OPERATIONS RESEARCH
Vol. 58, No. 1, January–February 2010, pp. 214–228
ISSN 0030-364X | EISSN 1526-5463 | 10 | 5801 | 0214

informs

DOI 10.1287/opre.1090.0697

Partially Observable Markov Decision Processes: A Geometric Technique and Analysis

Hao Zhang
Marshall School of Business, University of Southern California, Los Angeles, California 90089,
zhanghao@usc.edu

This paper presents a novel framework for studying partially observable Markov decision processes (POMDPs) with finite state, action, observation sets, and discounted rewards. The new framework is solely based on future-reward vectors associated with future policies, which is more parsimonious than the traditional framework based on belief vectors. It reveals the connection between the POMDP problem and two computational geometry problems, i.e., finding the vertices of a convex hull and finding the Minkowski sum of convex polytopes, which can help solve the POMDP problem more efficiently. The new framework can clarify some existing algorithms over both finite and infinite horizons and shed new light on them. It also facilitates the comparison of POMDPs with respect to their degree of observability, as a useful structural result.

Subject classifications: dynamic programming: Markov; analysis of algorithms: computational complexity; mathematics: combinatorics; computers/computer science: artificial intelligence.
Area of review: Optimization.
History: Received January 2008; revision received July 2008; accepted November 2008. Published online in *Articles in Advance* July 29, 2009.

1. Introduction and Literature Review

Markov decision processes (MDPs) provide one of the fundamental models in operations research, in which a decision maker controls the evolution of a dynamic system. Although this model is mature, with well-developed theories, as in Puterman (1994), it is based on the assumption that the state of the system can be perfectly observed. Partially observable Markov decision processes (POMDPs) extend the MDPs by relaxing this assumption. The first explicit POMDP model is commonly attributed to Drake (1962), and it attracted the attention of researchers and practitioners in operations research, computer science, and beyond. However, this problem is well known for its computational complexity. The pioneering work of Sondik (1971) and Smallwood and Sondik (1973) addressed this issue first, and since then the study of POMDPs has mainly followed two directions: (1) finding computationally feasible algorithms for the general model over both finite and infinite horizons, and (2) finding structural results for special POMDP models with well-defined applications, such as machine replacement and quality control problems. On both fronts, the research encountered significant obstacles. Strictly speaking, the problem does not have an efficiently computable solution because computing an optimal policy is PSPACE-complete, as shown by Papadimitriou and Tsitsiklis (1987), and finding an ε-optimal policy is NP-hard, as pointed out by Lusena et al. (2001). However, because of its modeling power and potentially wide applications, the problem continues to draw attention from the operations research and computer science communities.

A central concept underlying most existing methods in the POMDP literature is the "belief vector," which is a distribution of the system state (assuming a finite state space). Because the system state cannot be observed, the decision maker maintains a belief about the state and updates it after each (imperfect) observation. It is known that the belief vector is a sufficient statistic of the complete information history (Aoki 1965, Astrom 1965, Bertsekas 1976). Thus, the problem can be formulated by dynamic programming based on belief vectors. The optimality condition implies that the part of an optimal policy from any period onward maximizes the expected value-to-go with respect to the belief vector at the beginning of that period. A belief vector of an n-state system belongs to the $(n-1)$-dimensional simplex of $\mathbb{R}^n$, which is the main source of difficulty in this approach. Much of the literature discusses techniques to replace this uncountable set with a finite or countable set.

In this paper, we propose a new perspective and framework for analyzing the POMDP problem, based on some strong geometric properties of the problem and free from the belief vectors. We show that the key step of the problem formulation is equivalent to an existing problem in computational geometry, the "Minkowski sum" of convex polytopes. This connection opens the door for using the state-of-the-art Minkowski-sum algorithms (e.g., Fukuda 2004) to solve the POMDP problem, which may improve the computational efficiency of the latter substantially.

Other steps of the problem formulation are also related to a computational geometry problem, i.e., identifying the vertices of the convex hull of a point set. Geometric intuitions can enhance our understanding of the problem and facilitate algorithm design and structure exploration.

The belief vectors are completely suppressed in this framework. We view the part of a policy from any period onward as a stand-alone object of its own, hereafter referred to as a "continuation policy." Because the system state is hidden, a continuation policy yields an expected value-to-go from each possible state, which gives rise to a vector of expected values-to-go, hereafter referred to as a "continuation-value vector." We present a dynamic programming formulation based on these continuation-value vectors. The backward induction yields a collection of continuation-value vectors in each period that form a so-called "continuation-value frontier," where no vector on the frontier dominates another (or, weakly greater than the other in all dimensions). For a finite-horizon problem, the backward induction ends when the continuation-value frontier in the first period is found, and each vector on that frontier is optimal for certain distributions of the initial system state. This is the only place where the (initial) belief vector comes into play in this framework. For an infinite-horizon problem with discounting, there exists a unique continuation-value frontier, which is the same in every period and can be approximated by successive value iterations or policy iterations. We note in the paper that the traditional framework based on belief vectors and the new framework based on continuation-value vectors can be reconciled.

Some researchers have used the notion of continuation policies and continuation-value vectors in the POMDP literature before. However, this is the first time these concepts have been systematically treated and embedded in a rigorous framework. To the best knowledge of the author, the connection between the POMDP problem and the computational geometry problems, the concept of continuation-value frontier, the adoption of the Hausdorff metric, the new dynamic programming formulation, and the conceptual distinction between the so-called "finite-state controllers" and "finite-memory policies" are new to the literature. The new framework also facilitates a rigorous pursuit of the notion of observability. For a class of POMDPs sharing the same underlying MDP, we define a partial order through their observation matrices, which translates into a partial order of their value frontiers.

In the remainder of this section, we provide a brief literature review to offer a bird's-eye view of the rich history of the problem. Monahan (1982), Lovejoy (1991b), and White (1991) provide excellent surveys of the solution techniques and applications of the model up to 1991, and Poupart (2005) includes a review of the literature up to 2005. Note that we only discuss the POMDP problem with discounted rewards in this paper. For analyses of the average-reward version of the problem, we refer the reader to Platzman (1980), Fernández-Gaucherand et al. (1991), Hsu et al. (2006), and Yu and Bertsekas (2008).

Based on researchers' affiliations, the literature can be broadly divided into two parts—that in operations research and that in computer science—and we begin with the first part. A significant stream of research aims at the solution and structure of the general POMDP problem. Sondik (1971) and Smallwood and Sondik (1973) prove that the value functions (over belief vectors) are piecewise linear and convex and can be found through a value iteration algorithm that establishes the foundation for many existing algorithms today. Monahan (1982) improves this algorithm by systematically removing the redundant elements in defining the value functions. Sondik (1978) presents a policy iteration algorithm for the infinite-horizon model, focusing on a special type of policies called "finitely transient policies." Platzman (1977) presents another policy iteration algorithm based on finite-memory policies. However, these policy iteration algorithms are not easy to comprehend (using the criteria of Lovejoy 1991b, they are not "transparent"), which explains the lack of follow-up research. White and Scherer (1989) present three successive approximation algorithms to solve the infinite-horizon problem, and in (1994) they also investigate finite-memory policies. White (1979, 1980) provides sufficient conditions for the value functions and optimal policies to be monotone for some special POMDPs, which are strengthened by Lovejoy (1987). Lovejoy (1991a) approximates the belief space by a finite grid of points and constructs upper- and lower-bound solutions.

Another important stream of research focuses on special application-driven models, especially the machine maintenance and quality control problems. In such problems, the state of a machine is unobservable, but is partially reflected in its output. The state can be revealed by a costly inspection or replacement (the latter resets the machine state). The problem is to find a minimum-cost maintenance policy. We refer the reader to Monahan (1982) for a detailed description of this problem and a summary of the earlier work by Eckles (1968), Ross (1971), Ehrenfeld (1976), Wang (1977), and White (1977, 1979). Grosfeld-Nir (1996, 2007) finds the optimal control-limit policies for a two-state replacement model, and Anily and Grosfeld-Nir (2006) study a combined production and inspection problem in which the second stage, inspection, is modeled as a POMDP. Lane (1989) presents an application of the POMDP problem for fishermen, and Treharne and Sox (2002) examine several suboptimal policies for an inventory control problem in which the demand process is generated by a hidden Markov process.

The computer science community also has a long tradition of studying this problem. Because a POMDP can be considered as a special probabilistic automaton (Paz 1971), the study is related to the automata theory. Since the 1990s, there has been a surge of research activities when the model found applications in artificial intelligence (Cassandra et al. 1994, Kaelbling et al. 1998). To this date, the model has been applied to robot navigation (Littman et al. 1995,

Cassandra et al. 1996, Thrun 2000, Montemerlo et al. 2002), preference elicitation (Boutilier 2002), stochastic resource allocation (Meuleau et al. 1998, Marbach et al. 2000), and spoken-dialogue systems (Paek and Horvitz 2000, Zhang et al. 2001), among others.

A large part of the computer science literature on the POMDP problem focuses on the computational aspect of the problem, and numerous algorithms have been proposed, most of which take the value iteration approach. Sondik's (1971) algorithm is often referred to as the "one-pass" algorithm. Cheng (1988) proposes a "linear-support" algorithm that systematically builds up the value functions, and Cassandra et al. (1994) introduce a "witness" algorithm focusing on belief vectors that can identify a missing piece of a value function. Zhang and Liu (1996) present an "incremental pruning" algorithm that decomposes each value iteration into three steps and removes the redundancy in each step. Feng and Zilberstein (2004) propose a "region-based incremental pruning" algorithm that divides the belief space into smaller regions and performs independent pruning in each region. Some algorithms also take the policy iteration approach. Hansen (1998a, b) studies a special type of policy named "finite-state controller," and Meuleau et al. (1999) study "finite-memory policies." Poupart and Boutilier (2004) presents a "bounded policy-iteration" algorithm to tackle large-scale models.

This paper is organized as follows. Section 2 introduces the POMDP model and some basic facts. Section 3 focuses on the new framework for the problem, based on its geometric properties, and §4 reveals the connection of the key step of the formulation to the Minkowski-sum problem. Section 5 is dedicated to the basic properties of the infinite-horizon model and the policy iteration approach to solve it. We investigate the degree of observability in §6 and discuss some future research topics in the final section. Proofs of technical results are provided in the appendix.

2. Model Description and Decomposition

A partially observable Markov decision process (POMDP) describes the problem faced by a decision maker who controls an underlying Markov decision process, but can only observe imperfect signals of the system state. The model can be formally described below. We largely follow the notation of Lovejoy (1991b). In addition, throughout this paper, a vector will be in column format by default, whose transpose will be marked by a "$'$" symbol.

(1) The *state space* is a finite set, denoted by $X = \{1, 2, \ldots, n\}$. (2) The *action space* is a finite set, denoted by A. (3) The *observation space* is also a finite set, $\Theta = \{1, 2, \ldots, m\}$. (4) The state *transition matrices* are given by $\{P^a\}_{a\in A}$. For each action a, P^a is an $n \times n$ matrix $(p^a_{ij})_{i,\,j\in X}$, where each element $p^a_{ij} = \Pr(j \mid i, a)$ is the probability that the system will move from state i to state j following action a. (5) The *observation matrices* are given by $\{R^a\}_{a\in A}$. For each a, R^a is an $n \times m$ matrix $(r^a_{j\theta})_{j\in X,\,\theta\in\Theta}$, where each element $r^a_{j\theta} = \Pr\{\theta \mid j, a\}$ gives the probability that signal θ will be observed after action a is taken and the system moves to the new state j. Each row of R^a sums to one. For convenience in the subsequent analysis, the diagonal matrix formed from the θth column of R^a will be denoted by $R^a(\theta) = \operatorname{diag}(r^a_{1\theta}, r^a_{2\theta}, \ldots, r^a_{n\theta})$. (6) The *reward vectors* are given by $\{g^a\}_{a\in A}$. Each g^a is an n-dimensional vector $(g^a_i)_{i\in X}$, where each element g^a_i is the expected reward in a single period given state i and action a. (7) The *terminal reward vector* (for $T < \infty$) is g_{T+1}, whose element $g_{T+1,\,i}$ is the final reward of the system in state i at the end of period T. (8) The *distribution of the initial state* is given by an n-dimensional vector π_0. (9) The *horizon* of the model is $T \leqslant \infty$, which can be finite or infinite, and the periods are indexed by $t = 1, \ldots, T$. (10) The *discount factor* is $\beta \in [0, 1]$ if $T < \infty$ or $\beta \in [0, 1)$ if $T = \infty$.

