European Journal of Operational Research 282 (2020) 936–944

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Decision Support

Monotonicity properties for two-action partially observable Markov decision processes on partially ordered spaces

Erik Miehling[a,*], Demosthenis Teneketzis[b]

[a] Coordinated Science Lab, University of Illinois at Urbana–Champaign, Urbana, IL 61801, United States
[b] Department of Electrical and Computer Engineering, University of Michigan, Ann Arbor, MI 48109, United States

ARTICLE INFO

Article history:
Received 24 August 2018
Accepted 2 October 2019
Available online 9 October 2019

Keywords:
Dynamic programming
Decision analysis
Partially observable Markov decision processes
Partially ordered sets

ABSTRACT

This paper investigates monotonicity properties of optimal policies for two-action partially observable Markov decision processes when the underlying (core) state and observation spaces are partially ordered. Motivated by the desirable properties of the monotone likelihood ratio order in imperfect information settings, namely the preservation of belief ordering under conditioning on any new information, we propose a new stochastic order (a generalization of the monotone likelihood ratio order) that is appropriate for when the underlying space is partially ordered. The generalization is non-trivial, requiring one to impose additional conditions on the elements of the beliefs corresponding to incomparable pairs of states. The stricter conditions in the proposed stochastic order reflect a conservation of structure in the problem – the loss of structure from relaxing the total ordering of the state space to a partial order requires stronger conditions with respect to the ordering of beliefs. In addition to the proposed stochastic order, we introduce a class of matrices, termed generalized totally positive of order 2, that are sufficient for preserving the order. Our main result is a set of sufficient conditions that ensures existence of an optimal policy that is monotone on the belief space with respect to the proposed stochastic order.

1. Introduction

Partially observable Markov decision processes (POMDPs) model settings in which decisions must be made over time subject to imperfect information of the underlying status of the system. They have found applications in a multitude of practical settings including robotics (Hadfield-Menell, Russell, Abbeel, & Dragan, 2016), computer security (Miehling, Rasouli, & Teneketzis, 2018), and spoken dialogue systems (Young, Gašić, Thomson, & Williams, 2013). Unfortunately, obtaining a solution to a POMDP, that is, solving for an optimal decision rule (termed an *optimal policy*), is a computationally difficult process, particularly for the high-dimensional problems found in realistic decision environments.

Structural results for POMDPs investigate conditions under which optimal policies possess desirable properties. One such structural result involves determining conditions under which the optimal policy is increasing in the information state (belief), termed a *monotone policy*. Establishing such structure not only simplifies the search for an optimal policy (often a set of numbers is sufficient for characterizing monotone policies, as argued by Lovejoy (1987), but also provides insight into the nature of the solution, quantifying the relationship between optimal policy structure and the information pattern of the problem.

Questions concerning the structure of optimal policies are fundamental to decision analysis, spanning back to the seminal works of Alchian (1952), Girshick and Rubin (1952), and Bellman (1955). Early work in the area, such as that of Derman and Sacks (1960) and Derman (1963), focused on completely observable settings with the goal of determining the optimal time to replace a system that is probabilistically degrading over time, so-called *replacement rules*. In particular, Derman (1963) studied replacement rules for a completely-observable problem on a totally-ordered state space and derived "monotonicity-preserving" conditions on the transition matrix (*i.e.*, increasing failure rate or IFR) ensuring that the optimal decision rule takes a *control-limit* form.[1]

Investigating structural properties in problems of imperfect information represent a significant complication, primarily due to the requirement to (partially) order beliefs. The work of Ross (1971), one of the first to investigate such properties (in the context of POMDPs), largely avoids this requirement by considering a

* Corresponding author.
E-mail addresses: miehling@illinois.edu (E. Miehling), teneket@umich.edu (D. Teneketzis).

[1] Recent work of Zhuang and Li (2012) has derived an alternative set sufficient conditions, based on multimodularity, to ensure monotone optimal control policies in Markov decision processes.

https://doi.org/10.1016/j.ejor.2019.10.003
0377-2217/

two-state core process, resulting in a total order among beliefs (the belief becomes a single number, *e.g.*, the probability of being in the first state). In his work, Ross introduced an additional action, *inspect*, serving to reveal the true state of the system, and derived conditions that ensured the optimal policy takes an *at-most-four-region* (AM4R) structure.[2] Albright (1979) considered a two-state process similar to that of Ross, but restricted attention to actions that transition the system to improved states, rather than reveal information. Instead, information is revealed to the decision-maker via a finite set of observations, generated probabilistically via an observation matrix as a function of the underlying state. Under monotonicity conditions on the transition matrix and reward functions, as well as the assumption that the observation matrix is *totally positive of order 2* (TP_2), see Karlin (1968) for the definition, the optimal policy is monotone in the belief. Albright illustrates the difficulties associated with considering more than two core states, demonstrating that one loses important monotonicity properties of the belief update when first-order stochastic dominance is used to order beliefs. Nevertheless, building upon the structural results of Porteus (1975), White (1979) managed to derive sufficient conditions to ensure that optimal replacement policies are monotone under first-order stochastic dominance (complementing the completely observable and unobservable cases studied in White (1980). While a significant contribution to the field, White's conditions are fairly restrictive, requiring an upper bound on the discount factor in addition to monotonicity conditions on the model parameters. Lovejoy (1987) circumvented the issues raised by Albright by ordering beliefs using the monotone likelihood ratio, a stronger partial order than first-order stochastic dominance. Lovejoy presents natural sufficient conditions (monotonicity conditions and TP_2 transition matrices) that ensure monotone optimal replacement policies, avoiding the requirement to bound the discount factor.

The strength of the partial order used to compare beliefs is intimately related to the restrictiveness of the conditions involved in establishing the structural result. The conditions of White (1979) involve comparing beliefs in a first-order stochastic dominance sense, resulting in more restrictive conditions than those obtained when beliefs are compared using the stronger monotone likelihood ratio of Lovejoy (1987). The reason for this disparity arises directly from the fact that the monotone likelihood ratio, unlike first-order stochastic dominance, is preserved under conditioning on new information, as demonstrated in Lovejoy (1987). This property illustrates that the monotone likelihood ratio order is a more fitting stochastic order than first-order stochastic dominance for problems of imperfect information.

Lovejoy's results have found many applications (e.g., including use in defense Krishnamurthy & Djonin, 2009, wind turbine maintenance (Byon, Ntaimo, & Ding, 2010), and medical decision-making Sandıkçı, Maillart, Schaefer, & Roberts, 2013) and have formed a basis for various extensions. Namely, Rieder (1991) investigated an extension for settings with state and observation spaces that take the form of *ordered product spaces*, that is, spaces that decompose into the product of totally ordered spaces. Fernández-Gaucherand, Arapostathis, and Marcus (1991) derived structural properties of average cost optimal policies for settings with countably infinite and totally ordered state and observation spaces. More recently, Maillart (2006) studied a partially observable machine maintenance problem where observations can be scheduled, that is, the decision maker can actively choose when to reveal information about the underlying state. Saghafian (2018) introduced the concept of ambiguous POMDPs, describing a setting where one is unsure of the model of the POMDP, and provided an extension of existing structural results to this new setting. Independently of Lovejoy's results, Grosfeld-Nir (2007) and Ben-Zvi, Chernonog, and Avinadav (2016) study structural properties in two-state POMDPs. Similar to Ross (1971), the assumption of two core states leads to a belief that is totally ordered.

In this paper, we extend Lovejoy's results Lovejoy (1987) to problems where the core state space is partially ordered. With the exception of White (1979), existing work (e.g., Fernández-Gaucherand et al., 1991; Maillart, 2006; Saghafian, 2018) considers settings where the core state space is totally ordered (or decomposes into an ordered product space as in Rieder, 1991). The motivation for considering a partially ordered state space is largely a practical one; many problems have state spaces where one cannot necessarily label every state as *better* or *worse* than other states. A key example, and one that serves as the primary motivation for this paper, lies in cyber security. As discussed in Miehling et al. (2018), an attacker's progression through a network can be described by the set of nodes in a directed acyclic graph that it possesses control over. Defining a state as the set of nodes currently under the attacker's control, the natural order relationship is that of inclusion which, under general graph structures, results in a partially ordered state space. Existing structural results in the literature do not apply to such settings. An additional feature of our model is its consideration of observations that are partially ordered, modeling the fact that the quality of signals received from the environment is not always comparable. In the context of the cyber security setting, the security alerts generated from the monitoring system (intrusion detection system) cannot always be ordered by severity. Under this general setting, we investigate a similar topic as that of White (1979) and Lovejoy (1987), namely monotonicity properties of optimal policies. Specifically, we consider two actions, one that lets the system operate uninterrupted and another that transitions the system back to a state with certainty and investigate conditions that ensure optimal replacement policies are monotone in the belief. As will be seen, the extension of Lovejoy's results to the case where the state space is partially ordered requires the introduction of a new stochastic order, which we term the *generalized monotone likelihood ratio*. Under the proposed stochastic order, we are able to derive natural monotonicity preserving conditions on the model parameters, ensuring monotone optimal policies in the two-action setting.

