Information Relaxation Bounds for Partially Observed Markov Decision Processes

Martin B. Haugh
Department of IE&OR
Columbia University
mh2078@columbia.edu

Octavio Ruiz Lacedelli
Department of IE&OR
Columbia University
or2200@columbia.edu

First Draft: June 18, 2017

Abstract

Partially observed Markov decision processes (POMDPs) are an important class of control problems that are ubiquitous in fields ranging from electrical engineering to machine learning and economics. Unfortunately these problems are generally intractable and so we must be satisfied with (hopefully) good sub-optimal policies. But how do we evaluate the quality of these sub-optimal policies? This question has been addressed in recent years in the Markov decision process (MDP) literature through the use of information relaxation based duality where the nonanticipativity constraints are relaxed but a penalty is imposed for violations of these constraints. In this paper we extend the information relaxation approach to POMDPs. It is of course well known that a POMDP can be formulated as an MDP and so the previously developed theory for MDPs also applies to POMDPs. But constructing dual bounds in practice is challenging in most POMDP settings and the nature of the challenge depends on how the POMDP is formulated. If we work with explicit dynamics for the hidden state transitions and observations then the evaluation of so-called dual feasible penalties requires the evaluation of expectations that in general are not available explicitly and are strongly action-dependent. If instead we work with the belief-state formulation of the POMDP then there is no difficulty in evaluating these expectations but the inner problems that arise in the information relaxation framework become intractable. We circumvent both problems using a recently developed change-of-measure approach and show that it works well in applications from robotic navigation and personalized medicine. Our approach can easily be extended to other non-Markovian settings including, for example, dynamic zero-sum games where the players have asymmetric information and influence diagrams from the decision science literature.

Keywords: Partially observed Markov decision process, information relaxation, duality.
1. Introduction

Partially observed Markov decision processes (POMDPs) are an important class of control problems with wide-ranging applications in fields as diverse as electrical engineering, economics, machine learning and economics. The resulting problems are often very difficult to solve, however, due to the so-called curse of dimensionality. In general then, these problems are intractable and so we must make do with constructing sub-optimal policies that are (hopefully) close to optimal. But how can we evaluate a given sub-optimal policy? We can of course simulate it many times and obtain a primal bound, i.e. a lower (upper) bound in the case of a maximization (minimization) problem, on the true optimal value function. But absent a dual bound, i.e. an upper (lower) bound, there is no easy way in general to conclude that the policy is close to optimal.

In the case of Markov decision processes (MDPs), we can construct such dual bounds using the information relaxation approach that was developed independently by Brown, Smith and Sun [10] (hereafter BSS) and Rogers [34]. The information relaxation approach proceeds in two steps: (i) relax the nonanticipativity constraints that any feasible policy must satisfy and (ii) include a penalty that punishes violations of these constraints. In a finite horizon setting BSS showed how to construct a general class of dual feasible penalties and proved versions of weak and strong duality. In particular, they showed that if the dual feasible penalties were constructed using the optimal value function, then the resulting dual bound would be tight, i.e. it would equal the optimal value function. In practice of course, the optimal value function is unknown but the strong duality result suggests that a penalty constructed from a good approximate value function should lead to a good dual bound. If a good primal bound is also available, e.g. possibly by simulating the policy that is greedy with respect to the approximate value function, then the primal and dual bounds will be close and therefore yield a “certificate” of near-optimality for the policy.

The main goal of this work is to extend the information relaxation approach to POMDPs. It is well known of course that POMDPs can be formulated as MDPs and so the theoretical results established for MDPs go through. Nonetheless, there are important obstacles to implementing the approach in practice for POMDPs and the nature of these obstacles depends on how the POMDP is formulated. If we work with explicit dynamics for the hidden state transitions and observations (as in (1) and (2) or (3) and (4)) then the evaluation of so-called dual feasible penalties requires the evaluation of expectations that in general are not available explicitly and are strongly action-dependent. Indeed we need to be able to calculate these expectations efficiently for all possible action histories at each time point on each of the simulated inner paths (see (8)). The main contribution of this paper is to overcome this obstacle using a change-of-measure argument that limits dramatically the number of expectations that must be computed. The expectations that are required can then be computed using standard filtering techniques and so we can proceed to compute the corresponding dual bounds in the usual manner. Our change-of-measure approach builds on the recent work of Brown and Haugh [8] (hereafter BH) who use change-of-measure arguments to ensure the so-called inner problems are tractable in particularly challenging applications when these inner problems would otherwise be intractable. Indeed this is exactly the situation that arises in this paper when we work with the belief-state formulation of the POMDP as in (18). Regardless then of the formulation of the POMDP that we choose to work with, we can use change-of-measure arguments to ensure that dual bounds can be computed efficiently.

We show that our approach works well in applications from robotic navigation and personalized medicine. Each of these applications are maximization problems where we construct dual penalties from approximate
value functions that are known upper bounds on the optimal value function. In the language of BH we can
easily show that these upper bounds are super-solutions and therefore guarantee (using a result from BH)
that the resulting dual upper bound improves upon the original upper bound. Our numerical results are
consistent with these bound improvement results.

A further contribution of this work is the implication that the information relaxation approach can be
extended to other non-Markovian settings beyond POMDPs. The basic underlying probability structure of
a POMDP is a (controlled) hidden Markov model (HMM) where the filtered probability distributions that
we need can be computed efficiently. It should be clear from this work that other structures, specifically
(controlled) hidden singly-connected graphical models, would also be amenable to the information relaxation
approach since filtered probability distributions for these models can also be computed very quickly. More
generally, it should be possible to tackle control problems where the (controlled) hidden states form a
multiply-connected graphical model as long as the model is modestly sized so that the junction-tree algorithm
can be employed to compute the various probability distributions. A typical example of such a setting is the
influence diagram framework from the decision sciences literature.

1.1. Literature Review and Paper Outline
The work of BSS and [34] follows earlier work by [18] and [33] on the pricing of high-dimensional American
options. Other related work on American option pricing includes [13] and [2]. The pricing of swing options
with multiple exercise opportunities is an important problem in energy markets and the information relax-
atation approach was soon extended to this problem via the work of [29], [36], [1], [6] and [12] among others.
BSS were the the first to extend the information relaxation approach to general MDPs and demonstrate
the tractability of the approach on large-scale problems. Other notable developments include work by [11]
and [9] on the structure of dual feasible penalties, extensions by BH and [41] to infinite horizon settings,
the bound improvement guarantees of BH who also use change-of-measure arguments (building in part on
Rogers [34]) to solve intractable inner problems. The approach has also been extended to continuous-time
stochastic control by [40], and dynamic zero sum-games by [20] and [7]. Recently [5] and [4] have shown how
information relaxations can be used to construct analytical bounds on the suboptimality of heuristic policies
for problems including the stochastic knapsack and scheduling.

The information relaxation methodology has now become well established in the operations research
and quantitative finance community with applications in revenue management, inventory control, portfolio
optimization, multi-class queuing control and finance. Other interesting applications and developments
include [26], [14], [24], [17], [19], [16] and [42].

Finally, we note that POMDPs are a well-established and important class of problems and doing justice
to the enormous literature on POMDPs is beyond the scope of this paper. Instead we refer the interested
reader to the recent text [25] for a detailed introduction to the topic as well as an extensive list of references.

The remainder of this paper is organized as follows. In Section 2 we formulate the POMDP problem and
in Section 3 we describe how the information relaxation methodology can be used to construct dual bounds
for the problem. We also identify the key practical challenge that arise in the POMDP setting and in Section
4 we develop a change-of-measure approach for resolving this challenge. In Section 5 we discuss the belief
state formulation of the POMDP and we will see that the change-of-measure methodology is again required
to compute dual bounds for this formulation. We provide an application to robotic navigation in Section 6
and an application to personalized medicine and mammography screening in Section 7. We briefly discuss
how our approach can be extended to infinite horizon problems in Section 8 and then conclude in Section 9. Derivations, proofs and various technical details are relegated to the appendices.

2. Discrete-Time POMDPs

We begin with a standard POMDP formulation where we explicitly model the hidden state transitions and observations. We consider a discrete-time setting with a finite horizon $T$ and time indexed by $t \in \{0, 1, \ldots, T\}$. At each time $t$ there is a hidden state, $h_t \in \mathcal{H}$, as well as a noisy observation, $o_t \in \mathcal{O}$, of $h_t$. After observing $o_t$ at time $t > 0$, the decision maker (DM) chooses an action $a_t \in \mathcal{A}$. We also assume a known prior distribution, $\pi_0$, on the initial hidden state, $h_0$, and the initial action $a_0$ is based on $\pi_0$. For ease of exposition we assume that $\mathcal{H}, \mathcal{O}$ and $\mathcal{A}$ are all finite. It is standard to describe the dynamics\(^1\) for $t = 1, \ldots, T$ via the following:

- A $|\mathcal{H}| \times |\mathcal{H}|$ matrix, $P(a)$, of transition probabilities for each action $a \in \mathcal{A}$ with

$$P_{ij}(a) := \mathbb{P}(h_t = j \mid h_{t-1} = i, a_{t-1} = a), \quad i, j \in \mathcal{H}. \quad (1)$$

- A $|\mathcal{H}| \times |\mathcal{O}|$ matrix, $B(a)$, of observation probabilities for each action $a \in \mathcal{A}$ with

$$B_{ij}(a) := \mathbb{P}(o_t = j \mid h_t = i, a_{t-1} = a), \quad i \in \mathcal{H}, j \in \mathcal{O}. \quad (2)$$

Our POMDP formulation is therefore time-homogeneous but there is no difficulty extending our results to the time-inhomogeneous setting where $P$ and $B$ may also depend on $t$. Rather than using (1) and (2), however, we will find it more convenient to use the following alternative, but equivalent, dynamics. In particular, we assume the hidden state and observation dynamics satisfy

$$h_{t+1} = f_h(h_t, a_t, w_{t+1}), \quad (3)$$
$$o_{t+1} = f_o(h_{t+1}, a_t, v_{t+1}) \quad (4)$$

for $t = 0, 1, \ldots, T - 1$ and where the $v_t$’s and $w_t$’s are IID U$(0, 1)$ random variables for $t = 1, \ldots, T$. We can interpret the $v_t$’s and $w_t$’s as being the IID uniform random variables that are required by the inverse transform approach to generate the state transitions and observations of (1) and (2), respectively. At each time $t$, we assume the DM obtains a reward, $r_t(h_t, a_t)$, which is a function of the hidden state, $h_t$, and the action, $a_t$. As rewards depend directly on hidden states, but not the observations, the DM does not have perfect knowledge of the rewards obtained.

A policy $\mu = (\mu_0, \mu_1, \ldots, \mu_T)$ is non-anticipative if it only depends on past and current observations (as well as on the initial distribution, $\pi_0$, over $h_0$). For such a policy we can therefore write the time $t$ action $a_t$ as $a_t = \mu_t(o_{1:t})$ where $o_{1:t} := (o_1, \ldots, o_t)$ and where we have omitted the implicit dependence on $\pi_0$. We define a filtration $\mathcal{F} = (\mathcal{F}_0, \ldots, \mathcal{F}_T)$ to be the filtration generated by the observations so that $\mathcal{F}_t$ is the $\sigma$-algebra generated by $o_{1:t}$. A non-anticipative policy is therefore $\mathcal{F}$-adapted. We also define $\mathcal{F} := \mathcal{F}_T$. We denote the class of all non-anticipative policies by $\mathcal{U}_\mathcal{F}$. The objective of the DM is to find an $\mathcal{F}$-adapted policy, $\mu^*$, that

\(^1\)It may be the case that an initial observation, $o_0$, is also available and this presents no difficulty as long as its distribution conditional on $h_0$ is known.
maximizes the expected total reward. The POMDP problem is therefore to solve for

$$V_0^*(\pi_0) = \max_{\mu \in \mathcal{U}_f} \mathbb{E}\left\{ \sum_{t=0}^{T} r_t(h_t, \mu_t) \mid \mathcal{F}_0 \right\}. \quad (5)$$

### 3. Information Relaxations

Solving (5) is generally an intractable problem so the best we can hope for is to construct a good sub-optimal policy. In order to evaluate the quality of such a policy, however, we need to know how far its value is from the (unknown) optimal value function, $V_0^*(\pi_0)$. If we could somehow bound $V_0^*(\pi_0)$ with a lower bound, $V_0^{lower}$, and an upper bound, $V_0^{upper}$, satisfying $V_0^{lower} \leq V_0^*(\pi_0) \leq V_0^{upper}$ with $V_0^{lower} \neq V_0^{upper}$ then we can answer this question by simulating the policy in question and comparing its value to $V_0^{upper}$. In practice, we take $V_0^{lower}$ to be the value of our best $F$-adapted policy which can typically be estimated to any required accuracy via Monte-Carlo. The goal then is to construct $V_0^{upper}$ and if it is sufficiently close to $V_0^{lower}$ then we have a "certificate" of near-optimality for the policy in question.

Towards this end we will use the concept of information relaxations to do this and our development will follow that of BSS which can be consulted for additional details and proofs. An information relaxation $\mathcal{G}$ of the filtration $\mathcal{F}$ is a filtration $\mathcal{G} = (\mathcal{G}_0, \mathcal{G}_1, \ldots, \mathcal{G}_T)$, where $\mathcal{F}_t \subseteq \mathcal{G}_t$ for each $t$. We denote by $\mathcal{U}_G$ the set of $\mathcal{G}$-adapted policies. Then, $\mathcal{U}_G \subseteq \mathcal{U}_F$. Note that a $\mathcal{G}$-adapted policy is generally not feasible for the original primal problem in (5) as such a policy can take advantage of information that is not available to an $\mathcal{F}$-adapted policy.

Before proceeding we also need the concept of dual penalties. Penalties, like rewards, depend on states and actions and are incurred in each period. Specifically, for each $t$, we define a dual penalty, $c_t$, according to

$$c_t := \mathbb{E}[\vartheta_{t+1}(o_{1:t+1}) \mid \mathcal{F}_t] - \mathbb{E}[\vartheta_{t+1}(o_{1:t+1}) \mid \mathcal{G}_t] \quad (6)$$

where $\vartheta_{t+1}(o_{1:t+1})$ is a bounded real-valued function of the time $t + 1$ state $o_{1:t+1}$. It is straightforward to see that $\mathbb{E}[c_t \mid \mathcal{F}_t] = 0$ for all $t$ and any $\mathcal{F}$-adapted policy. (In general this is not the case for a $\mathcal{G}$-adapted policy.) This in turn implies $\mathbb{E}[\sum_{t=0}^{T} c_t \mid \mathcal{F}_0] = 0$ for any $\mathcal{F}$-adapted policy. Beginning with (5) we now obtain

$$V_0^*(\pi_0) = \max_{\mu \in \mathcal{G}_f} \mathbb{E}\left\{ \sum_{t=0}^{T} r_t(h_t, \mu_t) \mid \mathcal{F}_0 \right\} \\ = \max_{\mu \in \mathcal{U}_G} \mathbb{E}\left\{ \sum_{t=0}^{T} r_t(h_t, \mu_t) + c_t \mid \mathcal{F}_0 \right\} \\ \leq \max_{\mu \in \mathcal{U}_G} \mathbb{E}\left\{ \sum_{t=0}^{T} r_t(h_t, \mu_t) + c_t \mid \mathcal{F}_0 \right\}. \quad (7)$$

We can use (7) to construct upper bounds on $V_0^*(\pi_0)$ for general information relaxations $\mathcal{G}$ but it is perhaps easier to understand how to do this when we use the perfect information relaxation, which is the most common choice in applications.