It has been shown (Sondik 1971) that the problem has a dynamic programming formulation based on the belief of the underlying system state. A belief is a distribution of the system state, denoted by an n-dimensional vector π. Let $V_t(\pi)$ be the expected future reward from period t onward given belief π. It satisfies the following equation:

$$V_t(\pi) = \max_{a\in A}\left\{\pi' g^a + \beta \sum_{\theta\in\Theta} \Pr\{\theta \mid \pi, a\} V_{t+1}(\Lambda(\pi, a, \theta))\right\}. \quad (1)$$

In the expression, $\Pr\{\theta|\pi, a\} = \sum_i \sum_j \pi_i p^a_{ij} r^a_{j\theta}$ is the probability of observing θ after state distribution (belief) π and action a. The function $\Lambda(\pi, a, \theta)$ gives the posterior belief $\widetilde{\pi}$ after prior belief π, action a, and observation θ. According to Bayes' rule, $\widetilde{\pi}_j = \sum_i \pi_i p^a_{ij} r^a_{j\theta} / \Pr\{\theta \mid \pi, a\}$. Recall that $R^a(\theta) = \operatorname{diag}(r^a_{1\theta}, r^a_{2\theta}, \ldots, r^a_{n\theta})$. Then, in matrix form,

$$\widetilde{\pi}' = \frac{\pi' P^a R^a(\theta)}{\Pr\{\theta \mid \pi, a\}}. \quad (2)$$

Sondik (1971) proves that for any finite t, $V_t(\pi)$ is piecewise linear and convex and thus admits the form $V_t(\pi) = \max_{\omega\in\Omega_t}\{\pi'\omega\}$, where Ω_t is a finite set of n-dimensional vectors. The point set Ω_t can be recursively determined as follows. Substituting $V_{t+1}(\widetilde{\pi}) = \max_{\omega\in\Omega_{t+1}}\{\widetilde{\pi}'\omega\}$ into Equation (1), we obtain

$$\begin{aligned} V_t(\pi) &= \max_{a\in A}\left\{\pi' g^a + \beta \sum_{\theta\in\Theta} \Pr\{\theta \mid \pi, a\} V_{t+1}\left(\frac{\pi' P^a R^a(\theta)}{\Pr\{\theta \mid \pi, a\}}\right)\right\} \\ &= \max_{a\in A}\left\{\pi' g^a + \beta \sum_{\theta\in\Theta} \Pr\{\theta \mid \pi, a\} \max_{\omega\in\Omega_{t+1}}\left\{\frac{\pi' P^a R^a(\theta)\omega}{\Pr\{\theta \mid \pi, a\}}\right\}\right\} \\ &= \max_{a\in A}\left\{\pi' g^a + \beta \sum_{\theta\in\Theta} \max_{\omega\in\Omega_{t+1}} \pi' P^a R^a(\theta)\omega\right\}, \\ &\qquad\qquad t = 1, 2, \ldots, T, \quad (3) \end{aligned}$$

$$V_{T+1}(\pi) = \pi' g_{T+1}, \qquad \Omega_{T+1} = \{g_{T+1}\}. \quad (4)$$

Because $V_t(\pi)$ can be expressed as $\max_{\omega\in\Omega_t}\{\pi'\omega\}$, the above expression determines Ω_t from Ω_{t+1} implicitly. However, the iteration is a difficult task—there seems to be no natural way of enumerating the members of Ω_t, and the size of Ω_t grows exponentially fast with iterations.

The value function $V_t(\pi)$ can be decomposed, as in Zhang and Liu (1996):

$$V_t^a(\pi;\theta) = \max_{\omega\in\Omega_{t+1}} \pi' P^a R^a(\theta)\omega, \tag{5}$$

$$V_t^a(\pi) = \sum_{\theta\in\Theta} V_t^a(\pi;\theta), \tag{6}$$

$$V_t(\pi) = \max_{a\in A}\{\pi' g^a + \beta V_t^a(\pi)\}. \tag{7}$$

The intermediate functions $V_t^a(\pi;\theta)$ and $V_t^a(\pi)$ are also piecewise linear and convex. Therefore, the problem becomes one of finding the minimum sets of ω-vectors that describe $V_t^a(\pi;\theta)$, $V_t^a(\pi)$, and $V_t(\pi)$, respectively. We will refer to this three-step iteration (essentially any iteration derived from Equation (1)) as a *belief-value iteration*, indicating the role of belief vectors in the formulation. As noted in the literature, the first and third steps in the iteration are relatively easy, whereas the second step, finding the minimum ω-set for $V_t^a(\pi)$, is the most time consuming. Significant efforts have been devoted to this step, and various existing algorithms mainly differ in this step.

3. A Geometric Framework

Although many researchers have recognized the central role of the ω-vectors, few have studied them independently from the belief vectors. In this section, we take a dual perspective and generate the ω-vectors in a systematic way, free from the belief vectors. We show that an ω-vector is essentially the expected value-to-go (or continuation-value) vector associated with a continuation policy (part of a complete policy from any period t onward). We propose a new dynamic programming formulation based on this view. The development in this section leads the way to a reduction of the time complexity of the critical step, to be discussed in §4.

3.1. Duality Between Value Functions and Convex Hulls

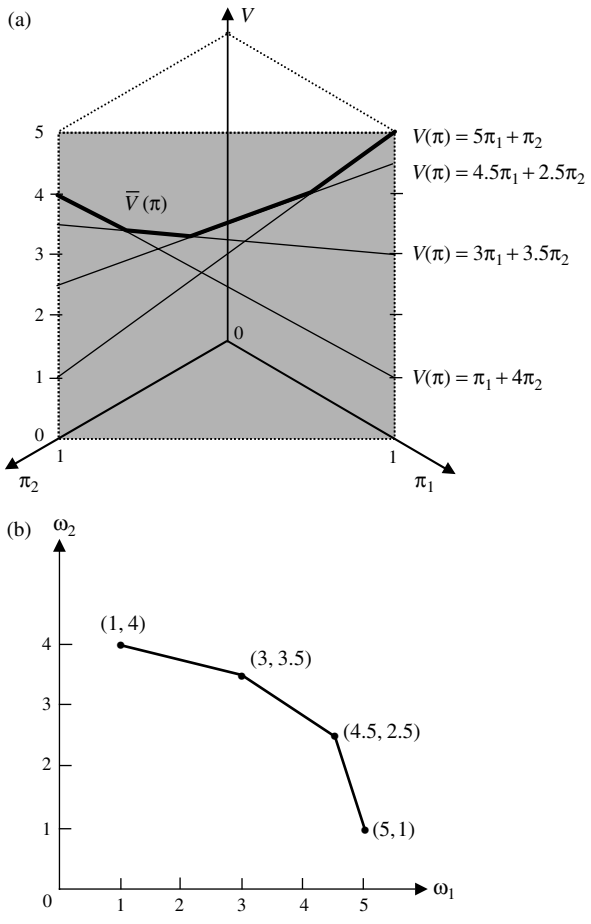
Various dual relationships between hyperplanes and points have been investigated in the convex analysis literature (Rockafellar 1970) and computational geometry literature (Berg et al. 2000, Boissonnat and Yvinec 1998). In a typical setting, a *hyperplane* in the primal $\mathbb{R}^n$ space is an affine function $\phi(u) = u'h + d$, where u is an $(n-1)$-dimensional vector, h is a given vector, and d is a given scalar. Such a hyperplane corresponds to a point (h, d) in the dual $\mathbb{R}^n$ space. The *upper envelope* (with respect to the ϕ-axis) of a set of hyperplanes $\{\phi^k(u) = u'h^k + d^k\}_{k\in K}$ defines a piecewise-linear and convex function $\bar{\phi}(u) = \max_{k\in K}\{u'h^k + d^k\}$. The minimum set of hyperplanes that completely determine the upper envelope are called *supporting hyperplanes*. One can show that (Berg et al. 2000) there is a one-to-one correspondence between the supporting hyperplanes of an upper envelope in the primal space and the vertices of an *upper convex hull* (with respect to the d-axis) in the dual space.

However, this result must be modified for the POMDP problem because the primal space is $\Pi \times \mathbb{R}$, where $\Pi = \{\pi \in \mathbb{R}^n\colon \sum_{i=1}^n \pi_i = 1 \text{ and } \pi_i \geqslant 0 \text{ for all } i\}$ is an $(n-1)$-dimensional simplex, and a hyperplane in this space is of the form $V(\pi) = \pi'\omega$, given an n-dimensional vector ω. We define the corresponding dual space as the $\mathbb{R}^n$ space of those ω-vectors. Figure 1 illustrates the dual relationship: panel (a) depicts a piecewise-linear and convex function $\bar{V}(\pi) = \max\{\pi_1 + 4\pi_2, 3\pi_1 + 3.5\pi_2, 4.5\pi_1 + 2.5\pi_2, 5\pi_1 + \pi_2\}$ in the primal space, and panel (b) depicts the corresponding points $\{(1,4), (3,3.5), (4.5,2.5), (5,1)\}$ in the dual space.

To provide a precise description of this duality, we introduce the following definitions.

Definition 1. Given a point set $S \subset \mathbb{R}^n$, the generated *convex hull* is the set $Co(S) \equiv \{\sum_{j=1}^{n+1} \lambda_j v_j\colon \sum_{j=1}^{n+1} \lambda_j = 1$

Figure 1. Duality between (a) a piecewise-linear and convex function $\bar{V}(\pi) = \max_{\omega\in\Omega}\{\pi'\omega\}$ in the primal space, and (b) the positive convex hull generated by the point set Ω in the dual space.



and $v_j \in S, \lambda_j \geqslant 0, \forall j\}$; the surface of the convex hull with positive outernormal directions, or simply the *positive convex hull* (*PCO*), is the set $PCo(S) \equiv \text{cl}(\{\omega \in Co(S)$: $\exists \pi \in \Pi^+, \pi'\omega \geqslant \pi' v, \forall v \in Co(S)\})$, where $\Pi^+ = \{\pi \in \mathbb{R}^n$: $\sum_{i=1}^n \pi_i = 1$ and $\pi_i > 0, \forall i\}$, and $\text{cl}(B)$ is the closure of B.

Definition 2. The *pairwise addition* of two point sets Ω_1 and Ω_2 is the set $\Omega_1 + \Omega_2 \equiv \{\omega_1 + \omega_2$: $\omega_1 \in \Omega_1, \omega_2 \in \Omega_2\}$.

In the definition of PCO, the closure is redundant if S is a finite set. It may be needed when S has a smooth surface, e.g., a sphere. However, it is still an open question whether smooth surfaces can arise in the POMDP problem with finite states, finite actions, and finite observations over the infinite horizon. With the above definitions, the dual relationships pertinent to the POMDP problem can be formally stated, as follows.

Lemma 1. *Suppose that $\Omega \subset \mathbb{R}^n$ is closed and bounded.* (1) *A piecewise-linear and convex function $\overline{V}(\pi) = \max_{\omega\in\Omega}\{\pi'\omega\}$, $\pi \in \Pi$, is dual to the set $PCo(\Omega)$. More precisely, for any $\widehat{\pi} \in \Pi$, there is a $\widehat{\omega} \in PCo(\Omega)$ such that $\overline{V}(\widehat{\pi}) = \widehat{\pi}'\widehat{\omega}$, and conversely, for any $\widehat{\omega} \in PCo(\Omega)$, there is a $\widehat{\pi} \in \Pi$ such that $\overline{V}(\widehat{\pi}) = \widehat{\pi}'\widehat{\omega}$.* (2) *Given two piecewise-linear and convex functions $\overline{V}_1(\pi) = \max_{\omega\in\Omega_1}\{\pi'\omega\}$ and $\overline{V}_2(\pi) = \max_{\omega\in\Omega_2}\{\pi'\omega\}$, the function $\overline{V}_1(\pi) + \overline{V}_2(\pi)$ is dual to $PCo(\Omega_1 + \Omega_2)$.* (3) *Given the above $\overline{V}_1(\pi)$ and $\overline{V}_2(\pi)$, the function $\max\{\overline{V}_1(\pi), \overline{V}_2(\pi)\}$ is dual to $PCo(\Omega_1 \cup \Omega_2)$.*

The generality of set Ω introduces some technical subtlety at the boundary of $PCo(\Omega)$, as can be seen in the proof (in the appendix). If Ω is a finite set, the duality can be obtained more directly. We can easily see that there is a one-to-one correspondence between the linear pieces of $\overline{V}(\pi)$, if any, and the vertices of $PCo(\Omega)$. Parts (1) and (2) of the lemma are illustrated in Figures 1 and 2, respectively.