2. The partially observable sequential decision model

Consider a finite time-horizon of length T. At each time t, the state of the system takes on one of finitely many states from the set $\mathcal{S} = \{s_1, \dots, s_n\}$. The controller has access to two actions, $\mathcal{A} = \{a_0, a_1\}$, where a_0 lets the system evolve uninterrupted and a_1 transitions the system to state s_1 with certainty. Actions are costly – for a given state-action pair (s_i, a), an instantaneous cost $c(s_i, a)$ is incurred. Let $c(s_i)$ denote the action-independent cost for state s_i incurred at the terminal stage. Given the current state $x_t = s_i$ and current action $u_t = a$, the system evolves probabilistically as dictated by the conditional transition probability matrices P^a, $a \in \mathcal{A}$, with elements $p^a_{ij} = P(X_{t+1} = s_j \mid X_t = s_i, U_t = a)$.[3] The controller does not observe the underlying state perfectly, instead it receives an observation $o_k \in \mathcal{O} = \{o_1, \dots, o_m\}$ as dictated by the conditional observation (emission) matrix R, with elements $r_{jk} = P(Y_{t+1} = o_k \mid X_{t+1} = s_j)$. Notice for our model that, without loss of generality, the conditional observation probabilities are assumed to be independent of the control action.

[2] Rosenfield (1976a,b) also derived conditions to ensure the AM4R property under a slightly different paradigm in which the state consists of the pair (i, k), representing that it has been k time-steps since the system was known to be in state i.

[3] Random variables are represented by uppercase letters (e.g., X, Y, U) with their realizations given by lowercase letters (e.g., x, y, u).

The state space $\mathcal{S}$ is assumed to be partially ordered by $\succeq_s$, resulting in the partially ordered set, or *poset*, $(\mathcal{S}, \succeq_s)$. It is further assumed that s_1 is the *least* (or *minimal*) element in $\mathcal{S}$ in the sense that $s_i \succeq_s s_1$ for all $i = 2, \dots, n$. Pairs of states $s_i, s_j \in \mathcal{S}$ that are not comparable with respect to $\succeq_s$, that is, neither $s_i \succeq_s s_j$ nor $s_j \succeq_s s_i$, are said to be *incomparable* with respect to $\succeq_s$ and denoted by $s_i \parallel_s s_j$. It is additionally assumed that the observation space $\mathcal{O}$ is partially ordered by $\succeq_o$, forming the poset $(\mathcal{O}, \succeq_o)$, where incomparable observations $o_k, o_l \in \mathcal{O}$ are denoted by $o_k \parallel_o o_l$. Lastly, it is assumed that the action space $\mathcal{A}$ is totally ordered by $\geq$, such that $a_1 \geq a_0$. Without loss of generality, assume that states and observations are indexed according to their respective partial orders, that is, if $s_i \succeq_s s_j$ ($o_k \succeq_o o_l$) then we index s_i and s_j (o_k and o_l) such that $i \geq j$ ($k \geq l$).

For $t = 0, 1, \dots, T-1$, events unfold in the following order:

(1) A control action, $u_t = a \in \mathcal{A}$, is specified.
(2) A cost $c(s_i, a)$ for action a, dependent upon the true state $x_t = s_i \in \mathcal{S}$, is incurred.
(3) The state transitions to $x_{t+1} = s_j \in \mathcal{S}$ as dictated by the transition probabilities $p^a_{ij} = P(X_{t+1} = s_j \mid X_t = s_i, U_t = a)$.
(4) An observation $y_{t+1} = o_k \in \mathcal{O}$ is received as dictated by the conditional observation probabilities $r_{jk} = P(Y_{t+1} = o_k \mid X_{t+1} = s_j)$.

At the final time, $t = T$, a terminal cost $c(x_T)$ is incurred based on the final state of the system.

The information available to the controller at any given time t, represented by the history of actions and observations (as well as the distribution π_0 over the initial state), can be summarized by a probability mass function over the state space $\mathcal{S}$, termed an *information state* or *belief* (see Åström, 1965, Kumar & Varaiya, 1986), denoted by $\pi \in \Delta(\mathcal{S})$. The i'th component of the belief, denoted by π_i, is the conditional probability that the system is in state $s_i \in \mathcal{S}$ given the realized history. Given new information, consisting of the current action $u_t = a$ and the observation $y_{t+1} = o_k$, the updated belief is $\tau(\pi, a, o_k) = \big(\tau_1(\pi, a, o_k), \dots, \tau_n(\pi, a, o_k)\big)$ where the belief update function $\tau : \Delta(\mathcal{S}) \times \mathcal{A} \times \mathcal{O} \to \Delta(\mathcal{S})$ is defined elementwise as

$$\tau_j(\pi, a, o_k) = \frac{\sum_{i=1}^n \pi_i p^a_{ij} r_{jk}}{\sigma(\pi, a, o_k)} \tag{1}$$

where

$$\sigma(\pi, a, o_k) = \sum_{i=1}^n \sum_{j=1}^n \pi_i p^a_{ij} r_{jk}. \tag{2}$$

For later convenience, define $\sigma(\pi, a) \in \Delta(\mathcal{O})$ as a probability mass function consisting of elements $\sigma(\pi, a, o_k)$ over all $o_k \in \mathcal{O}$ for a fixed (π, a), and $r_i \in \Delta(\mathcal{O})$ as a probability mass function consisting of elements r_{ik} over all $o_k \in \mathcal{O}$ for a fixed state s_i.

The goal of the controller is to specify a control action at each time in order to minimize the expected total discounted cost over the time horizon, given by

$$\mathbb{E}\left[\sum_{t=0}^{T-1} \beta^t c(X_t, U_t) + \beta^T c(X_T)\right]$$

where $\beta \in (0, 1]$ is a discount factor. The rule designating this choice is termed a *control policy*, denoted by $g = \big(g_0, g_1, \dots, g_{T-1}\big) \in \mathcal{G}$, where each $g_t : \Delta(\mathcal{S}) \to \mathcal{A}$ is a function mapping an element of the probability simplex over $\mathcal{S}$ to a control action in $\mathcal{A}$ and $\mathcal{G}$ represents the space of admissible control policies. The optimal control policy, denoted by $g^* \in \mathcal{G}$, is the control policy that achieves the minimum expected total discounted cost.

2.1. Objective of the paper

The aim of this paper is to obtain sufficient conditions for two-action POMDPs such that optimal policies are monotone in the belief when the state and observation spaces do not admit total orders but are instead only partially ordered. Formally, we seek to derive an appropriate partial order $\succeq_p$ on $\Delta(\mathcal{S})$ and conditions on the POMDP model such that optimal policies exist in the set $\mathcal{G}^* = \{g = (g_0, g_1, \dots, g_{T-1}) \mid \pi \succeq_p \pi' \Rightarrow g_t(\pi) \geq g_t(\pi') \text{ for } t = 0, 1, \dots, T-1\}$.

3. Preliminary definitions

Achieving the above objective requires one to demonstrate certain properties of the value function and the dynamic programming recursion. To this end, following standard notation from Porteus (1975), White (1979), and Lovejoy (1987), we define the function $\eta : \Delta(\mathcal{S}) \times \mathcal{A} \times \mathscr{B}(\mathcal{S}) \to \mathbb{R}$ as

$$\eta(\pi, a, V) = \sum_{i=1}^n \pi_i c(s_i, a) + \beta \sum_{k=1}^m \sigma(\pi, a, o_k) V(\tau(\pi, a, o_k)) \tag{3}$$

where $\mathscr{B}(\mathcal{S})$ is the set of bounded, real functions on $\Delta(\mathcal{S})$. The value function at any time t, denoted by V^*_t, maps each belief $\pi \in \Delta(\mathcal{S})$ to a value representing the best that one can do from the given belief. Using the definition of η, the value function at any time $t = 1, \dots, T-1$ is given by

$$V^*_t(\pi) = \min_{a \in \mathcal{A}} \eta(\pi, a, V^*_{t+1})$$

with terminal value function $V^*_T(\pi) = \sum_{i=1}^n \pi_i c(s_i)$. An optimal control policy, denoted by $g^* = (g^*_0, g^*_1, \dots, g^*_{T-1})$, is computed from the value functions as follows

$$g^*_t(\pi) = \operatorname*{argmin}_{a \in \mathcal{A}} \eta(\pi, a, V^*_{t+1}) \tag{4}$$

for $t = 0, 1, \dots, T-1$.

Obtaining monotonicity properties of optimal policies in POMDPs requires one to be able to compare beliefs, that is, to say when one belief π is *larger* than another belief π'. This necessitates the use of stochastic orders. Two such stochastic orders that will be useful for later discussion are first-order stochastic dominance and the stronger[4] monotone likelihood ratio order, defined below.

Definition 1 (First-order Stochastic Dominance)**.** Given $\mathcal{Z} = \{z_1, \dots, z_n\}$ and elements $\delta, \delta' \in \Delta(\mathcal{Z})$, δ is said to be greater than δ' with respect to *first order stochastic dominance (FOSD)*, written $\delta \succeq_{st} \delta'$, if $\sum_{j \geq i} \delta_j \geq \sum_{j \geq i} \delta'_j$ for all $i = 1, \dots, n$.

Definition 2 (Monotone Likelihood Ratio)**.** Given $\mathcal{Z} = \{z_1, \dots, z_n\}$ and elements $\delta, \delta' \in \Delta(\mathcal{Z})$, δ is said to be greater than δ' with respect to the *monotone likelihood ratio (MLR)*, written $\delta \succeq_r \delta'$, if $\delta_i \delta'_j \geq \delta_j \delta'_i$ for every $i \geq j$.

The above definitions apply only in the case where the underlying space $\mathcal{Z} = \{z_1, \dots, z_n\}$ is totally ordered, that is, for any two $z_i, z_j \in \mathcal{Z}$, one can write either $z_i \preceq z_j$ or $z_i \succeq z_j$. First-order stochastic dominance has been generalized to the case where the underlying space is partially ordered. Let I_K denote the indicator vector, containing a one for all elements in the set K and a zero otherwise. The definition below, which we refer to as *generalized first-order stochastic dominance* (GFOSD), is courtesy of White (1979).[5]

[4] In the sense that if π dominates π' with respect to the monotone likelihood ratio, then π also dominates π' with respect to first-order stochastic dominance (Whitt, 1979).

