Information Relaxation Bounds for Partially Observed Markov Decision Processes

Martin B. Haugh
Imperial College Business School
Imperial College London
martin.b.haugh@gmail.com

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

First Draft: June 18, 2017
Second draft: August 8, 2018
This draft: April 15, 2019

Abstract

Partially observed Markov decision processes (POMDPs) are an important class of control problems that are ubiquitous in a wide range of fields. Unfortunately these problems are generally intractable and so in general we must be satisfied with sub-optimal policies. But how do we evaluate the quality of these 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 non-anticipativity 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 the belief-state formulation of a POMDP is an MDP and so the previously developed results for MDPs also apply to POMDPs. Under the belief-state formulation, we use recently developed change-of-measure arguments to solve the so-called inner problems and we use standard filtering arguments to identify the appropriate Radon-Nikodym derivatives. We also show, however, that dual bounds can also be constructed without resorting to the belief-state formulation. In this case, change-of-measure arguments are required for the evaluation of so-called dual feasible penalties rather than for the solution of the inner problems. We compare dual bounds for both formulations and argue that in general the belief-state formulation provides tighter bounds. The second main contribution of this paper is to show that several value function approximations for POMDPs are in fact *supersolutions*. This is of interest because it can be particularly advantageous to construct penalties from supersolutions since absolute continuity (of the change-of-measure) is no longer required and so significant variance reduction can be achieved when estimating the duality gap directly. Dual bounds constructed from supersolution based penalties are also guaranteed to provide tighter bounds than the bounds provided by the supersolutions themselves. We use applications from robotic navigation and telecommunication to demonstrate our results.

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

1. Introduction

Partially observed Markov decision processes (POMDPs) are an important class of control problems with wide-ranging applications in fields as diverse as engineering, 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 [32]. The information relaxation approach proceeds in two steps: (i) relax the non-anticipativity 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 (AVF) 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 AVF, 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 by working with the belief-state formulation of the POMDP and so the results established for MDPs therefore also apply to POMDPs. Under the belief-state formulation, we use the recently developed change-of-measure arguments of Brown and Haugh [8] (hereafter BH) to solve the so-called inner problems and we use standard filtering arguments to identify the appropriate Radon-Nikodym derivatives. We also show that information relaxation bounds can also be constructed without resorting to the belief-state formulation of the POMDP. In particular, we can still construct these bounds if we work with the *non*-belief-state formulation of the POMDP, i.e. with the explicit dynamics for the hidden state transitions and observations. If we work with the non-belief-state formulation, however, 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 problems (see (21)). We show that this obstacle can be overcome by again 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.

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. It is perhaps worth emphasizing, however, that the motivation for using a change-of-measure depends on the POMDP formulation that we work with. With the belief-state formulation evaluating the dual penalties is easy but solving the inner

problems is hard. In contrast, when we work with the explicit dynamics for the hidden state transitions and observations, then evaluating the dual penalties is hard but solving the inner problems is easy.

We compare the perfect-information (PI) relaxation bounds that arise from the belief-state and non-belief-state formulation of the POMDP. We argue that the two bounds will be identical under the *same* absolutely continuous change-of-measure. In practice, however, we never use the same change-of-measure for the PI and BSPI bounds although they will be closely related. In particular, when calculating the belief-state bound we can use a filtered version of the change-of-measure that we used for the non-belief-state formulation. In that case we argue that the resulting information relaxation bound for the belief-state formulation should be tighter than the information relaxation bound for the non-belief-state formulation.

The second main contribution of this paper is to show that several standard value function approximations for POMDPs are in fact *supersolutions*. Supersolutions are feasible solutions for the linear programming formulation of an MDP and are therefore upper bounds (in the case of a maximization problem) on the unknown optimal value function. Desai, Farias and Moallemi [15] and BH showed information relaxation bounds constructed from supersolution based penalties are also guaranteed to provide tighter bounds than the bounds provided by the supersolutions themselves. A further advantage of constructing penalties from supersolutions is that absolute continuity (of the change-of-measure) is no longer required and so significant variance reduction can be achieved when estimating the duality gap directly. These advantages were identified by BH although perhaps not emphasized sufficiently. We therefore believe that the information relaxation approach is particularly valuable in the context of POMDPs. One of the standard AVFs we consider is the so-called fast informed bound update AVF [21]. We extend this approach in a natural way to construct what we call the Lag-2 AVF. We show the Lag-2 AVF is a supersolution and prove that it is a tighter upper bound than that provided by the fast informed bound update AVF.

We demonstrate our results in applications from robotic navigation and telecommunications. The robotic navigation application requires controlling the movements of a robot in a maze with the goal of reaching a desired state within a finite number of time-steps. Our telecommunications application concerns packet transmissions in a multi-access communication setting that uses the *slotted aloha* protocol. In both cases we use the aforementioned supersolutions to construct penalties for the dual bounds. We also use them to construct primal bounds by simulating the policies that are greedy with respect to them. We demonstrate the bound improvement results of BH and also show that tight duality gaps can be achieved in these applications. In particular, the duality gap can be as much as 85% smaller than the gap given by the primal bound and the corresponding supersolution. (This reduction in duality gap under-estimates the upper bound improvement since the duality gap includes the gap from the primal lower bound to the unknown optimal value function.) In our robotic navigation application, for example, we will see that the tightest duality gap, i.e. the gap between our best lower bound and our best information relaxation-based upper bound, is obtained using the Lag-2 AVF. Moreover, the duality gap is so small that we could argue that we have essentially succeeded in solving the problem.

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 is often the case with influence diagrams in the decision sciences literature. In this latter case, we suspect that the non-belief-state formulation is the more natural approach to take.

1.1. Literature Review and Paper Outline

The work of BSS and [32] follows earlier work by [18] and [31] on the pricing of high-dimensional American options. Other related work on American option pricing includes [14] and [2]. The pricing of swing options with multiple exercise opportunities is an important problem in energy markets and the information relaxation approach was soon extended to this problem via the work of [29], [33], [1], [5] and [13] 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 [35] to infinite horizon settings, the bound improvement guarantees of BH who also use change-of-measure arguments (building in part on Rogers [32]) to solve intractable inner problems. The approach has also been extended to continuous-time stochastic control by [34], and dynamic zero sum-games by [20] and [6]. Recently [4] and [3] 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 among others. Other interesting applications and developments include [27], [15], [24], [17], [19], [16] and [36].

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 [26] 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 our discrete-time, discrete-state POMDP and also discuss its belief-state formulation there. In Section 3 we review information relaxations and the change-of-measure approach of BH for solving the difficult inner problems that arise in the belief-state formulation of POMDPs. In Section 4 we consider information relaxations for the *non*-belief-state formulation and then compare information relaxation bounds from the belief-state and non-belief state formulations in Section 5. We construct several standard value function approximations for POMDPs in Section 6. We also introduce our Lag-2 AVF there and prove that all of these AVFs are in fact supersolutions. We describe our applications to robotic navigation and multiaccess communication in Sections 7 and 8, respectively. We conclude in Section 9. Derivations, proofs and various technical details including how to extend our approach to the infinite horizon setting are relegated to the appendices.

2. Discrete-Time POMDPs

We begin with the 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, π_0, on the initial hidden state, h_0, and the initial action a_0 is based on π_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

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

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

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

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 always using (1) and (2), however, we will sometimes find it more convenient to use the following alternative, but equivalent, dynamics. In particular, we assume the hidden state and observation dynamics satisfy

$$h_{t+1} = f_h(h_t, a_t, w_{t+1}), \tag{3}$$

$$o_{t+1} = f_o(h_{t+1}, a_t, v_{t+1}) \tag{4}$$

for $t = 0, 1, \ldots, T-1$ and where the v_t's and w_t's are IID $\mathrm{U}(0,1)$ 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. We will assume, however, that the final observation satisfies $o_T = h_T$ so that $r_T(h_T) = r_T(o_T)$. This is without loss of generality since the DM cannot act at time T and so there is no benefit to receiving any information at time T.