3.2. POMDP Problem in the Dual Space

Lemma 1 implies that the ω-vectors representing the functions $V_t^a(\pi; \theta)$, $V_t^a(\pi)$, and $V_t(\pi)$ in expressions (5)–(7) are the vertices of some positive convex hulls. Thus, the belief-value iteration (5)–(7) has the following counterpart in the dual space, obtaining the set Ω_t from the set Ω_{t+1}:

Continuation-Value Iteration

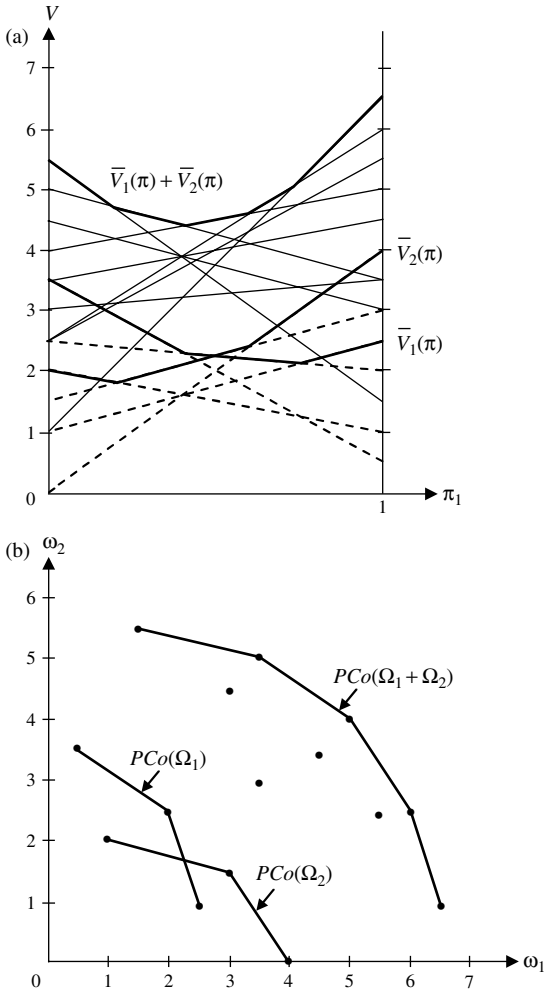
(1) $\Omega_t^a(\theta) = PCoEx(\{P^a R^a(\theta)\omega$: $\omega \in \Omega_{t+1}\})$, $a \in A$, $\theta \in \Theta$;

(2) $\Omega_t^a = PCoEx(\Omega_t^a(1) + \Omega_t^a(2) + \cdots + \Omega_t^a(m))$, $a \in A$;

(3) $\Omega_t = PCoEx(\bigcup_{a\in A}\{g^a + \beta\omega$: $\omega \in \Omega_t^a\})$.

We call this three-step iteration a *continuation-value iteration* because the ω-vectors are the "continuation-value vectors" defined in the next subsection. The set $PCoEx(\Omega)$ contains the extreme points of the positive convex hull generated from the point set Ω. Viewed as an operator,

Figure 2. Duality between $\overline{V}_1(\pi) + \overline{V}_2(\pi)$ and $PCo(\Omega_1 + \Omega_2)$.



PCoEx corresponds to a standard problem in computational geometry—removing redundant points from the convex hull of Ω. Recall that x is an *extreme point* of a convex set C if $x \neq \lambda y + (1 - \lambda)z$ for any $\lambda \in (0, 1)$ and $y, z \in C\backslash\{x\}$, and that extreme points and vertices are the same in the special case of convex polytopes. Lemma 1 implies the following result immediately:

Theorem 1. *The belief-value iteration and continuation-value iteration are equivalent.*

Like the belief-value iteration, the continuation-value iteration can be aggregated, as follows.

Lemma 2. *The continuation-value iteration is equivalent to the following iteration*:

$$\Omega_t = PCoEx\left(\bigcup_{a\in A}\left\{g^a + \beta\sum_{\theta\in\Theta} P^a R^a(\theta)\omega_{t+1}^a(\theta)\colon \omega_{t+1}^a(\theta) \in \Omega_{t+1},\ \forall\, \theta \in \Theta\right\}\right). \tag{8}$$

In the above expression, the matrix P^a can be moved outside of the sum operator. Thus, the continuation-value iteration has an alternative form in which P^a only appears in the last step.

Alternative Continuation-Value Iteration

(1) $\widetilde{\Omega}_t^a(\theta) = PCoEx(\{R^a(\theta)\omega\colon\ \omega \in \Omega_{t+1}\})$, $a \in A$, $\theta \in \Theta$;

(2) $\widetilde{\Omega}_t^a = PCoEx(\widetilde{\Omega}_t^a(1) + \widetilde{\Omega}_t^a(2) + \cdots + \widetilde{\Omega}_t^a(m))$, $a \in A$;

(3) $\Omega_t = PCoEx(\bigcup_{a\in A}\{g^a + \beta P^a\omega\colon\ \omega \in \widetilde{\Omega}_t^a\})$.

We note that the first two steps of the alternative iteration are independent of the action if the observations are, which is an advantage of this procedure. Now, we briefly compare the first step of the two continuation-value iterations, because their last two steps are similar in terms of computation. Let Ω_{t+1} be the vertex set of some positive convex hull. In step 1 of the continuation-value iteration, after a linear transformation $P^aR^a(\theta)$, the set of points $\{P^aR^a(\theta)\omega\colon \omega \in \Omega_{t+1}\}$ still forms a convex hull, but it may not have positive outernormal directions or even full dimensions. In either case, the operator *PCoEx* removes the redundant points. In step 1 of the alternative continuation-value iteration, if $R^a(\theta)$ has a full rank, the points $\{R^a(\theta)\omega\colon \omega \in \Omega_{t+1}\}$ still form a positive convex hull with no redundancy; otherwise, the resulting point set has reduced dimensions and typically contains redundant points.

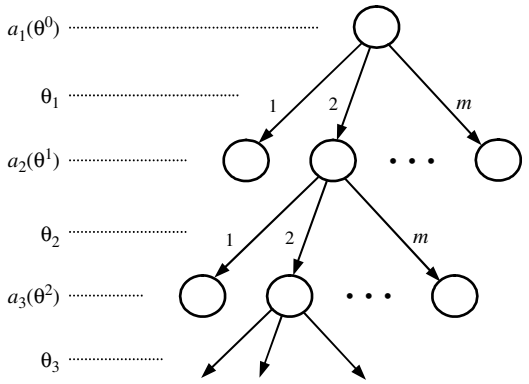
Similar to the belief-value iteration, the most time-consuming step in the two continuation-value iterations is the second one. As illustrated in Figure 2(b), the pairwise addition of two point sets is exponentially large, i.e., $|\Omega_1 + \Omega_2| = |\Omega_1| \cdot |\Omega_2|$. The corresponding situation in the primal space is illustrated in Figure 2(a). There are as many redundant points in $\Omega_1 + \Omega_2$ as there are redundant hyperplanes to define $\overline{V}_1(\pi) + \overline{V}_2(\pi)$. Almost all algorithms that tackle a general POMDP problem strive to trim down the number of hyperplanes or dual points. Although there seems to be no difference between the belief- and continuation-value iterations regarding algorithm efficiency at this moment, we will see in §4 that it is more convenient to explore the geometric properties in the dual space.

3.3. Continuation Policies and Continuation-Value Vectors

In this subsection, we connect the ω-vectors with policies. To perform backward induction, we define a policy recursively. Although recursive representations of POMDP policies exist in the literature, such as the "policy graph" in Cassandra et al. (1994), the "finite-state controller" in Hansen (1998a), and the "conditional plan" in Poupart (2005), a systematic treatment as provided here is useful.

Definition 3. We refer to the beginning of period t as *time* t. In a POMDP problem with $T \leqslant \infty$ periods, the *observation history* up to time t is the sequence of observations $(\theta_1, \ldots, \theta_{t-1})$, denoted by θ^{t-1}, where θ^0 is the empty sequence by default. Let Θ^{t-1} be the set of θ^{t-1}. A (*deterministic*) *policy* σ is a sequence of mappings $\{a_t\colon \Theta^{t-1} \to A\}_{t=1,\ldots,T}$, or $\{a_1(\theta^0), \ldots, a_T(\theta^{T-1})\}_{\theta^{T-1}\in\Theta^{T-1}}$. A (*deterministic*) *continuation policy* from time t is the part of σ from time t onward, with recursive representation $\sigma_t(\theta^{t-1}) = \{a_t(\theta^{t-1}), \{\sigma_{t+1}(\theta^{t-1}, \theta_t)\}_{\theta_t\in\Theta}\}$.

Figure 3. Tree representation of a deterministic policy.



A policy can be represented as a tree, as in Figure 3. Every node of the tree has two interpretations: Viewed from the top down, it represents the history of observations θ^{t-1} and is associated with an action $a_t(\theta^{t-1})$; viewed from the bottom up, it corresponds to a continuation policy.

The observation history is the only information needed to define a policy because the state history is unavailable and the action history is jointly determined by the observation history and the policy itself. This definition deviates from the tradition of including actions in the history as well (Monahan 1982, Lovejoy 1991b, White 1991). The actions are needed to update the belief of the system state from the initial belief π_0. If the belief process is suppressed, however, the actions no longer need to be recorded. The detachment of the belief process from the core formulation is the key to creating a more parsimonious dynamic programming expression, as in Theorem 2 below.

Because the system state is unknown, the expected value-to-go under a continuation policy σ_t is not a scalar but a vector, namely, the *continuation-value vector*, denoted by u_t. Each component $u_{t,x}$ is the expected value-to-go when the system starts in state x at time t, and the policy σ_t is followed from then on. Those u_t vectors generated by legitimate continuation policies are denoted by a set U_t. To make U_t a convex set (if needed), randomization should be introduced in continuation policies. In a *randomized policy*, the action plan in period t is a function $a_t\colon (A \times \Theta)^{t-1} \to \Delta(A)$, where $(A \times \Theta)^{t-1} = \{(a_1, \theta_1, \ldots, a_{t-1}, \theta_{t-1})\}$ is the set of action and observation histories up to time t and $\Delta(A) = \{\delta\colon \sum_{a\in A}\delta_a = 1 \text{ and } \delta_a \geqslant 0 \text{ for all } a\}$ is the set of distributions over A. In general, there is no $u_t^* \in U_t$ that dominates all others in all dimensions, so the entire frontier of U_t as defined below must be considered, which can be found by backward induction in the next theorem.

DEFINITION 4. A continuation-value vector is *feasible* if it can be generated by a randomized continuation policy. Let U_t denote the set of feasible continuation-value vectors at time t. The *continuation-value frontier* at time t is the set $PCo(U_t)$, denoted by $\overline{U}_t$. The *extreme points* of the time-t continuation-value frontier are denoted by set U_t^*.

THEOREM 2. *The feasible continuation-value sets U_t, the continuation-value frontiers $\overline{U}_t$, and the extreme-point sets U_t^* can be determined recursively*:

$$U_t = Co\Bigg(\bigcup_{a\in A}\Big\{g^a + \beta P^a \sum_{\theta\in\Theta} R^a(\theta)u_{t+1}^a(\theta):\ u_{t+1}^a(\theta)\in U_{t+1},\ \forall\,\theta\in\Theta\Big\}\Bigg), \tag{9}$$

$$\overline{U}_t = PCo\Bigg(\bigcup_{a\in A}\Big\{g^a + \beta P^a \sum_{\theta\in\Theta} R^a(\theta)u_{t+1}^a(\theta):\ u_{t+1}^a(\theta)\in \overline{U}_{t+1},\ \forall\,\theta\in\Theta\Big\}\Bigg), \tag{10}$$

$$U_t^* = PCoEx\Bigg(\bigcup_{a\in A}\Big\{g^a + \beta P^a \sum_{\theta\in\Theta} R^a(\theta)u_{t+1}^a(\theta):\ u_{t+1}^a(\theta)\in U_{t+1}^*,\forall\,\theta\in\Theta\Big\}\Bigg). \tag{11}$$

Furthermore, each $u_t^*\in U_t^*$ can be generated by a deterministic continuation policy.