#### 3.1. The Perfect Information Relaxation

The perfect information (PI) relaxation corresponds to the filtration $\mathcal{I} = (\mathcal{I}_0, \mathcal{I}_1, \ldots, \mathcal{I}_T)$, with $\mathcal{I}_t = \mathcal{F}$ for all $t$. In particular, the DM gets to observe all of the $w_t$’s, $v_t$’s and $h_0$ at time 0 under $\mathcal{I}$. It is worth noting that

\footnote{In practice we will take $\vartheta_{t+1}(o_{1:t+1})$ to be an approximation to the time $t + 1$ optimal value function. We note that dual feasible penalties are essentially action-dependent control variates, a standard variance reduction technique in the simulation literature. Recall also that $o_{1:t+1}$ is a function of the actions $a_{0:t}$ as well as exogenous noise as described in (3) and (4).}
knowledge of the \(w_i\)'s, \(v_i\)'s and \(h_0\) implies knowledge of the observations \(o_{1:T}\) corresponding to all possible action sequences. It therefore follows that \(\mathcal{F}_t \subseteq \mathcal{I}_t\) for all \(t\) so that \(\mathcal{I}\) is indeed a relaxation of \(\mathcal{F}\). Under the PI relaxation, we can move the max operator of (7) inside the expectation to obtain

\[
V^*_0(\pi_0) \leq \mathbb{E} \left[ \max_{a_t, \pi_{t+1}} \sum_{t=0}^{T} r_t(h_t, a_t) + c_t \mid \mathcal{F}_0 \right]. \tag{8}
\]

In principle we can evaluate the right-hand-side of (8) by simulating \(J\) sample paths, \((h_0^{(j)}, w_1^{(j)}, v_1^{(j)})\), for \(j = 1, \ldots, J\), and then solving the deterministic maximization problem inside the expectation in (8) (sometimes known as the inner problem) for each such path. If we let \(V^{(j)}\) denote the optimal value of the \(j^{th}\) inner problem, then \(\sum_j V^{(j)}/J\) provides an unbiased estimator of an upper bound, \(V^*_0^{\text{upper}}\), on the optimal value function, \(V^*_0(\pi_0)\). Moreover standard methods can be used to construct approximate confidence intervals for \(V^*_0^{\text{upper}}\).

BSS also showed that strong duality holds. Specifically, if we could take \(\vartheta_{t+1}(o_{1:t+1}) = V_t^*(o_{1:t+1})\), i.e. use the (unknown) optimal value function as our generating function in (6), then we would have equality in (7) and (8). Indeed a simple inductive proof that works backwards from time \(T\) establishes strong duality and also shows that equality holds in (8) almost surely. That is, if we could use the optimal value function to construct the dual penalties then the optimal value of the inner problem (inside the expectation in (8)) would equal \(V^*_0(\pi_0)\) almost surely. This result has two implications when we have a good approximation, \(\hat{V}_t\), to \(V^*_t\) and we take \(\vartheta_{t+1}(o_{1:t+1}) = \hat{V}_{t+1}(o_{1:t+1})\). First it suggests that (8) should yield a good upper bound on \(V^*_0\) and second, the almost sure property of the preceding paragraph suggests that relatively few sample paths should be needed to estimate \(V^*_0^{\text{upper}}\) to any given accuracy.

It is perhaps worth emphasizing that strong duality follows directly from the results of BSS. This follows because of the well known result that a POMDP can be formulated as an MDP. Under this MDP formulation the state of the system is the belief state, \(\pi_t\), a probability distribution on the \(|H|\)-dimensional simplex with state transitions calculated via standard filtering methods. While the time \(t\) optimal value function is then a function of \(\pi_t\), there is no inconsistency when we write \(V_t^*(o_{1:t})\) above since \(\pi_t\) is in fact a function of \(o_{1:t}\) (and of course the chosen actions \(a_{0:t-1}\)). This observation leads naturally to an alternative information relaxation, namely the belief state perfect information (BSPI) relaxation which arises naturally from the belief state formulation of the POMDP. We will introduce the belief state formulation and the corresponding BSPI relaxation in Section 5.

### 3.2. Solving the Inner Problem in (8)

We would therefore like to use the PI relaxation to construct an upper bound on \(V^*_0\) by solving the inner problem in (8) as a deterministic dynamic program. The main obstacle we will encounter under the PI relaxation, however, is computing the \(c_t\)'s which now take the form

\[
c_t := \mathbb{E}[\vartheta_{t+1}(o_{1:t+1}) \mid \mathcal{F}_t] - \vartheta_{t+1}(o_{1:t+1}). \tag{9}
\]

More specifically, the \(\mathbb{E}[\vartheta_{t+1}(o_{1:t+1}) \mid \mathcal{F}_t]\) term in (9) is challenging to handle. In typical MDP applications of the information relaxation approach, we can either compute these expectations analytically or numerically but still very efficiently. In the POMDP setting, however, this is not possible because the probability distribution required to compute \(\mathbb{E}[\vartheta_{t+1}(o_{1:t+1}) \mid \mathcal{F}_t]\) depends on the entire history of actions, \(a_{0:t}\), up to
time $t$. Moreover, this probability distribution is not available explicitly and must be calculated via a filtering algorithm. This means that in solving the inner problem in (8) as a deterministic dynamic program, we would need to compute $\mathbb{E}[\varrho_{t+1}(o_{1:t+1}) | \mathcal{F}_t]$ at each time $t$ via a filtering algorithm for all possible action histories, $a_{o:t}$. This is clearly impractical for any realistic application.

Building on the work of BH, we will develop a change-of-measure approach in Section 4 for resolving this problem. Before developing this change-of-measure approach, however, it is worth noting there is one exception where evaluating the $\mathbb{E}[\varrho_{t+1}(o_{1:t+1}) | \mathcal{F}_t]$ terms might be analytically tractable. This is the classic linear-quadratic setting with Gaussian noise. In the uncontrolled case the Kalman filter yields the mean and covariance matrix of the required normal distribution. Depending on how the controls influence the hidden state dynamics, e.g. linearly on the drift, then the Kalman filter should still apply and allow for the analytic evaluation of penalties when, for example, the $\varrho_{t+1}$'s are quadratic.

4. The Uncontrolled Formulation

Instead of computing $\mathbb{E}[\varrho_{t+1}(o_{1:t+1}) | \mathcal{F}_t]$ at each time $t$ via a filtering algorithm for all possible action histories when solving the inner problems in (8), we use a change-of-measure approach to ensure that $\mathbb{E}[\varrho_{t+1}(o_{1:t+1}) | \mathcal{F}_t]$ only needs to be computed once for each time $t$ inside each inner problem. This results in a dramatic decrease in the computations required to compute the upper bound. This approach is motivated by BH who used change-of-measure arguments to reduce the size of the state space in each inner problem in an MDP setting. By reducing the size of the state space they were able to ensure that the inner problem was tractable and easily solved. In our POMDP setting, the size of the state space in the inner problem with a PI relaxation is $|\mathcal{H}|$ which is typically quite reasonable. Our motivation for the change-of-measure approach under the PI relaxation is therefore not to reduce the size of the state space that needs to be considered in an inner problem but instead to make explicit calculation of the penalties in (9) tractable.

Towards this end, we simply define a hidden Markov model (HMM) on the same hidden state and observation spaces as our POMDP. Specifically, for $t = 1, \ldots, T$ we define:

- A $|\mathcal{H}| \times |\mathcal{H}|$ matrix, $Q$, of transition probabilities, with
  \[
  Q_{ij} := \mathbb{P}(h_t = j \mid h_{t-1} = i), \quad i, j \in \mathcal{H}. \tag{10}
  \]

- A $|\mathcal{H}| \times |\mathcal{O}|$ matrix, $E$, of observation probabilities with
  \[
  E_{ij} := \mathbb{P}(o_t = j \mid h_t = i), \quad i \in \mathcal{H}, j \in \mathcal{O}. \tag{11}
  \]

Note that both $Q$ and $E$ are action independent although in general they could depend on time in which case we would write $Q^t_{ij}$ and $E^t_{ij}$. In general\(^3\) we will also require them to satisfy the following absolute continuity conditions:

(i) $Q_{ij} > 0$ for any $i, j \in \mathcal{H}$ for which there exists an action $a \in \mathcal{A}$ such that $P_{ij}(a) > 0$

(ii) $E_{ij} > 0$ for any $i \in \mathcal{H}$ and $j \in \mathcal{O}$ for which there exists an action $a \in \mathcal{A}$ such that $B_{ij}(a) > 0$.

\(^3\)We will see later in Section 6.2 that we can ignore these absolutely continuity conditions when we use supersolutions to define the $\varrho_t$'s.
A trivial way to ensure these conditions is to have \( Q_{ij} > 0 \) and \( E_{ik} > 0 \) for all \( i, j \in H \) and \( k \in O \). We let \( \tilde{P} \) denote the probability measure induced by \( Q \) and \( E \) with \( \tilde{E} \) denoting expectations under \( \tilde{P} \). We now proceed by reformulating our POMDP under \( \tilde{P} \) and adjusting rewards (and penalties) with appropriate Radon-Nikodym (RN) derivatives. These RN derivatives are of the form

\[
\phi(i, j, k, a) := \frac{P_{ij}(a)}{Q_{ij}} \cdot \frac{B_{jk}(a)}{E_{jk}}. \tag{12}
\]

\[
\Phi_t(h_0, a_{0:t}, a_{0:t-1}) := \prod_{s=0}^{t-1} \phi(h_s, h_{s+1}, o_{s+1}, a_s). \tag{13}
\]

It is then straightforward to see that

\[
V_0^*(\pi_0) = \max_{\mu \in \mathcal{D}_\theta} \mathbb{E} \left[ \sum_{t=0}^{T} r_t(h_t, \mu_t) \mid \mathcal{F}_0 \right] = \max_{\mu \in \mathcal{D}_\theta} \tilde{E} \left[ \sum_{t=0}^{T} \Phi_t r_t(h_t, \mu_t) \mid \mathcal{F}_0 \right]. \tag{14}
\]

Following BH, we refer to (14) as an uncontrolled formulation of the POMDP. The “uncontrolled” terminology reflects the fact that the policy, \( \mu \), does not influence the dynamics of the system which are now determined by the action independent transition and observation distributions in \( Q \) and \( E \), respectively. The impact of the policy instead manifests itself via the \( \Phi_t \)'s. With this uncontrolled formulation the analog of (8), i.e. weak duality for the PI relaxation, is given by

\[
V_0^*(\pi_0) \leq \tilde{E} \left[ \max_{\mu \in \mathcal{D}_\theta} \sum_{t=0}^{T} \Phi_t [r_t(h_t, a_t) + c_t] \mid \mathcal{F}_0 \right]. \tag{15}
\]

with

\[
c_t := \mathbb{E} [\vartheta_{t+1}(o_{1:t+1}) \mid \mathcal{F}_t] - \phi(h_t, h_{t+1}, o_{t+1}, a_t) \vartheta_{t+1}(o_{1:t+1}) \tag{16}
\]

and where we acknowledge\(^4\) a slight abuse of notation in (15) since there is no time \( T \) action \( a_T \). We also note (following BH) that strong duality continues to hold for these uncontrolled formulations. Returning to the penalty in (9) we recall that we need to compute \( \mathbb{E} [\vartheta_{t+1}(o_{1:t+1}) \mid \mathcal{F}_t] \) but note that we no longer need to compute it for all possible action histories, \( a_{0:t} \), when solving an inner problem in (15). This is because the action histories under \( \tilde{P} \) influence neither the dynamics of the hidden states nor the observations. This means we only need to compute \( \mathbb{E} [\vartheta_{t+1}(o_{1:t+1}) \mid \mathcal{F}_t] \) once for each time \( t \) in each inner problem. This is a straightforward calculation and the expectation can be computed as

\[
\mathbb{E} [\vartheta_{t+1}(o_{1:t+1}) \mid \mathcal{F}_t] = \sum_{o \in O, h, h' \in H} \pi_t(h) P_{hh'}(a_t) B_{h'o}(a_t) \vartheta_{t+1}(o_{1:t}, o) \tag{17}
\]

where \( \pi_t(h) := \tilde{P}(h_t = h \mid o_{1:t}) \) can be calculated efficiently using standard HMM filtering methods. As discussed in Section 3.1, we can now calculate an unbiased upper bound on \( V_0^* \) by solving \( J \) instances of the inner problems in (15) and averaging their optimal objective values. Note that an inner problem can be

\(^4\)This abuse is also found elsewhere in this article but we can resolve it by simply assuming the existence of a dummy action at time \( T \) which has no impact on the time \( T \) reward.
solved recursively according to

\[ V_t^i = \max_a \{ r_t(h_t, a) + c_t + \phi(h_t, h_{t+1}, o_{t+1}, a) V_{t+1}^i \} \]

for \( t = 0, \ldots, T - 1 \) and where \( h_{0:T} \) and \( o_{1:T} \) are the hidden states and observations that were generated for that specific inner problem. We also have the terminal condition \( V_T^i = r_T(h_T) \) since \( c_T = 0 \) as each \( \vartheta_{T+1} \) can be assumed to be identically zero. Each of these \( J \) inner problem instances should be independently generated via \( \tilde{\mathbb{P}} \) and they can be solved as deterministic dynamic programs. Strong duality suggests that if \( \vartheta_t \) is a “good” approximation to the optimal value function, \( V_t^* \), then we should obtain tight upper bounds on \( V_0^* \). We will see that this is indeed the case in the robot navigation and personalized medicine applications of Sections 6 and 7, respectively.

5. The Belief State Formulation of the POMDP

Rather than use the hidden state and observation dynamics of (3) and (4), we can instead define the POMDP state dynamics in terms of the belief state process, \( \pi_t \), which lies in the \( |\mathcal{H}| \)-dimensional simplex. Specifically we can equivalently write the POMDP dynamics as

\[ \pi_{t+1} = f_x(\pi_t, a_t, u_{t+1}), \quad t = 0, 1, \ldots, T - 1 \]  

(18)

where the \( u_t \)'s are IID \( U(0,1) \) random variables and \( f_x \) is the state transition function which is only defined implicitly via the filtering\(^5\) algorithm. We now define the filtration \( \mathbb{F} = (\mathcal{F}_t^0, \ldots, \mathcal{F}_T^0) \) where \( \mathcal{F}_t^0 \) is the \( \sigma \)-algebra generated by \( \pi_{0:t} \). We note that the filtrations \( \mathbb{F} \) and \( \mathbb{F}^* \) are not identical though they are of course closely related. We can also write the time \( t \) reward as a function of the belief state by setting\(^6\)

\[ r(\pi_t, a_t) := E[r(h_t, a_t) \mid \mathcal{F}_t^0]. \]

5.1. The Belief State Perfect Information Relaxation

Using this equivalent belief-state formulation of the POMDP, we can now define the BSPI relaxation. Specifically, the BSPI information relaxation is given by the filtration \( \mathbb{B} = (\mathcal{B}_0, \ldots, \mathcal{B}_T) \) where \( \mathcal{B}_0 = \mathcal{B}_1 = \cdots = \mathcal{B}_T := \sigma(u_{1:T}) \) so that the DM gets to observe \( u_{1:T} \) at time 0. Moreover, knowledge of \( u_{1:T} \) implies knowledge of the belief states \( \pi_{0:T} \) corresponding to all possible action sequences, which implies that \( \mathcal{F}_t^0 \subseteq \mathcal{B}_t \) for all \( t \) so that \( \mathbb{B} \) is indeed a relaxation of \( \mathbb{F}^* \). The same argument that led to (8) now yields

\[ V_0^*(\pi_0) \leq E \left[ \max_{\vartheta_{0:T-1}} \sum_{t=0}^T r(\pi_t, a_t) + c_t \mid \mathcal{F}_0^0 \right] \]

(19)

where \( c_t \) now takes the form

\[ c_t := E[\vartheta_{t+1}(\pi_{t+1}) \mid \mathcal{F}_t^0] - \vartheta_{t+1}(\pi_{t+1}). \]

(20)

As was the case with (8), we could in principle evaluate the right-hand-side of (19) by simulating \( J \) sample paths, \( (u_{1:t}^{(j)}) \), for \( j = 1, \ldots, J \), solving the deterministic maximization problem inside the expectation in (19) for each path and then averaging the optimal objective functions.