[5] It is worth noting that GFOSD reduces to FOSD (Definition 1) in the case where the underlying space is totally ordered; the set $\mathcal{K}$ reduces to contain sets of the form $\{z_1, \dots, z_n\}, \{z_2, \dots, z_n\}, \dots, \{z_n\}$ (since all states are comparable).

Definition 3 (Generalized First-order Stochastic Dominance)**.** Given a poset $(\mathcal{Z}, \succeq_z)$ and elements $\delta, \delta' \in \Delta(\mathcal{Z})$, δ is said to be greater than δ' with respect to *generalized first order stochastic dominance (GFOSD)*, written $\delta \succeq_{gst} \delta'$, if $\delta I_K \geq \delta' I_K$ for all $K \in \mathcal{K} = \{K \subseteq \mathcal{Z} \mid \forall z_i \in K, \forall z_j \in \mathcal{Z}, (z_j \succeq_z z_i \Rightarrow z_j \in K)\}$.

Useful characterizations exist for both FOSD and GFOSD. A common characterization for FOSD, courtesy of Stoyan (1983), is as follows: π is said to dominate π' with respect to $\succeq_{st}$ if and only if $\sum_i \pi_i f(z_i) \geq \sum_i \pi'_i f(z_i)$ for all increasing functions $f : \mathcal{Z} \to \mathbb{R}$. An analogous characterization for GFOSD is courtesy of Kamae, Krengel, and O'Brien (1977). Let us first define the notion of $\succeq_z$-increasing functions: a function $f : \mathcal{Z} \to \mathbb{R}$ is said to be $\succeq_z$-increasing if for any $z_i, z_j \in \mathcal{Z}$ such that $z_i \succeq_z z_j$ we have that $f(z_i) \geq f(z_j)$. The characterization of GFOSD, restated in terms of the notation of our paper, is summarized by the following lemma:

Lemma 1 (Kamae et al., 1977)**.** *Given a poset $(\mathcal{Z}, \succeq_z)$ and elements $\delta, \delta' \in \Delta(\mathcal{Z})$, δ is said to dominate δ' with respect to $\succeq_{gst}$ if and only if $\sum_i \delta_i f(z_i) \geq \sum_i \delta'_i f(z_i)$ for all $\succeq_z$-increasing functions f.*

4. Generalization of the monotone likelihood ratio order to partially ordered spaces

Motivated by the desirable properties of the monotone likelihood ratio order in imperfect information settings – namely the preservation of order under conditioning upon new information – we wish to generalize the MLR order, Definition 2, to the case where the core state space $\mathcal{Z}$ is partially ordered. The proposed order, which we term the generalized monotone likelihood ratio (GMLR), is defined below.

Definition 4 (Generalized Monotone Likelihood Ratio)**.** Given a poset $(\mathcal{Z}, \succeq_z)$ and elements $\delta, \delta' \in \Delta(\mathcal{Z})$, δ is said to be greater than δ' with respect to the *generalized monotone likelihood ratio (GMLR)*, written $\delta \succeq_{gr} \delta'$, if

$$\delta_i \delta'_j \geq \delta_j \delta'_i \quad \text{for } z_i \succeq_z z_j, \tag{5}$$

$$\delta_i \delta'_j = \delta_j \delta'_i \quad \text{for } z_i \parallel_z z_j. \tag{6}$$

An intuitively analogous result to the totally ordered case, and a technical requirement of our main result, is that the GMLR order implies GFOSD. At first glance, one may think that an appropriate generalization of the MLR order to the partially ordered case would be to simply restrict the inequality conditions of the MLR order to comparable pairs of states, as expressed by (5), and not impose any restrictions on the elements of the beliefs that correspond to incomparable states (6). Unfortunately, such a generalization does not ensure comparability of the beliefs with respect to GFOSD, as demonstrated by the following example:

Example 1. Consider a poset $(\mathcal{Z}, \succeq_z)$ where $\mathcal{Z} = \{z_1, z_2, z_3, z_4\}$ such that $z_2 \succeq_z z_1$, $z_3 \succeq_z z_1$, $z_4 \succeq_z z_1$, $z_4 \succeq_z z_2$, $z_4 \succeq_z z_3$, and $z_2 \parallel_z z_3$. First consider the pair of beliefs $\delta = (0, \frac{1}{6}, \frac{1}{3}, \frac{1}{2})$ and $\delta' = (0, \frac{3}{10}, \frac{3}{10}, \frac{2}{5})$ which satisfy the conditions on comparable states (5), $\delta_2\delta'_1 \geq \delta_1\delta'_2$, $\delta_3\delta'_1 \geq \delta_1\delta'_3$, $\delta_4\delta'_1 \geq \delta_1\delta'_4$, $\delta_4\delta'_2 \geq \delta_2\delta'_4$, $\delta_4\delta'_3 \geq \delta_3\delta'_4$, but not the condition on incomparable states (6), $\delta_2\delta'_3 = \delta_3\delta'_2$. Testing dominance with respect to GFOSD (see Definition 3), yields $\mathcal{K} = \{\{z_1, z_2, z_3, z_4\}, \{z_2, z_4\}, \{z_3, z_4\}, \{z_4\}, \varnothing\}$ resulting in the (non-trivial [6]) conditions,

$$\delta_2 + \delta_4 = \frac{1}{6} + \frac{1}{2} = \frac{2}{3} < \frac{7}{10} = \frac{3}{10} + \frac{4}{10} = \delta'_2 + \delta'_4,$$

$$\delta_3 + \delta_4 = \frac{1}{3} + \frac{1}{2} = \frac{5}{6} > \frac{7}{10} = \delta'_3 + \delta'_4,$$

$$\delta_4 = \frac{1}{2} > \frac{2}{5} = \delta'_4.$$

Due to the conflicting directions of the inequalities, δ and δ' are not comparable with respect to GFOSD. Now consider a pair of beliefs, $\delta = (0, \frac{1}{6}, \frac{1}{3}, \frac{1}{2})$ and $\delta' = (\frac{1}{6}, \frac{1}{6}, \frac{1}{3}, \frac{1}{3})$, that satisfy both (5) and (6). Carrying out the same exercise as before yields

$$\delta_2 + \delta_4 = \frac{2}{3} > \frac{1}{2} = \frac{1}{6} + \frac{1}{3} = \delta'_2 + \delta'_4,$$

$$\delta_3 + \delta_4 = \frac{5}{6} > \frac{2}{3} = \frac{1}{3} + \frac{1}{3} = \delta'_3 + \delta'_4,$$

$$\delta_4 = \frac{1}{2} > \frac{1}{3} = \delta'_4,$$

illustrating that the desired implication, $\delta \succeq_{gr} \delta' \Rightarrow \delta \succeq_{gst} \delta'$, now holds.

In summary, the generalization of the MLR order to the partially ordered case not only requires conditions on the beliefs corresponding to comparable pairs of states, but also calls for additional conditions due to the incomparable pairs.[7] The implication illustrated by Example 1 holds in general, and is summarized by the following lemma:

Lemma 2. *For a given poset $(\mathcal{Z}, \succeq_z)$, $\delta \succeq_{gr} \delta'$ implies $\delta \succeq_{gst} \delta'$.*

Proof. Let $\delta \succeq_{gr} \delta'$, so $\delta_i\delta'_j \geq \delta_j\delta'_i$ if $z_i \succeq_z z_j$ and $\delta_i\delta'_j = \delta_j\delta'_i$ if $z_i \parallel_z z_j$. Recall the definition of generalized first-order stochastic dominance (Definition 3). For each $K \in \mathcal{K}$, define $\bar{K} = \mathcal{Z} \setminus K$. As a result of the definition of the set K, and the fact that $\delta \succeq_{gr} \delta'$, for each $(z_i, z_j) \in K \times \bar{K}$ there exists either an expression $\delta_i\delta'_j \geq \delta_j\delta'_i$ if $z_i \succeq_z z_j$ or $\delta_i\delta'_j = \delta_j\delta'_i$ if $z_i \parallel_z z_j$. For a given $K, \bar{K}$ pair, sum the corresponding expressions over all $(z_i, z_j) \in K \times \bar{K}$, yielding

$$\sum_{(z_i,z_j)\in K\times\bar{K}} \delta_i\delta'_j \geq \sum_{(z_i,z_j)\in K\times\bar{K}} \delta_j\delta'_i$$

due to the fact that $\delta \succeq_{gr} \delta'$. The above inequality can be factored into the form $\delta I_K \delta' I_{\bar{K}} \geq \delta I_{\bar{K}} \delta' I_K$. Now,

$$\begin{aligned}\delta I_K \delta' I_{\bar{K}} \geq \delta I_{\bar{K}} \delta' I_K &\equiv (\delta I_K)(1 - \delta' I_K) \geq (1 - \delta I_K)(\delta' I_K)\\ &\equiv \delta I_K - \delta I_K \delta' I_K \geq \delta' I_K - \delta I_K \delta' I_K\\ &\equiv \delta I_K \geq \delta' I_K\end{aligned}$$

for each $K \in \mathcal{K}$, thus $\delta \succeq_{gst} \delta'$. □

An important step in establishing the desired monotonicity properties is characterizing the class of matrices that preserve the GMLR order, that is, given $\delta \succeq_{gr} \delta'$, finding the class of matrices Q such that $\delta Q \succeq_{gr} \delta' Q$. In the case where the underlying space is totally ordered, it is known that the MLR order is preserved by a class of matrices termed totally positive of order 2 (TP_2), that is, if $\delta \succeq_r \delta'$ and Q is a stochastic, TP_2 matrix then $\delta Q \succeq_r \delta' Q$ (see Karlin, 1968; Karlin & Rinott, 1980). We define a generalized notion of TP_2 matrices (in Definition 5) for the case where the underlying space is partially ordered, which we term *generalized totally positive of order 2* (GTP_2), and show (in Proposition 1) that matrices of this type are sufficient for preserving the GMLR order.