The theorem shows that randomization is unnecessary if we focus on the extreme-point sets U_t^*. However, sets U_t and $\overline{U}_t$ are still useful in our later developments when convex or connected sets are more convenient. We note that the above expressions can be easily generalized such that A, P^a, and R^a are time dependent and Θ depends upon both the time and the action.

Clearly, expression (11) coincides with iteration (8), and the u_t vectors coincide with the ω_t vectors. The belief process is completely implicit in the picture except that, the initial state distribution is used to determine optimal solution through $\max\{\pi_0' u_1\colon u_1\in U_1^*\}$. This "mystery" will be resolved in Theorem 4 in §5.

4. Value Iteration Through Minkowski Sums

Value iteration is a general approach to solving the POMDP problem over both finite and infinite horizons. We have seen three value iteration procedures in previous sections. In this section, we show that the geometric interpretation presented in last section connects the POMDP problem with an existing problem in computational geometry, the Minkowski-sum problem.

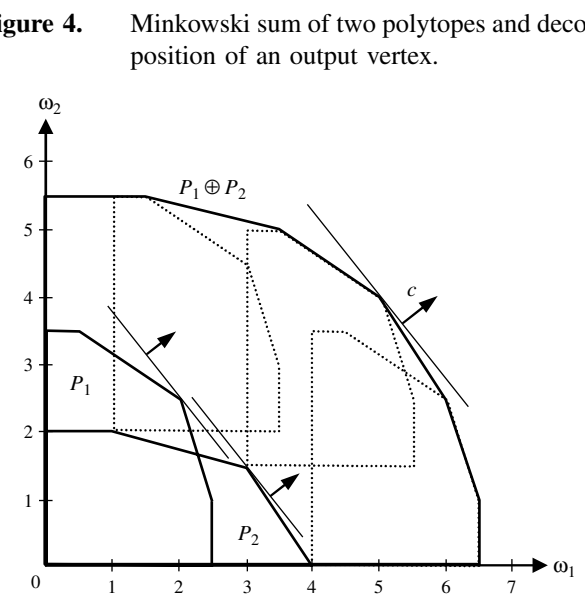
4.1. Minkowski Sum of Convex Polytopes

A polytope is the convex hull of a finite set of points. The *Minkowski sum* of two polytopes P_1 and P_2 is defined as $P_1\oplus P_2=\{u+v\colon u\in P_1, v\in P_2\}$, i.e., the pairwise addition of convex sets P_1 and P_2. The result is also a polytope. For convenience, we call the positive convex hull of $P_1\oplus P_2$ the *positive Minkowski sum*. The second step of the two continuation-value iterations can be compactly written as $\Omega_t^a=\bigoplus_{\theta=1}^m \Omega_t^a(\theta)$ and $\widetilde{\Omega}_t^a=\bigoplus_{\theta=1}^m \widetilde{\Omega}_t^a(\theta)$, respectively. Fukuda (2004) presents an algorithm that finds the vertices of a Minkowski sum in time linear to the number of output vertices. The key idea is to arrange the vertices of the output polytope as a minimum spanning tree. The algorithm can be directly used in the value iterations to save computational time. In this subsection, we discuss some basic properties of the Minkowski sum and main features of Fukuda's algorithm, which is valuable for understanding the geometry behind the POMDP problem and developing more efficient algorithms in the future.

A subset F is a *face* of a polytope P if there exists a vector $c\in\mathbb{R}^n$ such that $F=\arg\max_{v\in P}\{c'v\}$, denoted by $F(P;c)$ more precisely. An n-dimensional polytope has n types of proper faces, with dimensions $0,1,\ldots,$ and $n-1$, respectively. The 0- and 1-dimensional faces are also called *vertices* and *edges*, respectively. For the Minkowski sum of a set of polytopes, it can be easily shown that the faces of the output polytope can be constructed from those of the input polytopes: $F(\bigoplus_{i=1}^k P_i;c)=\bigoplus_{i=1}^k F(P_i;c)$ (Gritzmann and Sturmfels 1993). Figure 4 illustrates the Minkowski sum of two polytopes in the positive orthant and the decomposition of one output vertex. Compared with Figure 2(b), the vertices of the output polytope are exactly the ω-vectors on the value frontier.

The vertex/edge decomposition is the core of Fukuda's algorithm. *Let $P=\bigoplus_{i=1}^k P_i$.* It is shown that: (1) *a point $v\in P$ is a vertex of P if and only if $v=\sum_{i=1}^k v_i$ for some vertex v_i of P_i and there exists $c\in R^n$ with $\{v_i\}=F(P_i;c)$ for all i*; (2) *a subset E of P is an edge of P if and only if $E=\bigoplus_{i=1}^k F_i$, where each F_i is a vertex or edge of P_i* (*all*

Figure 4. Minkowski sum of two polytopes and decomposition of an output vertex.



such edges must be parallel) *and there exists* $c \in R^n$ *such that* $F_i = F(P_i; c)$ *for all* i.

Fukuda's algorithm also needs the adjacency information of the input polytopes, i.e., the pairs of vertices linked by edges. Let δ_i be the *maximum degree* of P_i, i.e., the maximum number of vertices adjacent to any vertex of P_i. Then, $\delta = \sum_{i=1}^{k} \delta_i$ is the upper bound of the maximum degree of $\bigoplus_{i=1}^{k} P_i$ due to the decomposition property. It is shown that there is a compact polynomial algorithm for the Minkowski sum of k polytopes that runs in time $O(z\delta LP(n, \delta))$ and space linear in the input size, where z denotes the number of vertices of P, and $LP(n, \delta)$ denotes the time needed to solve a linear program with n variables and δ constraints.

By today's standards, those LPs in Fukuda's algorithm are small, and their sizes grow with δ rather than the total number of vertices. The worst-case complexity of the algorithm is linear in the output size z, and is polynomial in that sense. It is not necessarily polynomial with respect to the input size $\sum_{i=1}^{k} z_i$, where z_i is the number of vertices of P_i. In comparison, the same step in the state-of-the-art POMDP algorithms (Cassandra et al. 1997, Feng and Zilberstein 2004) needs $O(z(\sum_{i=1}^{k} z_i)LP(n, z))$ time. Because the maximum degree of a polytope is typically much smaller than the number of vertices, the advantage of Fukuda's algorithm seems clear.

It is noteworthy that Fukuda's algorithm requires the adjacency information of the input polytopes. For an input polytope P_i, a straightforward method to obtain the adjacency information can take $O(z_i^2 LP(n, z_i))$ time, which is still better than the existing POMDP algorithms theoretically, because z may not be polynomially bounded by z_i. Further, because linear transformations do not alter adjacency relationships, in the entire iteration t, we only need to acquire the adjacency information for a single input polytope, corresponding to the set Ω_{t+1}. This is especially beneficial for systems with large action and observation spaces.

4.2. Numerical Examples

In this subsection, we illustrate some properties of the Minkowski sum through two examples. The first example demonstrates that the output size of a Minkowski sum is much smaller than the number of combinations of the input vertices. In the context of the POMDP problem, we focus on the first two steps of the alternative continuation-value iteration for a single action.

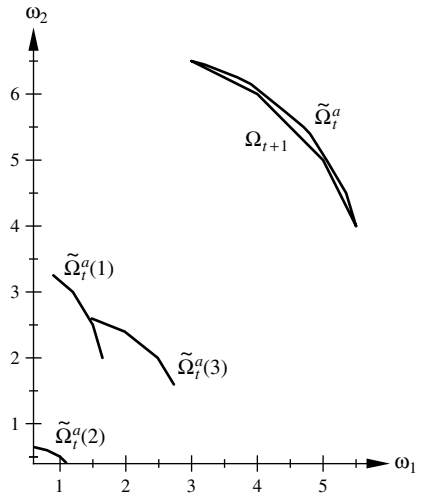
EXAMPLE 1. The parameters are: $n = 2$, $m = 3$,

$$R^a = \begin{pmatrix} 0.3 & 0.2 & 0.5 \\ 0.5 & 0.1 & 0.4 \end{pmatrix}$$

for a given action a, and

$$\Omega_{t+1} = \begin{pmatrix} 3 & 4 & 5 & 5.5 \\ 6.5 & 6 & 5 & 4 \end{pmatrix},$$

Figure 5. An alternative continuation-value iteration given a single action–positive Minkowski sum of three polytopes determined by three observations.



where each column of Ω_{t+1} represents a continuation-value vector. The first two steps of the alternative continuation-value iteration generates the following point set, in matrix form:

$$\widetilde{\Omega}_t^a = \begin{pmatrix} 3.0 & 3.2 & 3.7 & 3.9 & 4.4 & 4.7 & 4.8 & 5.05 & 5.35 & 5.5 \\ 6.5 & 6.45 & 6.25 & 6.15 & 5.75 & 5.5 & 5.4 & 5.0 & 4.5 & 4.0 \end{pmatrix}.$$

Although $|\widetilde{\Omega}_t^a| = 10$ is much larger than $|\Omega_{t+1}| = 4$, it is still much smaller than $|\Omega_{t+1}|^3 = 64$. The example is illustrated in Figure 5.

In the second example, we also focus on a single action and perform the alternative continuation-value iteration repeatedly. From an initial point set Ω_T, the output Ω_{t+1} of iteration $t + 1$ is used as the input of iteration t. The example demonstrates that the maximum degree δ of a Minkowski sum is much smaller than the output size z and grows much more slowly.

EXAMPLE 2. The parameters are: $n = m = 3$,

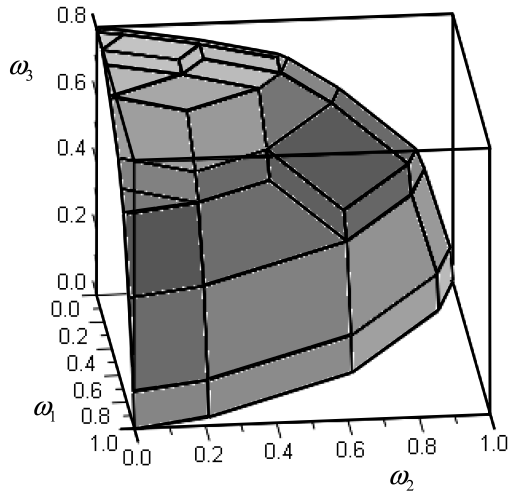
$$R^a = \begin{pmatrix} 0.3 & 0.3 & 0.4 \\ 0.2 & 0.2 & 0.6 \\ 0.5 & 0.1 & 0.4 \end{pmatrix}$$

for a given action a, and

$$\Omega_T = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0.8 \end{pmatrix}$$

(for demonstration purposes, we start with a representative Ω_T instead of a singleton set Ω_{T+1}). The following table

Figure 6. An output polytope after three alternative continuation-value iterations given a single action.



lists the maximum degree δ and the number of output vertices z after each iteration. The output polytope of iteration $T-3$ is illustrated in Figure 6.

t	T	$T-1$	$T-2$	$T-3$	$T-4$
δ	2	4	5	5	6
z	3	9	22	46	86

In last subsection, we argued the theoretical advantage of adopting Fukuda's algorithm in the value iteration of the POMDP problem, whereas in this subsection we have presented two simple numerical examples. A thorough comparison of the new POMDP algorithm (incorporating Fukuda's Minkowski-sum algorithm) with the existing POMDP algorithms requires extensive numerical studies, which is beyond the scope of this paper and is a valuable topic for future research.

5. Policy Iteration Over an Infinite Horizon

In the infinite-horizon setting, besides value iterations, the POMDP problem can also be tackled by the policy iteration approach that was first introduced by Howard (1960) in the MDP problem. According to limited results in the POMDP literature, the policy iteration approach appears to outperform the value iteration approach (Sondik 1978, Hansen 1998a). In this section, we provide a systematic treatment of the infinite-horizon POMDP problem based on the new framework. We first show some basic properties of the problem, then examine a special type of policy made up of finitely many components, and finally, we discuss a policy iteration algorithm proposed by Hansen (1998a). This section also unifies some existing concepts and algorithms in the literature.