\(^5\) The filtering algorithm takes \( \pi_t \) and \( o_{t+1} \) (which is a function of \( \pi_t, a_t \) and \( u_{t+1} \)) as inputs and outputs \( \pi_{t+1} \).

\(^6\) Indeed, when simulating a policy to compute a primal bound using the original POMDP formulation of Section 2, we can use \( r_t(\pi_t, a_t) \) instead of \( r_t(h_t, a_t) \) to compute the rewards. Using \( r_t(\pi_t, a_t) \) instead of \( r_t(h_t, a_t) \) to estimate a primal bound amounts to performing a conditional Monte-Carlo which is a standard variance reduction technique.
Note the only difference between the penalties in (9) and (20) is that the expectation in (9) is taken conditional on $F_t$ whereas the expectation in (20) is taken conditional on $F^*_t$. But this difference is significant. Indeed the expectation $E[\vartheta_{t+1}(\pi_{t+1}) \mid F^*_t]$ in (20) only depends on $\pi_t$ and the action $a_t$ rather than the sequence of actions, $a_{0:t}$, as was the case with the expectation in (9). We therefore only need to compute the expectation in (20) for all feasible time $t$ actions and this is easy to do in most POMDP settings with moderately sized action spaces. The difficulty we encountered with calculating the penalties in the PI setting therefore does not arise in the BSPI setting.

However, in the BSPI setting a new problem arises, namely the dimensionality of the state space which in this case is the $|\mathcal{H}|$-dimensional simplex. In particular, solving the inner problem in (19) amounts to solving a deterministic DP with a $|\mathcal{H}| - 1$-dimensional state space. For all but the smallest problems, these deterministic DPs will in generally be intractable. This of course is the curse of dimensionality. Once again, however, the change-of-measure approach comes to rescue and using an uncontrolled formulation results in a dramatic reduction of the state space that needs to be considered in solving the inner problems in (19). Indeed the desire to (dramatically) reduce the inner problem’s state space was the motivation behind the development of the uncontrolled formulations in BH.

5.2. The Uncontrolled Formulation
The analog of (19), i.e. weak duality under the uncontrolled BSPI relaxation, is given by

$$V^*_0(\pi_0) \leq \tilde{E}\left[ \max_{a_{0:T-1}} \sum_{t=0}^{T} \Phi_t^*[r_t(\pi_t, a_t) + c_t] \mid F_0 \right],$$

and

$$c_t := E[\vartheta_{t+1}(\pi_{t+1}) \mid F^*_t] - \phi(\pi_t, \pi_{t+1}, a_t) \vartheta_{t+1}(\pi_{t+1}).$$

where the $\Phi_t^*$'s are appropriately defined RN derivatives. In this setting more care is required in calculating these RN derivatives for a given uncontrolled formulation. While there are many ways to define an uncontrolled formulation, one convenient approach is to use (10) and (11) but with a suitable adjustment to guarantee the absolute continuity condition is satisfied. We explain in detail how to do this in Appendix A where we also provide an explicit expression for the corresponding $\Phi_t^*$'s and $\phi$'s.

6. An Application to Robotic Navigation
We now consider one of the benchmark applications of POMDP modeling, namely the robot navigation problem, and in our specific application we follow [27, 22, 31]. A robot is placed randomly in one of the 22 white squares (excluding the goal state) inside the maze depicted in Figure 1. The robot must navigate the maze, one space at a time, with the objective of reaching the goal state in 10 movements and only traversal along white squares is possible. The exact position within the maze is not directly known to the robot. Sensors placed on the robot provide noisy information on whether or not a wall (depicted as grey squares and edges of the maze) is present on the neighboring space for each of the four compass directions. After taking these readings, the robot must choose one of five possible actions: (attempt to) move north, east, south or west, or stay in the current position.

The sensors have a noise factor of $\alpha \in [0, 1]$. This factor represents two types of errors: A wall will fail to
be recognized with probability \( \alpha \) when a wall exists, and a wall is incorrectly observed with probability \( \alpha/2 \) when it does not exist. A second source of uncertainty results from the imperfect movements of the robot. Specifically, after a decision to move has been made, the robot will move in the opposite direction with probability 0.001, the +90 degree direction with probability 0.01, the -90 degree direction with probability 0.01 and it will fail to move at all with probability 0.089. The robot therefore succeeds in moving in the desired direction with probability 0.89. These movement probabilities are normalized in the event that a particular direction is not possible due to the presence of a wall. The robot may also choose to stay in its current location and such a decision is successful with probability 1.

We formulate the control problem as a POMDP with horizon \( T = 10 \) periods, 23 hidden states including the goal state \( h_{\text{goal}} \), five actions and 16 possible observations. The hidden state, \( h_t \), at time \( t \) is the current position of the robot and is 1 of the 23 white squares in the maze. The observation at time \( t < T \) is a \( 4 \times 1 \) binary vector of sensor readings indicating whether or not a wall was observed in each compass direction. The possible actions are the direction of desired movement or the decision to stay. Note the observation probabilities are action-independent conditional on the current hidden state. That is, \( B_{ij} \) in (2) (or equivalently \( f_o \) in (4)) does not depend on the current action \( a \) given the current hidden state \( h \).

At time \( t = 0 \) the robot is allowed to take an initial sensor reading \( o_0 \), with the distribution of \( o_0 \) as described above. Prior to this initial observation, the robot has a prior distribution over the initial hidden state \( h_0 \) that is uniform over the 22 non-goal states.

There is a reward function at \( t = T \) which is defined as \( r_T(h_T) = 1 \) if \( h_T = h_{\text{goal}} \), and zero otherwise. All intermediate rewards are zero. Finally, we define \( \alpha_T = h_T \) so that we know for certain whether or not the terminal reward was earned or not at the end of the horizon.

### 6.1. Value Function Approximations

We now discuss several standard approaches for obtaining approximations to the optimal value function. We can use each such approximation, \( \hat{V}_t \), to:

1. Construct a lower bound, \( V_0^{lower} \), on \( V_T^* \), by simulating the policy that is greedy\(^8\) with respect to \( \hat{V}_t \).

\(^8\)Recall that a policy is said to be greedy with respect to \( \hat{V}_t \) if the action, \( a_t \), chosen by the policy at time \( t \) is an action that maximizes the current time \( t \) reward plus the expected discounted value of \( \hat{V}_{t+1} \), i.e. \( a_t = \arg\max_a (r_t(\pi, a) + \mathbb{E}[\hat{V}_{t+1}(o_{t+1})]) \).
(ii) Construct an upper bound, \( V_{\text{upper}} \), via our uncontrolled information relaxation approach by setting 
\[ \vartheta_t = \tilde{V}_t. \]

If our best lower bound is close to our best upper bound then we will have a certificate of near-optimality for the policy that yielded the best lower bound. We now describe four different value function approximations, namely the \( MDP \), \( QMDP \) and \( \text{Fast Informed Lag-1} \) and \( \text{Lag-2} \) approximations. Other approximate solution approaches can be found in [25], for example. Before proceeding further, we note that the optimal value function \( V^*_T(o_0:T) \) is known at time \( T \) and satisfies 
\[ V^*_T(o_0:T) = 1_{\{h_T = h_{\text{goal}}\}} \]  because of our earlier assumption that \( o_T = h_T \). This means that each of our approximate value functions can also be assumed to satisfy 
\[ \tilde{V}_T(o_0:T) = 1_{\{h_T = h_{\text{goal}}\}}. \]

The \textbf{MDP approximation} uses the optimal value function, \( V^{MDP}_t(h) \), from the corresponding fully observable MDP formulation where the hidden state, \( h_t \), is actually observed at each time \( t \). It is generally easy to solve for \( V^{MDP}_t(h) \) in typical POMDP settings and we can use it to construct an approximate value function according to 
\[ \tilde{V}_t(o_0:t) := E[V^{MDP}_t(h_t) | \mathcal{F}_t] = \sum_{h \in \mathcal{H}} \pi_t(h) V^{MDP}_t(h) \]  where 
\[ V^{MDP}_t(h) = \max_{a_t \in \mathcal{A}} E[V^{MDP}_{t+1}(h_{t+1}) | h_t = h] = \max_{a_t \in \mathcal{A}} \sum_{h' \in \mathcal{H}} P_{hh'}(a_t) V^{MDP}_{t+1}(h') \]  for \( t \in \{0, \ldots, T-1\} \) and with \( V^{MDP}_T(h) := 1_{\{h = h_{\text{goal}}\}} \).

To describe the \textbf{QMDP approximation} we first need to define the so-called Q-functions, \( V^{Q}_t(h,a) \), which were introduced in [27] and are defined according to 
\[ V^{Q}_t(h,a) := \sum_{h' \in \mathcal{H}} P_{hh'}(a) V^{MDP}_{t+1}(h') \]  for \( t \in \{0, \ldots, T-1\} \). The QMDP approximate value function is then defined as 
\[ \tilde{V}_t(o_0:t) := \max_{a_t} \sum_{h \in \mathcal{H}} \pi_t(h) V^{Q}_t(h,a_t) \]  Note that by exchanging the order of the expectation and max operators in (26) and then applying Jensen’s inequality, we easily obtain that the QMDP value function is less than or equal to the MDP value function in (23).

The \textbf{Lag-1 approximation} was first proposed in [21] as the \textit{fast informed bound update}. This approximation uses the optimal value function, \( V^{L1}_t(h_{t-1}, a_{t-1}, o_t) \), from the corresponding lag-1 formulation of the POMDP where the hidden state, \( h_{t-1} \), is observed before making decision \( a_t \) at time \( t \) for all \( t < T \). As was the case
with the MDP formulation in (24), we can calculate $V^{L_1}$ recursively via

$$V^{L_1}_t(h_{t-1}, a_{t-1}, o_t) = \max_{a_t} \mathbb{E}[V^{L_1}_{t+1}(h_t, a_t, o_{t+1}) \mid h_{t-1}, \mathcal{F}_t]$$

(27)

for $t \in \{1, \ldots, T-1\}$, and with terminal condition $V^{L_1}_T(h_{T-1}, a_{T-1}, o_T) := 1_{\{h_T = h_{goal}\}}$ (since $o_T = h_T$). The corresponding approximate value function is then defined according to

$$\hat{V}_t(o_{0:t}) := \max_{a_t} \mathbb{E}[V^{L_1}_{t+1}(h_t, a_t, o_{t+1}) \mid \mathcal{F}_t]$$

(28)

where the expectation is taken with respect to $o_{t+1}$ and $h_t$, given the current belief state, $\pi_t$. Further details on calculating $V^{L_1}_t$ can be found in Appendix B.

The **Lag-2 approximation** is derived by first constructing the optimal value function $V^{L_2}_t(h_{t-2}, a_{t-2}, o_{t-1}, a_{t-1}, o_t)$ corresponding to the POMDP where the hidden state, $h_{t-2}$, is observed before taking the decision $a_t$ at time $t$ for all $t < T-1$. The calculation of $V^{L_2}_t$ is more demanding than the calculation of $V^{L_1}_t$ since we have to follow a specific order in the marginalization and maximization of the variables. Further details on these calculations may be found in Appendix B. The corresponding approximate value function is then defined according to

$$\hat{V}_t(o_{0:t}) := \max_{a_t} \mathbb{E}[\max_{a_{t+1}} \mathbb{E}[V^{L_2}_{t+2}(h_t, a_t, o_{t+1}, a_{t+1}, o_{t+2}) \mid \mathcal{F}_{t+1}] \mid \mathcal{F}_t]$$

(29)

for $t \in \{0, \ldots, T-2\}$, with the understanding that when $t = T-1$, the Lag-2 approximation is equal to the Lag-1 approximation, as there is only one remaining time period at that point.

More generally, we could define a **Lag-d approximation** following a similar construction. The nesting of maximizations and expectations required to calculate the approximate value functions, however, implies that the complexity of the calculations scale exponentially in the number of lags, $d$. We note that $d = 0$ corresponds to the MDP approximation while the original POMDP is recovered when we take $d = T$.

In addition to being used in the calculations of penalties, we use the QMDP, Lag-1 and Lag-2 value functions to obtain lower bounds by simulating the policies that are greedy w.r.t each of them. We generate $J$ independent sample paths $(h_0^{(j)}, o_0^{(j)}, w_1^{(j)}, v_1^{(j)})$, for $j = 1, \ldots, J$, where we recall the $w$'s and $v$'s are used for generating the hidden and observation states in equations (3) and (4) in Section 2. For each sample path $j$ and using standard filtering techniques, we calculate at time $t$ the corresponding belief state $\pi_t$ and take the action that obtains the maximum in the chosen approximate value function from each of (26), (28) or (29). If we denote by $V^{(j)}_{lower}$ the reward obtained from following one of these policies on the $j^{th}$ sample path, then an unbiased estimator of a lower bound on the true optimal value function is given by $\sum_j V^{(j)}_{lower} / J$.

### 6.2. Supersolutions

Before proceeding to the dual problem, we note that each of the MDP, QMDP, Lag-1 and Lag-2 approximate value functions is a *supersolution*, a concept that we now define.

**Definition 6.1.** Let $\vartheta_t$ be any approximate value function that satisfies

$$\vartheta_t(o_{0:t}) \geq \max_{a_t \in A} \{r_t(\pi_t, a_t) + \mathbb{E}[\vartheta_{t+1}(o_{0:t+1}) \mid \mathcal{F}_t]\}$$

(30)
for all possible observation paths, \( o_{0:t} \), and all \( t \in \{0, \ldots, T\} \). Then we say that \( \vartheta_t \) is a supersolution.

We recall that the belief state \( \pi_t \) is a function of \( o_{0:t} \) and so there is no notational inconsistency in Definition 6.1. We have the following proposition whose proof is straightforward and can be found in Appendix C.

**Proposition 6.1.** The MDP, QMDP, Lag-1 and Lag-2 approximate value functions are all supersolutions.

It is well-known\(^9\) that a supersolution, \( \vartheta_t \), is an upper bound on the optimal value function, \( V^*_t(o_{0:t}) \). It therefore follows from Proposition 6.1 that the MDP, QMDP, Lag-1 and Lag-2 approximate value functions are all upper bounds on \( V^*_t(o_{0:t}) \). The significance of Proposition 6.1, however, is that a dual upper bound (as given by (15) or (21)) based on a penalty constructed from a supersolution is guaranteed to improve on the original upper bound provided by the supersolution itself. This is Proposition 4.1 in BH and we will see this result in action in our numerical results below.

### 6.3. The Uncontrolled Formulation

To set up the uncontrolled formulation of Section 4, we must define an action-independent\(^10\) transition function \( Q \) as in (10). We considered two approaches:

(i) Use the policy that is greedy w.r.t. the QMDP approximate value function in (25). Specifically, we would like to set the uncontrolled probability of a transition from \( h_i \) to \( h_j \) at time \( t \) according to

\[
Q^t_{ij} \equiv P_{ij} \left( \arg \max_{a \in A} V_t^Q(i,a) \right)
\]

for \( t \in \{0, \ldots, T-1\} \). Recall, however, that the \( Q^t \) transition matrices must also satisfy the absolute continuity condition and this is not true of the \( Q^t \) matrices in (31). In fact it is easy to see that the absolute continuity condition breaks down when the robot is in the goal state, \( h_{goal} \), and a move is attempted. This could occur because the robot has imperfect information and therefore attempts to move resulting in a strictly positive probability of moving to the single neighboring square immediately to the left of \( h_{goal} \); see Figure 1. In contrast \( Q^t \) will put zero probability on such a move because \( \arg \max_{a \in A} V_t^Q(h_{goal},a) = \text{‘stay’ and ‘stay’ decisions are successful w.p.1} \). Clearly then we will have \( P_{h_{goal},j}(a_{move}) > 0 \) for any ‘move’ action, \( a_{move} \), whereas \( Q^t_{h_{goal},j} = 0 \) and where we use \( j \) to denote the neighboring state of \( h_{goal} \). But this issue is easy to resolve: we simply set \( Q^t_{h_{goal},j} \) to some small but strictly positive quantity, \( \epsilon \), and then re-normalize the appropriate probabilities so that they sum to 1.

