5059	8.1574	27.8741
		0.9	Fluid, $T = 1$	2.1639	2.5757	2.7914
			Fluid, $T = 5$	**2.1286**	**2.4244**	**2.6105**
			Fluid, $T = 10$	**2.1286**	**2.4244**	**2.6105**
			ALR	2.6231	4.3336	6.5445
		0.95	Fluid, $T = 1$	2.2758	2.4911	2.6111
			Fluid, $T = 5$	**2.2564**	**2.4104**	**2.5125**
			Fluid, $T = 10$	**2.2564**	**2.4104**	**2.5125**
			ALR	2.5039	3.3414	4.3872
		0.99	Fluid, $T = 1$	2.3645	2.4090	2.4350
			Fluid, $T = 5$	2.3604	2.3927	2.4140
			Fluid, $T = 10$	**2.3604**	**2.3927**	**2.4140**
			ALR	2.4099	2.5735	2.7749
RSTLS.DET.SBR	1	0.5	Fluid, $T = 1$	0.5910	3.2775	7.7156
			Fluid, $T = 5$	0.0403	0.2744	0.6560
			Fluid, $T = 10$	**0.0152**	**0.0664**	**0.6012**
			ALR	1.7255	7.8925	23.0435
		0.9	Fluid, $T = 1$	0.6112	2.2488	3.5360
			Fluid, $T = 5$	0.0975	0.4078	**1.4104**
			Fluid, $T = 10$	**0.0816**	**0.3973**	**1.4104**
			ALR	1.0275	3.3645	4.6934
		0.95	Fluid, $T = 1$	0.4020	1.3776	2.1960
			Fluid, $T = 5$	0.0662	0.3673	**0.9590**
			Fluid, $T = 10$	**0.0609**	**0.3446**	**0.9590**
			ALR	0.5910	1.8808	2.6698
		0.99	Fluid, $T = 1$	0.1102	0.3352	0.5066
			Fluid, $T = 5$	0.0223	0.0995	**0.2475**
			Fluid, $T = 10$	**0.0158**	**0.0984**	**0.2475**
			ALR	0.1497	0.3972	0.5404

Note. In each instance, value of β and metric, the best value is indicated in bold.

• RSTLS.DET.SBR, consisting of restless bandits, where the reward structure is the same as RSTLS.SBR. Each active and passive transition probability matrix was generated by permuting the rows of the n-dimensional identity matrix uniformly at random (i.e., transition matrices are still randomly generated, but the transitions that they govern are now *deterministic*).

The reason for considering the types of reward structures in sets REG.SAR, RSTLS.SAR, RSTLS.SBR, and RSTLS.DET.SBR is that in these sets of instances, the reward structures of any two bandits are identical, but they are different in their probabilistic structure. In order for a method to be successful, therefore, it must be able to recognize that the bandits are different in their probabilistic structure, which will directly affect the long-term expected reward that the method could possibly garner from each bandit. We would expect that the greedy method, which only uses reward information, would perform rather poorly on these instances. RSTLS.SAR and RSTLS.SBR are interesting to consider together because passive rewards are zero in the former and nonzero in the latter. RSTLS.DET.SBR is interesting as it is not stochastic and thus constitutes a potentially pathological instance set.

To compare the bounds, we define three different metrics as follows. Given a method h for solving the problem that is based on an optimization formulation, let $Z_h(\mathbf{s})$ be the objective value (upper bound) generated at $\mathbf{s}$. Define $\text{RD}_h(\mathbf{s}) = 100\% \times (Z_h(\mathbf{s}) - J^*(\mathbf{s}))/J^*(\mathbf{s})$ as the relative difference between $Z_h(\mathbf{s})$ and the optimal value function $J^*(\mathbf{s})$. Then, define the metrics $\mathcal{O}_{\text{mean},h}$, $\mathcal{O}_{P,h}$, and $\mathcal{O}_{\max,h}$ as the mean, Pth percentile and maximum of $\{\text{RD}_h(\mathbf{s})\}_{\mathbf{s}\in\mathcal{S}}$. In general, the lower the values of the $\mathcal{O}$ metrics, the closer the bound is to the true optimal value function; a value of zero for $\mathcal{O}_{\text{mean},h}$ implies that the bound/objective value is exactly equal to the true optimal value function. We will consider these metrics for the fluid and the ALR formulations. We compute the optimal value function J^* using value iteration.