Definition 5 (Generalized Totally Positive of Order 2)**.** Given two posets $(\mathcal{B}, \succeq_b)$, $(\mathcal{D}, \succeq_d)$, where $\mathcal{B} = \{b_1, \ldots, b_n\}$ and

[6] The conditions corresponding to the full set, $\mathcal{Z}$, and the empty set are always satisfied.

[7] Notice that if $\mathcal{Z}$ were totally ordered, there would be no $z_i, z_j \in \mathcal{Z}$ such that $z_i \parallel_z z_j$, resulting in $\succeq_{gr}$ reducing to $\succeq_r$ (Definition 2).

$\mathcal{D} = \{d_1, \dots, d_m\}$, a matrix Q is said to be *generalized totally positive of order 2 (GTP_2)* on $(\mathcal{B}, \succeq_b) \times (\mathcal{D}, \succeq_d)$ if for every $b_k \succeq_b b_l$,

$$q_{lj}q_{ki} - q_{kj}q_{li} \geq 0 \quad \text{for } d_i \succeq_d d_j,$$
$$q_{lj}q_{ki} - q_{kj}q_{li} = 0 \quad \text{for } d_i \parallel_d d_j.$$

The above definition states that for every pair of rows corresponding to comparable elements, the determinant of the second order minor formed by choosing pairs of columns must satisfy the appropriate conditions (depending on whether or not the columns correspond to comparable elements). There are no restrictions on pairs of rows corresponding to incomparable elements. Notice that when the spaces $(\mathcal{B}, \succeq_b)$ and $(\mathcal{D}, \succeq_d)$ are totally ordered, GTP_2 reduces to TP_2 (see Karlin, 1968). For notational simplicity, if the two posets are the same, that is, $(\mathcal{B}, \succeq_b) = (\mathcal{D}, \succeq_d)$, then the matrix is square and we refer to the matrix as being GTP_2 on $(\mathcal{B}, \succeq_b)$.

Proposition 1. *Given a poset $(\mathcal{Z}, \succeq_z)$, elements $\delta, \delta' \in \Delta(\mathcal{Z})$, and a stochastic, GTP_2 matrix Q on $(\mathcal{Z}, \succeq_z)$, $\delta \succeq_{gr} \delta'$ implies $\delta Q \succeq_{gr} \delta' Q$.*

Proof. Let Q be GTP_2 and $\delta \succeq_{gr} \delta'$. Denoting $Q_{\circ,i}$ as the ith column of matrix Q, we wish to show that $\delta Q_{\circ,i} \delta' Q_{\circ,j} \geq \delta Q_{\circ,j} \delta' Q_{\circ,i}$ for $z_i \succeq_z z_j$ and $\delta Q_{\circ,i} \delta' Q_{\circ,j} = \delta Q_{\circ,j} \delta' Q_{\circ,i}$ for $z_i \parallel_z z_j$. Equivalently, defining $\rho_{ij}(\delta, \delta') = \delta Q_{\circ,i} \delta' Q_{\circ,j} - \delta Q_{\circ,j} \delta' Q_{\circ,i}$, we wish to show that

$$\rho_{ij}(\delta, \delta') \geq 0 \quad \text{for } z_i \succeq_z z_j$$
$$\rho_{ij}(\delta, \delta') = 0 \quad \text{for } z_i \parallel_z z_j.$$

Observe that

$$\begin{aligned}\rho_{ij}(\delta, \delta') &= \delta Q_{\circ,i} \delta' Q_{\circ,j} - \delta Q_{\circ,j} \delta' Q_{\circ,i} \\ &= \delta (Q_{\circ,i} Q_{\circ,j}^T - Q_{\circ,j} Q_{\circ,i}^T) \delta'^T.\end{aligned}$$

Define $B^{ij} = Q_{\circ,i} Q_{\circ,j}^T - Q_{\circ,j} Q_{\circ,i}^T$ and notice that B^{ij} is skew-symmetric, that is, $(B^{ij})^T = -B^{ij}$. The (k, l)'th element of matrix B^{ij}, denoted by b_{kl}^{ij}, is given by

$$b_{kl}^{ij} = q_{lj}q_{ki} - q_{kj}q_{li}$$

where $b_{kl}^{ij} = 0$ for $k = l$. The function $\rho_{ij}(\delta, \delta') = \delta B^{ij} \delta'^T$ can then be written as

$$\rho_{ij}(\delta, \delta') = \sum_{l=1}^{n} \sum_{k=l+1}^{n} (q_{lj}q_{ki} - q_{kj}q_{li})(\delta_k \delta'_l - \delta_l \delta'_k). \tag{7}$$

Recall our objective of showing that $\rho_{ij}(\delta, \delta') \geq 0$ for $z_i \succeq_z z_j$ and $\rho_{ij}(\delta, \delta') = 0$ for $z_i \parallel_z z_j$. First, consider the case where $z_i \succeq_z z_j$. If $z_k \succeq_z z_l$, then by $\delta \succeq_{gr} \delta'$, $\delta_k \delta'_l - \delta_l \delta'_k \geq 0$, and since Q is assumed to be GTP_2, we have that $q_{lj}q_{ki} - q_{kj}q_{li} \geq 0$, and the corresponding term in the sum is positive (see Eq. (7)). Otherwise, if $z_k \parallel_z z_l$ then $\delta_k \delta'_l - \delta_l \delta'_k = 0$ and the corresponding term in the sum is zero, regardless of the sign of $q_{lj}q_{ki} - q_{kj}q_{li}$. Consequently $\rho_{ij}(\delta, \delta') \geq 0$ when $z_i \succeq_z z_j$. Second, consider the case where $z_i \parallel_z z_j$. As in the first case, if $z_k \succeq_z z_l$ then $\delta_k \delta'_l - \delta_l \delta'_k \geq 0$, but now since $z_i \parallel_z z_j$, we have that $q_{lj}q_{ki} - q_{kj}q_{li} = 0$ since Q is GTP_2, resulting in the corresponding term in the sum to be zero. If $z_k \parallel_z z_l$ then $\delta_k \delta'_l - \delta_l \delta'_k = 0$ and the corresponding term in the sum is zero, regardless of the sign of $q_{lj}q_{ki} - q_{kj}q_{li}$. Consequently $\rho_{ij}(\delta, \delta') = 0$ when $z_i \parallel_z z_j$. □

5. Main result: sufficient conditions for monotone optimal policies

Establishing monotonicity properties of optimal policies involves deriving the appropriate conditions on the state dynamics, observation dynamics, and structure of the instantaneous and terminal cost functions. Our main result, stated below in Theorem 1, provides sufficient conditions for optimal policies to be monotone in the belief with respect to the GMLR order.

Theorem 1. *If $\pi \succeq_{gr} \pi'$ and the following conditions hold:*

(a) *$c(s)$ is increasing in s on $(\mathcal{S}, \succeq_s)$*
(b) *$c(s, a)$ is increasing in s on $(\mathcal{S}, \succeq_s)$ for each $a \in \mathcal{A}$*
(c) *$c(s, a_1) - c(s, a_0)$ is decreasing in s on $(\mathcal{S}, \succeq_s)$*
(d) *P^a is GTP_2 on $(\mathcal{S}, \succeq_s)$ for each $a \in \mathcal{A}$*
(e) *R is GTP_2 on $(\mathcal{S}, \succeq_s) \times (\mathcal{O}, \succeq_o)$ and R^T is GTP_2 on $(\mathcal{O}, \succeq_o) \times (\mathcal{S}, \succeq_s)$*

then $g_t^*(\pi) \geq g_t^*(\pi')$ for all $t = 0, 1, \dots, T-1$.

The remainder of Section 5 will be dedicated to proving the above theorem. The results proceed by demonstrating, in Section 5.1, that conditions (a), (b), and (d), (e) ensure that the value functions are increasing in the belief with respect to the GMLR order, that is, the value functions are increasing on the poset $(\Delta(\mathcal{S}), \succeq_{gr})$. This result is formally stated in Lemma 6. An additional condition, (c), on the instantaneous cost function (decreasing differences), along with a result from Topkis Topkis (1978), ensures that optimal policies are also monotone on the poset $(\Delta(\mathcal{S}), \succeq_{gr})$. The section concludes in Section 5.2 with the proof of Theorem 1.

5.1. Monotonicity of the value functions

Establishing monotonicity of the value functions on the poset $(\Delta(\mathcal{S}), \succeq_{gr})$, that is, showing $V_t^*(\pi) \geq V_t^*(\pi')$ for any $\pi \succeq_{gr} \pi'$, requires first establishing some properties of the information dynamics. Specifically, the lemmas below (Lemmas 3 and 4) characterize monotonicity properties of the belief update function τ in both the observation and the belief. Lemma 3 establishes equivalence between monotonicity of the belief update in o on the observation poset $(\mathcal{O}, \succeq_o)$, for a fixed belief and action, and a condition on the observation matrix R. Lemma 4 shows equivalence between monotonicity of the belief update in π on the poset $(\Delta(\mathcal{S}), \succeq_{gr})$, for a fixed action and observation, and preservation of the order between GMLR-comparable beliefs. Lemmas 3 and 4 are the partially ordered analogues to Lemma 1.2, parts (1) and (2), of the totally ordered setting found in Lovejoy (1987).