We also note that such an adjustment is only necessary for \( h_{goal} \) because for all other states \( i \neq h_{goal} \), we will have \( Q^t_{ij} > 0 \) for any feasible neighboring state \( j \).

\(^9\) A proof can be found in standard dynamic programming texts and is based on the linear-programming formulation of the Bellman equation.

\(^10\) The emission matrix, \( B \), for this problem is already action-independent given \( h \) and so we can continue to use this in our uncontrolled formulation.
(ii) The second alternative is based on the observation in BH that absolute continuity of \( P \) (w.r.t. \( \tilde{P} \)) is not required to obtain a valid dual bound when the penalties are constructed from approximate value functions that are supersolutions. This means we could in fact set \( \tilde{P} = Q^t \) as defined in (31) and indeed we did this when estimating the lower and upper bounds associated with the QMDP approximation. The advantage of this is that the same set of paths could be used to estimate both lower and upper bounds. This of course provides a computational saving but it also allows us to directly and very accurately estimate the duality gap between the QMDP lower bound and its corresponding upper bound. A similar approach applies for the Lag-1 and Lag-2 bounds. For example, when estimating the lower and upper bounds corresponding to the Lag-1 approximation, we took \( \tilde{P} \) to be the probability measure induced by following the policy that is greedy w.r.t. the Lag-1 approximate value function.

Considerably more details are provided in Appendix D.

In the numerical results of Section 6.4 below, we followed the second option above as it provided a computational saving over the first approach, provided slightly better\(^{12} \) dual bounds and allowed for the direct calculation of tight approximate confidence intervals for the duality gap with relatively few sample paths required.

The final building block we require are the Radon-Nikodym derivatives, which we calculate using (12) and (13) for the PI relaxation, and (46) and (47) (with \( \epsilon = 0 \)) for the BSPI relaxation. We can then solve the inner problems in (15) and (21) as simple deterministic dynamic programs with terminal value \( V_T(o_0,T) := 1_{(h_T=h_{goal})} \). Because the hidden states and observations on each simulated path are fixed, only one expectation needs to be computed at each time \( t \) to evaluate the penalty in (9). We can then calculate an unbiased upper bound on \( V_0^* \) by averaging the optimal values of each the \( J \) inner problem instances for the PI and BSPI relaxations, respectively. Moreover, since our penalties are constructed from supersolutions we are guaranteed to obtain dual upper bounds that improve on the upper bounds provided by the supersolutions themselves.

### 6.4. Numerical Results

Figures 2 and 3 display numerical results from our experiments. Specifically, Figure 2 displays\(^{13} \) the MDP, QMDP, Lag-1 and Lag-2 approximate value functions at time \( t = 0 \). Since these approximations are supersolutions we know they are also valid upper bounds on the true unknown optimal value function. We also display two dual upper bounds obtained from the uncontrolled PI relaxation (as given by the r.h.s. of (15)), where the penalties were constructed from the Lag-1 and Lag-2 approximate value functions, respectively. All of these bounds are displayed as a function of \( \alpha \) with the time horizon fixed at \( T = 10 \) periods. The best lower bound was obtained by simulating the policy that is greedy w.r.t the Lag-2 approximate value function.

Several observations are in order. We see that each of the dual upper bounds improves upon the respective

\(^{11}\)The “corresponding” upper bound is the dual upper bound that uses penalties constructed from the QMDP value function.

\(^{12}\)While the value of the primal, i.e. lower, bound, does not depend on \( \tilde{P} \), this is not true of the dual upper bounds. In fact the penalties and \( \tilde{P} \) interact in a non-separable manner and an interesting research question for the information relaxation approach is to develop a better understanding of this interaction with a view to possibly optimizing it.

\(^{13}\)We report the expected value \( \mathbb{E}[\hat{V}_0(o_0) | \pi_0] \) for each approximation. All of the numerical results in this section were obtained using MATLAB release 2016b on a MacOS Sierra with a 1.3 GHz Intel Core i5 processor and 4 GB of RAM.
Figure 2: Comparison of upper bounds as a function of the noise factor $\alpha$. The dotted lines correspond to the MDP, QMDP, Lag-1 and Lag-2 approximations. The solid red and blue lines correspond to the dual PI relaxation upper bounds resulting from penalties constructed using the Lag-1 and Lag-2 approximations, respectively. The solid black line displays the best lower bound which in this case is obtained by simulating the policy that is greedy w.r.t. the Lag-2 approximate value function.

The supersolution that was used to construct the dual penalty in each case. This of course is guaranteed by Proposition 6.1 above together with Proposition 4.1 in BH. When $\alpha = 0.2$, for example, we see the best lower and dual upper bounds are approximately 87.1% and 88.1%, respectively, corresponding to a duality gap of 1%. This implies the true unknown optimal value function, $V_0^*$, lies in the interval $[0.871, 0.881]$. (This of course neglects the statistical error arising from the Monte-Carlo simulation but we can see from Figure 3 and Table 1 below that these bounds were computed to a high degree of accuracy.) This represents an approximate 70% improvement in the gap between the best lower bound and the best upper bound provided by a supersolution, which in this case is $\approx 90.5\%$ corresponding to the Lag-2 approximation.

We also see from Figure 2 that the duality gap decreases as $\alpha$ decreases and this of course is to be expected. Indeed when $\alpha = 0$ all of the bounds coincide and the duality gap is zero. This is because at that point the robot has enough accuracy and time to be able to infer its position in the maze, essentially collapsing the POMDP into the MDP version of the problem where the hidden state, $h_t$, is correctly observed at each time $t$.

Figure 3a displays lower and upper bounds corresponding to each of the four approximate value functions with $\alpha = 0.10$ and $T = 10$. Approximate 95% confidence intervals are also provided and they justify our earlier statement that the various bounds are computed to a high degree of accuracy. Several observations are again in order. First, we note the lower and upper bounds improve as we go from the MDP approximation to the QMDP approximation to the Lag-1 and Lag-2 approximations. This is not surprising since each of these
approximations uses successively less information regarding the true hidden state at each time $t$. Second, we again see that each of the dual upper bounds improves upon its corresponding supersolution.

We also note that the best duality gap (approximately $92.09\% - 91.98\% = 0.11\%$) is approximately an 83% relative improvement over the gap between the Lag-2 supersolution and the best lower bound (which is given by the policy that is greedy w.r.t the Lag-2 supersolution). While these numbers may not appear very significant on an absolute (rather than relative) basis, in many applications these differences can be significant at the margin. Moreover, there are undoubtedly applications where the best available supersolution will not be close to its corresponding lower bound in which case the improvement provided by the best information relaxation dual bound could$^{14}$ be very significant.

Finally, we observe that the two dual bounds perform quite similarly to each other with the PI bound being marginally superior to the BSPI bound. While we also noticed that the PI bound was slightly superior to the BSPI bound in the medical application of Section 7, we don’t have an explanation for this nor do we know if this is true in general.

$^{14}$It’s also worth mentioning that it may be hard to even calculate supersolutions for POMDPs when the underlying MDP is itself too complex to solve exactly. (All of the supersolutions that we construct in this paper rely on our ability to solve the underlying MDP.) In such circumstances the information relaxation approach can still be used to obtain an upper bound on the optimal value function.
<table>
<thead>
<tr>
<th>Approx.</th>
<th>MDP</th>
<th>QMDP</th>
<th>Lag-1</th>
<th>Lag-2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LB°</td>
<td>DG</td>
<td>LB</td>
<td>DG</td>
</tr>
<tr>
<td>Mean</td>
<td>1.15</td>
<td>91.71</td>
<td>0.84</td>
<td>91.80</td>
</tr>
<tr>
<td>Std. dev.</td>
<td>0.017</td>
<td>0.039</td>
<td>0.009</td>
<td>0.038</td>
</tr>
<tr>
<td>Run time (in minutes)</td>
<td>0.66</td>
<td>13.92</td>
<td>1.55</td>
<td>14.01</td>
</tr>
<tr>
<td>Supersolution UB</td>
<td>94.08</td>
<td>93.82</td>
<td>93.04</td>
<td>92.64</td>
</tr>
<tr>
<td>DG reduction</td>
<td>51%</td>
<td>60%</td>
<td>66%</td>
<td>83%</td>
</tr>
</tbody>
</table>

*There is no greedy policy w.r.t. the MDP approximate value function.

Table 1: Numerical results for the maze application with $\alpha = 0.10$. We used 50,000 sample paths to estimate the lower bounds and their corresponding PI upper bounds and duality gaps (DG). All numbers are expressed as percentages.

The number of simulated paths that we used to generate the various PI bounds and duality gaps are reported in Table 1 together with corresponding run-times and mean standard errors. All of the numbers are reported as percentages so for example, the Lag-2 duality gap is a mere 0.11%. The most obvious feature of Table 1 is how little time was required to compute the dual bounds in comparison to the lower bounds. This comparison is a little misleading, however. In particular, the lower bounds were constructed using the penalties as (action-dependent) control variates, a standard variance reduction technique. Once these control variates were calculated on each simulated path, they could then be re-used as penalties when solving the inner problem along the same path. These control variates were quite expensive to compute, however, and in Table 1 this cost has been allocated to the run times for the lower bound. As mentioned earlier, the specific details regarding the calculation of both lower and upper bounds on the same set of paths is provided in Appendix D.

7. An Application to Personalized Medicine: Mammography Screening

To date there have been relatively few successful medical applications of POMDPs. The reasons for this include the difficulty of determining a suitable objective to optimize, the difficulty of estimating the POMDP parameters and the general difficulty of solving POMDP problems. Recently Ayer et al. [3] proposed a POMDP formulation with the goal of determining an optimal screening policy for breast cancer, the most common cancer among U.S. women according to the American Cancer Society (ACS). The recommendation guidelines provided by the ACS in 2015 [32] is for women with an average risk of breast cancer to take mammograms beginning at age 45, and to continue annually until age 54. Beginning at age 55, they are then recommended to undergo biannual screenings (but they have the opportunity to continue annually if desired) and to continue taking mammograms as long as their life expectancy is at least 10 years. In addition, the ACS indicates that women aged 40 to 44 may choose to begin mammogram screening if desired. In contrast, in 2016 the U.S. Preventive Services Task Force (USPSTF) [37] recommended that women aged 50 to 74 screen biannually using mammography, and they left open the decision for women aged 40-49. In addition, they did not find enough evidence to recommend taking mammograms beyond the age of 75.

In this section we apply the information relaxation approach to the POMDP formulation of Ayer et al. We will use the term decision-maker (DM) to refer to the woman or patient in question but the decision-maker could also refer to a doctor or some other medical professional. We assume the DM has the objective

---

*A review of applications of MDPs and POMDPs to medical decision problems can be found in [35].
of maximizing her total expected quality-adjusted life years (QALYs). We assume a finite-horizon discrete-time model where the time intervals correspond to six-month periods beginning at age 40 and ending at age 100 so that $t \in \{0, \ldots, 120\}$. The hidden state space represents the true health state of the patient with $\mathcal{H} = \{0, 1, 2, 3, 4, 5\}$. Specifically:

- **State 0** represents a cancer-free patient.
- **States 1 and 2** indicate the presence of *in situ* and *invasive* cancer, respectively.
- **States 3 and 4** represent fully observed absorbing states in which the patient has been diagnosed with *in situ* and *invasive* cancer, respectively, and has begun treatment.
- **State 5** is a fully observed absorbing state representing the death of the patient.

Clearly states 3, 4 and 5 can be explicitly observed and are therefore not actually *hidden*. We include them among the set of hidden states, however, to account for the possible transition dynamics of the other hidden states into these absorbing states. The knowledge of being in these hidden absorbing states can then be modeled correctly through noiseless observations of them. We will refer to the subset of hidden states $\{0, 1, 2\}$ as *pre-cancer* states and the absorbing states $\{3, 4, 5\}$ as *post-cancer* states.

At each time $t$, the DM can choose to either have a mammography screening ($M$) or wait ($W$). If the decision to wait is made, the patient may perform a self-detection screening which will have either a positive or negative result. That is, if through self-detection the patient has reason to be concerned about the presence of cancer, we say the self-test is positive. The possible results of a mammogram are also positive or negative. In the former case, an accurate procedure, e.g. a biopsy, is then prescribed to precisely determine the true cancer status of the patient. If the biopsy result is positive and cancer is found with certainty, the patient will then exit the screening process and move into one of the absorbing states, 3 or 4, to indicate that cancer treatment has commenced. To code this behavior, Ayer uses hidden state transitions that are functions of the observations. To model this behavior as a conventional POMDP (where hidden state transitions do not depend on observations), we introduce an exit action ($E$) as the only available action after a positive biopsy has been observed. The transition into absorbing state 3 or 4 will now only depend on the current hidden state and the exit action which must be taken if the biopsy result is positive and cancer is found with certainty. The set of possible observations is therefore $\mathcal{O} = \{R^-, R^+, B_1, B_2, D\}$ where:

- **$R^-$** is a negative test result (either from a mammography or self-detection).
- **$R^+$** is a positive test result (including a negative biopsy if the test was a positive mammogram).
- **$B_1$** and **$B_2$** represent *in situ* cancer and *invasive* cancer, respectively, and can be observed via a biopsy following a positive mammogram. If $B_1$ or $B_2$ are observed, the action space is then restricted to the exit action $E$ which transfers the patient to the corresponding absorbing state.
- **$D$** represents the death of the patient.

We assume a prior probability distribution, $\pi_0$, on the true health-state of the woman at age 40. The transition probabilities of the latent pre-cancer health states are assumed to be age-specific and therefore a function of time $t$. We assume that a screening decision does not influence the development of cancer and
Table 2: Sources of the demographic rates for the transition probabilities.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mortality $m_0$</td>
<td>SSA Period Life Table, 2013, Female mortality [38]</td>
</tr>
<tr>
<td>Mortality $m_1$, $m_2$</td>
<td>SEER [23] Table 4.13, all stages and all ages\textsuperscript{17}</td>
</tr>
<tr>
<td>Incidence $p_{01}$</td>
<td>SEER Table 4.12, all races</td>
</tr>
<tr>
<td>Incidence $p_{02}$</td>
<td>SEER Table 4.11, all races</td>
</tr>
<tr>
<td>Incidence $p_{12}$</td>
<td>Assumed equal to $p_{02}$</td>
</tr>
<tr>
<td>Initial risk $\pi_0$</td>
<td>SEER Table 4.24, female 40-49 \textsuperscript{18}</td>
</tr>
</tbody>
</table>

The observation probabilities are determined by the accuracy of the examinations, which are commonly referred to as specificity and sensitivity. The specificity of a test corresponds to the true negative rate, i.e. the probability that a cancer-free woman obtains a negative test result, while the sensitivity of a test is the true positive rate, i.e. the probability of a positive test result given that the woman has cancer. For each test we employ the age-specific sensitivity and specificity factors that were computed and reported by Ayer

\textsuperscript{16}Experts in the field of breast cancer could almost certainly provide superior estimates for those parameters where we could not find external estimates.

\textsuperscript{17}We approximated the invasive cancer mortality rate by inferring the 6-month mortality rate from the 5 year survival rate (0.897) and used the maximum of this 6-month rate and the average female 6-month mortality for a woman of that age. We assumed that in situ mortality is equal to the female mortality times 1.02 for women of the same age.