Tables 2 and 3 compare the objective values obtained from the fluid formulation and from the ALR problem (11) with the optimal objective value. We only show the first instance from each set, as the results for the other instances in each instance set were qualitatively similar. Recall that by

Table 4. Large-scale policy performance and runtime simulation results for $M \in \{5, 10\}$, $n \in \{5, 10, 20\}$ RSTLS.DET.SBR instances.

Instance	β	Method (h)	$G_{\text{mean},h}$	Std. err.	$U_{\text{mean},h}$	Std. err.	$T_{\text{mean},h}$	Std. err.
$M=5, n=5$	0.99	Fluid, $T=1$	10.3367	(0.352)	0.2319	(0.009)	0.002	(0.00)
		Fluid, $T=2$	6.4168	(0.016)	0.1870	(0.008)	0.002	(0.00)
		Fluid, $T=5$	4.0224	(0.008)	0.0976	(0.006)	0.006	(0.00)
		Fluid, $T=10$	4.2265	(0.062)	0.0000	(0.000)	0.016	(0.00)
		Greedy	24.9822	(0.255)	—	—	—	—
		ALR	9.2136	(0.079)	0.2858	(0.009)	0.003	(0.00)
		BNMPD	31.6512	(1.407)	0.2858	(0.009)	0.001	(0.00)
$M=5, n=10$	0.99	Fluid, $T=1$	4.6314	(0.117)	0.1898	(0.008)	0.003	(0.00)
		Fluid, $T=2$	4.4024	(0.120)	0.1530	(0.007)	0.003	(0.00)
		Fluid, $T=5$	2.7667	(0.039)	0.0664	(0.004)	0.009	(0.00)
		Fluid, $T=10$	2.7258	(0.032)	0.0000	(0.000)	0.033	(0.00)
		Greedy	21.1001	(0.395)	—	—	—	—
		ALR	13.0486	(0.377)	0.2389	(0.010)	0.007	(0.00)
		BNMPD	31.9437	(0.690)	0.2389	(0.010)	0.002	(0.00)
$M=5, n=20$	0.99	Fluid, $T=1$	7.4533	(0.190)	0.2898	(0.011)	0.005	(0.00)
		Fluid, $T=2$	6.7666	(0.079)	0.2254	(0.009)	0.005	(0.00)
		Fluid, $T=5$	5.8754	(0.171)	0.1166	(0.006)	0.014	(0.00)
		Fluid, $T=10$	5.2519	(0.192)	0.0000	(0.000)	0.056	(0.00)
		Greedy	22.2220	(0.334)	—	—	—	—
		ALR	18.3783	(0.194)	0.3557	(0.012)	0.013	(0.00)
		BNMPD	36.8163	(0.724)	0.3557	(0.012)	0.004	(0.00)
$M=10, n=5$	0.99	Fluid, $T=1$	3.4065	(0.152)	0.0860	(0.005)	0.006	(0.00)
		Fluid, $T=2$	1.6663	(0.147)	0.0596	(0.004)	0.008	(0.00)
		Fluid, $T=5$	1.0393	(0.068)	0.0209	(0.002)	0.032	(0.00)
		Fluid, $T=10$	1.0984	(0.085)	0.0000	(0.000)	0.091	(0.00)
		Greedy	10.2405	(0.406)	—	—	—	—
		ALR	8.4630	(0.578)	0.1218	(0.007)	0.008	(0.00)
		BNMPD	34.9582	(1.601)	0.1218	(0.007)	0.002	(0.00)
$M=10, n=10$	0.99	Fluid, $T=1$	1.8662	(0.117)	0.1611	(0.007)	0.010	(0.00)
		Fluid, $T=2$	1.4467	(0.066)	0.1263	(0.006)	0.013	(0.00)
		Fluid, $T=5$	1.4348	(0.095)	0.0532	(0.003)	0.056	(0.00)
		Fluid, $T=10$	1.4386	(0.107)	0.0000	(0.000)	0.150	(0.00)
		Greedy	18.4851	(0.334)	—	—	—	—
		ALR	16.0600	(0.873)	0.1902	(0.007)	0.022	(0.00)
		BNMPD	34.3912	(1.470)	0.1902	(0.007)	0.003	(0.00)
$M=10, n=20$	0.99	Fluid, $T=1$	2.8752	(0.130)	0.3040	(0.010)	0.018	(0.00)
		Fluid, $T=2$	2.5045	(0.090)	0.2722	(0.011)	0.023	(0.00)
		Fluid, $T=5$	2.1928	(0.060)	0.1579	(0.009)	0.088	(0.00)
		Fluid, $T=10$	1.8246	(0.061)	0.0000	(0.000)	0.313	(0.00)
		Greedy	22.6095	(0.223)	—	—	—	—
		ALR	26.9484	(0.416)	0.3488	(0.011)	0.068	(0.00)
		BNMPD	40.5795	(0.381)	0.3488	(0.011)	0.004	(0.00)