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, π_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 π_0. We define a filtration $\mathbb{F} = (\mathcal{F}_0, \ldots, \mathcal{F}_T)$ to be the filtration generated by the observations so that $\mathcal{F}_t$ is the σ-algebra generated by $o_{1:t}$. A non-anticipative policy is therefore $\mathbb{F}$-adapted. We also define $\mathcal{F} := \mathcal{F}_T$. We denote the class of all non-anticipative policies by $\mathcal{U}_{\mathbb{F}}$. The objective of the DM is to find an $\mathbb{F}$-adapted policy, μ^*, that maximizes the expected total reward. The POMDP problem is therefore to solve for

$$V_0^*(\pi_0) = \max_{\mu \in \mathcal{U}_{\mathbb{F}}} \mathbb{E}\left\{ \sum_{t=0}^{T} r_t(h_t, \mu_t) \,\middle|\, \mathcal{F}_0 \right\} \tag{5}$$

and where we acknowledge[2] a slight abuse of notation in (5) since there is no time T action μ_T.

[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.

[2] 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.

2.1. 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, π_t, which lies in the $|\mathcal{H}|$-dimensional simplex. Specifically we can equivalently write the POMDP dynamics as

$$\pi_{t+1} = f_\pi(\pi_t, a_t, u_{t+1}), \qquad t = 0, 1, \ldots, T-1 \tag{6}$$

where the u_t's are IID U(0,1) random variables and f_π is the state transition function which is only defined implicitly via the filtering[3] algorithm. We now define the filtration $\mathbb{F}^\pi = (\mathcal{F}_0^\pi, \ldots, \mathcal{F}_T^\pi)$ where $\mathcal{F}_t^\pi$ is the σ-algebra generated by $\pi_{0:t}$. We note that the filtrations $\mathbb{F}$ and $\mathbb{F}^\pi$ are not identical although they are of course related. We can also write the time t reward as a function of the belief state by setting[4] $r(\pi_t, a_t) \coloneqq \mathbb{E}[r(h_t, a_t) \mid \mathcal{F}_t^\pi]$. The analog of (5) under the belief-state formulation is then

$$V_0^*(\pi_0) = \max_{\mu \in \mathcal{U}_{\mathbb{F}^\pi}} \mathbb{E}\left\{ \sum_{t=0}^{T} r_t(\pi_t, \mu_t) \,\middle|\, \mathcal{F}_0^\pi \right\} \tag{7}$$

where we use $\mathcal{U}_{\mathbb{F}^\pi}$ to denote the class of $\mathbb{F}^\pi$-adapted policies. The advantage of formulating the POMDP via the belief-state is that the problem becomes an MDP albeit a potentially high-dimensional one.

3. A Review of Information Relaxations

We now briefly describe the information relaxation approach for obtaining dual bounds. Because this theory has been developed for MDPs, we will focus on the belief-state formulation of (7). Solving (7) 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 $V_0^{\text{lower}} \le V_0^*(\pi_0) \le V_0^{\text{upper}}$ with $V_0^{\text{lower}} \approx V_0^{\text{upper}}$ 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 $\mathbb{F}^\pi$-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 and our development will follow that of BSS which can be consulted for additional details and proofs. An information relaxation $\mathbb{G}^\pi$ of the filtration $\mathbb{F}^\pi$ is a filtration $\mathbb{G}^\pi = (\mathcal{G}_0^\pi, \mathcal{G}_1^\pi, \ldots, \mathcal{G}_T^\pi)$, where $\mathcal{F}_t^\pi \subseteq \mathcal{G}_t^\pi$ for each t. We denote by $\mathcal{U}_{\mathbb{G}^\pi}$ the set of $\mathbb{G}^\pi$-adapted policies. Then, $\mathcal{U}_{\mathbb{F}^\pi} \subseteq \mathcal{U}_{\mathbb{G}^\pi}$. Note that a $\mathbb{G}^\pi$-adapted policy is generally not feasible for the original *primal* problem in (7) as such a policy can take advantage of information that is not available to an $\mathbb{F}^\pi$-adapted policy.

Before proceeding we also need the concept of dual penalties. Penalties, like rewards, depend on states

[3] The filtering algorithm takes π_t, a_t and o_{t+1} (which is a function of π_t, a_t and u_{t+1}) as inputs and outputs π_{t+1}. It might therefore seem more natural to write $\pi_{t+1} = f_\pi(\pi_t, a_t, o_{t+1})$ in (6) but the information relaxation approach requires the uncertainty to be exogenous, e.g. via a sequence of IID U(0,1) random variables u_t, rather than endogenous which would be the case if we took the o_t's to be the basic source of uncertainty. There is no loss of generality in working with the u_t's, however, since we can then generate each o_{t+1} from u_{t+1} (given π_t and a_t) via the inverse transform approach.

[4] 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.

and actions and are incurred in each period. Specifically, for each t, we define a dual penalty, c_t, according to

$$c_t := \mathbb{E}[\vartheta_{t+1}(\pi_{t+1}) \mid \mathcal{F}_t^{\pi}] - \mathbb{E}[\vartheta_{t+1}(\pi_{t+1}) \mid \mathcal{G}_t^{\pi}] \tag{8}$$

where $\vartheta_{t+1}(\pi_{t+1})$ is[5] a bounded real-valued function of the time $t+1$ state π_{t+1}. In practice we will take $\vartheta_{t+1}(\pi_{t+1})$ to be an approximation to the time $t+1$ optimal value function, i.e. an AVF. It is straightforward to see that $\mathbb{E}[c_t \,|\, \mathcal{F}_t^{\pi}] = 0$ for all t and any $\mathbb{F}^{\pi}$-adapted policy. (In general this is not the case for a $\mathbb{G}^{\pi}$-adapted policy.) This in turn implies $\mathbb{E}[\sum_{t=0}^{T} c_t \,|\, \mathcal{F}_0^{\pi}] = 0$ for any $\mathbb{F}^{\pi}$-adapted policy. Beginning with (7) we now obtain

$$\begin{aligned} V_0^*(\pi_0) = \max_{\mu \in \mathcal{U}_{\mathbb{F}^{\pi}}} \mathbb{E}\left[\sum_{t=0}^{T} r_t(\pi_t, \mu_t) \,\middle|\, \mathcal{F}_0^{\pi}\right] &= \max_{\mu \in \mathcal{U}_{\mathbb{F}^{\pi}}} \mathbb{E}\left[\sum_{t=0}^{T} r_t(\pi_t, \mu_t) + c_t \,\middle|\, \mathcal{F}_0^{\pi}\right] \\ &\leq \max_{\mu \in \mathcal{U}_{\mathbb{G}^{\pi}}} \mathbb{E}\left[\sum_{t=0}^{T} r_t(\pi_t, \mu_t) + c_t \,\middle|\, \mathcal{F}_0^{\pi}\right]. \end{aligned} \tag{9}$$

BSS also showed that *strong duality* holds. Specifically, if we could take $\vartheta_{t+1}(\pi_{t+1}) = V_{t+1}^*(\pi_{t+1})$, i.e. use the (unknown) optimal value function as our generating function in (8), then we would have equality in (9). Indeed a simple inductive proof that works backwards from time T establishes strong duality and also shows that equality holds in (9) *almost surely*. That is, if we could use the optimal value function V_t^* to construct the dual penalties then the optimal value of the expression inside the expectation on the r.h.s. of (9) would equal $V_0^*(\pi_0)$ almost surely. This result has two implications when we have a good approximation, $\widetilde{V}_t$, to V_t^* and we take $\vartheta_{t+1}(\pi_{t+1}) = \widetilde{V}_{t+1}(\pi_{t+1})$. First it suggests that (9) 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^{upper} to any given accuracy.

We can use (9) to construct upper bounds on $V_0^*(\pi_0)$ for general information relaxations $\mathbb{G}^{\pi}$ 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. We will actually refer to this relaxation as the belief-state perfect information relaxation (BSPI) as it is the perfect information relaxation for the belief-state formulation of the problem.