\textsuperscript{18}The initial risk for an average woman was taken from the breast cancer prevalence rate (0.9462\%) and split 80\% for invasive cancer and 20\% for in situ cancer, as discussed in [39].
et al. They are:

\[
spec_t(W) = 0.92, \quad \forall t \\
sens_t(W) = 0.44, \quad \forall t
\]

\[
spec_t(M) = \begin{cases} 
0.889, & \text{if } t \in \{0, \ldots, 19\} \\
0.893, & \text{if } t \in \{20, \ldots, 39\} \\
0.897, & \text{if } t \geq 40
\end{cases} \\
sens_t(M) = \begin{cases} 
0.722, & \text{if } t \in \{0, \ldots, 29\} \\
0.81, & \text{if } t \in \{30, \ldots, 59\} \\
0.862, & \text{if } t \geq 60.
\end{cases}
\]

Using these rates, we define the age-specific observation matrices according to

\[
B^t(W) = \begin{bmatrix}
spec_t(W) & 1 - spec_t(W) & 0 & 0 & 0 \\
1 - sens_t(W) & sens_t(W) & 0 & 0 & 0 \\
1 - sens_t(W) & sens_t(W) & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 1
\end{bmatrix}
\]

and

\[
B^t(M) = \begin{bmatrix}
spec_t(M) & 1 - spec_t(M) & 0 & 0 & 0 \\
1 - sens_t(M) & sens_t(M) & 0 & 0 & 0 \\
1 - sens_t(M) & 0 & sens_t(M) & 0 & 0 \\
0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 1
\end{bmatrix}
\]

where \(B^t_{ij}(a)\) is the probability of observation \(j \in \mathcal{O}\) when action \(a\) is taken and the hidden state is \(i \in \mathcal{H}\). Note that the observability of the hidden absorbing states 3, 4 and 5 is made evident through these matrices.

It is worth pointing out that once action \(E\) has been chosen, the DM immediately transitions to an absorbing fully-observable state, and therefore there is no need to define \(B^t(E)\).

A characteristic of medical decision problems, as pointed out in Ayer et al., is that the observation at time \(t\) is a function of the current action, \(B^t_{ij}(a) := \text{P}(a_t = j \mid h_t = i, a_t = a)\), as opposed to a conventional POMDP where the observation is a function of the prior action; see (2). This means that events take place in the following order: given a belief state the DM first takes an action, then immediately observes the result of the action and updates the belief, then a transition takes place and the belief is “carried forward”. This technicality results in a different version of the standard filtering update in which the transition occurs prior to the observation. Nonetheless, filtering in this non-standard form of the POMDP is still\(^{19} \) a straightforward task. And for the same reason, the natural filtration for the medical decision problem is one where \(\mathcal{F}_t\) is defined to be the \(\sigma\)-algebra generated by \(o_{0:t-1}\), for \(t \geq 1\), and with \(\mathcal{F}_0\) defined to be the \(\sigma\)-algebra generated by \(\pi_0\), the prior distribution on the initial hidden state.

We define the reward obtained at time \(t\), \(r_t(h_t, a_t, o_t)\), as the expected QALYs between times \(t\) and \(t + 1\) that a person in true health-state \(h_t\) would accrue after making decision \(a_t\) and obtaining observation \(o_t\).

\(^{19}\)Further technical details are provided in Appendix E.
Note that although the reward is a function of the as yet unseen observation (see previous paragraph), \( o_t \), we can instead use\(^{20}\) its expected value

\[
 r_t(h_t, a_t) := \mathbb{E}[r_t(h_t, a_t, o_t) \mid h_t]
\]

which is easy to calculate and is now in the standard form for a POMDP.

We follow the same calculations as Ayer et al. to define the reward functions. If the patient is in a pre-cancer state \( i \in \{0, 1, 2\} \), the wait action reward is given by \( r_t(i, W, o_t) = 0.25 m_i^t + 0.5(1 - m_i^t) \) where the \( m_i^t \)'s are the (semi-annual) mortality rates given in Table 2. This is the reward for a woman in period \( t \) and true pre-cancer health state, \( i \), and in fact does not depend on the observation \( o_t \). Specifically, if death occurs in the next six months (which occurs w.p. \( m_i^t \)), it is assumed to happen exactly at the three month mark and so the woman will therefore obtain 0.25 years of lifetime. In contrast, if she survives (which occurs w.p. \( 1 - m_i^t \)) she obtains the 0.5 half-years of lifetime in that period.

For the mammography screening action, we subtract a disutility function, \( du(h_t, o_t) \), from the reward so that

\[
 r_t(h_t, M, o_t) := r_t(h_t, W, o_t) - du(h_t, o_t).
\]

The disutility is given a value of 0.5 days for a negative mammogram, two weeks for a true positive mammogram and four weeks for a false positive mammogram. True positive mammograms will in addition force the DM to exit the system in the next period, and provide a lump-sum reward of \( R_t(i) = r_t(i, E) \) for \( i = 1, 2 \). Recall that a true positive mammogram followed by an exit action refers to a woman being accurately diagnosed with cancer and then going into treatment immediately. We expect that a patient under treatment would have a lower remaining expected lifetime than the remaining expected lifetime, \( e_t(0) \) say, of a healthy woman of the same age, but higher than the remaining expected lifetime, \( e_t(i) \) say, of a woman with cancer \( i \in \{1, 2\} \) who is undiagnosed and of the same age. (Note that the expected remaining lifetimes can be calculated using the corresponding mortality rates from times \( t \) to \( T \).) We therefore assume \( e_t(0) < R_t(i) < e_t(i) \) and in our numerical example, we set \( R_t(i) = 0.5 e_t(0) + 0.5 e_t(i) \) for \( i = 1, 2 \). We also assume that the absorbing states provide no rewards.

It is perhaps worth noting how the benefit of mammography screening is modelled in our POMDP setting. Specifically, it arises from the possibility of identifying a cancer early and therefore entering treatment and having an expected remaining lifetime that is greater than if the cancer went undiagnosed. The reduced expected lifetime of a woman with an undiagnosed cancer will be reflected via the specific values of the transition and mortality rates of the second and third rows (corresponding to undiagnosed cancer states 1 and 2) in (32). There is a cost to mammography screening, however, which is reflected via the disutility function and so the ultimate goal is to find a policy that trades the benefits of mammography screening off against its disutility.

### 7.1. Value Function Approximations

Two methods were used to obtain value function approximations: a QMDP approximation, adapted from the robot navigation problem to include intermediate rewards, and a grid-based approximation. The QMDP approximation is given by

\[
 \hat{V}_t(o_{0:t-1}) := \max_{a_t} \sum_{h \in H} \pi_t(h) V_t^Q(h, a_t)
\]

\(^{20}\)We acknowledge a slight abuse of notation here in that we are using the same \( r_t \) to denote time \( t \) rewards \( r_t(h_t, a_t) \), \( r_t(h_t, a_t, o_t) \) and \( r_t(\pi_t, a_t) \). It should be clear from the context what version of the reward we have in mind.
with the understanding that at \( t = 0 \), \( \hat{V}_0 := \hat{V}_0(\pi_0) \), and where \( V_t^{Q} \) is the Q-function of the corresponding fully observable MDP formulation, i.e.

\[
V_t^{Q}(h, a) := r_t(h, a) + \sum_{h' \in H} P_{hh'}(a)V_{t+1}^{MDP}(h')
\]

(35)

\[
V_t^{MDP}(h) := \max_{a_t \in A} V_t^{Q}(h, a_t)
\]

(36)

for \( t \in \{0, \ldots, T\} \) with terminal condition \( V_{T+1}^{MDP} := 0 \). Note that the only difference between these definitions and those given for the robot navigation application is the inclusion here of intermediate rewards.

The grid approximation corresponds to a point-based value iteration method using a fixed and finite grid approximation of the belief space, \( \Pi \) (see \[28, 21\]). A standard approximation tool in dynamic programming is to represent an infinite state space as a finite grid of points, \( P \subset \Pi \), and obtain an approximate value function by linear interpolation for points not in \( P \). Specifically, the approximate value function is obtained by solving a dynamic program with terminal condition \( \hat{V}_{T+1} = 0 \) and Bellman equation

\[
\hat{V}_t(\pi) = \max_{a_t} \left[ r_t(\pi, a_t) + \sum_{o} P_{\pi a}(o | \pi) \hat{V}_{t+1}(f(\pi, a_t, o)) \right]
\]

(37)

for \( t \in \{0, \ldots, T\} \), \( \pi \in P \) and where \( r_t(\pi, a) := \sum_h r_t(h, a)(h) f(\pi, a, o) \) is the belief update function, and \( P_{\pi a}(o | \pi) := \sum_h P_{\pi a}(o | h) \pi(h) \). Note that in general \( f(\pi, a_t, o) \) will not be an element in \( P \) and so we use linear interpolation to evaluate the approximate value function at those points. To tie in the grid approximation with our application, we take the 3-dimensional subspace corresponding to the pre-cancer states

\[
\tilde{\Pi} := \{ \pi \in \Pi \mid \pi = (\pi_0, \pi_1, \pi_2, 0, 0, 0), \pi_0 + \pi_1 + \pi_2 = 1 \}
\]

of the 6-dimensional simplex \( \Pi \). We call \( \tilde{\Pi} \) the pre-cancer belief space simplex\(^{21}\) and form a finite grid \( P \subset \tilde{\Pi} \). We then solve the dynamic program (37) for all elements of \( P \) union the elements \((0, 0, 0, 1, 0, 0), (0, 0, 0, 0, 1, 0) \) and \((0, 0, 0, 0, 0, 1) \). For our application, we use a grid \( P \) with elements \( 0.05 \times (i_1, i_2, i_3) \) with \( i_1, i_2, i_3 \) integer valued and such that they lie on \( \tilde{\Pi} \), i.e. \( 0.05 \times (i_1 + i_2 + i_3) = 1 \).

We can now generate lower bounds on the optimal value function, \( V_{\pi_0}^{*} \), by simulating the policies that are greedy w.r.t. each of the value function approximations. We will compare the performance of these greedy policies to the official policies recommended by ACS and USPSTF.

7.2. The Uncontrolled Formulation

The action-independent transition and emission matrices are built using different approaches for each approximate value function. First, using the fact that the QMDP approximation is a supersolution, we can drop the absolute continuity requirement and set the transition matrices, \( Q^t \), using (31) and, similarly, we set the uncontrolled emission matrices according to

\[
E_{ij}^{t} \equiv B_{ij} \{ \arg\max_{a \in A} V_{t}^{Q}(i, a) \}.
\]

(38)

\(^{21}\)Although the dimension of the hidden state space is 6, in reality the uncertainty in the process is entirely restricted to the 3 pre-cancer states. We can therefore reduce our analysis to the 3-dimensional pre-cancer belief space simplex.
In contrast, there is no guarantee that the approximate value function based on the grid approximation is a supersolution and so we must satisfy the absolute continuity conditions. To achieve this, we add a small positive quantity $\epsilon = 0.001$ to each $Q^t_{ij}$ if $j$ can be reached from $i$ under some action, and then normalize the probabilities. Similarly, we add $\epsilon$ to $B^t_{ik}$ only if $k$ can be observed from state $i$ under some action and again we then normalize the probabilities. This approach allows our transition and emission probabilities to satisfy absolute continuity for the PI relaxation. For the BSPI relaxation we would need to make an additional adjustment (as described in Appendix A.2) but the BSPI results were slightly inferior to the PI results (as was the case with the maze application) and so we don’t report them in our numerical results.

7.3. Numerical Results

We consider two different test cases: case 1 represents a woman at age 40 with an average risk of having cancer and therefore an initial distribution over hidden states given by $\pi_0 = [0.9905, 0.0019, 0.0076, 0, 0, 0]$. Case 2 represents a woman at age 40 with a high-risk of having cancer; she has an initial distribution of $\pi_0 = [0.96, 0.02, 0.02, 0, 0, 0]$. In Figure 4a we display the lower bounds obtained by simulating each of the four policies, namely the policies recommended USPSTF and ACS, as well as the policies that are greedy w.r.t. the QMDP and grid-based approximate value functions. We note that the latter two policies outperform the official recommendations of USPSTF and ACS, with the best lower bound coming from the grid approximation.

![Figure 4a](image)

Figure 4a: Lower bounds on the optimal value function obtained from simulating the USPSTF and ACS recommended policies as well as policies that are greedy w.r.t. the QMDP and grid-based approximate value functions. Case 1 corresponds to an average risk 40-year old woman while case 2 corresponds to a high risk 40-year old woman. The vertical lines on each bar represent 95% confidence intervals.

![Figure 4b](image)

Figure 4b: Upper bounds on the optimal value function compared to the best lower bound which was obtained by simulating the policy that is greedy w.r.t. the grid-based approximate value function. The best upper bound was also obtained by constructing penalties for the PI relaxation from the grid-based approximate value function. The optimal duality gap is displayed in each case.

Figure 4b displays the upper bounds obtained with the PI relaxation using penalties constructed from each of the two approximate value functions. Since the QMDP approximate value function is a supersolution and therefore also an upper bound we also plot its value in the figure. As a reference, we also display the value of the best lower bound to obtain a visual representation of the duality gap. The duality gap reduction of the best dual bound with respect to the supersolution is 57% in case 1, and 51% in case 2, or equivalently,
### Table 3: Summary statistics for the lower and upper bounds for the Case 1 scenario.

<table>
<thead>
<tr>
<th>Bound</th>
<th>Expected value</th>
<th>Standard dev.</th>
<th>Number of paths</th>
<th>Running time</th>
</tr>
</thead>
<tbody>
<tr>
<td>USPSTF (Lower)</td>
<td>41.15</td>
<td>0.0188</td>
<td>400,000</td>
<td>6.75 mins.</td>
</tr>
<tr>
<td>ACS (Lower)</td>
<td>41.14</td>
<td>0.0185</td>
<td>400,000</td>
<td>6.79 mins.</td>
</tr>
<tr>
<td>Greedy QMDP (Lower)</td>
<td>41.56</td>
<td>0.0160</td>
<td>500</td>
<td>11.1 secs.*</td>
</tr>
<tr>
<td>Greedy Grid (Lower)</td>
<td><strong>41.72</strong></td>
<td>0.0128</td>
<td>800,000</td>
<td>35.05 mins.</td>
</tr>
<tr>
<td>Grid PI (Upper)</td>
<td><strong>41.77</strong></td>
<td>0.0001</td>
<td>100</td>
<td>9.9 secs.</td>
</tr>
<tr>
<td>QMDP PI (Upper)</td>
<td>41.83</td>
<td>0.00004</td>
<td>500</td>
<td>8.1 secs.</td>
</tr>
<tr>
<td>QMDP Supersolution (Upper)</td>
<td>41.84</td>
<td>-</td>
<td>1</td>
<td>0.02 secs.</td>
</tr>
</tbody>
</table>

*Lower bound for QMDP greedy strategy was estimated using the penalties as control variates - see Appendix D

In Ayer et al., the authors were able to solve the POMDP to optimality using Monahan’s algorithm [30] with Eagle’s reduction [15]. The authors used an Intel Xeon 2.33 GHz processor with 16 GB RAM for their computations, and were able to solve the problem in 55.95 hours. As with our robot navigation application, we used MATLAB Release 2016b, and a MacOS Sierra with 1.3 GHz Intel Core i5 processor with 4 GB RAM. The numerical results in Table 3 display the running times and other statistics for the various Case 1 bounds as well as the best bounds in bold font. As we have noted, our bound approximations result in a very tight duality gap (19.7 days or 0.054 QALYs for an average woman) and we were able to obtain the best lower and upper bounds in 35.05 minutes and 9.9 seconds, respectively, with narrow confidence intervals. So while Ayer et al. were able to solve the problem to optimality, we were able to get provably close to optimality using a slower processor and less RAM with a total runtime that was approximately 2 orders of magnitude smaller. We also note that even tighter bounds information relaxation bounds should be attainable here if so desired using a partially controlled formulation as introduced in BH.