5.1. Basic Properties of the POMDP Problem

It is convenient to express the continuation-value iteration (11) by an operator Γ^*, as $U_t^* = \Gamma^* U_{t+1}^*$. Similarly, the recursive Equations (9) and (10) can be expressed as $U_t = \Gamma U_{t+1}$ and $\overline{U}_t = \overline{\Gamma}\ \overline{U}_{t+1}$, through two operators Γ and $\overline{\Gamma}$, respectively. These three operators are closely related and are useful in different contexts.

To study the convergence of continuation-value frontiers, we equip the space of continuation-value vectors with the *Hausdorff metric*, which measures the distance between two compact sets $U \subset \mathbb{R}^n$ and $V \subset \mathbb{R}^n$:

$$d_H(U,V) \equiv \max\left\{\sup_{u\in U}\inf_{v\in V}\|u-v\|, \sup_{v\in V}\inf_{u\in U}\|u-v\|\right\}.$$

In this definition, $\|\cdot\|$ denotes the maximum norm (which is chosen for convenience), i.e., $\|u\| = \max(|u_1|, |u_2|, \ldots, |u_n|)$ for any $u \in \mathbb{R}^n$. It is known that the induced Hausdorff-metric set space is complete if the underlying metric space ($\mathbb{R}^n$) is complete. We can show the following result (the Γ operator is more convenient here):

Lemma 3. *The operator Γ is a contraction mapping with modulus β in the continuation-value set space with respect to the Hausdorff metric.*

The contraction-mapping theorem immediately implies the existence and uniqueness of the continuation-value frontier over an infinite horizon:

Theorem 3. *If $T = \infty$, the POMDP problem has a unique continuation-value set U_∞, a unique continuation-value frontier $\overline{U}_\infty \subset U_\infty$, and a unique extreme point set $U_\infty^* \subset \overline{U}_\infty$.*

Lemma 3 also guarantees that any sequence of continuation-value frontiers resulting from successive $\overline{\Gamma}$ operations converges to a unique limit, which proves the convergence of value iterations. The next result unveils the coherence between the value frontier and the belief process that has been suppressed in our analysis.

Because the value frontier $\overline{U}_\infty$ is a positive convex hull, for any extreme point $u^* \in U_\infty^*$, the unnormalized belief set $\Pi(u^*) \equiv \{\pi \in \mathbb{R}_+^n\colon\ \pi' u^* \geqslant \pi' u,\ \forall\, u \in \overline{U}_\infty\}$ is a nonempty cone, where $\mathbb{R}_+^n$ is the nonnegative orthant of $\mathbb{R}^n$. This set of belief vectors can testify that u^* is on the continuation-value frontier. If $\overline{U}_\infty$ is smooth (differentiable) at u^*, $\Pi(u^*)$ is a single ray along the outernormal direction of $\overline{U}_\infty$ at u^*; if $\overline{U}_\infty$ is nonsmooth at u^*, $\Pi(u^*)$ is a convex set. The collection of sets $\{\Pi(u^*)\colon\ u^* \in U_\infty^*\}$ constitutes a partition of the unnormalized belief space $\mathbb{R}_+^n$, except that the boundaries of these sets may overlap. According to expression (2), the unnormalized posterior belief following prior belief π, action a, and observation θ is $(\pi' P^a R^a(\theta))'$. Let $\Pi^a(u^*;\theta) \equiv \{(\pi' P^a R^a(\theta))'\colon\ \pi \in \Pi(u^*)\}$ be the set of posterior beliefs updated from the set of prior beliefs $\Pi(u^*)$.

By expression (11), any extreme point of the continuation-value frontier can be generated from an action a^* and

a set of continuation-value vectors $\{u(\theta)\}_{\theta\in\Theta}$ that are also extreme points of the value frontier. We have the following result:

THEOREM 4. *If an extreme point on the continuation-value frontier, $u^* \in U_\infty^*$, is generated from an action a^* and extreme points $\{u(\theta) \in U_\infty^*\}_{\theta\in\Theta}$, $\Pi^{a^*}(u^*;\theta) \subset \Pi(u(\theta))$ for any $\theta \in \Theta$.*

The theorem suggests that if we start with a belief set that supports an optimal policy, the posterior-belief set shrinks (relatively) as time elapses and always supports a single continuation policy at any given point in time. Using the terminology of Sondik (1978), the collection $\{\Pi(u^*)\colon u^* \in U_\infty^*\}$ forms a Markov partition of the unnormalized belief space, although he only states this property for the so-called "finitely transient policies." In view of this result, the belief process can be kept implicit in the analysis because there always exists a belief process consistent with any policy derived from the continuation-value frontier. This result is also true in the finite-horizon setting, which can be shown by modifying the proof of Theorem 4.

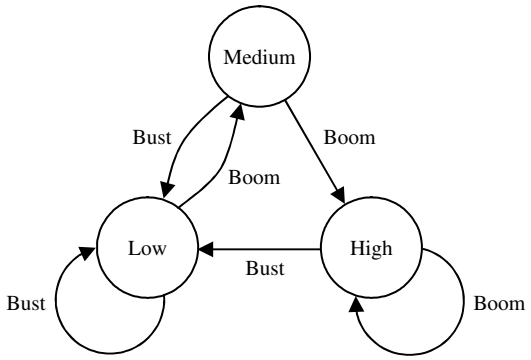
5.2. Stationary Policies: Finite-State Controllers

For the MDP problem with finite state and action sets, policy iteration can find an optimal policy in finite time because there always exists an optimal policy that is deterministic and stationary, and there are only finitely many such policies. However, this is an intriguing issue for the POMDP problem because there are infinitely many history-dependent deterministic policies.

Sondik's algorithm is deemed impractical by computer scientists because of its complexity (Hansen 1998a). However, within the new framework, Sondik's and Hansen's algorithms are related. The key concept is the *finite-state controller* (also called "plan graph" in Kaelbling et al. 1998), denoted by σ, which consists of a finite set K of control states, an action rule $\{a(k) \in A\}_{k\in K}$, and a transition rule $\{s(k,\theta) \in K\}_{k\in K,\,\theta\in\Theta}$. In control state k, action $a(k)$ is taken, and observation θ switches the controller to state $s(k,\theta)$. A finite-state controller simplifies an infinite policy tree, as shown in Figure 3, to a cyclic policy graph. An example is given next.

EXAMPLE 3. The true status of a simplistic economy is unobservable. However, the performance of a representative market in the economy can be observed, as either "boom" or "bust," which may serve as an indicator of the economy. A typical firm can produce at three quantity levels: low, medium, and high. The firm's action affects the underlying economy and the representative market. Thus, this economy can be modeled as a POMDP with hidden states, two observations, and three actions. A production policy of the firm in general depends upon the entire observation history, but simple policies like the one below may be particularly interesting: if the indicator was "bust" in the last period, choose the low production level; if the indicator was "boom" in the last period but "bust" one period earlier, choose medium; if the indicator was "boom" in the last two periods or more, choose high. This policy can be described by a three-state controller, as illustrated in Figure 7.

Figure 7. A three-state production policy in a simplistic economy.



A state of the controller $k \in K$ (a node in the policy graph) corresponds to a continuation policy and is associated with a continuation-value vector, denoted by $u_\sigma^*(k)$, which can be computed from a system of equations:

$$u_\sigma^*(k) = g^{a(k)} + \beta P^{a(k)} \sum_{\theta\in\Theta} R^{a(k)}(\theta) u_\sigma^*(s(k,\theta)), \quad k \in K. \quad (12)$$

Any deterministic policy can be approximated arbitrarily closely by a finite-state controller by increasing the number of control states. A finite-state controller may be optimal sometimes, which can be directly verified as follows (a corollary of Theorem 3).

COROLLARY 1. *If the set of continuation-value vectors generated by a finite-state controller is invariant under the Γ^* operation, the controller defines an optimal policy.*

Another type of policy determined by a finite number of objects is the finite-memory (or finite-history) policy, in which the current action depends only upon a finite number of recent observations and actions (Platzman 1977, White and Scherer 1994). Finite-memory policies and finite-state controllers have major differences that have not been fully elaborated on in the literature. Conceptually, a control state in the finite-state controller represents a continuation policy extending into the infinite future, whereas a memory state in the finite-memory policy represents a history of the past. Mathematically, any finite-memory policy with finitely many memory states can be represented by a finite-state controller, but not vice versa. In fact, the finite-state controller in Example 3 is also a finite-memory policy. However, it is easy to find examples in which a finite-state controller cannot be converted into a finite-memory policy (Yu and Bertsekas 2008).

5.3. Policy Iteration Through Finite-State Controllers

The policy-iteration algorithms by Sondik (1978) and Hansen (1998a) both center on finite-state controllers. However, Sondik's algorithm relies on both belief vectors and continuation-value vectors—essentially operating in both primal and dual spaces—and is inevitably more involved. For completeness, we present Hansen's algorithm here, with minor modifications in the description. This algorithm is extended by Poupart and Boutilier (2004) to incorporate randomized policies. The control state corresponding to a continuation-value vector u is denoted by $k(u)$ below.

Hansen's Policy Iteration Algorithm

Step 1. Initialization: define an initial finite-state controller σ, and select precision level $\varepsilon > 0$.

Step 2. Policy evaluation: calculate σ's value vectors (Equation (12)), denoted by set U.

Step 3. Policy improvement: perform one-step value iteration, $\widehat{U} = \Gamma^* U$ (Equation (11)), and modify σ to $\widehat{\sigma}$ as follows.

(a) For each $\hat{u} \in \widehat{U}$: (i) If $\hat{u} = u$ for some $u \in U$, *keep* $k(u)$ unchanged. (ii) Else, if $\hat{u} \geqslant u$ for some $u \in U$, *replace* the action and successor links of $k(u)$ by those used to create $\hat{u}$. If there are more than one such u, the corresponding control states can be combined into a single one. (iii) Otherwise, *add* a new control state $k(\hat{u})$ with the action and successor links used to create $\hat{u}$.

(b) For each $u \in U \backslash \widehat{U}$, if it is not used to create any vector in $\widehat{U}$, *remove* $k(u)$; otherwise, *keep* $k(u)$ unchanged.

Step 4. Termination test: if $d_H(PCo(\widehat{U}), PCo(U)) \leqslant \varepsilon(1-\beta)/\beta$, exit with an ε-optimal policy. Otherwise, change σ to $\widehat{\sigma}$ and return to step 2.

It can be shown that in a policy improvement step, if σ is not optimal, $\widehat{\sigma}$ generates a weakly improved value frontier that is strictly better at some control states, and the policy iteration algorithm converges to an ε-optimal policy after a finite number of iterations. Hansen tests the above algorithm on 10 POMDP problems and observes a convergence rate 40 to 50 times faster than that of the value iteration method on average. Because the Γ^* operation is still needed in the policy improvement step, which is in fact the bottleneck of the algorithm, Fukuda's Minkowski-sum algorithm may still be used to improve the efficiency of this algorithm.

6. Partial Order of Observability

Partial observability is the defining characteristic of POMDPs, but except for the two extreme cases, perfect observability and no observability (Satia and Lave 1973, White 1980), little can be found in the literature that compares POMDPs from the perspective of observability. In this section, we extend Blackwell's (1953) notion of informativeness to define a partial order of POMDPs and show that it leads to a partial order of continuation-value frontiers. The same notion can also be found in White (1979), but it has not been pursued rigorously since then, to the best knowledge of the author. In what follows, we start from the two extreme cases and then proceed to the middle.

Case 1: *Perfect Observability*. If an observation matrix R^a is the identity matrix I, the action a perfectly reveals the system state, and each diagonal matrix $R^a(\theta)$ contains only one nonzero element, the θth main-diagonal element. Then, the positive Minkowski sum $\bigoplus_{\theta\in\Theta}\{R^a(\theta)u\colon u \in U\}$ given any point set U is a single point, $\max\{u\colon u \in U\}$, where the maximum is taken componentwise. If all R^a, $a \in A$, are identity matrices, the value frontier $\overline{U}_t$ reduces to a singleton $\{u_t^*\}$, and the recursive expression (11) reduces to the standard dynamic programming formulation: $u_t^* = \max_{a\in A}\{g^a + \beta P^a u_{t+1}^*\}$. This MDP problem serves as an upper bound for the POMDP problem.