3.1. The BSPI Relaxation

The BSPI information relaxation is given by the filtration $\mathbb{B}^{\pi} := (\mathcal{B}_0^{\pi}, \ldots, \mathcal{B}_T^{\pi})$ where $\mathcal{B}_0^{\pi} = \mathcal{B}_1^{\pi} = \cdots = \mathcal{B}_T^{\pi} := \sigma(u_{1:T})$ where the u_t's are as in (6). The DM therefore gets to observe $u_{1:T}$ at time 0 under the BSPI relaxation. Moreover, knowledge of $u_{1:T}$ implies knowledge of the belief states $\pi_{0:T}$ corresponding to all possible action sequences, which implies that $\mathcal{F}_t^{\pi} \subseteq \mathcal{B}_t^{\pi}$ for all t so that $\mathbb{B}^{\pi}$ is indeed a relaxation of $\mathbb{F}^{\pi}$. The upper bound of (9) now yields

$$V_0^*(\pi_0) \leq \mathbb{E}\left[\max_{a_{0:T-1}} \sum_{t=0}^{T} r(\pi_t, a_t) + c_t \,\middle|\, \mathcal{F}_0^{\pi}\right] \tag{10}$$

where c_t now takes the form

$$c_t := \mathbb{E}[\vartheta_{t+1}(\pi_{t+1}) \mid \mathcal{F}_t^{\pi}] - \vartheta_{t+1}(\pi_{t+1}). \tag{11}$$

[5] We note that dual feasible penalties are essentially *action-dependent* control variates, a standard variance reduction technique in the simulation literature. Recall also that π_{t+1} is a function of the actions $a_{0:t}$ as well as exogenous noise as described in (6).

In principle we can evaluate the right-hand-side of (10) by simulating J sample paths, $\left(u_{1:T}^{(j)}\right)$, for $j = 1, \ldots, J$, and solving the deterministic maximization problem inside the expectation in (10) (the *inner* problem) for each 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^{upper}, on the optimal value function, $V_0^*(\pi_0)$. Moreover standard methods can be used to construct approximate confidence intervals for V_0^{upper}.

In the BSPI setting, however, the state space is the $|\mathcal{H}|$-dimensional simplex. As a result, solving the inner problem in (10) 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.

3.2. The Uncontrolled Formulation

Building on ideas from Rogers [32], BH showed how this problem could be solved using a change-of-measure approach. In particular they reformulated the primal problem of (7) using an equivalent probability measure under which the chosen actions do not influence the state transition dynamics. Instead, the actions are accounted for by the Radon-Nikodym (RN) derivatives which adjust for the change-of-probability measure. BH called this an *uncontrolled formulation* and showed that the weak and strong duality results continued to hold under such a formulation. In this case the analog of (10), i.e. weak duality under the uncontrolled BSPI relaxation, is given by

$$V_0^*(\pi_0) \leq \widetilde{\mathbb{E}}\left[\max_{a_{0:T-1}} \sum_{t=0}^{T} \Phi_t^{\pi}[r_t(\pi_t, a_t) + c_t] \,\middle|\, \mathcal{F}_0^{\pi}\right] \tag{12}$$

where

$$c_t := \mathbb{E}[\vartheta_{t+1}(\pi_{t+1}) \mid \mathcal{F}_t^{\pi}] - \phi(\pi_t, \pi_{t+1}, a_t)\, \vartheta_{t+1}(\pi_{t+1}) \tag{13}$$

$$\Phi_t^{\pi}(\pi_{0:t}, a_{0:t-1}) := \prod_{s=0}^{t-1} \phi(\pi_s, \pi_{s+1}, a_s) \tag{14}$$

and where $\widetilde{\mathbb{E}}[\cdot]$ denotes an expectation under the new probability[6] measure, $\widetilde{\mathbb{P}}$. In general, the measure $\mathbb{P}$ is required to be absolutely continuous with respect to $\tilde{\mathbb{P}}$ but we will see later why we do not need to impose this due to our choice of ϑ_t's. The $\phi(\pi_t, \pi_{t+1}, a_t)$ terms in (13) and (14) are one-step RN derivative terms and they will take the form

$$\phi(\pi, \pi', a) := \frac{\sum_{i,j,k} \pi(i) P_{ij}(a) B_{jk}(a) \mathbf{1}_{\{\pi' = f(\pi, a, k)\}}}{\sum_{i,j,k} \pi(i) Q_{ij} E_{jk} \mathbf{1}_{\{\pi' = \tilde{f}(\pi, k)\}}} \tag{15}$$

where:

- Q_{ij} and E_{jk} are action-independent transition and emission matrices, respectively. Specifically, we define a $|\mathcal{H}| \times |\mathcal{H}|$ matrix, Q, of transition probabilities, with

$$Q_{ij} := \mathbb{P}(h_t = j \mid h_{t-1} = i), \quad i, j \in \mathcal{H} \tag{16}$$

[6] Throughout the paper we will use $\mathbb{P}$ to denote the probability measure for the original controlled POMDP formulation such as (6) or (3) and (4). We will use $\tilde{\mathbb{P}}$ to denote the probability measure for any uncontrolled POMDP formulation. The particular controlled or uncontrolled formulation should be clear from the context.

and a $|\mathcal{H}| \times |\mathcal{O}|$ matrix, E, of observation probabilities with

$$E_{ij} := \mathbb{P}(o_t = j \mid h_t = i), \quad i \in \mathcal{H}, j \in \mathcal{O}. \tag{17}$$

The change-of-measure $\widetilde{\mathbb{P}}$ is determined by these matrices.

- $f(\pi, a, k)$ lies in the $|\mathcal{H}|$-dimensional simplex and is the new filtered belief-state that results under $\mathbb{P}$ from taking action a and observing k when the current belief-state is π. Specifically, if we use $f(j; \pi, a, k)$ to denote the j^{th} component of $f(\pi, a, k)$ then

$$f(j; \pi, a, k) = \frac{\sum_i \pi(i) P_{ij}(a) B_{jk}(a)}{\sum_{i,l} \pi(i) P_{il}(a) B_{lk}(a)}. \tag{18}$$

- $\tilde{f}(\pi, k)$ lies in the $|\mathcal{H}|$-dimensional simplex and is the new filtered belief-state that results under $\tilde{\mathbb{P}}$ after observing k when the current belief-state is π. Specifically, if we use $\tilde{f}(j; \pi, k)$ to denote the j^{th} component of $\tilde{f}(\pi, k)$ then

$$\tilde{f}(j; \pi; k) = \frac{\sum_i \pi(i) Q_{ij} E_{jk}}{\sum_{i,l} \pi(i) Q_{il} E_{lk}}. \tag{19}$$

Note that both Q and E are *action independent* and in general they will depend on time t so we often write Q^t_{ij},E^t_{ij}, $f_t(\pi, a, k)$, $\tilde{f}_t(j; \pi; k)$, ϕ_t, etc. Further details are provided and justified in Appendix A.2.

Using an uncontrolled formulation results in a dramatic reduction of the state space that needs to be considered in solving the inner problem in (12). In particular, when we solve the inner problem as a deterministic dynamic program, we do not need to solve this DP for all possible states π_t in the $|\mathcal{H}|$-dimensional simplex. This is because the sequence of states $\pi_0, \ldots, \pi_T$ is fixed inside the inner problem of (12) due to the uncontrolled nature of the formulation where the history of actions does not influence the state transition dynamics. As such, the deterministic DP that solves the inner problem only needs to be solved along the state path $\pi_0, \ldots, \pi_T$. Of course this state path will vary across inner problem instances. The deterministic DP that is the inner problem in (12) can be solved recursively according to

$$V_t^{\mathbb{B}^\pi} = \max_a \{ r_t(\pi_t, a) + c_t + \phi(\pi_t, \pi_{t+1}, a)\, V_{t+1}^{\mathbb{B}^\pi} \} \tag{20}$$

for $t = 0, \ldots, T-1$ where c_t is given by (13) and $\pi_{0:T}$ is the sequence of belief states that were generated for that specific inner problem. We also have the terminal condition $V_T^{\mathbb{B}^\pi} = r_T(h_T)$ since h_T is assumed to be observed at time T and since $c_T = 0$ as each ϑ_{T+1} can be assumed to be identically zero.