### 8. Extension to Infinite Horizon Problems

We can extend these techniques to the infinite horizon class of POMDPs with discounted rewards following the approach of BH and [41]. Let the discount factor be denoted by $\delta \in [0,1)$, indicating that rewards received at a later time contribute less than rewards received earlier. The corresponding infinite-horizon POMDP can be stated as solving the following optimization problem

$$V_0^* := \max_{\mu \in \mathcal{U}_0} E \left[ \sum_{t=0}^{\infty} \delta^t r(h_t, \mu_t) \bigg| F_0 \right]$$  \hspace{1cm} (39)

In order to solve the dual problem using a PI relaxation, we would have to simulate an infinite sequence of random variables $\{w_t, v_t\}_{t \geq 0}$, which is not possible in practice. An equivalent formulation, however, is to replace the discounting by a costless, absorbing state $h^a$ which can be reached from every state and feasible action with probability $1 - \delta$, at each $t$. The state transition distribution remains as before, conditional on not reaching the absorbing state. The equivalent absorbing state formulation is then given by

$$V_0^* := \max_{\mu \in \mathcal{U}_0} E \left[ \sum_{t=0}^{\tau} r(h_t, \mu_t) \bigg| F_0 \right]$$  \hspace{1cm} (40)

\[\text{We do not know what software Ayer et al. used}\]
where $\tau = \inf \{ t : h_t = h^a \}$ is the absorption time, distributed as a geometric random variable with parameter $1 - \delta$. In (40) the expected value is calculated over the modified state transition function that accounts for the presence of the absorbing state. In the dual problem formulation, knowledge of the absorption time should be included in the relevant information relaxation. For example, under the PI relaxation, the dual upper bound can be expressed as

$$V^*_0(\pi_0) \leq \mathbb{E} \left[ \max_{a_0, \tau \cdots} \sum_{t=0}^{\tau} \Phi_t[r_t(h_t, a_t) + c_t] \bigg| \mathcal{F}_0 \right].$$

(41)

An inner problem inside the expectation on the r.h.s of (41) can be generated by first simulating the absorption time $\tau \sim \text{Geom}(1 - \delta)$, and then generating the hidden states and observations, $h_{0:t}$ and $o_{1:t}$, respectively, using an action-independent transition and observation distribution as before. A lower bound can be obtained of course by simply simulating many paths of some feasible policy.

One concern with the bound of (41) is that the optimal objective of the inner problem in (41) might have an infinite variance. This was not a concern in the finite horizon setting with finite state and action spaces. It is a concern, however, in the infinite horizon setting where $\tau$ is now random and the presence of the RN derivative terms $\Phi_t$ might now cause the variance to explode. BH resolved this issue through the use of supersolutions to construct dual penalties. In that case their bound improvement result and other considerations allowed them to conclude that the variance of the upper bound estimator in (41) would remain bounded.

Of course an alternative approach to guarantee finite variance estimators is to truncate the infinite horizon to some large fixed value, $T$, and then add $\delta^T \bar{r} / (1 - \delta)$ as a terminal reward where $\bar{r} := \max_{h, a} r(h, a)$. Because the terminal reward is an upper bound on the total discounted remaining reward after time $T$ in the infinite horizon problem, we are guaranteed that a dual upper bound for the truncated problem will also be a valid upper bound on the infinite horizon problem. By choosing $T$ suitably large we can minimize the effect of truncation on the quality of the dual bound for the infinite horizon problem.

9. Conclusions and Further Research

We have shown how change of measure arguments and an uncontrolled problem formulation can be used to extend the information relaxation approach to POMDP settings where the calculation of dual penalties would otherwise be impossible except in the smallest of problem instances. Numerical applications to robotic control and personalized medicine have demonstrated that significant bound improvements can be obtained using information relaxations when the penalties are constructed from supersolutions.

There are several possible directions for future research. One direction would be to extend the approach to other non-Markovian control problems where the difficulty associated with calculating dual feasible penalties would also be problematic. A particularly interesting application would be to dynamic zero-sum games (ZSG’s) where the players have asymmetric information. Following [20], dual bounds on the optimal value of the game can be computed by fixing one player’s strategy and bounding the other player’s best response. In the case of asymmetric information (which was not considered by [20]), bounding the other player’s best response amounts to finding a dual bound on a POMDP and so the techniques developed in this paper also apply in that setting. Moreover, due to Shapley’s seminal results strong duality continues to hold in the ZSG

$^{23}$See also the discussion immediately following our Proposition 6.1.
framework so the dual bounds can be used to construct a certificate of near-optimality when each player has close-to-optimal strategies. Another interesting non-Markovian setting is the influence diagram framework which is popular in the decision science literature.

A second direction would be to explore the relationship between the quality of the dual bound and the action-independent transition and observation distributions. While the primal, i.e. lower bound, does not depend on the action-independent distributions of the uncontrolled problem formulation, this is not true for the dual bound. Indeed as pointed out in BH, the specific value of the dual bound will depend on the quality of the penalties and the action-independent distributions. It would therefore be of interest to explore this dependence further. Moreover, because of the abundance of supersolutions in the POMDP setting, absolute continuity of the action-independent distributions is not a requirement and so, as discussed in Appendix D, we would be free to explore dual bounds when the action-independent distributions are defined by good feasible policies.

Acknowledgements

The authors are very grateful to Yuan Zhong for helpful comments and conversations. All errors are our own.
References


A. RN Derivative Calculations

A.1. The Uncontrolled Original POMDP Formulation

To show that the RN derivatives in (12) and (13) are correct under the PI relaxation framework, it suffices to prove that

$$\mathbb{E}\left[\sum_{t=0}^{T} r_{t}(h_{t}, a_{t}) \mid F_{0}\right] = \tilde{\mathbb{E}}\left[\sum_{t=0}^{T} \Phi_{t} r_{t}(h_{t}, a_{t}) \mid F_{0}\right]$$

or equivalently that

$$\mathbb{E}\left[ r_{t}(h_{t}, a_{t}) \mid F_{0}\right] = \tilde{\mathbb{E}}\left[ \Phi_{t} r_{t}(h_{t}, a_{t}) \mid F_{0}\right]$$  \hspace{1cm} (42)

for all $t \in \{0, \ldots, T\}$, where we recall that $\mathbb{E}$ and $\tilde{\mathbb{E}}$ correspond to expectations under $\mathbb{P}$ and $\tilde{\mathbb{P}}$, respectively.

We first write the expectation on the r.h.s. of (42) explicitly to obtain

$$\sum_{a_{0:t}, h_{0:t}} \Phi_{t}(h_{0:t}, a_{1:t}, a_{0:t-1}) r_{t}(h_{t}, a_{t}) \tilde{\mathbb{P}}(a_{1:t}, h_{0:t} \mid \pi_{0})$$  \hspace{1cm} (43)

where $\pi_{0}$ is the initial hidden state distribution. Recalling (12) and (13) we have

$$\Phi_{t}(\cdot) = \prod_{s=0}^{t-1} \frac{P_{h_{s+1}}(a_{s}) B_{h_{s+1}, a_{s+1}}(a_{s})}{Q_{h_{s+1}, a_{s+1}} E_{h_{s+1}, a_{s+1}}}$$

$$= \prod_{s=0}^{t-1} \frac{P_{h_{s+1}}(a_{s}) \mathbb{P}_{a_{s}}(a_{s+1} \mid h_{s+1})}{\mathbb{P}(h_{s+1} \mid h_{s}) \mathbb{P}(a_{s+1} \mid h_{s+1})}$$

$$= \frac{\mathbb{P}_{a_{0:t-1}}(a_{1:t}, h_{0:t} \mid h_{0}) \pi_{0}(h_{0})}{\mathbb{P}(a_{1:t}, h_{0:t} \mid \pi_{0})}$$

$$\equiv \mathbb{P}_{a_{0:t-1}}(a_{1:t}, h_{0:t} \mid \pi_{0}) \equiv \mathbb{P}(a_{1:t}, h_{0:t} \mid \pi_{0})$$  \hspace{1cm} (44)

where $\mathbb{P}_{a_{0:t-1}}$ and $\mathbb{P}_{a_{s}}$ explicitly recognize the dependence of the given probabilities on $a_{0:t-1}$ and $a_{s}$, respectively. If we substitute (44) into (43) we obtain

$$\sum_{a_{0:t}, h_{0:t}} r_{t}(h_{t}, a_{t}) \mathbb{P}_{a_{0:t-1}}(a_{1:t}, h_{0:t} \mid \pi_{0}) = \mathbb{E}\left[ r_{t}(h_{t}, a_{t}) \mid F_{0}\right]$$  \hspace{1cm} (45)

which establishes the correctness of the RN derivatives in (12) and (13).

A.2. The Uncontrolled Belief State POMDP Formulation

In order to compute the RN derivatives for the uncontrolled belief state POMDP formulation we must first (of course) define the uncontrolled belief-state dynamics for $\pi_{t}$ which lies in the $|\mathcal{H}|$-dimensional simplex. Note that while there are infinitely many points in the simplex only a finite number of these points will have a strictly positive probability under $\mathbb{P}$ conditional on $\pi_{0}$ which is assumed known. These points with strictly positive $\mathbb{P}$-probability arise from the various possible combinations of hidden-state and observation sequences which are finite in number by assumption. It therefore makes sense to define the uncontrolled belief-state dynamics for $\pi_{t}$ by generating action-independent hidden state and observation sequences.

An obvious approach would be to use (10) and (11) to generate such sequences and then simply use the
generated observations to update the belief state appropriately, beginning with $\pi_0$. The only problem with this is that $\mathbb{P}$ will not be absolutely continuous w.r.t $\tilde{\mathbb{P}}$ even if $Q$ and $E$ as defined in (10) and (11) do satisfy their absolute continuity conditions.

We can resolve this issue by instead assuming that under $\tilde{\mathbb{P}}$ the current belief state, $\pi$, transitions w.p. $(1 - \epsilon)$ to a belief state consistent with (10) and (11) as described above, and w.p. $\epsilon$ transitions with equiprobability to any belief state, $\pi'$, which is feasible for some available action $a \in A$ given $\pi$. In this case the RN derivatives take the form

$$
\phi(\pi, \pi', a) := \frac{\sum_{i,j,k} \pi(i) P_{ij}(a) B_{jk}(a) 1_{\{\pi' = f(\pi, a, k)\}}}{\epsilon g(\pi, \pi') + (1 - \epsilon) \sum_{i,j,k} \pi(i) Q_{ij} E_{jk} 1_{\{\pi' = \tilde{f}(\pi, k)\}}}
$$

(46)

$$
\Phi_t^\pi(\pi_{0:t}, a_{0:t-1}) := \prod_{s=0}^{t-1} \phi(\pi_s, \pi_{s+1}, a_s)
$$

(47)

where $\pi_t(h) = \tilde{\mathbb{P}}(h_t = h \mid o_{1:t})$ is the filtered conditional distribution of $h_t$ under $\tilde{\mathbb{P}}$, $f(\pi, a, k)$ and $\tilde{f}(\pi, k)$ are the belief state update functions under $\mathbb{P}$ and $\tilde{\mathbb{P}}$, respectively, and $\epsilon \in (0, 1)$ is a given constant typically chosen to be close to 0. Note that $f(\pi, a, k)$ and $\tilde{f}(\pi, k)$ are implicitly defined via appropriate filtering algorithms so that the indicator functions in (46) are easily evaluated. Finally we define the function

$$
g(\pi, \pi') := \frac{1}{|A| \times |O|} \sum_{(a,k) \in A \times O} 1_{\{\pi' = f(\pi, a, k)\}}
$$

(48)

which ensures absolute continuity of $\mathbb{P}$ w.r.t $\tilde{\mathbb{P}}$. In order to justify (46) and (47) we first express the time $t$ reward as a function of the belief state according to

$$
r_t(\pi_t, a_t) := \mathbb{E}[r_t(h_t, a_t) \mid \mathcal{F}_t^\pi] = \sum_{h_t} r_t(h_t, a_t) \pi_t(h_t).
$$

(49)

The RN derivatives must then satisfy

$$
\mathbb{E} \left[ r_t(\pi_t, a_t) \right. \mid \mathcal{F}_0] = \tilde{\mathbb{E}} \left[ \Phi_t r_t(\pi_t, a_t) \right. \mid \mathcal{F}_0]$

or, writing the expectations explicitly,

$$
\sum_{\pi_{1:t}} r_t(\pi_t, a_t) \mathbb{P}_{a_{0:t-1}}(\pi_{1:t}) = \sum_{\pi_{1:t}} \Phi_t r_t(\pi_t, a_t) \tilde{\mathbb{P}}(\pi_{1:t}).
$$

It is clear then that the RN derivatives satisfy

$$
\Phi_t := \frac{\mathbb{P}_{a_{0:t-1}}(\pi_{1:t})}{\tilde{\mathbb{P}}(\pi_{1:t})}
$$

(50)
where we can calculate the numerator as

\[
\mathbb{P}_{a_{0:t-1}}(\pi_{1:t}) = \prod_{s=0}^{t-1} \mathbb{P}_{a_{s}}(\pi_{s+1} | \pi_{s})
\]

\[
= \prod_{s=0}^{t-1} \sum_{o_{s+1}} \mathbb{P}_{a_{s}}(o_{s+1} | \pi_{s}) \mathbb{P}_{a_{s}}(\pi_{s+1} | o_{s+1}, \pi_{s})
\]

\[
= \prod_{s=0}^{t-1} \sum_{h, h', o_{s+1}} \pi_{s}(h) \mathbb{P}_{a_{s}}(h' | h) \mathbb{P}_{a_{s}}(o_{s+1} | h') 1_{\{\pi_{s+1} = f(\pi_{s}, a_{s}, o_{s+1})\}}
\]

\[
= \prod_{s=0}^{t-1} \sum_{h, h', o} \pi_{s}(h) P_{hh'}(a_{s}) B_{h'o}(a_{s}) 1_{\{\pi_{s+1} = f(\pi_{s}, a_{s}, o)\}}
\]

(51)

In a similar manner, we can factorize the denominator and calculate each factor as

\[
\tilde{\mathbb{P}}(\pi_{s+1} | \pi_{s}) = \epsilon g(\pi_{s}, \pi_{s+1}) + (1 - \epsilon) \sum_{h, h', o} \pi_{s}(h) Q_{hh'} E_{h'o} 1_{\{\pi_{s+1} = f(\pi_{s}, o)\}}
\]

(52)

where \(g(\pi_{s}, \pi_{s+1})\) is defined in (48). This establishes then the resulting RN derivatives given in equations (46) and (47).