Lemma 3. *For any $\pi \in \Delta(\mathcal{S})$ and $a \in \mathcal{A}$,*

$$\tau(\pi, a, o_k) \succeq_{gr} \tau(\pi, a, o_l)$$

for all $o_k \succeq_o o_l$ in $\mathcal{O}$ if and only if R^T is GTP_2 on $(\mathcal{O}, \succeq_o) \times (\mathcal{S}, \succeq_s)$.

Proof. For any $\pi \in \Delta(\mathcal{S})$, $a \in \mathcal{A}$, and $o_k, o_l \in \mathcal{O}$ such that $o_k \succeq_o o_l$, $\tau(\pi, a, o_k) \succeq_{gr} \tau(\pi, a, o_l)$ if and only if (by Definition 4)

$$\tau_i(\pi, a, o_k)\tau_j(\pi, a, o_l) \geq \tau_j(\pi, a, o_k)\tau_i(\pi, a, o_l) \quad \text{for } s_i \succeq_s s_j$$
$$\tau_i(\pi, a, o_k)\tau_j(\pi, a, o_l) = \tau_j(\pi, a, o_k)\tau_i(\pi, a, o_l) \quad \text{for } s_i \parallel_s s_j$$

for all $o_k \succeq_o o_l$. Using the definition of $\tau_i(\pi, a, o)$, Eq. (1), we can expand the above expressions to obtain

$$\begin{aligned}&\left(\frac{r_{ik}\sum_{w=1}^n \pi_w p_{wi}^a}{\sigma(\pi, a, o_k)}\right)\left(\frac{r_{jl}\sum_{w=1}^n \pi_w p_{wj}^a}{\sigma(\pi, a, o_l)}\right) \\ &\geq \left(\frac{r_{jk}\sum_{w=1}^n \pi_w p_{wj}^a}{\sigma(\pi, a, o_k)}\right)\left(\frac{r_{il}\sum_{w=1}^n \pi_w p_{wi}^a}{\sigma(\pi, a, o_l)}\right) \quad \text{for } s_i \succeq_s s_j \\ &\left(\frac{r_{iv}\sum_{w=1}^n \pi_w p_{wi}^a}{\sigma(\pi, a, o_k)}\right)\left(\frac{r_{jl}\sum_{w=1}^n \pi_w p_{wj}^a}{\sigma(\pi, a, o_l)}\right) \\ &= \left(\frac{r_{jk}\sum_{w=1}^n \pi_w p_{wj}^a}{\sigma(\pi, a, o_k)}\right)\left(\frac{r_{il}\sum_{w=1}^n \pi_w p_{wi}^a}{\sigma(\pi, a, o_l)}\right) \quad \text{for } s_i \parallel_s s_j\end{aligned}$$

for all $o_k \succeq_o o_l$. Multiplying both sides of the expressions by $\sigma(\pi, a, o_k)\sigma(\pi, a, o_l)$, defined in Eq. (2), we obtain

$$\left(r_{ik}\sum_{w=1}^n \pi_w p_{wi}^a\right)\left(r_{jl}\sum_{w=1}^n \pi_w p_{wj}^a\right)$$

$$\geq \left(r_{jk}\sum_{w=1}^{n}\pi_w p^a_{wj}\right)\left(r_{il}\sum_{w=1}^{n}\pi_w p^a_{wi}\right) \quad \text{for } s_i \succeq_s s_j$$
$$\left(r_{ik}\sum_{w=1}^{n}\pi_w p^a_{wi}\right)\left(r_{jl}\sum_{w=1}^{n}\pi_w p^a_{wj}\right)$$
$$= \left(r_{jk}\sum_{w=1}^{n}\pi_w p^a_{wj}\right)\left(r_{il}\sum_{w=1}^{n}\pi_w p^a_{wi}\right) \quad \text{for } s_i \parallel_s s_j$$

for all $o_k \succeq_o o_l$. Rearranging, the expressions can be equivalently written as

$$(r_{ik}r_{jl} - r_{jk}r_{il})\left(\sum_{w=1}^{n}\pi_w p^a_{wi}\right)\left(\sum_{w=1}^{n}\pi_w p^a_{wj}\right) \geq 0 \quad \text{for } s_i \succeq_s s_j$$
$$(r_{ik}r_{jl} - r_{jk}r_{il})\left(\sum_{w=1}^{n}\pi_w p^a_{wi}\right)\left(\sum_{w=1}^{n}\pi_w p^a_{wj}\right) = 0 \quad \text{for } s_i \parallel_s s_j$$

for all $o_k \succeq_o o_l$. The above expressions are true if and only if

$$r_{ik}r_{jl} \geq r_{jk}r_{il} \quad \text{for } s_i \succeq_s s_j$$
$$r_{ik}r_{jl} = r_{jk}r_{il} \quad \text{for } s_i \parallel_s s_j$$

for all $o_k \succeq_o o_l$ or, equivalently, R^T is GTP$_2$ on $(\mathcal{O}, \succeq_o) \times (\mathcal{S}, \succeq_s)$. □

Lemma 4. *For any $a \in \mathcal{A}$ and $o_k \in \mathcal{O}$,*

$$\tau(\pi, a, o_k) \succeq_{gr} \tau(\pi', a, o_k)$$

for all $\pi \succeq_{gr} \pi'$ in $\Delta(\mathcal{S})$ if and only if $\pi P^a \succeq_{gr} \pi' P^a$ for all $\pi \succeq_{gr} \pi'$ in $\Delta(\mathcal{S})$.

Proof. We need to show that the information state update preserves the GMLR order, for a fixed action and observation, if and only if the transition matrices preserve the GMLR order. For any $a \in \mathcal{A}$, $o_k \in \mathcal{O}$, and $\pi, \pi' \in \Delta(\mathcal{S})$ such that $\pi \succeq_{gr} \pi'$, $\tau(\pi, a, o_k) \succeq_{gr} \tau(\pi', a, o_k)$ if and only if

$$\tau_i(\pi, a, o_k)\tau_j(\pi', a, o_k) \geq \tau_j(\pi, a, o_k)\tau_i(\pi', a, o_k) \quad \text{for all } s_i \succeq_s s_j$$
$$\tau_i(\pi, a, o_k)\tau_j(\pi', a, o_k) = \tau_j(\pi, a, o_k)\tau_i(\pi', a, o_k) \quad \text{for all } s_i \parallel_s s_j$$

for $\pi \succeq_{gr} \pi'$. The above can be shown to be equivalent to

$$r_{ik}r_{jk}\left(\sum_{w=1}^{n}\pi_w p^a_{wi}\right)\left(\sum_{w=1}^{n}\pi'_w p^a_{wj}\right)$$
$$\geq r_{jk}r_{ik}\left(\sum_{w=1}^{n}\pi_w p^a_{wj}\right)\left(\sum_{w=1}^{n}\pi'_w p^a_{wi}\right) \quad \text{for } s_i \succeq_s s_j$$
$$r_{ik}r_{jk}\left(\sum_{w=1}^{n}\pi_w p^a_{wi}\right)\left(\sum_{w=1}^{n}\pi'_w p^a_{wj}\right)$$
$$= r_{jk}r_{ik}\left(\sum_{w=1}^{n}\pi_w p^a_{wj}\right)\left(\sum_{w=1}^{n}\pi'_w p^a_{wi}\right) \quad \text{for } s_i \parallel_s s_j$$

for $\pi \succeq_{gr} \pi'$. Let $P^a_{\circ,i}$ denote the ith column of P^a. The above is equivalent to

$$\pi P^a_{\circ,i}\pi' P^a_{\circ,j} \geq \pi P^a_{\circ,j}\pi' P^a_{\circ,i} \quad \text{for } s_i \succeq_s s_j$$
$$\pi P^a_{\circ,i}\pi' P^a_{\circ,j} = \pi P^a_{\circ,j}\pi' P^a_{\circ,i} \quad \text{for } s_i \parallel_s s_j$$

for $\pi \succeq_{gr} \pi'$, which is equivalent to $\pi P^a \succeq_{gr} \pi' P^a$ for $\pi \succeq_{gr} \pi'$. □

Before showing monotonicity of the value function, the following result regarding stochastic ordering of the pmf's $\sigma(\pi, a) \in \Delta(\mathcal{O})$, $\pi \in \Delta(\mathcal{S}), a \in \mathcal{A}$, will be useful. This result, shown in Lemma 5 below, follows from the conditions that the transition matrices P^a, for each $a \in \mathcal{A}$, and the observation matrix, R, are GTP$_2$.

Lemma 5. *If $\pi \succeq_{gr} \pi'$ and the following conditions hold:*

1. *P^a is GTP$_2$ on $(\mathcal{S}, \succeq_s)$ for each $a \in \mathcal{A}$*
2. *R is GTP$_2$ on $(\mathcal{S}, \succeq_s) \times (\mathcal{O}, \succeq_o)$*

then $\sigma(\pi, a) \succeq_{gst} \sigma(\pi', a)$ for each $a \in \mathcal{A}$.