Case 2: *No Observability*. If a column of an observation matrix R^a is proportional to the vector e (consisting of all 1s), the corresponding observation reveals no information about the system state. If every column of R^a is proportional to e, i.e., $R^a = (\lambda_1^a e, \ldots, \lambda_m^a e)$ for $\lambda_\theta^a \geqslant 0$ with $\sum_{\theta\in\Theta}\lambda_\theta^a = 1$, no observation carries any information about the system state. In this case, we have $R^a(\theta) = \lambda_\theta^a I$ for all θ. For any positive convex hull U, the positive Minkowski sum $\bigoplus_{\theta\in\Theta}\{R^a(\theta)u\colon u \in U\}$ equals U exactly. Thus, expression (11) reduces to $U_t^* = PCoEx(\bigcup_{a\in A}\{g^a + \beta P^a u_{t+1}^a\colon u_{t+1}^a \in U_{t+1}^*\})$. This hidden-state MDP problem is significantly simpler than the POMDP problem because the most time-consuming step, the Minkowski sum, is absent from the formulation. This problem provides a lower bound for the POMDP problem.

Case 3: *Partial Observability*. The above upper and lower bounds depend purely upon the underlying Markov decision process. If the observation matrices $\{R^a\}$ satisfy neither of the above conditions, the model falls into the partial-observability case, and its continuation-value frontier falls between the two bounding frontiers. To compare POMDPs that share the same underlying MDP, we start with a partial order of observation matrices, following Blackwell (1953). Recall that an n-state observation matrix is a nonnegative stochastic matrix with n rows, all summing to one.

Definition 5. Among the n-state observation matrices, A is *more informative* than B, denoted by $A \succeq B$, if there exists a nonnegative stochastic matrix X such that $AX = B$. If $A \succeq B$ and $B \succeq A$, A is *equivalent* to B, denoted by $A \approx B$.

Notice that permuting the columns of any observation matrix creates an equivalent observation matrix. For a study of the equivalence classes, see Sulganik (1995). Next, we show that a more informative observation matrix generates greater continuation values.

Lemma 4. *Consider two n-state observation matrices A and B, with m^A and m^B columns, respectively. Let $A(\theta)$ and $B(\theta)$ denote the diagonal matrices formed from the*

θth column of A and B, respectively. Then, if $A \succeq B$, for any point set $U \subset \mathbb{R}^n$, the positive convex hull $W^A(U) = \bigoplus_{\theta=1}^{m^A}\{A(\theta)u_\theta\colon\ u_\theta \in U\}$ dominates $W^B(U) = \bigoplus_{\theta=1}^{m^B}\{B(\theta)u_\theta\colon\ u_\theta \in U\}$. If $A \approx B$, $W^A(U) = W^B(U)$ for any $U \subset \mathbb{R}^n$.

The class of n-state observation matrices contains a *least* equivalence subclass and a *greatest* equivalence subclass, consistent with our previous discussion of the two extreme cases. The proof follows from the definition of the $\succeq$ relation and is omitted.

THEOREM 5. (1) *All n-state observation matrices with full row rank and one nonzero element in each column are equivalent, forming the perfect-observability class.* (2) *All n-state observation matrices of the form $(\lambda_1 e, \ldots, \lambda_m e)$ for any $\lambda_\theta \geqslant 0$ with $\sum_{\theta=1}^{m} \lambda_\theta = 1$ are equivalent, forming the no-observability class.* (3) *If A and C belong to the perfect- and no-observability classes, respectively, and B is an arbitrary n-state observation matrix, we have $A \succeq B \succeq C$.*

Finally, the informativeness relation of the observation matrices induces a partial order of the POMDPs, which in turn leads to a partial order of their value frontiers.

DEFINITION 6. For two POMDPs sharing the same underlying MDP, one is *more observable* than the other if, for every action, the observation matrix in the former is more informative than that in the latter.

THEOREM 6. *Suppose that one POMDP is more observable than another. Then,* (1) *if $T < \infty$, starting with the same singleton set $\{g_{T+1}\}$, the continuation-value frontier of the former dominates that of the latter in every period*; *and* (2) *if $T = \infty$, the continuation-value frontier of the former dominates that of the latter.*

7. Conclusion

In this paper, we have proposed a novel framework for the POMDP problem, based on continuation policies and continuation-value vectors, with natural geometric interpretations. The framework is more parsimonious than the traditional framework based on belief vectors. It unveils the relationship between the POMDP problem and two existing computational geometry problems, which can help solve the POMDP problem more efficiently. The framework can clarify some existing POMDP algorithms over both finite and infinite horizons and sheds new light on them. It also facilitates the comparison of POMDPs in terms of observability, which is a useful structural result.

We conclude the paper with a brief discussion of possible future research topics. An important topic not addressed in this paper is the structural properties of optimal policies for some special POMDPs. This is a well-known challenging task, even in the two-state case. For example, Ross (1971) shows that the optimal policy for a two-state machine replacement problem does not have the control-limit property. Other examples of structural results for the optimal policies are White (1977, 1979) and Grosfeld-Nir (1996, 2007). The geometry underlying the new framework provides a handy tool for exploring policy structures, which can be a fruitful future research direction.

Another important topic absent from the paper is the approximation of optimal solutions. Because of the computational burden inherent to the POMDP problem, approximation may be unavoidable in solving practical problems. The finite-state controller discussed in §5 can generate lower bounds for the continuation-value frontier in the infinite-horizon case. The number of control states can be judiciously chosen to balance the quality of approximation and cost of computation. To generate upper bounds, we can conduct successive value iterations starting with the perfect-observability solution. A common approximation approach in the primal space is to focus on belief vectors on a finite grid (Kakalik 1965, Eckles 1966, Cheng 1988, Lovejoy 1991a). The same idea can be applied to the dual space, which is another valuable topic for future studies.

Last but not least, the belief vectors can be reintroduced into the geometric framework. In this paper, we have implicitly aimed to solve the POMDP problem for all initial state distributions simultaneously. However, many applications of the problem only require the solution for a given initial state distribution and may not need the complete continuation-value frontiers. Some (approximation) algorithms in the literature are tailored to such solutions (e.g., Satia and Lave 1973, Hansen 1998b, Pineau et al. 2003). Starting from an initial belief vector and moving forward, in each period, we can directly compute a set of feasible beliefs that can be reached under an arbitrary policy. An extreme point of the continuation-value frontier at time t need not be created if it is not supported by any feasible belief at time t. A main implication from this paper is that the feasible belief sets and continuation-value frontiers can be disentangled and independently generated (forward and backward, respectively). Acquiring and utilizing both types of information prudently may save computational time substantially for problems stressing the initial state distribution. The details are left for future investigations.

Appendix. Proofs of Lemmas and Theorems

PROOF OF LEMMA 1. (1) We first show that for any $\widehat{\pi} \in \Pi$, there is a $\widehat{\omega} \in PCo(\Omega)$ such that $\widehat{\pi}'\widehat{\omega} \geqslant \widehat{\pi}'\omega$ for all $\omega \in \Omega$. If $\widehat{\pi} \in \Pi^+$, by the definition of PCO, such a $\widehat{\omega}$ must exist, and then we are done. Suppose that $\widehat{\pi} \in \Pi\backslash\Pi^+$. By the definitions of Π and Π^+, there exists a sequence of belief vectors $\pi_k \in (\widehat{\pi} + (1/k)B) \cap \Pi^+$, where B is the unit open ball and $\widehat{\pi} + (1/k)B$ is the open ball centered at $\widehat{\pi}$ with radius $1/k$. Clearly, $\lim_{k\to\infty} \pi_k = \widehat{\pi}$. For each π_k, there exists a $\omega_k \in PCo(\Omega)$ such that $\pi_k'\omega_k \geqslant \pi_k'\omega$ for all $\omega \in \Omega$. Because $\Omega \subset \mathbb{R}^n$ is bounded, by the Bolzano-Weierstrass theorem, the sequence $\{\omega_k\}$ has a convergent subsequence $\{\omega_{k_j}\}$. Let

$\hat{\omega} = \lim_{j\to\infty} \omega_{k_j}$. Because $PCo(\Omega)$ is closed, $\hat{\omega} \in PCo(\Omega)$. We write $\hat{\pi}'(\hat{\omega} - \omega) = \hat{\pi}'(\hat{\omega} - \omega_{k_j}) + (\hat{\pi} - \pi_{k_j})' \cdot (\omega_{k_j} - \omega) + \pi'_{k_j}(\omega_{k_j} - \omega)$. Because (i) $\lim_{j\to\infty} \hat{\pi}'(\hat{\omega} - \omega_{k_j}) = 0$, (ii) $\lim_{j\to\infty} (\hat{\pi} - \pi_{k_j})'(\omega_{k_j} - \omega) = 0$ for all $\omega \in \Omega$ (by the boundedness of Ω), and (iii) $\pi'_{k_j}(\omega_{k_j} - \omega) \geqslant 0$ for all $\omega \in \Omega$ and j, we have $\hat{\pi}'(\hat{\omega} - \omega) \geqslant 0$ for all $\omega \in \Omega$.

Next, we show that for any $\hat{\omega} \in PCo(\Omega)$, there is a $\hat{\pi} \in \Pi$ such that $\hat{\pi}'\hat{\omega} \geqslant \hat{\pi}'\omega$ for all $\omega \in \Omega$. If $\hat{\omega} \in \arg\max\{\hat{\pi}'\omega: \omega \in \Omega\}$ for some $\hat{\pi} \in \Pi^+$, we are done. Otherwise, by the definition of PCO, $\hat{\omega}$ must be the limit of a point sequence $\{\omega_k \in PCo(\Omega)\}$ associated with a belief sequence $\{\pi_k \in \Pi^+\}$ such that $\pi'_k\omega_k \geqslant \pi'_k\omega$ for all $\omega \in \Omega$. Again, by the Bolzano-Weierstrass theorem, $\{\pi_k\}$ has a convergent subsequence $\{\pi_{k_j}\}$. Let $\hat{\pi} = \lim_{j\to\infty} \pi_{k_j}$. We have $\hat{\pi} \in \Pi$ by the closeness of Π. We write

$$\hat{\pi}'(\hat{\omega} - \omega) = (\hat{\pi} - \pi_{k_j})'(\hat{\omega} - \omega) + \pi'_{k_j}(\hat{\omega} - \omega_{k_j}) + \pi'_{k_j}(\omega_{k_j} - \omega).$$

Because

$$\text{(i)} \lim_{j\to\infty} \pi'_{k_j}(\hat{\omega} - \omega_{k_j}) = 0, \quad \text{(ii)} \lim_{j\to\infty} (\hat{\pi} - \pi_{k_j})'(\hat{\omega} - \omega) = 0$$

for all $\omega \in \Omega$ (by the boundedness of Ω), and (iii) $\pi'_{k_j}(\omega_{k_j} - \omega) \geqslant 0$ for all $\omega \in \Omega$ and j, we have $\hat{\pi}'(\hat{\omega} - \omega) \geqslant 0$ for all $\omega \in \Omega$.

(2) Because

$$\begin{aligned}\overline{V}_1(\pi) + \overline{V}_2(\pi) &= \max_{\omega\in\Omega_1}\{\pi'\omega\} + \max_{\omega\in\Omega_2}\{\pi'\omega\} \\ &= \max_{\omega_1\in\Omega_1,\, \omega_2\in\Omega_2}\{\pi'(\omega_1 + \omega_2)\} \\ &= \max_{\omega\in\Omega_1+\Omega_2}\{\pi'\omega\},\end{aligned}$$

the function is dual to $PCo(\Omega_1 + \Omega_2)$, by part (1).