A.3. An Alternative POMDP Formulation

An alternative POMDP formulation is one where the state at time \(t\) is the observation sequence \(o_{1:t}\). This is closely related to the belief-state formulation where the state at time \(t\) is the belief state, \(\pi_{t}\), which of course is a function of \(o_{1:t}\). Under this alternative formulation the dynamics take the form \(o_{t+1} = f_{o}(o_{1:t}, a_{t}, u_{t+1})\) where \(u_{t+1} \sim U(0,1)\) is exogenous and \(f_{o}\) is only defined implicitly via a filtering algorithm which, as an intermediate step, would calculate \(\pi_{t}\). We can redefine the time \(t\) reward to be a function of the time \(t\) state, \(o_{1:t}\), according to

\[
r_{t}(o_{1:t}, a_{0:t}) := \mathbb{E}[r_{t}(h_{t}, a_{t}) | \mathcal{F}_{0}] = \sum_{h_{t}} r_{t}(h_{t}, a_{t}) \pi_{t}(h_{t}).
\]

(53)

The RN derivatives for an uncontrolled version of this alternative formulation of the POMDP must satisfy

\[
\mathbb{E} \left[ \sum_{t=0}^{T} r_{t}(o_{1:t}, a_{0:t}) \big| \mathcal{F}_{0} \right] = \mathbb{E} \left[ \sum_{t=0}^{T} \Phi_{t} r_{t}(o_{1:t}, a_{0:t}) \big| \mathcal{F}_{0} \right]
\]

or equivalently, for each \(t \in \{0, \ldots, T\}\) the RN derivatives must satisfy

\[
\mathbb{E} \left[ r_{t}(o_{1:t}, a_{0:t}) \big| \mathcal{F}_{0} \right] = \mathbb{E} \left[ \Phi_{t} r_{t}(o_{1:t}, a_{0:t}) \big| \mathcal{F}_{0} \right]
\]

or, writing the expectations explicitly, we must have

\[
\sum_{o_{1:t}} r_{t}(o_{1:t}, a_{0:t}) \mathbb{P}_{a_{0:t-1}}(o_{1:t}) = \sum_{o_{1:t}} \Phi_{t} r_{t}(o_{1:t}, a_{0:t}) \tilde{\mathbb{P}}(o_{1:t}).
\]
It follows immediately that
\[ \Phi_t = \frac{P_{a_{t-1}}(a_{t})}{P(o_{t})} \quad (54) \]

We can calculate the numerator in (54) as
\[
P_{a_{t-1}}(o_{t}) = \prod_{s=0}^{t-1} P_{a_{s}}(o_{s+1} | o_{s})
\]
\[= \prod_{s=0}^{t-1} \sum_{h_s,a_{s+1}} P_{a_{s-1}}(h_s | o_{1:s}) P_{a_{s}}(h_s | a_{s}) P_{a_{s+1}}(o_{s+1} | h_s) \]
\[= \prod_{s=0}^{t-1} \sum_{h_s,a_{s+1}} \pi_s(h_s) P_{h_s} a_{s+1}(a_s) B_{h_{s+1} o_{s+1}}(a_s) \quad (55)\]

where we take \( P_{a_{0,s}}(o_{s+1} | o_{1:s}) \) to be \( P_{a_{0}}(o_{s} | \pi_0) \) when \( s = 0 \) on the first line of the preceding equation. Similarly, we calculate the denominator of (54) as
\[
\tilde{P}(o_{t}) = \prod_{s=0}^{t-1} \sum_{h_s,a_{s+1}} \tilde{P}(h_s | o_{1:s}) \tilde{P}(h_{s+1} | h_s) \tilde{P}(o_{s+1} | h_s) \]
\[= \prod_{s=0}^{t-1} \sum_{h_s,a_{s+1}} \tilde{\pi}_s(h_s) Q_{h_s} a_{s+1} E_{h_{s+1} o_{s+1}} \quad (56)\]

where \( \tilde{\pi}_s \) denotes the time \( s \) belief state under \( \tilde{E} \).

Although this formulation of the POMDP is valid, the resulting RN derivatives in (54) for the uncontrolled version of the problem bring an additional difficulty. Specifically, the \( \pi_s(h_s) \) term in (55) depends on the history of actions \( a_{0:s-1} \) and so it would not be possible to separate the actions into different factors in the RN derivatives, which would be required when solving inner problems for the perfect information of the uncontrolled formulation of the problem. This is not the case with the other two relaxations, and so the uncontrolled version of this alternative formulation of the POMDP does not lend itself to inner problems that are easy to solve.

**B. Value Function Approximations**

**B.1. Computing the Value Function for the Lag-1 Approximation**

The Lag-1 formulation corresponds to the relaxed problem in which decision \( a_t \) has information on the true value of state \( h_{t-1} \), as well as the observation path \( o_{0:t} \) and all past decisions. Given \( h_{t-1} \), there is no information gained from observations up to that time, therefore \( a_t \) effectively depends only on \( \{ h_{t-1}, o_t, a_{t-1} \} \). We recall the terminal value function is \( V^L_T(h_{T-1}, a_{T-1}, h_T) := 1_{\{h_T = h_{goal}\}} \) and note that the optimal value function for the Lag-1 formulation for earlier times is computed iteratively according to

\[ V^L_t(h_{t-1}, a_{t-1}, o_t) := \max_{a_t} \mathbb{E}[V^L_{t+1}(h_t, a_t, o_{t+1}) | h_{t-1}, o_t] \]
\[ = \max_{a_t} \sum_{h_t, o_{t+1}} P_{a_{t-1}, a_{t+1}}(h_t, o_{t+1} | h_{t-1}, o_t) V^L_{t+1}(h_t, a_t, o_{t+1}) \]
for $t \in \{1, \ldots, T-1\}$ and where $P_{a_{t-1}^t}$ recognizes the implicit dependency of the conditional PMF on the actions $a_t$ and $a_{t-1}$. These probabilities can be calculated explicitly using standard manipulations. In particular, we have

$$P_{a_{t-1}^t}(h_t, a_{t+1} | h_{t-1}, o_t) = \frac{P_{a_{t-1}^t}(h_t, a_t, o_{t+1} | h_{t-1})}{P_{a_{t-1}^t}(o_t | h_{t-1})}$$

$$= \frac{\sum_{h_{t+1}} P_{a_{t-1}^t}(h_t, a_t, h_{t+1}, o_{t+1} | h_{t-1})}{\sum_{h_t} P_{a_{t-1}^t}(h_t, o_t | h_{t-1})}$$

$$= \frac{P_{h_{t-1}^t} h_{t} B_{t o_t} \sum_{h_{t+1}} P_{h_{t+1}^t} h_{t+1} B_{t+1 o_{t+1}}}{\sum_{h_t} P_{h_{t-1}^t} h_{t} B_{t o_t}}$$

(57)

where the dependence of the various quantities in (57) on the actions $a_{t-1}$ and $a_t$ is implicit and should24 be clear from the context. We can calculate $V_0^{L_1}$ in a similar fashion by noting that

$$V_0^{L_1}(o_0) := \max_{a_0} \mathbb{E}[V_1^{L_1}(h_0, o_0, o_1) | \mathcal{F}_0]$$

$$= \max_{a_0} \sum_{h_0, o_0} P_{a_0}(h_0, o_1 | o_0) V_1^{L_1}(h_0, a_0, o_1)$$

where $P_{a_0}(h_0, o_1 | o_0)$ can be calculated as in (57) but with $P_{h_{t-1}^t} h_{t} (a_{t-1})$ replaced by $P(h_0)$.

B.2. Computing the Value Function for the Lag-2 Approximation

The Lag-2 formulation corresponds to the relaxed problem in which the decision $a_t$ is based on knowledge of $h_{t-2}$ as well as the observation path $o_{t-1}$ and all past actions. Similar to the Lag-1 formulation, $a_t$ therefore depends only on $\{h_{t-2}, a_{t-2}, o_{t-1}, a_{t-1}, o_t\}$, since the history of actions and observations prior to $t-2$ is irrelevant after observing $h_{t-2}$. Again the terminal value function is $V_t^{L_2}(h_{T-2}, a_{T-2}, o_{T-1}, a_{T-1}, h_T) := 1_{\{h_T = h_{\text{goal}}\}}$ and the optimal value function, $V_t^{L_2}$, at earlier times is computed iteratively according to

$$V_t^{L_2}(h_{t-2}, a_{t-2}, o_{t-1}, a_{t-1}, o_t) := \max_{a_t} \mathbb{E}[V_{t+1}^{L_2}(h_{t-2}, a_{t-1}, o_t, a_t, o_{t+1}) | h_{t-2}, \mathcal{F}_t]$$

(58)

$$= \max_{a_t} \sum_{h_{t+1}} P_{a_{t-2}^t}(h_{t-1}, o_{t+1} | h_{t-2}, o_{t-1}, o_t) V_t^{L_2}$$

for $t \in \{2, \ldots, T-1\}$ and where we use $P_{a_{t-2}^t}$ to denote a probability conditional on inputs $\{a_{t-2}, a_{t-1}, o_t\}$. For the sake of clarity, we have omitted the arguments $(h_{t-1}, a_{t-1}, o_t, a_{t-1}, o_t)$ of $V_t^{L_2}$ from the r.h.s. of (58).

We note it is straightforward to calculate $P_{a_{t-2}^t}$ using standard probabilistic arguments. Specifically, we have

$$P_{a_{t-2}^t}(h_{t-1}, o_{t+1} | h_{t-2}, a_{t-1}, o_t) = \frac{P_{a_{t-2}^t}(h_{t-1}, o_t, o_{t+1} | h_{t-2})}{P_{a_{t-2}^t}(o_t | h_{t-2})}$$

$$= \frac{\sum_{h_{t+1}, a_{t+1}} P_{a_{t-2}^t}(h_{t-1}, o_{t-1}, h_t, o_t, h_{t+1}, o_{t+1} | h_{t-2})}{\sum_{h_{t+1}, a_{t+1}} P_{a_{t-2}^t}(h_{t-1}, o_{t-1}, h_t, o_t | h_{t-2})}$$

$$= \frac{P B_{t-2} \sum_{h_{t+1}, a_{t+1}} P B_{t-1} P B_t}{\sum_{h_{t+1}, a_{t+1}} P B_{t-2} P B_{t-1}}$$

(59)

---

24In this Appendix we will often not write the transition and observation probabilities explicitly as a function of the relevant actions but this dependence is there nonetheless.
where we denote by $PB_t \equiv P_{h_{t+1}}(a_t)B_{h_{t+1},a_{t+1}}(a_t)$. A slightly different calculation is required for each of $V_0^{L_2}$ and $V_1^{L_2}$ as there is no hidden state information available at times $t = 0$ and $t = 1$. For $t = 1$ we have

$$V_1^{L_2}(a_0, a_0) := \max_{a_1} \mathbb{E}[V_2^{L_2}(h_0, a_0, a_1, a_2) \mid \mathcal{F}_1] = \max_{a_1} \sum_{h_0, o_2} P_{a_0, a_1}(h_0, o_1) V_2^{L_2}(h_0, a_0, a_1, a_2)$$

where $P_{a_0, a_1}(h_0, o_1)$ is calculated as in (59) but where we replace $P_{h_{t-2}h_{t-1}}(a_{t-2})$ in $PB_{t-2}$ with the initial distribution $P(h_0)$. Similarly, at $t = 0$ we have

$$V_0^{L_2}(a_0) := \max_{a_0} \mathbb{E}[V_1^{L_2}(a_0, a_0) \mid \mathcal{F}_0] = \max_{a_0} \sum_{a_1} \mathbb{P}_{a_0}(o_1 \mid o_0) V_1^{L_2}(o_1, a_0)$$

and where

$$\mathbb{P}_{a_0}(o_1 \mid o_0) = \frac{\sum_{h_0, h_1} P(h_0)B_{h_0, o_0}P_{B_0}}{\sum_{h_0} P(h_0)B_{h_0, o_0}}.$$

C. Proof of Proposition 6.1

Recall that we define a supersolution as an approximate value function $\vartheta$ that, for all possible observation paths $o_{0:t}$ and times $t$, satisfies

$$\vartheta_t(o_{0:t}) \geq \max_{a_t \in \mathcal{A}} \{r_t(\pi_t, a_t) + \mathbb{E}[\vartheta_{t+1}(o_{0:t+1}) \mid \mathcal{F}_t]\}.$$  \hspace{1cm} (60)

Because a reward is only obtainable at time $T$ in the maze example and (60) holds with equality at that time for all of our approximate value functions, we only need to show (60) for times $t < T$ when it reduces to

$$\vartheta_t(o_{0:t}) \geq \max_{a_t \in \mathcal{A}} \mathbb{E}[\vartheta_{t+1}(o_{0:t+1}) \mid \mathcal{F}_t].$$

We also note that the belief state at time $t+1$, $\pi_{t+1}$, given the current belief state, $\pi_t$, and the next observation, $o_{t+1} = o \in \mathcal{O}$, can be computed according to

$$\pi_{t+1}(h') = \frac{\sum_h \pi_t(h)P_{h', o}B_{h', o}}{\sigma(o, \pi_t)} \hspace{1cm} (61)$$

where

$$\sigma(o, \pi_t) := P(o_{t+1} \mid \pi_t) = \sum_{h, h'} \pi_t(h)P_{h', o}B_{h', o}$$

for $t \in \{0, \ldots, T - 1\}$ where once again, the dependence of the various quantities on actions is implicit rather than explicit.
Proof that the MDP Approximation (23) is a Supersolution

We have that

\[ \hat{V}_t(o_{0:t}) := \sum_h \pi_t(h) V^\text{MDP}_t(h) \]

\[ = (a) \sum_h \pi_t(h) \max_{a_t} \sum_{h'} P_{hh'}(a_t) V^\text{MDP}_{t+1}(h') \]

\[ \geq (b) \max_{a_t} \sum_h \pi_t(h) \sum_{h'} P_{hh'}(a_t) V^\text{MDP}_{t+1}(h') \]

\[ = (c) \max_{a_t} \sum_{h', o_{t+1}} \left[ \sum_h \pi_t(h) P_{hh'}(a_t) \right] B_{h', o_{t+1}}(a_t) V^\text{MDP}_{t+1}(h') \]

\[ = (d) \max_{a_t} \sum_{h', o_{t+1}} P(o_{t+1} | a_t) \pi_{t+1}(h') V^\text{MDP}_{t+1}(h') \]

\[ = (e) \max_{a_t} \mathbb{E}[\hat{V}_{t+1}(o_{0:t+1}) | \mathcal{F}_t] \]

where (a) results from using the definition of \( V^\text{MDP}_t(h) \), which is calculated recursively using the MDP formulation from equation (24). Inequality (b) results from Jensen’s inequality after changing the order of \( \max_{a_t} \). In equality (c) we have re-ordered the terms and included the factor \( \sum_{o_{t+1}} B_{h', o_{t+1}} = 1 \). Equality (d) follows from (61) while we have used the definition of \( \hat{V}_{t+1}(o_{0:t+1}) \) to obtain (e).

Proof that the QMDP Approximation (26) is a Supersolution

The proof for the QMDP approximation follows a similar argument. We have that

\[ \hat{V}_t(o_{0:t}) := \max_{a_t} \sum_h \pi_t(h) V^Q_t(h, a_t) \]

\[ = \max_{a_t} \sum_{h, h'} \pi_t(h) P_{hh'}(a_t) V^\text{MDP}_{t+1}(h') \]

\[ = (a) \max_{a_t} \sum_{h, h'} P(o_{t+1} | a_t) \pi_{t+1}(h') V^\text{MDP}_{t+1}(h') \]

\[ \geq (b) \max_{a_t} \sum_{h, h'} P(o_{t+1} | a_t) \pi_{t+1}(h') \max_{a'_t} V^Q_{t+1}(h', a'_t) \]

\[ = (c) \max_{a_t} \sum_{h, h'} P(o_{t+1} | a_t) \pi_{t+1}(h') \max_{a'_t} \sum_{h''} \pi_{t+1}(h'') V^Q_{t+1}(h'', a'_t) \]

\[ = (d) \max_{a_t} \mathbb{E}[\hat{V}_{t+1}(o_{0:t+1}) | \mathcal{F}_t] \]

where (a) follows from following steps (b) to (d) of the MDP proof above and (b) then follows from the definition of both \( V^\text{MDP}_{t+1} \) and \( V^Q_{t+1} \). Inequality (c) follows from Jensen’s inequality after changing the order of \( \max_{a'_t} \) and the marginalization of \( h' \). Finally (d) follows from the definition of \( \hat{V}_{t+1}(o_{0:t+1}) \).
Proof that the Lag-1 Approximation is a Supersolution

Rather than writing out probabilities explicitly, we will write the relevant quantities as expectations and use $\mathbb{E}_x$ to denote an expectation taken over the random variable, $x$. The Lag-1 approximation satisfies

$$\tilde{V}_t(o_{0:t}) := \max_{a_t} \mathbb{E}_{h_t, o_{t+1}} [V_{t+1}^{L_1}(h_t, a_t, o_{t+1}) | \mathcal{F}_t]$$

(a) $\max_{a_t} \mathbb{E}_{h_t, o_{t+1}} [\max_{a_{t+1}} \mathbb{E}_{h_{t+1}, o_{t+2}} [V_{t+2}^{L_1}(h_{t+1}, a_{t+1}, o_{t+2}) | h_t, \mathcal{F}_{t+1}] | \mathcal{F}_t]$

(b) $\max_{a_t} \mathbb{E}_{o_{t+1}} [\max_{a_{t+1}} \mathbb{E}_{h_{t+1}, o_{t+2}} [V_{t+2}^{L_1}(h_{t+1}, a_{t+1}, o_{t+2}) | h_t, \mathcal{F}_{t+1}] | \mathcal{F}_t]$

(c) $\max_{a_t} \mathbb{E}_{o_{t+1}} [\max_{a_{t+1}} \mathbb{E}_{h_{t+1}, o_{t+2}} [V_{t+2}^{L_1}(h_{t+1}, a_{t+1}, o_{t+2}) | \mathcal{F}_{t+1}] | \mathcal{F}_t]$

(d) $\max_{a_t} \mathbb{E}[\tilde{V}_{t+1}(o_{0:t+1}) | \mathcal{F}_t]$

where (a) follows from using the definition of $V_{t+1}^{L_1}$ in (27), inequality (b) follows from Jensen’s inequality after changing the order of $\max_{a_{t+1}}$ and the marginalization of $h_t$. Note that after bringing the expectation w.r.t. $h_t$ inside the maximization, we are still conditioning on the value of $o_{t+1}$ outside, i.e. the expectation is conditional on $\mathcal{F}_{t+1}$. Equality (c) follows from applying the tower property to the nested expectations and finally (d) follows from the definition of the approximate value function, $\tilde{V}_{t+1}(o_{0:t+1})$.