4. Information Relaxations for the Non-Belief-State Formulation

Until now we have followed the approach of BSS and BH to outline how information relaxation dual bounds can be computed for POMDPs using the belief-state (and hence MDP) formulation of these problems. In this section we will show that information relaxation bounds for POMDPs can also be obtained using the non-belief-state formulation of the problem as described in the first part of Section 2. This leads to a very different form of inner problem which in principle is much simpler to solve. We will still need to use an

uncontrolled formulation, however, in order to evaluate the dual penalties. This is in contrast to the inner problems of the BSPI relaxation where, as discussed in Section 3.2, an uncontrolled formulation was required to reduce the effective dimension of the inner problem.

In Section 5 we will argue that the information relaxation bounds provided by a certain version of the non-belief state formulation of this section will coincide with the corresponding bounds provided by the belief-state formulation of Section 3.2. This will no longer hold, however, when different changes-of-measure are used to construct the two bounds and (for the measure changes we propose), we will generally expect the BSPI approach to yield tighter bounds than the PI approach. Nonetheless, we believe the non-belief state formulation (and the resulting PI relaxation) may be potentially useful for other non-Markovian control problems where a belief-state formulation doesn't arise as naturally as it does in the case of POMDPs. Influence diagrams, for example, is one such class of problems. See [23] or Chapter 23 of [25] for an introduction to influence diagrams.

4.1. The Perfect Information Relaxation

We now assume that the POMDP is formulated using the hidden state and observation dynamics of (3) and (4). We recall that the filtration $\mathbb{F} = (\mathcal{F}_0, \ldots, \mathcal{F}_T)$ is the filtration generated by the observations so that $\mathcal{F}_t$ is the σ-algebra generated by $o_{1:t}$ and π_0. The perfect information (PI) relaxation corresponds to the filtration $\mathbb{I} = (\mathcal{I}_0, \mathcal{I}_1, \ldots, \mathcal{I}_T)$, with $\mathcal{I}_t = \sigma(h_0, w_{1:T}, v_{1:T})$ 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 $\mathbb{I}$. It is worth noting that knowledge of the w_t's, v_t'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 $\mathbb{I}$ is indeed a relaxation of $\mathbb{F}$. Under the PI relaxation, the equivalent of (10), i.e. weak duality for the non-belief-state formulation, corresponds to

$$V_0^*(\pi_0) \leq \mathbb{E}\left[\max_{a_{0:T-1}} \sum_{t=0}^{T} r_t(h_t, a_t) + c_t \,\middle|\, \mathcal{F}_0\right] \tag{21}$$

where the c_t's now take the form

$$c_t := \mathbb{E}[\vartheta_{t+1}(o_{1:t+1}) \mid \mathcal{F}_t] - \vartheta_{t+1}(o_{1:t+1}). \tag{22}$$

We note that the same ϑ_{t+1}'s that we use in (11) can also be used in (22). This follows because π_{t+1} is in fact a function of $o_{1:t+1}$ and so it is perfectly fine to write $\vartheta_{t+1}(o_{1:t+1})$ instead of $\vartheta_{t+1}(\pi_{t+1})$.

In principle we can again compute an unbiased estimate of the right-hand-side of (21) by first simulating J sample paths, $\left(h_0^{(j)}, w_{1:T}^{(j)}, v_{1:T}^{(j)}\right)$, for $j = 1, \ldots, J$. We solve the inner problem inside the expectation in (21) for each such path and then average the corresponding optimal objective functions.

4.2. Solving the Inner Problem in (21)

We would therefore like to use the PI relaxation to construct an upper bound on V_0^* by solving the inner problem in (21) as a deterministic dynamic program. The main obstacle we will encounter under the PI relaxation, however, is computing the c_t's as defined in (22). We can see this most clearly if we consider the zero-penalty case where we set $\vartheta_{t+1} \equiv 0$. In that case $c_t \equiv 0$ for all t and the inner problem in (21) is a simple deterministic DP with just $|\mathcal{H}|$ states. In contrast, when $c_t \equiv 0$ in (10), we see that the inner problem in (10) is still a deterministic DP but now the state space lies in the $|\mathcal{H}|$-dimensional simplex. The inner problems in

(10) for the BSPI relaxation are therefore in principle considerably more challenging than the inner problems in (21) and this is why the uncontrolled formulation of (12) was required for the BSPI relaxation.

Unfortunately, if we want to use a non-zero ϑ_{t+1} (as is typically the case), then evaluating the term $\mathbb{E}[\vartheta_{t+1}(o_{1:t+1}) \mid \mathcal{F}_t]$ in (22) is challenging. With the PI relaxation of the non-belief-state formulation of (3) and (4), 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 (21) as a deterministic dynamic program, we would need to compute $\mathbb{E}[\vartheta_{t+1}(o_{1:t+1}) \mid \mathcal{F}_t]$ at each time t for *all* possible action histories, $a_{0:t}$. In fact this is also true for the second term in (22), $\vartheta_{t+1}(o_{1:t+1})$. Evaluating the penalties c_t for all possible action histories is therefore clearly impractical for any realistic application. Once again, however, we can use an uncontrolled formulation to resolve this problem.

Before proceeding to the uncontrolled formulation, however, it is worth emphasizing why the calculation of these penalty terms is straightforward for the BSPI relaxation. Consider the term $\mathbb{E}[\vartheta_{t+1}(\pi_{t+1}) \mid \mathcal{F}_t^{\pi}]$ that arises in the calculation of the penalty in (11) in the case of the BSPI relaxation. Because we are conditioning on $\mathcal{F}_t^{\pi}$ the calculation of $\mathbb{E}[\vartheta_{t+1}(\pi_{t+1}) \mid \mathcal{F}_t^{\pi}]$ depends on π_t (which is known given $\mathcal{F}_t^{\pi}$) and the time t action a_t. In particular, it does *not depend* on the action history $a_{0:t-1}$ which is in contrast to the term $\mathbb{E}[\vartheta_{t+1}(o_{1:t+1}) \mid \mathcal{F}_t]$ that arises in the PI penalty of (22). Therefore under the BSPI relaxation the penalties are easy to calculate for any state π_t. Of course, what is really happening here is that the complexity of evaluating penalties for the inner problems of the PI relaxation is transferred to the complexity of working with a much higher dimensional state-space when solving inner problems for the BSPI relaxation. Either way then, we must use an uncontrolled formulation.

4.3. The Uncontrolled Formulation

In order to define an action-independent change-of-probability-measure, we simply define a hidden Markov model (HMM) on the same hidden state and observation spaces as our POMDP. Specifically, we simply define action-independent transition and emission matrices Q and E as in (16) and (17), respectively. In general Q and E will depend on time t in which case we might prefer to write Q_{ij}^t and E_{ij}^t, and[7] 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 $B_{ij}(a) > 0$.

A trivial way to ensure these conditions is to have $Q_{ij} > 0$ and $E_{ik} > 0$ for all $i, j \in \mathcal{H}$ and $k \in \mathcal{O}$. As mentioned earlier, we let $\tilde{\mathbb{P}}$ denote the probability measure induced by Q and E with $\tilde{\mathbb{E}}$ denoting expectations under $\tilde{\mathbb{P}}$. We now proceed by reformulating our POMDP under $\tilde{\mathbb{P}}$ and adjusting rewards (and penalties) with appropriate Radon-Nikodym (RN) derivatives. In Appendix A.1 we show that these RN derivatives are of

[7] We will see later in Section 6.2 that we can ignore these absolutely continuity conditions when we take the ϑ_t's to be *supersolutions*. This also applies to the $\tilde{\mathbb{P}}$ discussed in Section 3.2 for the belief-state formulation. Indeed this is the approach we will take for the numerical experiments of Sections 7 and 8.

the form $d\mathbb{P}/d\tilde{\mathbb{P}} = \Phi_T(h_{0:T}, o_{1:T}, a_{0:T-1})$ with

$$\phi(i,j,k,a) := \frac{P_{ij}(a)}{Q_{ij}} \cdot \frac{B_{jk}(a)}{E_{jk}} \tag{23}$$

$$\Phi_t(h_{0:t}, o_{1:t}, a_{0:t-1}) := \prod_{s=0}^{t-1} \phi(h_s, h_{s+1}, o_{s+1}, a_s). \tag{24}$$