(3) Because

$$\begin{aligned}\max\{\overline{V}_1(\pi), \overline{V}_2(\pi)\} &= \max\left\{\max_{\omega\in\Omega_1}\{\pi'\omega\}, \max_{\omega\in\Omega_2}\{\pi'\omega\}\right\} \\ &= \max_{\omega\in\Omega_1\cup\Omega_2}\{\pi'\omega\},\end{aligned}$$

the function is dual to $PCo(\Omega_1 \cup \Omega_2)$, by part (1). □

Proof of Lemma 2. The proof is by induction. Consider period t and suppose that the two iterations start with the same Ω_{t+1} set. For the sake of clarity, we dedicate the label Ω_t to the Ω_t set that results from the three-step iteration and use the label Ω_t^* for the one that results from the aggregated iteration. It suffices to show that the PCOs underlying Ω_t and Ω_t^* are identical. For convenience, define $W_t^* = \bigcup_{a\in A}\{g^a + \beta\sum_{\theta\in\Theta} P^a R^a(\theta)\omega^a_{t+1}(\theta)\colon \omega^a_{t+1}(\theta) \in \Omega_{t+1}, \forall\, \theta \in \Theta\}$; hence, $\Omega_t^* = PCoEx(W_t^*)$ and $PCo(\Omega_t^*) = PCo(W_t^*)$. Clearly, $\Omega_t \subset Co(W_t^*)$, so Ω_t is dominated by $PCo(\Omega_t^*)$ (with respect to the directions in Π). It remains to show the reverse, i.e., Ω_t^* is dominated by $PCo(\Omega_t)$.

Consider any $\hat{\omega} \in \Omega_t^*$. By the definition of PCO, one of the following must be true: (a) there exists $\hat{\pi} \in \Pi^+$ such that $\hat{\pi}'\hat{\omega} \geqslant \hat{\pi}'\omega$ for all $\omega \in W_t^*$, or (b) $\hat{\omega}$ is the limit of a point sequence $\{\omega_k \in W_t^*\}$ associated with a direction sequence $\{\pi_k \in \Pi^+\}$ such that $\pi'_k\omega_k \geqslant \pi'_k\omega$ for all $\omega \in W_t^*$ and for all k. In case (a), because $\hat{\omega}$ is an extreme point of $Co(W_t^*)$, there must exist $\hat{a} \in A$ and $\{\omega^{\hat{a}}_{t+1}(\theta) \in \Omega_{t+1}\}_{\theta\in\Theta}$ such that $\hat{\omega} = g^{\hat{a}} + \beta\sum_{\theta\in\Theta} P^{\hat{a}} R^{\hat{a}}(\theta)\omega^{\hat{a}}_{t+1}(\theta)$. For any $\theta \in \Theta$, $P^{\hat{a}} R^{\hat{a}}(\theta)\omega^{\hat{a}}_{t+1}(\theta)$ is dominated by $PCo(\Omega_t^{\hat{a}}(\theta))$, and hence $\sum_{\theta\in\Theta} P^{\hat{a}} R^{\hat{a}}(\theta)\omega^{\hat{a}}_{t+1}(\theta)$ is dominated by $PCo(\Omega_t^{\hat{a}})$. As a result, $\hat{\omega}$ is dominated by $PCo(\Omega_t)$. In case (b), because $PCo(\Omega_t)$ is closed and every ω_k is dominated by $PCo(\Omega_t)$, $\hat{\omega}$ must be dominated by $PCo(\Omega_t)$ as well. Thus, Ω_t^* is dominated by $PCo(\Omega_t)$. Therefore, we have $PCo(\Omega_t^*) = PCo(\Omega_t)$ and $\Omega_t^* = \Omega_t$. □

Proof of Theorem 2. (1) We show that Equation (9) is true. Given the time-$(t+1)$ continuation-value set U_{t+1}, a time-t continuation-value vector

$$g^a + \beta P^a \sum_{\theta\in\Theta} R^a(\theta) u^a_{t+1}(\theta)$$

can be obtained by taking action a and selecting time-$(t+1)$ continuation-value vector $u^a_{t+1}(\theta)$ after observation θ. The convex hull of

$$\bigcup_{a\in A}\left\{g^a + \beta P^a \sum_{\theta\in\Theta} R^a(\theta) u^a_{t+1}(\theta)\colon u^a_{t+1}(\theta) \in U_{t+1}, \forall\, \theta \in \Theta\right\}$$

contains all time-t continuation-value vectors that can be obtained by randomization.

(2) We show that (10) is true. Note that: (a) for any $\pi \in \Pi$, $a \in A$, and $\theta \in \Theta$, if $\|\pi' P^a R^a(\theta)\|_2 > 0$, $\pi' P^a R^a(\theta)/\|\pi' P^a R^a(\theta)\|_2 \in \Pi$, where $\|\cdot\|_2$ denotes the Euclidean norm; (b) by Lemma 1(1), for any compact convex set $U_{t+1} \subset \mathbb{R}^n$ and any $\pi \in \Pi$, there exists $\hat{u} \in PCo(U_{t+1})$ such that $\pi'\hat{u} \geqslant \pi' u$ for any $u \in U_{t+1}$. These facts, combined with Equation (9) and the definition of PCO, imply that any point in $PCo(U_t)$ can be created from $PCo(U_{t+1})$. Thus, Equation (10) follows.

(3) Now we show that (11) is true. Consider any *extreme point* $u_t^* \in \overline{U}_t$. By expression (10), there must exist $a^* \in A$ and $\{u^{a^*}_{t+1}(\theta) \in \overline{U}_{t+1}\}_{\theta\in\Theta}$ such that $u_t^* = g^{a^*} + \beta P^{a^*}\sum_{\theta\in\Theta} R^{a^*}(\theta) u^{a^*}_{t+1}(\theta)$. If $u^{a^*}_{t+1}(\theta)$ is an extreme point of $\overline{U}_{t+1}$ for all $\theta \in \Theta$, we are done. For any $\theta \in \Theta$, if $u^{a^*}_{t+1}(\theta)$ is not an extreme point of $\overline{U}_{t+1}$, it must lie on a face $F \subset U_{t+1}$ (with at least one dimension) or in the interior of U_{t+1} (define $F = U_{t+1}$ in that case). It follows that $P^{a^*} R^{a^*}(\theta) v = P^{a^*} R^{a^*}(\theta) u^{a^*}_{t+1}(\theta)$ for all $v \in F$; otherwise, $P^{a^*} R^{a^*}(\theta) u^{a^*}_{t+1}(\theta)$ can be expressed as the convex combination of $P^{a^*} R^{a^*}(\theta) v'$ and $P^{a^*} R^{a^*}(\theta) v''$ for some $v' \neq v'' \in F$, and hence u_t^* can be expressed as the convex combination of two distinct points in U_t, a contradiction. Thus, $u^{a^*}_{t+1}(\theta)$ can be replaced by any extreme point of $\overline{U}_{t+1}$ on F without altering u_t^*. Therefore, u_t^* can always be constructed from the extreme points of $\overline{U}_{t+1}$, and expression (11) holds.

(4) From part (3) and by induction, it is clear that every $u_t^* \in U_t^*$ can be generated by a deterministic continuation policy. □

PROOF OF LEMMA 3. We show that for any two sets U, $V \subset \mathbb{R}^n$ such that $d_H(U,V) = d$, ΓU and ΓV are also in $\mathbb{R}^n$ and $d_H(\Gamma U, \Gamma V) \leqslant \beta d$. Consider any $u^* \in Ex(\Gamma U)$ (i.e., an extreme point of ΓU). By expression (10), there exist an action a^* and a set of vectors $\{u(\theta) \in U\}_{\theta\in\Theta}$ such that $u^* = g^{a^*} + \beta P^{a^*} \sum_{\theta\in\Theta} R^{a^*}(\theta)u(\theta)$. Because $d_H(U,V) = d$, for each $u(\theta)$, there exists $v(\theta) \in V$ such that $\|u(\theta) - v(\theta)\| \leqslant d$. Let e denote the vector of 1s with a proper dimension. We have

$$\begin{aligned} u^* - \beta de &= g^{a^*} + \beta P^{a^*} \sum_{\theta\in\Theta} R^{a^*}(\theta)(u(\theta) - de) \\ &\leqslant g^{a^*} + \beta P^{a^*} \sum_{\theta\in\Theta} R^{a^*}(\theta)v(\theta) \equiv v^* \\ &\leqslant g^{a^*} + \beta P^{a^*} \sum_{\theta\in\Theta} R^{a^*}(\theta)(u(\theta) + de) = u^* + \beta de. \end{aligned}$$

The first and last equations above follow from $P^{a^*}\sum_{\theta\in\Theta} R^{a^*}(\theta)e = e$. Clearly, $v^* \in \Gamma V$ and $\|u^* - v^*\| \leqslant \beta d$. The result can be generalized to nonextreme $u^* \in \Gamma U$ by convex combinations. Similarly, we can show that for any $v^* \in \Gamma V$, there exists $u^* \in \Gamma U$ such that $\|u^* - v^*\| \leqslant \beta d$. Thus, $d_H(\Gamma U, \Gamma V) \leqslant \beta d$, and the operator Γ is a contraction mapping with modulus β. □

PROOF OF THEOREM 4. Consider any $\pi^* \in \Pi(u^*)$. By definition, $(\pi^*)'u^* \geqslant (\pi^*)'u$ for all $u \in \bar{U}_\infty$. By expression (10), $u^* = g^{a^*} + \beta P^{a^*}\sum_{\theta\in\Theta} R^{a^*}(\theta)u(\theta)$. Thus, for any $\theta \in \Theta$, we must have $(\pi^*)'P^{a^*}R^{a^*}(\theta)u(\theta) \geqslant (\pi^*)'P^{a^*}R^{a^*}(\theta)u$ for all $u \in \bar{U}_\infty$, i.e., $(\pi^*)'P^{a^*}R^{a^*}(\theta) \in \Pi(u(\theta))$. By definition, $\Pi^{a^*}(u^*;\theta) \subset \Pi(u(\theta))$. □

PROOF OF LEMMA 4. Suppose that $A \succeq B$, i.e., there exists nonnegative matrix X with row sum 1 such that $AX = B$. Then, the θth column of B is given by $b_{\cdot\theta} = Ax_{\cdot\theta}$, where $x_{\cdot\theta}$ is the θth column of X. Let $x_{k\theta}$ be the (k,θ)th element of X. For any point $w^B \in W^B(U)$, suppose that $w^B = \sum_{\theta=1}^{m^B} B(\theta)u_\theta^B$ for some $\{u_\theta^B \in U\}_{\theta=1,\ldots,m^B}$. Then, $w^B = \sum_{\theta=1}^{m^B}\sum_{k=1}^{m^A} x_{k\theta}A(k)u_\theta^B = \sum_{k=1}^{m^A}\sum_{\theta=1}^{m^B} x_{k\theta}A(k)u_\theta^B$. For each $k = 1,\ldots,m^A$, $\sum_{\theta=1}^{m^B} x_{k\theta}A(k)u_\theta^B$ is a convex combination of the set of points $\{A(k)u_\theta^B\}_{\theta=1,\ldots,m^B}$ and is hence weakly dominated by the positive convex hull of $\{A(k)u_\theta^B\}_{\theta=1,\ldots,m^B}$. By the definition of positive Minkowski sum, $w^B = \sum_{k=1}^{m^A}(\sum_{\theta=1}^{m^B} x_{k\theta}A(k)u_\theta^B)$ is dominated by $\bigoplus_{k=1}^{m^A}\{A(k)u_\theta^B\}_{\theta=1,\ldots,m^B}$. The latter is in turn dominated by $W^A(U)$ because $\{u_\theta^B\}_{\theta=1,\ldots,m^B} \subset U$. Therefore, the entire set $W^B(U)$ is dominated by $W^A(U)$. If $A \approx B$, we have $A \succeq B$ and $B \succeq A$, and the above result implies $W^A(U) = W^B(U)$ for any $U \subset \mathbb{R}^n$. □

PROOF OF THEOREM 6. For clarity, we label the second POMDP by a "~" symbol. Thus, $R^a \succeq \tilde{R}^a$ for all $a \in A$. By Lemma 4, for any set U and any action a, the positive Minkowski sum $\bigoplus_{\theta\in\Theta}\{R^a(\theta)u_\theta^a\colon\ u_\theta^a \in U\}$ dominates $\bigoplus_{\theta\in\tilde{\Theta}}\{\tilde{R}^a(\theta)u_\theta^a\colon\ u_\theta^a \in U\}$. Thus, $\bar{\Gamma}U = PCo(\bigcup_{a\in A}\{g^a + \beta P^a\sum_{\theta\in\Theta} R^a(\theta)u_\theta^a\colon\ u_\theta^a \in U, \forall\,\theta \in \Theta\})$ dominates $\tilde{\bar{\Gamma}}U = PCo(\bigcup_{a\in A}\{g^a + \beta P^a\sum_{\theta\in\tilde{\Theta}} \tilde{R}^a(\theta)u_\theta^a\colon\ u_\theta^a \in U, \forall\,\theta \in \Theta\})$, for any point set U. If $T < \infty$, $\bar{\Gamma}^k U$ dominates $\tilde{\bar{\Gamma}}^k U$ for any set U and any $k = 1, 2, \ldots, T$; if $T = \infty$, the sequences $\{\bar{\Gamma}^k U\}_{k=1,2,\ldots}$ and $\{\tilde{\bar{\Gamma}}^k U\}_{k=1,2,\ldots}$ converge to the continuation-value frontiers $\bar{U}_\infty$ and $\tilde{\bar{U}}_\infty$ of the two POMDPs, respectively, and hence $\bar{U}_\infty$ dominates $\tilde{\bar{U}}_\infty$. □

Acknowledgments

The author thanks the associate editor and two anonymous referees for their constructive suggestions that improved the exposition of this paper, and Mahesh Nagarajan for his helpful comments on this work.