Proof. Let $P^a_{i,\circ}$ denote the ith row of matrix P^a. By assumption 1,

$$P^a_{i,\circ} \succeq_{gst} P^a_{j,\circ}$$

for $s_i \succeq_s s_j$. This can be seen by recognizing that, for any $s_i \succeq_s s_j$, the degenerate beliefs $e_i, e_j \in \Delta(\mathcal{S})$ (where e_i is a pmf with all mass on element i) satisfy $e_i \succeq_{gr} e_j$ and noticing that $e_i P^a = P^a_{i,\circ} \succeq_{gr} P^a_{j,\circ} = e_j P^a$ by Proposition 1 and thus $P^a_{i,\circ} \succeq_{gst} P^a_{j,\circ}$ by Lemma 2. By assumption 2, $r_i \succeq_{gr} r_j$ for all $s_i \succeq_s s_j$ and thus $r_i \succeq_{gst} r_j$ for all $s_i \succeq_s s_j$ by Lemma 2. That is

$$r_i I_J \geq r_j I_J$$

for all $J \in \mathcal{J} = \{J \subseteq \mathcal{O} \mid \forall o_l \in J, \forall o_k \in \mathcal{O}, (o_k \succeq_o o_l \Rightarrow o_k \in J)\}$. Using the aforementioned Lemma 1 of stochastic dominance on a poset, we have that $\sum_{j=1}^{n} p^a_{ij} r_j I_J$ is increasing in i on $(\mathcal{S}, \succeq_s)$ for all $J \in \mathcal{J}$. Now, since $\pi \succeq_{gr} \pi'$ by assumption, $\pi \succeq_{gst} \pi'$ by Lemma 2, and again by Lemma 1 we have

$$\sum_{i=1}^{n}\pi_i \sum_{j=1}^{n} p^a_{ij} r_j I_J \geq \sum_{i=1}^{n}\pi'_i \sum_{j=1}^{n} p^a_{ij} r_j I_J$$

for all $J \in \mathcal{J}$. Recall $\sigma(\pi, a, o_k) = \sum_{i=1}^{n}\pi_i \sum_{j=1}^{n} p^a_{ij} r_{jk}$, so the above inequality is equivalent to $\sigma(\pi, a) \succeq_{gst} \sigma(\pi', a)$. □

Using Lemmas 3 through 5 and imposing monotonicity conditions on the instantaneous and terminal cost functions enable one to show that the optimal value function is increasing on the poset $(\Delta(\mathcal{S}), \succeq_{gr})$.

Lemma 6. *Let $\pi \succeq_{gr} \pi'$ and assume the following conditions hold:*

1. *$c(s)$ is increasing in s on $(\mathcal{S}, \succeq_s)$*
2. *$c(s, a)$ is increasing in s on $(\mathcal{S}, \succeq_s)$ for each $a \in \mathcal{A}$*
3. *P^a is GTP$_2$ on $(\mathcal{S}, \succeq_s)$ for each $a \in \mathcal{A}$*
4. *R is GTP$_2$ on $(\mathcal{S}, \succeq_s) \times (\mathcal{O}, \succeq_o)$ and R^T is GTP$_2$ on $(\mathcal{O}, \succeq_o) \times (\mathcal{S}, \succeq_s)$*

then $V^*_t(\pi) \geq V^*_t(\pi')$ for all $t = 1, 2, \ldots, T$.

Proof. The proof proceeds by induction. By assumption $\pi \succeq_{gr} \pi'$ and thus, by Lemma 2, $\pi \succeq_{gst} \pi'$. Under the assumption that $c(s)$ is increasing in s on $(\mathcal{S}, \succeq_s)$, Lemma 1 yields

$$V^*_T(\pi) = \sum_{i=1}^{n}\pi_i c(s_i) \geq \sum_{i=1}^{n}\pi'_i c(s_i) = V^*_T(\pi').$$

Now, assume that $V^*_{t+1}(\pi)$ is increasing on $(\Delta(\mathcal{S}), \succeq_{gr})$, that is, $V^*_{t+1}(\pi) \geq V^*_{t+1}(\pi')$ for $\pi \succeq_{gr} \pi'$ (induction hypothesis). Also, let action a' be optimal in π', that is $a' = g^*_t(\pi')$, so

$$V^*_t(\pi') = \sum_{i=1}^{n}\pi'_i c(s_i, a') + \beta \sum_{k=1}^{m}\sigma(\pi', a', o_k) V^*_{t+1}(\tau(\pi', a', o_k))$$
$$\leq \sum_{i=1}^{n}\pi'_i c(s_i, a) + \beta \sum_{k=1}^{m}\sigma(\pi', a, o_k) V^*_{t+1}(\tau(\pi', a, o_k)) \quad (8)$$

where $a = g^*_t(\pi)$. By Lemma 3 and assumption 4, $\tau(\pi, a, o)$ is increasing in o on $(\mathcal{O}, \succeq_o)$ for any $\pi \in \Delta(\mathcal{S})$, $a \in \mathcal{A}$, and by the induction hypothesis, $V^*_{t+1}(\tau(\pi', a, o))$ is also increasing in o on $(\mathcal{O}, \succeq_o)$. Now by Lemmas 1 and 5, we have that

$$\sum_{i=1}^{n}\pi_i' c(s_i,a)+\beta\sum_{k=1}^{m}\sigma(\pi',a,o_k)V_{t+1}^*(\tau(\pi',a,o_k))$$
$$\leq\sum_{i=1}^{n}\pi_i' c(s_i,a)+\beta\sum_{k=1}^{m}\sigma(\pi,a,o_k)V_{t+1}^*(\tau(\pi',a,o_k)). \tag{9}$$

Next, note that since $\pi\succeq_{gst}\pi'$ and by assumption 2, $\sum_{i=1}^{n}\pi_i' c(s_i,a)\leq\sum_{i=1}^{n}\pi_i c(s_i,a)$, follows by Lemma 1. Furthermore, by Lemma 4 and assumption 3, $\tau(\pi,a,o)$ is increasing in π on $(\Delta(\mathcal{S}),\succeq_{gr})$ for any $a\in\mathcal{A}$, $o_k\in\mathcal{O}$, and using the induction hypothesis, we have

$$\sum_{i=1}^{n}\pi_i' c(s_i,a)+\beta\sum_{k=1}^{m}\sigma(\pi,a,o_k)V_{t+1}^*(\tau(\pi',a,o_k))$$
$$\leq\sum_{i=1}^{n}\pi_i c(s_i,a)+\beta\sum_{k=1}^{m}\sigma(\pi,a,o_k)V_{t+1}^*(\tau(\pi,a,o_k))=V_t^*(\pi). \tag{10}$$

The result holds by transitivity of Eqs. (8)–(10) and induction. □

It is worth noting that, when the state and observation spaces are totally ordered, the two requirements in condition (4) of Lemma 6 collapse to the requirement that R is TP_2.

5.2. Proof of the main result

Recall the function $\eta:\Delta(\mathcal{S})\times\mathcal{A}\times\mathscr{B}(\mathcal{S})\to\mathbb{R}$ of Eq. (3) and consider Lemma 7 below, a special case of Lemma 6.1 from Topkis (1978), restated using the notation of our model.

Lemma 7 (Topkis, 1978)**.** *If $\eta(\pi,a,V_{t+1}^*)$ has decreasing differences in (π,a) on $(\Delta(\mathcal{S}),\succeq_{gr})\times\mathcal{A}$, then there exists a function $g_t^*(\pi)=\operatorname{argmin}_{a\in\mathcal{A}}\{\eta(\pi,a,V_{t+1}^*)\}$ that is nondecreasing in π on $(\Delta(\mathcal{S}),\succeq_{gr})$.*

Under condition (c) of Theorem 1, Lemma 7 allows us to translate monotonicity of the value function into monotonicity of optimal policies. The proof of Theorem 1 can now be stated.

Proof of Theorem 1. First, we show that $\eta(\pi,a,V_{t+1}^*)$ has decreasing differences in (π,a) on $(\Delta(\mathcal{S}),\succeq_{gr})\times\mathcal{A}$, that is, $\eta(\pi,a_1,V_{t+1}^*)-\eta(\pi,a_0,V_{t+1}^*)$ is decreasing in π on $(\Delta(\mathcal{S}),\succeq_{gr})$. Then, application of Lemma 7 proves the result. Recall that $\tau(\pi,a_1,o_k)=e_1$ for any $\pi\in\Delta(\mathcal{S})$, $o_k\in\mathcal{O}$, that is, action a_1 causes the system state to transition to s_1 with certainty. Using this fact, along with the definition of η, see Eq. (3), we can write the following:

$$\eta(\pi,a_1,V_{t+1}^*)-\eta(\pi,a_0,V_{t+1}^*)$$
$$=\sum_{i=1}^{n}\pi_i(c(s_i,a_1)-c(s_i,a_0))$$
$$+\beta\left(V_{t+1}^*(e_1)-\sum_{k=1}^{m}\sigma(\pi,a_0,o_k)V_{t+1}^*(\tau(\pi,a_0,o_k))\right).$$

Thus, for $\pi\succeq_{gr}\pi'$, we wish to show

$$\sum_{i=1}^{n}\pi_i(c(s_i,a_1)-c(s_i,a_0))$$
$$+\beta\left(V_{t+1}^*(e_1)-\sum_{k=1}^{m}\sigma(\pi,a_0,o_k)V_{t+1}^*(\tau(\pi,a_0,o_k))\right)$$
$$\leq\sum_{i=1}^{n}\pi_i'(c(s_i,a_1)-c(s_i,a_0))$$
$$+\beta\left(V_{t+1}^*(e_1)-\sum_{k=1}^{m}\sigma(\pi',a_0,o_k)V_{t+1}^*(\tau(\pi',a_0,o_k))\right). \tag{11}$$