Proof that the Lag-2 Approximation is a Supersolution

The proof for the Lag-2 approximation is similar but slightly more complicated than the corresponding Lag-1 proof above. We have

$$\tilde{V}_t(o_{0:t}) = \max_{a_t} \mathbb{E} [\max_{a_{t+1}} \max_{a_{t+2}} \mathbb{E}_{h_t, o_{t+3}} [V_{t+3}^{L_2}(h_t, a_t, o_{t+1}, a_{t+1}, o_{t+2}) | \mathcal{F}_{t+1}] | \mathcal{F}_t]$$

(a) $\max_{a_t} \mathbb{E} [\max_{a_{t+1}} \max_{a_{t+2}} \mathbb{E}_{h_{t+1}, o_{t+3}} [V_{t+3}^{L_2}(h_{t+1}, a_{t+1}, o_{t+2}, a_{t+2}, o_{t+3}) | h_t, \mathcal{F}_{t+2}, \mathcal{F}_{t+1}] | \mathcal{F}_t]$

(b) $\max_{a_t} \mathbb{E} [\max_{a_{t+2}} \max_{a_{t+1}} \mathbb{E}_{h_t, o_{t+3}} [V_{t+3}^{L_2}(h_t, a_{t+1}, o_{t+2}, a_{t+2}, o_{t+3}) | h_t, \mathcal{F}_{t+2}] | \mathcal{F}_{t+1}] | \mathcal{F}_t]$

(c) $\max_{a_t} \mathbb{E} [\max_{a_{t+2}} \max_{a_{t+1}} \mathbb{E}_{h_{t+1}, o_{t+3}} [V_{t+3}^{L_2}(h_{t+1}, a_{t+1}, o_{t+2}, a_{t+2}, o_{t+3}) | \mathcal{F}_{t+2}] | \mathcal{F}_{t+1}] | \mathcal{F}_t]$

(d) $\max_{a_t} \mathbb{E}[\tilde{V}_{t+1}(o_{0:t+1}) | \mathcal{F}_t]$

where (a) follows from the definition of $V_{t+3}^{L_2}$ in (58), inequality (b) follows from Jensen’s inequality after changing the order of the $\max_{a_{t+2}}$ operator and the marginalization of $h_t$, (c) follows from the tower property and (d) uses the definition of the approximate value function $\tilde{V}_{t+1}(o_{0:t+1})$. □

D. Dropping the Absolute Continuity Condition

D.1. Dual Bounds Without Absolute Continuity of $\mathbb{P}$ w.r.t. $\bar{\mathbb{P}}$

We explain here why we do not require $\mathbb{P}$, the probability measure for the controlled formulation, to be absolutely continuous w.r.t. $\bar{\mathbb{P}}$ (the probability measure for the original uncontrolled formulation), when the penalties in (9) are constructed from supersolutions. This result was originally shown in [8] but we outline the details here in the finite horizon case for the sake of completeness. We will work with the PI relaxation but it should be clear that the result is general and holds for MDP’s (and therefore POMDP’s) and general
We therefore assume the penalty function, \( c_t := \mathbb{E}[\vartheta_{t+1}(o_{0:t+1}) \mid \mathcal{F}_t] - \vartheta_{t+1}(o_{0:t+1}) \), is such that \( \vartheta_t \) is a supersolution satisfying\(^{25}\) \( \vartheta_{T+1} \equiv 0 \). From Definition 6.1, it follows that for each \( t \in \{0, \ldots, T\} \) we have
\[
\vartheta_t(o_{0:t}) \geq r_t(\pi_t, a_t) + \mathbb{E}[\vartheta_{t+1}(o_{0:t+1}) \mid \mathcal{F}_t] \quad \forall a_t \in \mathcal{A}.
\] (62)

Subtracting \( \vartheta_t(o_{0:t}) \) from both sides of (62), and summing over \( t \), we obtain
\[
0 \geq \sum_{t=0}^{T} \left\{ r_t(\pi_t, a_t) + \mathbb{E}[\vartheta_{t+1}(o_{0:t+1}) \mid \mathcal{F}_t] - \vartheta_t(o_{0:t}) \right\}
= \sum_{t=0}^{T} \left( r_t(\pi_t, a_t) + c_t \right) - \vartheta_0(o_0).
\] (63)

We now obtain
\[
V_0^* - \vartheta_0 = \max_{\mu \in \mathcal{G}_c} V_0^\mu - \vartheta_0
\leq (a) \max_{\mu \in \mathcal{G}_c} \mathbb{E} \left[ \sum_{t=0}^{T} (r_t + c_t) - \vartheta_0 \mid \mathcal{F}_0 \right]
\leq (b) \max_{\mu \in \mathcal{G}_c} \mathbb{E} \left[ \sum_{t=0}^{T} \Phi_t(r_t + c_t) - \vartheta_0 \mid \mathcal{F}_0 \right]
\leq (c) \mathbb{E} \left[ \max_{\alpha \in T-1} \sum_{t=0}^{T} \Phi_t(r_t + c_t) \mid \mathcal{F}_0 \right] - \vartheta_0.
\] (64)

where we have omitted the arguments of \( r_t \) and \( \vartheta_0 \) for the sake of clarity. Equality (a) follows since \( \mathbb{E}[\sum_{t=0}^{T} c_t \mid \mathcal{F}_0] = 0 \) for any \( \mathcal{F} \)-adapted policy and since \( \vartheta_0(o_0) \) is \( \mathcal{F}_0 \)-adapted. In order to establish inequality (b), we first note that (63) implies the random quantity inside the expectation in (64) is non-negative w.p. 1. The inequality then follows\(^{26}\) for any probability measure, \( \hat{\mathbb{P}} \), regardless of whether or not it is absolutely continuous w.r.t. \( \mathbb{P} \). Inequality (c) follows from the usual weak duality argument. We also note that \( \Phi_0 \equiv 1 \) which explains why there is no RN term multiplying \( \vartheta_0(o_0) \).

We can now add \( \vartheta_0(o_0) \) across both sides of (65) to establish the result, i.e. weak duality continues to hold even if the probability measure, \( \mathbb{P} \), is not absolutely continuous w.r.t. \( \hat{\mathbb{P}} \) as long as the penalty is constructed from a supersolution. It is also interesting to note that inequality (b) will in fact be an equality if \( \hat{\mathbb{P}} \) is the measure induced by following an optimal policy for the primal problem since in that case \( \mathbb{P} \) and \( \hat{\mathbb{P}} \) will coincide. Strong duality will then also continue to hold. In particular, (c) will then also be an equality if \( \vartheta_t \) coincides with the optimal value function, \( V_0^* \), which is itself a supersolution.

### D.2. The Duality Gap and Variance Reduction

As pointed out in [8], the weak duality result of (65) (allowing for \( \mathbb{P} \) to not be absolutely continuous w.r.t. \( \hat{\mathbb{P}} \)) opens up the possibility of estimating the duality gap, \( V_0^{upper} - V_0^{lower} \), very efficiently. Specifically, suppose

\(^{25}\)There is no difficulty in assuming \( \vartheta_{T+1} \equiv 0 \) since \( \vartheta_t(o_{0:t}) \) represents an approximate value function and all of our approximate value functions naturally satisfy this assumption.

\(^{26}\)This result was stated as Lemma A.1 in [8] and we state it here for the sake of completeness. Consider a measurable space \((\Omega, \Sigma)\) and two probability measures \( P \) and \( Q \). Let \( \phi \) represent the Radon-Nikodym derivative of the absolutely continuous component of \( P \) with respect to \( Q \). If \( Y = Y(\omega) \) is a bounded random variable such that \( Y(\omega) \leq 0 \) for all \( \omega \notin \Omega_Q := \{ \omega \in \Omega : Q(\omega) > 0 \} \), then \( \mathbb{E}^P[Y] \leq \mathbb{E}^Q[\phi Y] \).
we have a good candidate \( \mathcal{F} \)-adapted policy, \( \mu \), and let \( \mathbb{P} \) be the probability measure induced by following this policy. If we set \( V_{0}^{\text{lower}} \) to be the value of this policy, we then have

\[
V_{0}^{\text{lower}} = \mathbb{E}\left[ \sum_{t=0}^{T} (r_{t}(\pi_{t}, \mu_{t}) + c_{t}) \mid \mathcal{F}_{0} \right]
\]

\[
= \mathbb{E}\left[ \sum_{t=0}^{T} \Phi_{t}(\mu)(r_{t}(\pi_{t}, \mu_{t}) + c_{t}) \mid \mathcal{F}_{0} \right]
\]

(66)

where the \( c_{t} \)'s now play the role of (action-dependent) control variates and where \( \Phi_{t} = \Phi_{t}(\mu) = 1 \) for all \( t \) in (66) because \( \mathbb{P} \) and \( \mathbb{P} \) coincide when the policy \( \mu \) is followed. We can use this same \( \mathbb{P} \) to estimate an upper bound

\[
V_{0}^{\text{upper}} = \mathbb{E}\left[ \max_{a_{0:T-1}} \sum_{t=0}^{T} \Phi_{t}(a_{0:t-1}; \mu)(r_{t}(\pi_{t}, \mu_{t}) + c_{t}) \mid \mathcal{F}_{0} \right]
\]

(67)

as long as \( \vartheta_{t} \) is constructed from a supersolution and where (67) now explicitly recognizes the dependence of the \( \Phi_{t} \)'s on \( a_{0:t-1} \) and \( \mu \). Since both lower and upper bounds (66) and (67) are simulated using the same measure, \( \mathbb{P} \), we may as well use the same set of paths to estimate each bound. This has an obvious computational advantage since the \( r_{t}(\pi_{t}, \mu_{t}) \)'s and \( c_{t} \)'s that were computed along each sample path for estimating (66) can now be re-used\(^{27} \) on the corresponding inner problem in (67).

There is a further benefit to this proposal, however. Because the actions of the policy, \( \mu \), are feasible for the inner problem in (67), it is clear the term inside the expectation in (66) will always be less than or equal to the optimal objective of the inner problem in (67) along each simulated path. In fact the difference, \( D \), between the two terms satisfies

\[
0 \leq D \approx \max_{a_{0:T-1}} \sum_{t=0}^{T} \Phi_{t}(a_{0:t-1}; \mu)(r_{t}(\pi_{t}, \mu_{t}) + c_{t}) - \sum_{t=0}^{T}(r_{t}(\pi_{t}, \mu_{t}) + c_{t}) \quad \mathbb{P} \text{ a.s.}
\]

(68)

and provides an unbiased estimate of the duality gap, \( V_{0}^{\text{upper}} - V_{0}^{\text{lower}} \). Finally, we expect that the variance of the random variable, \( D \), should be very small due to a strong positive correlation between each of the terms in (68). As a result, we anticipate that very few sample paths should be required to estimate the duality gap to a given desired accuracy as long as \( \mu \) is sufficiently close to optimal. This approach to evaluating a strategy, i.e. by estimating the duality gap, requires very little work over and beyond the work required to estimate \( V_{0}^{\text{lower}} \). And because the variance of \( D \) is often extremely small, we generally only need to estimate the duality gap and solve the inner problem on a small subset of the paths that were used to estimate \( V_{0}^{\text{lower}} \).

E. Technical Details for the Mammography Screening Application

The main difference between our mammography screening application and the usual POMDP framework is the timing of the observations. In particular, in the mammography screening application an observation occurs immediately after an action is taken and is therefore a function of the current hidden state and the \textit{current} action. In contrast, in the usual POMDP setting, an observation is a function of the current hidden state and the action from the previous period. It is therefore the case that the filtering approach for the belief update is different than the standard update as given in (61). The belief update for these new dynamics is

\(^{27}\text{Recall that with an uncontrolled formulation, the states, } \pi_{t}, \text{ are action-independent and that the actions, } a_{0:T}, \text{ only enter the objective of the inner problems via the RN terms, } \Phi_{t}, \text{ for } t = 0, \ldots, T.\)
then given by

$$\pi_{t+1}(h') \propto \sum_h \pi_t(h) B_{ho_t}(a_t) P_{hh'}(a_t)$$  \hspace{1cm} (69)$$

where it is straightforward to see (by summing (69) over \(h'\)) that the constant of proportionality is given by

$$\sum_h \pi_t(h) B_{ho_t}(a_t) := \mathbb{P}_{a_t}(o_t \mid \pi_t)$$  \hspace{1cm} (70)$$

for \(t \in \{0, \ldots, T-1\}\). It is worth emphasizing that the belief state for time \(t+1\) is a function of the time \(t\) action and observation.

These alternative dynamics also impact the form of the RN derivatives, which we now describe in the case of the hidden state formulation of the problem (as opposed to the belief state formulation). We first note that the RN derivatives must satisfy

$$E[r_t(h_t, a_t) \mid \mathcal{F}_0] = \mathbb{E}[\Phi_t r_t(h_t, a_t) \mid \mathcal{F}_0]$$

for all \(t \in \{0, \ldots, T\}\) and where \(r_t(h_t, a_t)\) was defined in (33). Writing the expectations explicitly we obtain

$$\sum_{o_{0:t-1}} r_t(h_t, a_t) \mathbb{P}_{a_{0:t-1}}(o_{0:t-1}, h_{0:t}) = \sum_{o_{0:t-1}} \Phi_t r_t(h_t, a_t) \tilde{\mathbb{P}}(o_{0:t-1}, h_{0:t})$$

so that the RN derivatives satisfy

$$\Phi_t := \frac{\mathbb{P}_{a_{0:t-1}}(o_{0:t-1}, h_{0:t})}{\tilde{\mathbb{P}}(o_{0:t-1}, h_{0:t})}. \hspace{1cm} (71)$$

The numerator and denominator of (71) satisfy

$$\mathbb{P}_{a_{0:t-1}}(o_{0:t-1}, h_{0:t}) = \pi_0(h_0) \prod_{s=0}^{t-1} \mathbb{P}_{a_s}(o_s \mid h_s) \mathbb{P}_{a_s}(h_{s+1} \mid h_s)$$

$$\tilde{\mathbb{P}}(o_{0:t-1}, h_{0:t}) = \pi_0(h_0) \prod_{s=0}^{t-1} \tilde{\mathbb{P}}(o_s \mid h_s) \tilde{\mathbb{P}}(h_{s+1} \mid h_s)$$

from which it immediately follows that the factorized form of the RN derivatives are given by

$$\phi(i, j, k, a) := \frac{B_{ik}(a) P_{ij}(a)}{E_{ik} Q_{ij}} \hspace{1cm} (72)$$

$$\Phi_t(h_{0:t}, o_{0:t-1}, a_{0:t-1}) := \prod_{s=0}^{t-1} \phi(h_s, h_{s+1}, o_s, a_s). \hspace{1cm} (73)$$