References

Anily, S., A. Grosfeld-Nir. 2006. An optimal lot-sizing and offline inspection policy in the case of nonrigid demand. *Oper. Res.* **54**(2) 311–323.

Aoki, M. 1965. Optimal control of partially observable Markovian systems. *J. Franklin Inst.* **280** 367–386.

Astrom, K. J. 1965. Optimal control of Markov decision processes with incomplete state estimation. *J. Math. Anal. Appl.* **10** 174–205.

Berg, M. de, M. van Kreveld, M. Overmars, O. Schwarzkopf. 2000. *Computational Geometry: Algorithms and Applications*. Springer-Verlag, Berlin.

Bertsekas, D. 1976. *Dynamic Programming and Stochastic Control*. Academic Press, New York.

Blackwell, D. 1953. Equivalent comparisons of experiments. *Ann. Math. Statist.* **24** 265–272.

Boissonnat, J-D., M. Yvinec. 1998. *Algorithmic Geometry*. Cambridge University Press, Cambridge, UK.

Boutilier, C. 2002. A POMDP formulation of preference elicitation problems. *Proc. Eighteenth National Conf. Artificial Intelligence* (AAAI-02). AAAI Press, Menlo Park, CA, 239–246.

Cassandra, A. R., L. P. Kaelbling, J. A. Kurien. 1996. Acting under uncertainty: Discrete Bayesian models for mobile robot navigation. *Proc. IEEE/RSJ Internat. Conf. Intelligent Robots and Systems*. IEEE, Piscataway, NJ, 963–972.

Cassandra, A. R., L. P. Kaelbling, M. L. Littman. 1994. Acting optimally in partially observable stochastic domains. *Proc. Twelfth National Conf. Artificial Intelligence* (AAAI-94). AAAI Press, Menlo Park, CA, 1023–1028.

Cassandra, A. R., M. Littman, N. L. Zhang. 1997. Incremental pruning: A simple, fast, exact method for partially observable Markov decision processes. *Proc. Thirteenth Conf. Uncertainty in Artificial Intelligence* (UAI-97). Morgan Kaufmann, San Francisco, 54–61.

Cheng, H.-T. 1988. Algorithms for partially observable Markov decision processes. Ph.D. dissertation, University of British Columbia, Vancouver, British Columbia, Canada.

Drake, A. 1962. Observation of a Markov process through a noisy channel. Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA.

Eckles, J. E. 1966. Optimum replacement of stochastically failing systems. Ph.D. dissertation, Stanford University, Stanford, CA.

Eckles, J. E. 1968. Optimum maintenance with incomplete information. *Oper. Res.* **16** 1058–1067.

Ehrenfeld, S. 1976. On a sequential Markovian decision procedure with incomplete information. *Comput. Oper. Res.* **3** 39–48.

Feng, Z., S. Zilberstein. 2004. Region-based incremental pruning for POMDPs. *Proc. 20th Conf. Uncertainty in Artificial Intelligence* (UAI-04). Morgan Kaufmann, San Francisco, 146–153.

Fernández-Gaucherand, E., A. Arapostathis, S. I. Marcus. 1991. On the average cost optimality equation and the structure of optimal policies for partially observable Markov decision processes. *Ann. Oper. Res.* **29** 439–470.

Fukuda, K. 2004. From the zonetope construction to the Minkowski addition of convex polytopes. *J. Symbolic Comput.* **38** 1261–1272.

Gritzmann, P., B. Sturmfels. 1993. Minkowski addition of polytopes: Computational complexity and applications to Grobner bases. *SIAM J. Discrete Math.* **6**(2) 246–269.

Grosfeld-Nir, A. 1996. A two-state partially observable Markov decision process with uniformly distributed observations. *Oper. Res.* **44**(3) 458–463.

Grosfeld-Nir, A. 2007. Control limits for two-state partially observable Markov decision processes. *Eur. J. Oper. Res.* **182** 300–304.

Hansen, E. A. 1998a. An improved policy iteration algorithm for partially observable MDPs. *Advances in Neural Inform. Processing Systems* **10** (NIPS-97). MIT Press, Cambridge, MA, 1015–1021.

Hansen, E. A. 1998b. Solving POMDPs by searching in policy space. *Proc. Fourteenth Conf. Uncertainty in Artificial Intelligence* (UAI-98). Morgan Kaufmann, San Francisco, 211–219.

Howard, R. 1960. *Dynamic Programming and Markov Processes*. MIT Press, Cambridge, MA.

Hsu, S.-P., D.-M. Chuang, A. Arapostathis. 2006. On the existence of stationary optimal policies for partially observed MDPs under the long-run average cost criterion. *Systems Control Lett.* **55** 165–173.

Kaelbling, L. P., M. Littman, A. R. Cassandra. 1998. Planning and acting in partially observable stochastic domains. *Artificial Intelligence* **101** 99–134.

Kakalik, J. S. 1965. Optimum policies for partially observable Markov systems. Technical Report 18, Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA.

Lane, D. 1989. A partially observable model of decision making by fishermen. *Oper. Res.* **37**(2) 240–254.

Littman, M. L., A. R. Cassandra, L. P. Kaelbling. 1995. Learning policies for partially observable environments: Scaling up. *Proc. Twelfth Internat. Conf. Machine Learning* (ICML-95). Morgan Kaufmann, San Francisco, 362–370.

Lovejoy, W. S. 1987. Some monotonicity results for partially observed Markov decision processes. *Oper. Res.* **35** (5) 736–743.

Lovejoy, W. S. 1991a. Computationally feasible bounds for partially observed Markov decision processes. *Oper. Res.* **39** 162–175.

Lovejoy, W. S. 1991b. A survey of algorithmic methods for partially observed Markov decision processes. *Ann. Oper. Res.* **28** 47–66.

Lusena, C., J. Goldsmith, M. Mundhenk. 2001. Nonapproximability results for partially observable Markov decision processes. *J. Artificial Intelligence Res.* **14** 83–103.

Marbach, P., O. Mihatsch, J. N. Tsitsiklis. 2000. Call admission control and routing in integrated service networks using neuro-dynamic programming. *IEEE J. Selected Areas Commun.* **18**(2) 197–208.

Meuleau, N., K.-E. Kim, L. P. Kaelbling, A. R. Cassandra. 1999. Solving POMDPs by searching the space of finite policies. *Proc. Fifteenth Conf. Uncertainty in Artificial Intelligence* (UAI-99). Morgan Kaufmann, San Francisco, 417–426.

Meuleau, N., M. Hauskrecht, K.-E. Kim, L. Peshkin, L. P. Kaelbling, T. Dean, C. Boutilier. 1998. Solving very large weakly coupled Markov decision processes. *Proc. Fifteenth National Conf. Artificial Intelligence* (AAAI-98). AAAI Press, Menlo Park, CA, 165–172.

Monahan, G. E. 1982. A survey of partially observable Markov decision processes: Theory, models and algorithms. *Management Sci.* **28** 1–16.

Montemerlo, M., J. Pineau, N. Roy, S. Thrun, V. Verma. 2002. Experiences with a mobile robotic guide for the elderly. *Proc. Eighteenth National Conf. Artificial Intelligence* (AAAI-02). AAAI Press, Menlo Park, CA, 587–592.

Paek, T., E. Horvitz. 2000. Conversation as action under uncertainty. *Proc. Sixteenth Conf. Uncertainty in Artificial Intelligence* (UAI-2000). Morgan Kaufmann, San Francisco, 455–464.

Papadimitriou, C. H., J. N. Tsitsiklis. 1987. The complexity of Markov decision processes. *Math. Oper. Res.* **12**(3) 441–450.

Paz, A. 1971. *Introduction to Probabilistic Automata*. Academic Press, New York.

Pineau, J., G. Gordon, S. Thrun. 2003. Point-based value iteration: An anytime algorithm for POMDPs. *Proc. Eighteenth Internat. Joint Conf. Artificial Intelligence* (IJCAI-03). Morgan Kaufmann, San Francisco, 1025–1030.

Platzman, L. K. 1977. Finite memory estimation and control of finite probabilistic systems. Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA.

Platzman, L. K. 1980. Optimal infinite-horizon undiscounted control of finite probabilistic systems. *SIAM J. Control Optim.* **18**(4) 362–380.

Poupart, P. 2005. Exploiting structure to efficiently solve large scale partially observable Markov decision processes. Ph.D. dissertation, University of Toronto, Toronto, Ontario, Canada.

Poupart, P., C. Boutilier. 2004. Bounded finite state controllers. *Advances in Neural Inform. Processing Systems* **16** (NIPS-03). MIT Press, Cambridge, MA, 823–830

Puterman, M. L. 1994. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, New York.

Rockafellar, R. T. 1970. *Convex Analysis*. Princeton University Press, Princeton, NJ.

Ross, S. 1971. Quality control under Markovian deterioration. *Management Sci.* **17** 587–596.

Satia, J. K., R. E. Lave. 1973. Markovian decision processes with probabilistic observation of states. *Management Sci.* **20** 1–13.

Smallwood, R. D., E. J. Sondik. 1973. The optimal control of partially observable Markov processes over a finite horizon. *Oper. Res.* **21** 1071–1088.

Sondik, E. J. 1971. The optimal control of partially observable Markov processes. Ph.D. dissertation, Stanford University, Stanford, CA.

Sondik, E. J. 1978. The optimal control of partially observable Markov processes over the infinite horizon: Discounted costs. *Oper. Res.* **26** 282–304.

Sulganik, E. 1995. On the structure of Blackwell's equivalence classes of information systems. *Math. Soc. Sci.* **29** 213–223.

Thrun, S. 2000. Monte Carlo POMDPs. *Advances in Neural Inform. Processing Systems* **12** (NIPS-99). MIT Press, Cambridge, MA, 1064–1070.

Treharne, J. T., C. R. Sox. 2002. Adaptive inventory control for nonstationary demand and partial information. *Management Sci.* **48**(5) 607–624.

Wang, R. 1977. Optimal replacement policy under unobservable states. *J. Appl. Probab.* **14** 340–348.

White, C. C. 1977. A Markov quality control process subject to partial observation. *Management Sci.* **23** 843–852.

White, C. C. 1979. Optimal control-limit strategies for a partially observed replacement problem. *Internat. J. Systems Sci.* **10** 321–331.

White, C. C. 1980. Monotone control laws for noisy, countable-state Markov chains. *Eur. J. Oper. Res.* **5** 124–132.

White, C. C. 1991. A survey of solution techniques for the partially observed decision process. *Ann. Oper. Res.* **32** 215–230.

White, C. C., W. T. Scherer. 1989. Solution procedures for partially observed Markov decision processes. *Oper. Res.* **37** 791–797.

White, C. C., W. T. Scherer. 1994. Finite-memory suboptimal design for partially observed Markov decision processes. *Oper. Res.* **42** 439–455.

Yu, H., D. Bertsekas. 2008. On near optimality of the set of finite-state controllers for average cost POMDP. *Math. Oper. Res.* **33**(1) 1–11.

Zhang, N. L., W. Liu. 1996. Planning in stochastic domains: Problem characteristics and approximation. Technical Report HKUST-CS96-31, Hong Kong University of Science and Technology, Hong Kong.

Zhang, B., Q. Cai, J. Mao, B. Guo. 2001. Planning and acting under uncertainty: A new model for spoken dialogue systems. *Proc. Seventeenth Conf. Uncertainty in Artificial Intelligence* (UAI-01). Morgan Kaufmann, San Francisco, 572–579.