By condition (c) of Theorem 1, we have that $c(s_i,a_1)-c(s_i,a_0)$ is decreasing in s_i on $(\mathcal{S},\succeq_s)$. Consequently, since $\pi\succeq_{gst}\pi'$, Lemma 1 ensures that $\sum_{i=1}^{n}\pi_i(c(s_i,a_1)-c(s_i,a_0))\leq\sum_{i=1}^{n}\pi_i'(c(s_i,a_1)-c(s_i,a_0))$. Now, to ensure the relationship in Eq. (11) holds, we need to show that

$$\sum_{k=1}^{m}\sigma(\pi',a_0,o_k)V_{t+1}^*(\tau(\pi',a_0,o_k))$$
$$\leq\sum_{k=1}^{m}\sigma(\pi,a_0,o_k)V_{t+1}^*(\tau(\pi,a_0,o_k)). \tag{12}$$

Eq. (12) follows directly from the arguments found in the proof of Lemma 6 (recall the arguments for establishing the inequalities in Eqs. (9) and (10)). Specifically, notice that by conditions (a), (b), and (d), (e), we have that $V_{t+1}^*(\pi)\geq V_{t+1}^*(\pi')$ for $\pi\succeq_{gr}\pi'$ by Lemma 6. Furthermore, by conditions (d) and (e), and Lemmas 1, 3, and 5, monotonicity of the value function ensures that

$$\sum_{k=1}^{m}\sigma(\pi',a_0,o_k)V_{t+1}^*(\tau(\pi',a_0,o_k))$$
$$\leq\sum_{k=1}^{m}\sigma(\pi,a_0,o_k)V_{t+1}^*(\tau(\pi',a_0,o_k)). \tag{13}$$

Additionally, by condition (d), Lemma 4, and monotonicity of the value function, we have

$$\sum_{k=1}^{m}\sigma(\pi,a_0,o_k)V_{t+1}^*(\tau(\pi',a_0,o_k))$$
$$\leq\sum_{k=1}^{m}\sigma(\pi,a_0,o_k)V_{t+1}^*(\tau(\pi,a_0,o_k)). \tag{14}$$

Eq. (12) follows by the transitivity of Eqs. (13) and (14), and thus $\eta(\pi,a,V_{t+1}^*)$ has decreasing differences in (π,a) on $(\Delta(\mathcal{S}),\succeq_{gr})\times\mathcal{A}$. Application of Lemma 7 ensures that the optimal policy $g_t^*(\pi)=\operatorname{argmin}_{a\in\mathcal{A}}\left(\eta(\pi,a,V_{t+1}^*)\right)$ is increasing in π on $(\Delta(\mathcal{S}),\succeq_{gr})$. □

6. An example

We now present an instance of a POMDP that satisfies the conditions of Theorem 1. Consider the same state space and partial order as in Example 1, that is, $\mathcal{S}=\{s_1,s_2,s_3,s_4\}$ and $\succeq_s$ such that $s_2\succeq_s s_1$, $s_3\succeq_s s_1$, $s_4\succeq_s s_1$, $s_4\succeq_s s_2$, $s_4\succeq_s s_3$, and $s_2\parallel_s s_3$. Given the poset $(\mathcal{S},\succeq_s)$, the following cost structure satisfies conditions (a) – (c) of Theorem 1: instantaneous costs of $c(\cdot,a_0)=(0,\frac{1}{4},\frac{1}{2},1)$, $c(\cdot,a_1)=(\frac{1}{2},\frac{1}{2},\frac{3}{4},1)$, and terminal cost $c(\cdot)=(0,\frac{1}{4},\frac{1}{2},1)$. Recalling the conditions of Definition 5, the following transition matrices are GTP_2.

$$P^{a_0}=\begin{bmatrix}\frac{1}{4}&\frac{1}{4}&\frac{1}{4}&\frac{1}{4}\\0&\frac{1}{3}&\frac{1}{3}&\frac{1}{3}\\0&\frac{1}{4}&\frac{1}{4}&\frac{1}{2}\\0&0&0&1\end{bmatrix}\qquad P^{a_1}=\begin{bmatrix}1&0&0&0\\1&0&0&0\\1&0&0&0\\1&0&0&0\end{bmatrix}$$

For purposes of this example, assume that the observation poset is equal to the state poset, $(\mathcal{O},\succeq_o)=(\mathcal{S},\succeq_s)$, that is, we see a noisy representation of the state. Under this assumption, condition (e) of Theorem 1 requires that both R and R^T are GTP_2 on $(\mathcal{O},\succeq_o)$ (or equivalently on $(\mathcal{S},\succeq_s)$). The following observation

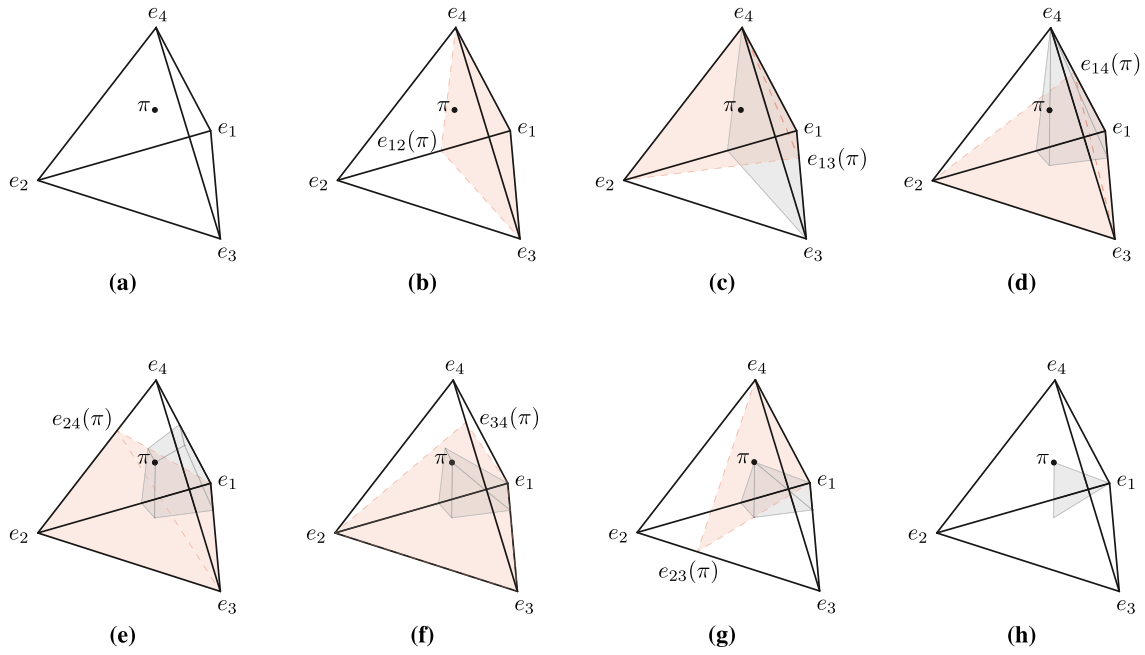

Fig. 1. Construction of comparable beliefs π', such that $\pi = (0.3, 0.2, 0.1, 0.4) \succeq_{gr} \pi'$, for the state space ordering $s_2 \succeq_s s_1$, $s_3 \succeq_s s_1$, $s_4 \succeq_s s_1$, $s_4 \succeq_s s_2$, $s_4 \succeq_s s_3$, and $s_2 \parallel_s s_3$. The belief point π is shown in **(a)**. Half-spaces corresponding to comparable states are **(b)** $\pi_2\pi_1' \geq \pi_1\pi_2'$ arising from $s_2 \succeq_s s_1$ with intersect $e_{12}(\pi) = \left(\frac{\pi_1}{\pi_1+\pi_2}, 1-\frac{\pi_1}{\pi_1+\pi_2}, 0, 0\right)$, **(c)** $\pi_3\pi_1' \geq \pi_1\pi_3'$ arising from $s_3 \succeq_s s_1$ with $e_{13}(\pi) = \left(\frac{\pi_1}{\pi_1+\pi_3}, 0, 1-\frac{\pi_1}{\pi_1+\pi_3}, 0\right)$, **(d)** $\pi_4\pi_1' \geq \pi_1\pi_4'$ from $s_4 \succeq_s s_1$ with $e_{14}(\pi) = \left(\frac{\pi_1}{\pi_1+\pi_4}, 0, 0, 1-\frac{\pi_1}{\pi_1+\pi_4}\right)$, **(e)** $\pi_4\pi_2' \geq \pi_2\pi_4'$ from $s_4 \succeq_s s_2$ with $e_{24}(\pi) = \left(0, \frac{\pi_2}{\pi_2+\pi_4}, 0, 1-\frac{\pi_2}{\pi_2+\pi_4}\right)$, **(f)**$\pi_4\pi_3' \geq \pi_3\pi_4'$ from $s_4 \succeq_s s_3$ with $e_{34}(\pi) = \left(0, 0, \frac{\pi_3}{\pi_3+\pi_4}, 1-\frac{\pi_3}{\pi_3+\pi_4}\right)$. The hyperplane corresponding to the incomparable states is **(g)** $\pi_3\pi_2' = \pi_2\pi_3'$ from $s_2 \parallel_s s_3$ with $e_{23}(\pi) = \left(0, \frac{\pi_2}{\pi_2+\pi_3}, 1-\frac{\pi_2}{\pi_2+\pi_3}, 0\right)$. The resulting set of comparable beliefs $\pi \succeq_{gr} \pi'$ is given by the plane in **(h)**.

matrix satisfies this condition:

$$R = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}$$

The given state poset $(\mathcal{S}, \succeq_s)$ induces conditions on the comparability of beliefs as per Definition 4. That is, for a given belief π, the conditions define the space of beliefs π' such that $\pi \succeq_{gr} \pi'$. In general, for a given π, each pair of comparable states induces a half-space in the probability simplex, as expressed by (5), whereas each pair of incomparable states induces a hyperplane, as per (6). For our example, the set of comparable beliefs π' to a given π, with respect to the GMLR order, is illustrated by Fig. 1.