Theorem 2, Propositions 8 and 9 that the ALR problem (11), the CLR problem (8), the ALO problem (5), and the performance region problem (15) all yield the same objective value for a fixed initial state $\mathbf{s}$. We can see that for every instance and every discount factor, the objective values from the fluid formulation are closer to the optimal objective than those from the ALR and, by extension, those from the ALO, CLR, and performance region formulations. We can also see that although part (b) of Theorem 3 indicates that the fluid bound does not worsen as T increases, there is negligible improvement beyond the $T=5$ to $T=10$ range. This suggests that we can obtain substantially better state-wise bounds than the ALR formulation by solving only a modestly larger LO problem.

5.5. Large-Scale Bandit Results

So far, we have focused on bandit problems that are relatively small, and we have compared the bounds from the different methods. In this section, we compare the *policy performance* of the methods on larger instances, where the optimal value function is unavailable to us and where we must resort to simulation. We consider instances that are generated in the same way as the RSTLS.DET.SBR set of the previous section, for values of M in $\{5, 10, 15, 20\}$ and values of n in $\{5, 10, 20\}$. We restrict ourselves to a discount factor of $\beta = 0.99$ and simulate the system for 500 steps. We consider the fluid heuristic with T values of 1, 2, 5, and 10; the ALR approach; the primal dual heuristic of Bertsimas and Niño-Mora (2000); and the greedy heuristic (which activates

Table 5. Large-scale policy performance and runtime simulation results for $\beta = 0.99$, $M \in \{15, 20\}$, $n \in \{5, 10, 20\}$ RSTLS.DET.SBR instances.