It is then straightforward to see that

$$V_0^*(\pi_0) = \max_{\mu \in \mathcal{U}_{\mathbb{F}}} \mathbb{E}\left[\sum_{t=0}^{T} r_t(h_t, \mu_t) \,\middle|\, \mathcal{F}_0\right] = \max_{\mu \in \mathcal{U}_{\mathbb{F}}} \widetilde{\mathbb{E}}\left[\sum_{t=0}^{T} \Phi_t r_t(h_t, \mu_t) \,\middle|\, \mathcal{F}_0\right]. \tag{25}$$

We refer to (25) as an uncontrolled formulation of the non-belief-state POMDP formulation. The "uncontrolled" terminology reflects the fact that the policy, μ, 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 Φ_t's. With this uncontrolled formulation the analog of (21), i.e. weak duality for the PI relaxation, is given by

$$V_0^*(\pi_0) \leq \widetilde{\mathbb{E}}\left[\max_{a_{0:T-1}} \sum_{t=0}^{T} \Phi_t[r_t(h_t, a_t) + c_t] \,\middle|\, \mathcal{F}_0\right] \tag{26}$$

with

$$c_t := \mathbb{E}[\vartheta_{t+1}(o_{1:t+1}) \mid \mathcal{F}_t] - \phi(h_t, h_{t+1}, o_{t+1}, a_t)\, \vartheta_{t+1}(o_{1:t+1}). \tag{27}$$

Returning to the penalty in (22) 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 (26). This is because the action histories under $\tilde{\mathbb{P}}$ 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

$$\mathbb{E}[\vartheta_{t+1}(o_{1:t+1}) \mid \mathcal{F}_t] = \sum_{o \in \mathcal{O}; h, h' \in \mathcal{H}} \pi_t(h) P_{hh'}(a_t) B_{h'o}(a_t) \vartheta_{t+1}(o_{1:t}, o) \tag{28}$$

where $\pi_t(h) := \tilde{\mathbb{P}}(h_t = h \mid o_{1:t})$ can be calculated efficiently using standard HMM filtering methods. As discussed in Section 4.1, we can now calculate an unbiased upper bound on V_0^* by solving J instances of the inner problems in (26) and averaging their optimal objective values. Note that an inner problem can be solved recursively according to

$$V_t^{\mathbb{I}} = \max_a \{r_t(h_t, a) + c_t + \phi(h_t, h_{t+1}, o_{t+1}, a)\, V_{t+1}^{\mathbb{I}}\} \tag{29}$$

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 $V_T^{\mathbb{I}} = r_T(h_T)$ since $c_T = 0$ as each ϑ_{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 ϑ_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 robotic navigation and multi-access communication applications of Sections 7 and 8, respectively.

5. Comparing the BSPI and PI Dual Bounds

Consider now the primal problems in (5) and (7) corresponding to the non-belief-state and belief-state formulations, respectively. In (5) the rewards are $r_t(h_t, a_t)$ and the optimization is over $\mathbb{F}$-adapted policies. In contrast, the rewards are $r_t(\pi_t, a_t)$ and the optimization is over $\mathbb{F}^\pi$-adapted policies in (7). Of course the two objectives are equal since $r(\pi_t, a_t) := \mathbb{E}[r(h_t, a_t) \mid \mathcal{F}_t^\pi]$ and because $\mathcal{F}_t$ contains no relevant information beyond what is in $\mathcal{F}_t^\pi$ (even though $\mathcal{F}_t^\pi \subset \mathcal{F}_t$). Consider now a third equivalent formulation where the rewards are $r(\pi_t, a_t)$ but the optimization is over $\mathbb{F}$-adapted policies. In this case we have

$$V_0^*(\pi_0) = \max_{\mu \in \mathcal{U}_\mathbb{F}} \mathbb{E}\left\{ \sum_{t=0}^{T} r_t(\pi_t, \mu_t) \,\middle|\, \mathcal{F}_0^\pi \right\} \tag{30}$$

where we note the only difference between (7) and (30) is that the optimization is over $\mu \in \mathcal{U}_{\mathbb{F}^\pi}$ in the former and over $\mu \in \mathcal{U}_\mathbb{F}$ in the latter. Despite the presence of $r_t(\pi_t, \mu_t)$ in (30), this is also a *non-belief-state* formulation of the problem because $\mathbb{F} = (\mathcal{F}_0, \dots, \mathcal{F}_T)$ where $\mathcal{F}_t$ is the σ-algebra generated by $o_{1:t}$ (and π_0).

5.1. Which Non-Belief-State Formulation Leads to Tighter Dual Bounds?

We therefore have two non-belief-state formulations of the problem – the original with rewards $r_t(h_t, a_t)$ and the new one with rewards $r_t(\pi_t, a_t)$. Each of these formulations has a corresponding PI dual bound but it should not be at all surprising that the latter one (with rewards $r_t(\pi_t, a_t)$) leads to tighter dual bounds. The following simple example should suffice to see why this is the case.

Example 1. Consider a POMDP with just two periods, $t = 0$ and $t = 1$. There are two possible hidden states h_{good} and h_{bad} and the initial belief-state distribution π_0 puts equal probability on each of them. The only possible actions are a_{stay} and a_{switch}. If the chosen action at time $t = 0$ is a_{stay} then at time $t = 1$ you will stay in the same hidden state that you were in at time $t = 0$. If the chosen action is a_{switch} at time $t = 0$ then at time $t = 1$ you will move to the other hidden state. So for example, if $h_0 = h_{\mathsf{bad}}$ and you choose action a_{switch} then w.p.1 $h_1 = h_{\mathsf{good}}$. A reward of 1 is realised at $t = 1$ if $h_1 = h_{\mathsf{good}}$ and this is the only possible reward. The observations in this POMDP are completely uninformative.

Consider now a PI inner problem in the non-belief-state formulation with rewards $r_t(h_t, a_t)$ and zero penalties. In this case the DM is guaranteed to get a reward of 1 since she will see h_0. In particular, she will know which of a_{stay} and a_{switch} she should choose to guarantee she is in state h_{good} at time $t = 1$ and therefore earn the reward of 1. For the PI inner problem in the non-belief-state formulation with rewards $r_t(\pi_t, a_t)$ (and again zero penalties), the DM can again guarantee that $h_1 = h_{\mathsf{good}}$. This time, however, the reward is $r_1(\pi_1, a_1) = 1/2$ because the observations are non-informative and so π_1 puts equal weight on the two possible hidden states at time $t = 1$. So even though the PI decision-maker knows what the true state is at $t = 1$ she only receives a reward of $1/2$ for this.

More generally, suppose that the observations were informative although in general still noisy. With rewards $r_t(h_t, a_t)$ the DM can always guarantee a reward of 1 at time $t = 1$ in the PI relaxation. In contrast, with rewards $r_t(\pi_t, a_t)$, the DM would receive a reward of $r_t(\pi_t, a_t) \in (1/2, 1]$ at time $t = 1$ if she ensured $h_1 = h_{\mathsf{good}}$ since π_1 would then put more weight on $h_1 = h_{\mathsf{good}}$ given that the observations are informative.

It seems clear then that the averaging that results in $r_t(\pi_t, a_t)$ leads to tighter dual PI bounds than those

we'd obtain if we persisted with the use of $r_t(h_t, a_t)$. For this reason, whenever we refer to PI bounds in the sequel we will be referring (unless otherwise stated) to PI bounds where the rewards are the $r_t(\pi_t, a_t)$'s. In particular, the PI bounds of Sections 7 and 8 use the $r_t(\pi_t, a_t)$ form of the rewards. We now show that the PI and BSPI bounds (both of which are now based on rewards $r_t(\pi_t, a_t)$) are identical when there is no change-of-measure involved.