The construction in Fig. 1 yields a set of beliefs that forms a plane in the probability simplex. It is straightforward to see that a similar procedure yields the beliefs π' such that $\pi' \succeq_{gr} \pi$; visually this set is the extension of the plane in Fig. 1(h) that, instead of joining π, e_1, and the point on the facet opposite e_4, extends from π joining e_4 and a point on the facet opposite of e_1. By Theorem 1, one can restrict attention to policies of the form $g^* \in \mathcal{G}^* = \{g = (g_0, g_1, \ldots, g_{T-1}) \in \mathcal{G} \mid \pi \succeq_{gr} \pi' \Rightarrow g_t(\pi) \geq g_t(\pi') \text{ for } t = 0, 1, \ldots, T-1\}$, thus knowledge of an optimal action at any belief point on this plane is informative for knowing the optimal actions at other belief points on the plane.

Note that the shape of the region of beliefs that are comparable to a given π depends on the partial order structure. In the context of the above example, considering a total order instead of a partial order (by assuming $s_3 \succeq_s s_2$ in place of $s_2 \parallel_s s_3$) would result in a solid region of comparable beliefs passing through π, rather than a plane. On the other hand, further relaxing the ordering to the case where $s_3 \parallel_s s_4$ instead of $s_4 \succeq_s s_3$ would introduce an additional hyperplane constraint in the belief simplex, further constraining the set of orderable beliefs.

7. Discussion & conclusion

We have derived conditions to ensure monotone optimal policies in the case where the core state space is partially ordered, providing an extension to the conditions for the totally ordered case of Lovejoy (1987).[8] While an intuitive property, establishing the optimality of monotone policies is non-trivial, primarily due to the requirement to select an appropriate partial order for beliefs. To this end, we have introduced a new partial order, termed the GMLR order, that is appropriate for comparing beliefs when not all core states are comparable. Properties of the proposed order, as well as an associated class of order preserving matrices (termed GTP_2), have also been introduced.

The loss of structure resulting from relaxing the total ordering of the state space to a partial ordering results in stronger comparability conditions on the belief space, namely the introduction of conditions arising from incomparable pairs of states (see Definition 4). The need for additional conditions in the partially ordered case reflects a *conservation of structure* in the problem; the loss of structure with respect to the core state space must be compensated by additional structure with respect to the comparability of beliefs. Our results provide formal evidence of Lovejoy's intuition in the closing remark of Lovejoy (1987): "One can anticipate a natural trade-off between the strength of the assumptions one is willing to make regarding the state and information processes and the strength of the partial order invoked." While the generalization introduces somewhat strong conditions, the results provide insight into the challenges associated with ensuring monotone

[8] The proposed conditions reduce to those of Lovejoy when the core state and observation spaces are totally ordered.

optimal policies when one cannot assume a total ordering of the core states. Additionally, we have demonstrated that there exist POMDP instances where the proposed order is meaningful (see the example of Section 6).

Our monotonicity result has potentially useful implications for the design of efficient policy search algorithms. From the example of Section 6, one can see that for a given belief there is a plane of comparable beliefs through the probability simplex such that the optimal policy takes a threshold form on the plane. In general, by computing the optimal action at a collection of beliefs in the probability simplex, the application of Theorem 1 would allow one to infer the optimal action at all comparable beliefs in the simplex. Appropriate choice of a representative collection of beliefs would be informative for knowing optimal actions in regions of the simplex, allowing for heuristic search algorithms to perform more guided search of the policy space.

Acknowledgments

The authors are grateful to the funding provided by the National Science Foundation Grant number CNS-1238962) and the Army Research Office Multidisciplinary University Research Initiatives program (Grant number W911NF-13-1-0421).

References

Albright, S. C. (1979). Structural results for partially observable Markov decision processes. *Operations Research, 27*(5), 1041–1053.
Alchian, A. A. (1952). Economic replacement policy. *R-224*.
Åström, K. J. (1965). Optimal control of Markov processes with incomplete state information. *Journal of Mathematical Analysis and Applications, 10*(1), 174–205.
Bellman, R. (1955). Equipment replacement policy. *Journal of the Society for Industrial and Applied Mathematics, 3*(3), 133–136.
Ben-Zvi, T., Chernonog, T., & Avinadav, T. (2016). A two-state partially observable Markov decision process with three actions. *European Journal of Operational Research, 254*(3), 957–967.
Byon, E., Ntaimo, L., & Ding, Y. (2010). Optimal maintenance strategies for wind turbine systems under stochastic weather conditions. *IEEE Transactions on Reliability, 59*(2), 393–404.
Derman, C. (1963). On optimal replacement rules when changes of state are Markovian. In R. Bellman (Ed.), *Mathematical optimization techniques: 396* (pp. 201–210). University of California Press, Berkeley and Los Angeles, CA.
Derman, C., & Sacks, J. (1960). Replacement of periodically inspected equipment (An optimal optional stopping rule). *Naval Research Logistics, 7*(4), 597–607.
Fernández-Gaucherand, E., Arapostathis, A., & Marcus, S. I. (1991). On the average cost optimality equation and the structure of optimal policies for partially observable Markov decision processes. *Annals of Operations Research, 29*(1), 439–469.
Girshick, M. A., & Rubin, H. (1952). A Bayes approach to a quality control model. *The Annals of Mathematical Statistics, 23*(1), 114–125.
Grosfeld-Nir, A. (2007). Control limits for two-state partially observable Markov decision processes. *European Journal of Operational Research, 182*(1), 300–304.
Hadfield-Menell, D., Russell, S. J., Abbeel, P., & Dragan, A. (2016). Cooperative inverse reinforcement learning. In *Proceedings of the advances in neural information processing systems* (pp. 3909–3917).
Kamae, T., Krengel, U., & O'Brien, G. L. (1977). Stochastic inequalities on partially ordered spaces. *The Annals of Probability, 5*(6), 899–912.
Karlin, S. (1968). *Total positivity*: 1. Stanford, CA: Stanford University Press.
Karlin, S., & Rinott, Y. (1980). Classes of orderings of measures and related correlation inequalities, I. Multivariate totally positive distributions. *Journal of Multivariate Analysis, 10*(4), 467–498.
Krishnamurthy, V., & Djonin, D. V. (2009). Optimal threshold policies for multivariate POMDPs in radar resource management. *IEEE Transactions on Signal Processing, 57*(10), 3954–3969.
Kumar, P. R., & Varaiya, P. (1986). *Stochastic systems: estimation, identification, and adaptive control*. Englewood Cliffs, NJ: Prentice Hall.
Lovejoy, W. S. (1987). Some monotonicity results for partially observed Markov decision processes. *Operations Research, 35*(5), 736–743.
Maillart, L. M. (2006). Maintenance policies for systems with condition monitoring and obvious failures. *IEEE Transactions, 38*(6), 463–475.
Miehling, E., Rasouli, M., & Teneketzis, D. (2018). A POMDP approach to the dynamic defense of large-scale cyber networks. *IEEE Transactions on Information Forensics and Security, 13*(10), 2490–2505.
Porteus, E. L. (1975). On the optimality of structured policies in countable stage decision processes. *Management Science, 22*(2), 148–157.
Rieder, U. (1991). Structural results for partially observed control models. *Zeitschrift für Operations Research, 35*(6), 473–490.
Rosenfield, D. (1976a). Markovian deterioration with uncertain information. *Operations Research, 24*(1), 141–155.
Rosenfield, D. (1976b). Markovian deterioration with uncertain information – a more general model. *Naval Research Logistics, 23*(3), 389–405.
Ross, S. M. (1971). Quality control under Markovian deterioration. *Management Science, 17*(9), 587–596.
Saghafian, S. (2018). Ambiguous partially observable Markov decision processes: Structural results and applications. *Journal of Economic Theory, 178*, 1–35.
Sandıkçı, B., Maillart, L. M., Schaefer, A. J., & Roberts, M. S. (2013). Alleviating the patient's price of privacy through a partially observable waiting list. *Management Science, 59*(8), 1836–1854.
Stoyan, D. (1983). *Comparison methods for queues and other stochastic models*. New York, NY: John Wiley & Sons.
Topkis, D. M. (1978). Minimizing a submodular function on a lattice. *Operations Research, 26*(2), 305–321.
White, C. C. (1979). Optimal control-limit strategies for a partially observed replacement problem. *International Journal of Systems Science, 10*(3), 321–332.
White, C. C. (1980). Monotone control laws for noisy, countable-state Markov chains. *European Journal of Operational Research, 5*(2), 124–132.
Whitt, W. (1979). A note on the influence of the sample on the posterior distribution. *Journal of the American Statistical Association, 74*(366a), 424–426.
Young, S., Gašić, M., Thomson, B., & Williams, J. D. (2013). POMDP-based statistical spoken dialog systems: A review. *Proceedings of the IEEE, 101*(5), 1160–1179.
Zhuang, W., & Li, M. Z. (2012). Monotone optimal control for a class of Markov decision processes. *European Journal of Operational Research, 217*(2), 342–350.