Instance	β	Method (h)	$G_{\text{mean},h}$	Std. err.	$U_{\text{mean},h}$	Std. err.	$T_{\text{mean},h}$	Std. err.
$M = 15$, $n = 5$	0.99	Fluid, $T = 1$	0.7693	(0.030)	0.0359	(0.003)	0.020	(0.00)
		Fluid, $T = 2$	0.7681	(0.032)	0.0172	(0.002)	0.028	(0.00)
		Fluid, $T = 5$	0.7015	(0.013)	0.0039	(0.001)	0.143	(0.00)
		Fluid, $T = 10$	0.7001	(0.013)	0.0000	(0.000)	0.218	(0.00)
		Greedy	9.1309	(0.216)	—	—	—	—
		ALR	4.3748	(0.428)	0.0523	(0.004)	0.022	(0.00)
		BNMPD	28.2468	(0.800)	0.0523	(0.004)	0.002	(0.00)
$M = 15$, $n = 10$	0.99	Fluid, $T = 1$	2.2354	(0.054)	0.1176	(0.006)	0.025	(0.00)
		Fluid, $T = 2$	1.5200	(0.023)	0.0931	(0.005)	0.036	(0.00)
		Fluid, $T = 5$	1.3230	(0.006)	0.0355	(0.003)	0.137	(0.00)
		Fluid, $T = 10$	1.1694	(0.004)	0.0000	(0.000)	0.533	(0.00)
		Greedy	11.8774	(0.160)	—	—	—	—
		ALR	21.4492	(0.302)	0.1443	(0.006)	0.091	(0.00)
		BNMPD	34.4392	(0.535)	0.1443	(0.006)	0.004	(0.00)
$M = 15$, $n = 20$	0.99	Fluid, $T = 1$	2.1894	(0.081)	0.2589	(0.010)	0.046	(0.00)
		Fluid, $T = 2$	2.0305	(0.040)	0.2312	(0.011)	0.062	(0.00)
		Fluid, $T = 5$	1.7992	(0.032)	0.1106	(0.006)	0.217	(0.00)
		Fluid, $T = 10$	1.6754	(0.036)	0.0000	(0.000)	1.170	(0.00)
		Greedy	13.2699	(0.106)	—	—	—	—
		ALR	14.6558	(0.174)	0.2893	(0.011)	0.354	(0.00)
		BNMPD	37.7471	(0.626)	0.2893	(0.011)	0.005	(0.00)
$M = 20$, $n = 5$	0.99	Fluid, $T = 1$	1.1169	(0.045)	0.0554	(0.004)	0.056	(0.00)
		Fluid, $T = 2$	0.8508	(0.008)	0.0318	(0.003)	0.058	(0.00)
		Fluid, $T = 5$	0.7976	(0.004)	0.0065	(0.001)	0.148	(0.00)
		Fluid, $T = 10$	0.7953	(0.004)	0.0000	(0.000)	0.418	(0.00)
		Greedy	9.9393	(0.228)	—	—	—	—
		ALR	8.2415	(0.916)	0.0816	(0.005)	0.043	(0.00)
		BNMPD	30.1071	(0.724)	0.0816	(0.005)	0.003	(0.00)
$M = 20$, $n = 10$	0.99	Fluid, $T = 1$	2.6674	(0.039)	0.1360	(0.005)	0.076	(0.00)
		Fluid, $T = 2$	2.6459	(0.032)	0.1104	(0.005)	0.069	(0.00)
		Fluid, $T = 5$	1.5767	(0.025)	0.0416	(0.003)	0.278	(0.00)
		Fluid, $T = 10$	1.3553	(0.024)	0.0000	(0.000)	1.206	(0.01)
		Greedy	9.7231	(0.106)	—	—	—	—
		ALR	12.4764	(0.206)	0.1632	(0.006)	0.203	(0.00)
		BNMPD	33.3620	(0.786)	0.1632	(0.006)	0.004	(0.00)
$M = 20$, $n = 20$	0.99	Fluid, $T = 1$	5.1385	(0.105)	0.1918	(0.007)	0.086	(0.00)
		Fluid, $T = 2$	3.3105	(0.085)	0.1614	(0.006)	0.100	(0.00)
		Fluid, $T = 5$	1.5725	(0.039)	0.0822	(0.004)	0.487	(0.00)
		Fluid, $T = 10$	1.2369	(0.026)	0.0000	(0.000)	2.657	(0.01)
		Greedy	11.2028	(0.134)	—	—	—	—
		ALR	11.9748	(0.090)	0.2238	(0.007)	0.089	(0.00)
		BNMPD	36.2378	(0.759)	0.2238	(0.007)	0.005	(0.00)

the arm that leads to the highest immediate reward). For the ALR approach, we re-solve it at each new state $\mathbf{s}$ and take the action that is greedy with respect to the value function approximation $\mathbf{V}$. Note that we do not consider the policy that is greedy with respect to the value function approximation from the ALO formulation (5) since by part (b) of Theorem 2, any value function approximation that is optimal for the ALR (11) is a value function approximation that is optimal for the ALO (5), and vice versa. Similarly, we do not consider the policy that arises from the CLR formulation (8), since by Propositions 8 and 9, the CLR and ALR formulations are equivalent for this problem.