5.2. When Are the PI and BSPI Bounds Are Equal?

The PI relaxation bound corresponding to formulation (30) is given by

$$\begin{aligned}\mathbb{E}[V_0^{\mathbb{I}}] &:= \mathbb{E}\left[\max_{a_{0:T-1}} \sum_{t=0}^{T} r_t(\pi_t, a_t) + c_t \,\middle|\, \mathcal{F}_0^{\pi}\right] \\ &= \mathbb{E}_{h_{0:T}, o_{1:T}}\left[\max_{a_{0:T-1}} \sum_{t=0}^{T} \left[r_t(\pi_t, a_t) + \mathbb{E}[\vartheta_{t+1}(o_{1:t+1}) \mid \mathcal{F}_t] - \vartheta_{t+1}(o_{1:t+1})\right] \,\middle|\, \mathcal{F}_0^{\pi}\right] \end{aligned} \tag{31}$$

where we have substituted for c_t using (22) and where we have used $\mathbb{E}_x$ to denote an expectation taken w.r.t. the random vector x. As we shall see in Section 6 all our AVFs $\vartheta(o_{1:t})$ can be written equivalently as $\vartheta(\pi_t)$. Together with the fact that $\mathcal{F}_t$ contains no relevant information beyond what is in $\mathcal{F}_t^{\pi}$, this implies we can write (31) as

$$\begin{aligned}\mathbb{E}[V_0^{\mathbb{I}}] &= \mathbb{E}_{h_{0:T}, o_{1:T}}\left[\max_{a_{0:T-1}} \sum_{t=0}^{T} \left[r_t(\pi_t, a_t) + \mathbb{E}[\vartheta_{t+1}(\pi_{t+1}) \mid \mathcal{F}_t^{\pi}] - \vartheta_{t+1}(\pi_{t+1})\right] \,\middle|\, \mathcal{F}_0^{\pi}\right] \\ &= \mathbb{E}_{o_{1:T}}\left[\mathbb{E}_{h_{0:T}}\left[\max_{a_{0:T-1}} \sum_{t=0}^{T} \left[r_t(\pi_t, a_t) + \mathbb{E}[\vartheta_{t+1}(\pi_{t+1}) \mid \mathcal{F}_t^{\pi}] - \vartheta_{t+1}(\pi_{t+1})\right] \,\middle|\, o_{1:T}, \mathcal{F}_0^{\pi}\right] \,\middle|\, \mathcal{F}_0^{\pi}\right] \end{aligned} \tag{32}$$

where the second equality follows from the tower property of conditional expectations. Note that the π_t's appearing inside the inner expectation in (32) are deterministic functions of π_0, $o_{1:t}$ and $a_{0:t-1}$ and as such, are independent of $h_{0:T}$, given π_0, $o_{1:T}$ and $a_{0:T}$. It therefore follows that (32) becomes

$$\mathbb{E}[V_0^{\mathbb{I}}] = \mathbb{E}_{o_{1:T}}\left[\max_{a_{0:T-1}} \sum_{t=0}^{T} \left[r_t(\pi_t, a_t) + \mathbb{E}[\vartheta_{t+1}(\pi_{t+1}) \mid \mathcal{F}_t^{\pi}] - \vartheta_{t+1}(\pi_{t+1})\right] \,\middle|\, \mathcal{F}_0^{\pi}\right] \tag{33}$$

$$= \mathbb{E}_{\pi_{1:T}}\left[\max_{a_{0:T-1}} \sum_{t=0}^{T} \left[r_t(\pi_t, a_t) + \mathbb{E}[\vartheta_{t+1}(\pi_{t+1}) \mid \mathcal{F}_t^{\pi}] - \vartheta_{t+1}(\pi_{t+1})\right] \,\middle|\, \mathcal{F}_0^{\pi}\right] \tag{34}$$

$$= \mathbb{E}[V_0^{\mathbb{B}^{\pi}}] \tag{35}$$

where we recognize the right-hand-side of (34) as the BSPI relaxation bound in (10) with penalties given by (11) and we use $V_0^{\mathbb{B}^{\pi}}$ to denote the optimal value of a BSPI inner problem. We therefore have the following result.

Proposition 5.1. *Given penalties constructed from the same AVF, the BSPI information relaxation bound is equal to the PI information relaxation bound with rewards $r_t(\pi_t, a_t)$.*

Remark 5.1. *One direction of Proposition 5.1 is quite obvious and follows immediately from BSS. In particular we note that the BSPI relaxation is weaker than the PI relaxation, i.e. $\mathcal{B}_t^{\pi} \subseteq \mathcal{I}_t$ for all t. This follows because knowledge of $(v_{1:T}, w_{1:T})$ together with π_0 and the action history $a_{0:T-1}$ is sufficient to determine the π_t's. That the BSPI bound is at least as good as the PI bound (with rewards $r_t(\pi_t, a_t)$) now follows*

immediately from Prop. 2.3(i) of BSS since the rewards are identical in both formulations.

It's clear that Proposition 5.1 continues to hold under the *same* absolutely continuous change-of-measure. In particular, such a measure change will preserve equality in (33) to (35). That said, we never use the same change-of-measure for the PI and BSPI bounds. In general, it is difficult to compare bounds constructed via different changes-of-measure since (see Rogers [32] and BH) the dual bound depends[8] on the specific change-of-measure. In our POMDP setting, however, the change-of-measures that we propose to use for the PI and BSPI bounds will be closely related. In particular, in our numerical experiments of Sections 7 and 8 the change-of-measure we use for the PI and BSPI bounds will be the measure-change induced by following some feasible strategy, μ say. The change-of-measure for the uncontrolled BSPI formulation, however, will require a layer of filtering so that the corresponding RN derivatives will be a function of belief-states. In contrast, the RN derivatives for the uncontrolled PI formulation will be a function of the hidden states. This is best understood by comparing the RN derivative terms in (15) and (23). In our numerical experiments the Q_{ij}'s and E_{jk}'s that appear in both (15) and (23) will coincide and equal $P_{ij}(a_\mu)$ and $B_{jk}(a_\mu)$, respectively, where a_μ is the action induced by following μ (which of course will depend on the belief-state at that time). The RN derivative term in (15) can therefore be loosely viewed as a filtered version of the RN derivative term in (23). Moreover, a similar argument to that presented in Example 1 suggests that the BSPI bound should be tighter than the corresponding PI bound. We discuss in further detail the relationship between these two measure changes in Appendix A.3.

6. Approximate Value Functions and Supersolutions

We now discuss several standard approaches for obtaining approximations to the optimal value function in our POMDP setting. In general we can use each such approximation, $\widetilde{V}_t$, to:

(i) Construct a lower bound, V_0^{lower}, on V_0^*, by simulating the policy that is greedy[9] with respect to $\widetilde{V}_t$. Towards this end, we can generate J independent sample paths $\left(h_0^{(j)}, w_{1:T}^{(j)}, v_{1:T}^{(j)}\right)$, 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 we calculate at time t the corresponding belief state π_t using standard filtering techniques, and take the action a_t that obtains the maximum in the chosen AVF from each of (39), (41) or (43) below. If we denote by $V_{\text{lower}}^{(j)}$ 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 $\sum_j V_{\text{lower}}^{(j)}/J$.

(ii) Construct an upper bound, V_0^{upper}, via our BSPI and PI uncontrolled information relaxations by setting $\vartheta_t = \widetilde{V}_t$ in (13) and (27). This of course is motivated by the strong duality result of BSS which states that if we take $\vartheta_t = V_t^*$ then the dual bound will be tight and coincide with V_0^*.

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. Later in Section 6.1 we will discuss the concept of supersolutions

[8] Unless a perfect penalty is used in which case strong duality implies both bounds will coincide with the optimal value function.

[9] Recall that a policy is said to be greedy with respect to $\widetilde{V}_t$ 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 value of $\widetilde{V}_{t+1}$, i.e. $a_t = \text{argmax}_a\{r_t(\pi_t, a) + \mathbb{E}[\widetilde{V}_{t+1}(\pi_{t+1}) \mid \mathcal{F}_t^\pi]\}$.

and state a proposition asserting that the approximate-value functions that we define below are indeed supersolutions. The significance of supersolutions will then be discussed in Sections 6.1 and 6.2.

We now describe the *MDP*, *QMDP* and *Fast Informed (Lag-1)* value function approximations together with the *Lag-2* approximation which we propose as a natural extension of the Lag-1 approximation. More generally, we could define a Lag-d approximation but the computational requirements for calculating it scale exponentially in the number of lags d. Other approximate solution approaches can be found, for example, in [26]. Before proceeding further, we note that the optimal value function $V_T^*(\pi_T)$ is known at time T and satisfies $V_T^*(\pi_T) = r_T(o_T)$ because of our earlier w.l.o.g. assumption that $o_T = h_T$. This means that each of our AVFs can also be assumed to satisfy $\widetilde{V}_T(\pi_T) = r_T(o_T)$.