To compare the methods, for each pair (M, n), we generate $K = 100$ random initial states $\mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(K)}$ by uniformly selecting one of the n states for each component. We simulate each policy h from each initial state $\mathbf{s}^{(k)}$ to obtain a realized reward $J_{k,h}$. We also compute the initial objective value $Z_{k,h}$ of the policy h (where applicable) at each initial state. For each initial state $\mathbf{s}^{(k)}$ and method h, we thus obtain a gap value $G_{k,h}$, defined as

$$G_{k,h} = 100\% \times \frac{Z_k^* - J_{k,h}}{Z_k^*},$$

where $Z_k^* = \min_h Z_{k,h}$ is the lowest upper bound available (in this set of experiments, this is the fluid method with the largest value of T). We then consider the mean value of $\{G_{k,h}\}_{k=1}^K$ for each method h, which we report as $G_{\text{mean},h}$. In addition, for each initial state $\mathbf{s}^{(k)}$ and method h based on

a mathematical optimization formulation, we compute the relative difference $U_{k,h}$ between the upper bound from h and the best upper bound, defined as

$$U_{k,h} = 100\% \times \frac{Z_{k,h} - Z_k^*}{Z_k^*},$$

and we compute the mean over the K initial states as $U_{\text{mean},h}$. Finally, for each initial state $\mathbf{s}^{(k)}$ and each method h that is based on an optimization formulation, we compute $T_{k,h}$, which is the average solution time in seconds of the underlying formulation over all of the steps of the simulation. We then consider the mean value of $\{T_{k,h}\}_{k=1}^K$ for each applicable method h, which we report as $T_{\text{mean},h}$.

Tables 4 and 5 display the results from this collection of instances. With regard to policy performance, the results indicate that the fluid method delivers excellent performance, even in the most challenging instance ($M = 20$, $n = 20$), and significantly outperforms the greedy heuristic, the Lagrangian relaxation approach, and the primal dual heuristic. From a solution time perspective, the finite fluid formulation (3) does take considerably more time per action than either the performance region formulation (15) or the ALR formulation (11). However, even in the largest case ($M = 20$, $n = 20$) and for the largest value of T, the average time per action is on the order of 2.6 seconds; for certain applications, this amount of time may still be feasible.

6. Conclusion

In this paper, we have considered a fluid optimization approach for solving decomposable MDPs. The essential feature of the approach is that it models the transitions of the system at the level of individual components; in this way, the approach is tractable and scalable. We provided theoretical justification for this approach by showing that it provides tighter bounds on the optimal value than three state-of-the-art approaches. We showed computationally that this approach leads to strong performance in multiarmed bandit problems.

There are several promising directions for future research. It would be valuable to extend the approach to deal with situations where the data (e.g., the transition probabilities) are not known precisely and may become known more precisely with time. Problems of this kind fall in the domain of robust optimization (see Bertsimas et al. 2011), and it would seem that an adaptable robust version of the fluid formulation could be appropriate in this setting. At the same time, problems of this kind could also be viewed as reinforcement learning problems. One approach from this direction could involve combining the fluid approach with posterior sampling (see, e.g., Russo and Van Roy 2014): in this approach, one would maintain a distribution over the problem data, and at each period, one would take a sample from this distribution, solve the fluid problem corresponding to the sample to determine the action to take and update the distribution with the realized reward and transitions from that action. Exploring the benefits of such a scheme, as well as other ways of combining the fluid method with reinforcement learning methods, thus constitutes another interesting direction of future work.

Supplemental Material

Supplemental material to this paper is available at http://dx.doi.org/10.1287/opre.2016.1531.

Acknowledgments

The authors thank the area editor Costis Maglaras, the associate editor, and the three anonymous referees for their detailed and thoughtful reviews that have significantly improved the paper, particularly with regard to the connection between the fluid approach and the method of Adelman and Mersereau (2008). The second author was partially supported by a Postgraduate Scholarship–Doctoral (PGS-D) award from the Natural Sciences and Engineering Research Council (NSERC) of Canada.

References

Adelman D (2007) Dynamic bid prices in revenue management. *Oper. Res.* 55(4):647–661.

Adelman D, Mersereau AJ (2008) Relaxations of weakly coupled stochastic dynamic programs. *Oper. Res.* 56(3):712–727.

Anderson EJ, Nash P (1987) *Linear Programming in Infinite-Dimensional Spaces* (John Wiley & Sons, Chichester, UK).

Bellman R (1957) *Dynamic Programming* (Princeton University Press, Princeton, NJ).

Bellman R (1961) *Adaptive Control Processes: A Guided Tour*, Vol. 4 (Princeton University Press, Princeton, NJ).

Bertsekas DP (1995) *Dynamic Programming and Optimal Control*, Vol. 1 (Athena Scientific, Belmont, MA).

Bertsekas DP, Tsitsiklis JN (1996) *Neuro-Dynamic Programming* (Athena Scientific, Belmont, MA).

Bertsimas D (1995) The achievable region method in the optimal control of queueing systems; formulations, bounds and policies. *Queueing Systems: Theory Appl.* 21(3–4):337–389.

Bertsimas D, Niño-Mora J (1996) Conservation laws, extended polymatroids and multiarmed bandit problems; a polyhedral approach to indexable systems. *Math. Oper. Res.* 21(2):257–306.

Bertsimas D, Niño-Mora J (2000) Restless bandits, linear programming relaxations, and a primal-dual index heuristic. *Oper. Res.* 48(1): 80–90.

Bertsimas D, Brown DB, Caramanis C (2011) Theory and applications of robust optimization. *SIAM Rev.* 53(3):464–501.

Boyd S, Parikh N, Chu E, Peleato B, Eckstein J (2011) Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations Trends Machine Learn.* 3(1):1–122.

Coffman EG, Mitrani I (1980) A characterization of waiting time performance realizable by single-server queues. *Oper. Res.* 28(3):810–821.

de Farias DP, Van Roy B (2003) The linear programming approach to approximate dynamic programming. *Oper. Res.* 51(6):850–865.

de Farias DP, Van Roy B (2004) On constraint sampling in the linear programming approach to approximate dynamic programming. *Math. Oper. Res.* 29(3):462–478.

Federgruen A, Groenevelt H (1988) Characterization and optimization of achievable performance in general queueing systems. *Oper. Res.* 36(5):733–741.

Ghate A, Smith RL (2013) A linear programming approach to nonstationary infinite-horizon markov decision processes. *Oper. Res.* 61(2): 413–425.

Goldfarb D, Ma S (2012) Fast multiple-splitting algorithms for convex optimization. *SIAM J. Optim.* 22(2):533–556.

Hawkins JT (2003) A Lagrangian decomposition approach to weakly coupled dynamic optimization problems and its applications. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA.

Heyman DP, Sobel MJ (1984) *Stochastic Models in Operations Research. Vol. 2, Stochastic Optimization* (McGraw-Hill, New York).

Howard RA (1971) *Dynamic Probabilistic Systems, Volume* II*: Semi-Markov and Decision Processes* (Dover, Mineola, NY).

Lee I, Epelman MA, Romeijn HE, Smith RL (2013) A linear programming approach to constrained nonstationary infinite-horizon markov decision processes. Technical Report 13-01, Ann Arbor, MI: University of Michigan, Department of Industrial and Operations Engineering.

Powell WB (2007) *Approximate Dynamic Programming: Solving the Curses of Dimensionality* (Wiley-Interscience, Hoboken, NJ).

Puterman ML (1994) *Markov Decision Processes: Discrete Dynamic Stochastic Programming* (John Wiley & Sons, Chichester, UK).

Romeijn HE, Smith RL, Bean JC (1992) Duality in infinite dimensional linear programming. *Math. Programming* 53(1–3):79–97.

Russo D, Van Roy B (2014) Learning to optimize via posterior sampling. *Math. Oper. Res.* 39(4):1221–1243.

Shanthikumar JG, Yao DD (1992) Multiclass queueing systems: polymatroidal structure and optimal scheduling control. *Oper. Res.* 40(3):S293–S299.

Van Roy B (2002) Neuro-dynamic programming: Overview and recent trends. *Handbook of Markov Decision Processes* (Springer, New York), 431–459.

Dimitris Bertsimas is the Boeing Professor of Operations Research and codirector of the Operations Research Center at the Massachusetts Institute of Technology. His interests include analytics, optimization, and stochastics. This paper is part of his long-term interest in optimization under uncertainty.

Velibor V. Mišić is an assistant professor of Decisions, Operations, and Technology Management at the Anderson School of Management, University of California, Los Angeles. His research interests are in analytics, operations management, and optimization under uncertainty, of which this paper is a part.