The MDP Approximate Value Function

The MDP AVF is constructed from $V_t^{\text{MDP}}(h)$, the optimal value function 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_t^{MDP} in typical POMDP settings and we can use it to construct an AVF according to

$$\widetilde{V}_t^{\text{MDP}}(\pi_t) \coloneqq \mathbb{E}[V_t^{\text{MDP}}(h_t) \mid \mathcal{F}_t^{\pi}] = \sum_{h \in \mathcal{H}} \pi_t(h) V_t^{\text{MDP}}(h) \tag{36}$$

where $V_T^{\text{MDP}}(h) \coloneqq r_T(h)$ and for $t \in \{0, \dots, T-1\}$ we define

$$V_t^{\text{MDP}}(h) \coloneqq \max_{a_t \in \mathcal{A}} \big\{ r_t(h, a_t) + \mathbb{E}[V_{t+1}^{\text{MDP}}(h_{t+1}) \mid h_t = h] \big\}. \tag{37}$$

The QMDP Approximate Value Function

The QMDP AVF is constructed using the Q-values [28] which are defined as

$$V_t^{\text{Q}}(h, a) \coloneqq r_t(h, a) + \sum_{h' \in \mathcal{H}} P_{hh'}(a) V_{t+1}^{\text{MDP}}(h') \tag{38}$$

for $t \in \{0, \dots, T-1\}$. The QMDP AVF is then defined according to

$$\widetilde{V}_t^{\text{Q}}(\pi_t) \coloneqq \max_{a_t} \sum_{h \in \mathcal{H}} \pi_t(h) V_t^{\text{Q}}(h, a_t). \tag{39}$$

Note that by exchanging the order of the expectation and max operators in (39) 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 (36).

The Lag-1 Approximate Value Function

The Lag-1 approximation was first proposed in [21] as the *fast informed bound update*. This approximation uses the optimal value function, $V_t^{\text{L1}}(h_{t-1}, a_{t-1}, o_t)$, from the corresponding lag-1 formulation of the POMDP where the hidden state, h_{t-1}, is observed before deciding on the time t action a_t for all $t < T$. We can calculate V_t^{L1} recursively via

$$V_t^{\text{L1}}(h_{t-1}, a_{t-1}, o_t) = \max_{a_t} \mathbb{E}[r_t(h_t, a_t) + V_{t+1}^{\text{L1}}(h_t, a_t, o_{t+1}) \mid h_{t-1}, o_t] \tag{40}$$

for $t \in \{1, \ldots, T-1\}$ and with terminal condition $V_T^{\text{L}_1}(h_{T-1}, a_{T-1}, o_T) := r_T(h_T)$ (since $o_T = h_T$). The corresponding AVF is then defined according to

$$\widetilde{V}_t^{\text{L}_1}(\pi_t) := \max_{a_t} \mathbb{E}[r_t(h_t, a_t) + V_{t+1}^{\text{L}_1}(h_t, a_t, o_{t+1}) \mid \mathcal{F}_t^{\pi}] \tag{41}$$

where the expectation is taken with respect to o_{t+1} and h_t, given the current belief state, π_t. Further details on calculating $V_t^{\text{L}_1}$ can be found in Appendix B.

The Lag-2 Approximate Value Function

The Lag-2 approximation is derived by first constructing the optimal value function $V_t^{\text{L}_2}(h_{t-2}, a_{t-2:t-1}, o_{t-1:t})$ corresponding to the MDP where the hidden state, h_{t-2}, is observed before taking the decision a_t at time t for all $t < T$. Again the terminal value function is $V_T^{L_2}(h_{T-2}, a_{T-2:T-1}, o_{T-1:T}) := r_T(o_T) = r_T(h_T)$ and the optimal value function, $V_t^{L_2}$, at earlier times is computed iteratively according to

$$V_t^{\text{L}_2}(h_{t-2}, a_{t-2:t-1}, o_{t-1:t}) := \max_{a_t} \mathbb{E}[r_t(h_t, a_t) + V_{t+1}^{\text{L}_2}(h_{t-1}, a_{t-1:t}, o_{t:t+1}) \mid h_{t-2}, o_{t-1:t}] \tag{42}$$

for $t \in \{2, \ldots, T-1\}$. When $t = 0$ or 1 we must adjust (42) appropriately so that we only condition on o_0 and $o_{0:1}$, respectively. The calculation of $V_t^{L_2}$ is clearly more demanding than the calculation of $V_t^{\text{L}_1}$ since its state space is larger and since the expectation in (42) over (h_{t-1}, h_t, o_{t+1}) is more demanding to compute than the expectation in (40) which is over (h_t, o_{t+1}). We define the corresponding Lag-2 AVF according to

$$\widetilde{V}_t^{\text{L}_2}(\pi_t) := \max_{a_t} \mathbb{E}[\max_{a_{t+1}} \mathbb{E}[r_t(h_t, a_t) + r_{t+1}(h_{t+1}, a_{t+1}) + V_{t+2}^{L_2}(h_t, a_{t:t+1}, o_{t+1:t+2}) \mid \mathcal{F}_t^{\pi}, o_{t+1}] \mid \mathcal{F}_t^{\pi}] \tag{43}$$

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 time period remaining at that point. While more demanding to compute, we show in Appendix B.3 that the Lag-2 AVF is superior to the Lag-1 AVF in that $V_t^*(\pi_t) \leq \widetilde{V}_t^{\text{L}_2}(\pi_t) \leq \widetilde{V}_t^{\text{L}_1}(\pi_t)$. (The first inequality follows from the supersolution property of the AVFs as discussed in Section 6.1 below.) Before proceeding we mention that an alternative and perhaps more natural definition of the Lag-2 AVF is

$$\widetilde{V}_t^{\text{Alt}_2}(\pi_t) := \max_{a_t} \mathbb{E}[r_t(h_t, a_t) + V_{t+1}^{\text{L}_2}(h_{t-1}, a_{t-1:t}, o_{t:t+1}) \mid \mathcal{F}_t^{\pi}]. \tag{44}$$

However, it is straightforward to show that $\widetilde{V}_t^{\text{L}_2}(\pi_t) \leq \widetilde{V}_t^{\text{Alt}_2}(\pi_t)$ and so we prefer to use $\widetilde{V}_t^{\text{L}_2}(\pi_t)$ as our generalization of the Lag-1 AVF.

6.1. Supersolutions and Bound Guarantees

We begin by defining the concept of a supersolution.

Definition 6.1. *Let ϑ_t be any AVF that satisfies*

$$\vartheta_t(\pi_t) \geq \max_{a_t \in \mathcal{A}} \left\{ r_t(\pi_t, a_t) + \mathbb{E}[\vartheta_{t+1}(\pi_{t+1}) \mid \mathcal{F}_t^{\pi}] \right\} \tag{45}$$

for all belief states π_t, and all $t \in \{0, \ldots, T\}$. Then we say that ϑ_t is a supersolution.

It is well-known[10] that a supersolution ϑ_t is an upper bound on the optimal value function $V_t^*(\pi_t)$. Indeed the condition (45) is simply the feasibility condition for the linear programming formulation of the belief-state MDP. The supersolution property is particularly important in the context of information relaxations and there are two reasons for this, the first of which is Proposition 6.1 below from BH. A slightly less general version of this result was shown earlier by Desai et al. [15].

Proposition 6.1. *(Prop 4.1 in Brown & Haugh, 2017) An information relaxation upper bound based on a penalty constructed from a supersolution is guaranteed to be at least as good as the upper bound provided by the supersolution itself.*

We now state the main result of this section. The result itself is not surprising and a proof can be found in Appendix C.

Proposition 6.2. *The MDP, QMDP, Lag-1 and Lag-2 AVFs are all supersolutions.*

Propositions 6.1 and 6.2 imply that a dual upper bound (as given by (12)) based on a penalty constructed from a supersolution is guaranteed to be no worse than the original upper bound provided by the supersolution itself. We will see this result in action in the numerical results of Sections 7 and 8 when we see that the information relaxation upper bound is typically significantly better than the bound provided by the supersolution.

6.2. Using Supersolutions to Estimate the Duality Gap Directly

A second advantage of working with a supersolution AVF is that when the dual penalties are constructed using a supersolution then the requirement that $\mathbb{P} \ll \tilde{\mathbb{P}}$ can be ignored. This was shown by BH who then exploited[11] this fact by directly estimating the duality gap $V_0^{\text{upper}} - V_0^{\text{lower}}$. We describe their approach here and defer to Appendix D an explanation for why the absolute continuity condition, i.e. $\mathbb{P} \ll \tilde{\mathbb{P}}$, can be ignored when the dual penalties are constructed using a supersolution.

Specifically, suppose we have a good candidate $\mathcal{F}^\pi$-adapted policy, μ, and let $\tilde{\mathbb{P}}$ be the probability measure induced by following this policy. If we set V_0^{lower} to be the expected value of this policy, we then have

$$\begin{aligned} V_0^{\text{lower}} &= \mathbb{E}\Big[\sum_{t=0}^{T} \big(r_t(\pi_t, \mu_t) + c_t\big) \mid \mathcal{F}_0^\pi \Big] \\ &= \widetilde{\mathbb{E}}\Big[\sum_{t=0}^{T} \Phi_t(\mu)\big(r_t(\pi_t, \mu_t) + c_t\big) \mid \mathcal{F}_0^\pi \Big] \end{aligned} \tag{46}$$

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 (46) because $\mathbb{P}$ and $\tilde{\mathbb{P}}$ coincide when the policy μ is followed. We can use this same $\tilde{\mathbb{P}}$ to estimate an upper bound

$$V_0^{\text{upper}} = \widetilde{\mathbb{E}}\Big[\max_{a_{0:T-1}} \sum_{t=0}^{T} \Phi_t(a_{0:t-1}; \mu)\big(r_t(\pi_t, \mu_t) + c_t\big) \mid \mathcal{F}_0^\pi \Big] \tag{47}$$

as long as ϑ_t is constructed from a supersolution and where (47) now explicitly recognizes the dependence of the Φ_t's on $a_{0:t-1}$ and μ. Since both lower and upper bounds (46) and (47) are simulated using the

[10] A proof can be found in standard dynamic programming texts and is based on the linear-programming formulation of the Bellman equation.

[11] BH discussed this in their Section 4.3.1 but perhaps under-emphasized this practically important aspect of working with supersolutions.

same measure, $\tilde{\mathbb{P}}$, 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 (46) can now be re-used on the corresponding inner problem in (47).

There is a further benefit to this proposal, however. Because the actions of the policy, μ, are feasible for the inner problem in (47), it is clear the term inside the expectation in (46) will be less than or equal to the optimal objective of the inner problem in (47) along each simulated path. In fact the difference, D, between the two terms satisfies

$$0 \quad \leq \quad D \coloneqq \max_{a_{0:T-1}} \sum_{t=0}^{T} \Phi_t(a_{0:t-1}; \mu)\big(r_t(\pi_t, \mu_t) + c_t\big) - \sum_{t=0}^{T} \big(r_t(\pi_t, \mu_t) + c_t\big) \quad \tilde{\mathbb{P}} \text{ a.s.} \tag{48}$$

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 (48). 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 μ 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^{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 may have been used to estimate V_0^{lower} directly.

Propositions 6.1 and 6.2 together with the ability to focus directly on the duality gap in (48) highlight the importance of super-solutions in constructing bounds on the unknown optimal value function for POMDPs. The overall approach that we propose is:

1. Use a supersolution AVF $\vartheta_{t+1}(\pi_{t+1})$ to construct the penalties as in (11) / (13) for BSPI bounds or (22) / (27) for PI bounds.
2. Use the change-of-measure induced by following the best available feasible policy.
3. Use the penalties as control variates for the primal bound and therefore estimate the duality gap directly as in (48).

This is the approach we take in our numerical examples of Sections 7 and 8.

7. An Application to Robotic Navigation

We now apply our results to a well-known robotic navigation application and our problem formulation follows [28, 22, 30]. 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 α 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.

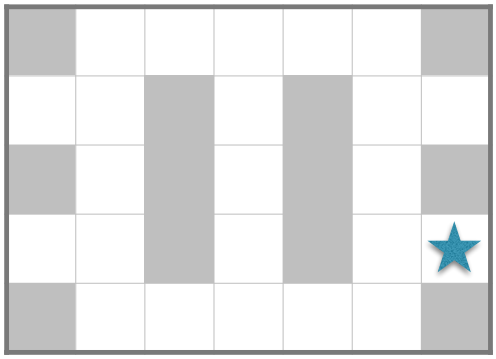


Figure 1: Maze representation for the robot navigation problem. The white spaces indicate the possible hidden states where the robot can be located. The star indicates the goal state.

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_{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 4×1 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 time T which is defined as $r_T(h_T) = 1$ if $h_T = h_{goal}$, and zero otherwise. All intermediate rewards are zero. Finally, we define $o_T \equiv h_T$ so that we know for certain whether or not the terminal reward was earned or not at the end of the horizon.

7.1. The Uncontrolled Formulation

Because all of our AVFs are supersolutions we were able to ignore the absolute continuity requirement when defining the change-of-measures for the uncontrolled formulations. Specifically we used the policies that were greedy w.r.t the QMDP, Lag-1 and Lag-2 AVFs to define uncontrolled-measure changes for the PI and BSPI bounds, respectively. The specific details for these uncontrolled measure changes and their RN derivatives are provided in Appendices A.1 and A.2. They are also described in slightly more general terms at the end of Section 5. The inner problems in (12) and (26) are solved as simple deterministic dynamic programs (see (20) and (29)) with terminal value $V_T(o_{0:T}) := 1_{\{h_T = h_{goal}\}}$. We can then calculate an unbiased upper bound on V_0^* by generating J inner problem instances and averaging their optimal values 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. Furthermore, we can use these penalties as control-variates for the primal problem and therefore estimate the duality gap directly as explained in Section 6.2.

7.2. Numerical Results

Figures 2 and 3 display numerical results from our experiments. Specifically, Figure 2 displays[12] the MDP, QMDP, Lag-1 and Lag-2 AVFs 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 the dual upper bounds obtained from the uncontrolled PI and BSPI relaxations when the penalties were constructed from the Lag-1 and Lag-2 AVFs, respectively. All of these bounds are displayed as a function of α 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 AVF.

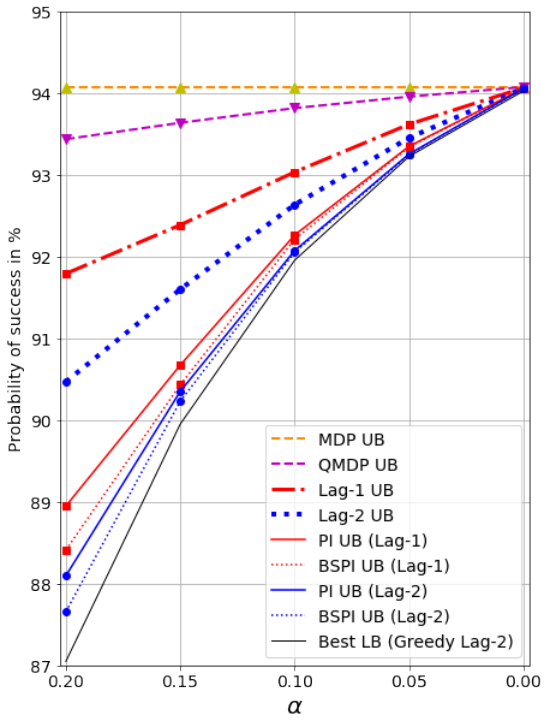


Figure 2: Comparison of upper bounds as a function of the noise factor α. The thick dotted lines correspond to the MDP, QMDP, Lag-1 and Lag-2 approximations. The solid (thin dotted) red and blue lines correspond to the dual PI (BSPI) relaxation upper bounds resulting from penalties constructed using 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 AVF.

Several observations are in order. We see that each of the dual upper bounds improves upon the respective supersolution that was used to construct the dual penalty in each case. We also see from Figure 2 that the duality gap decreases as α 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

[12]The figures actually report $\mathbb{E}[\widetilde{V}_0^{\mathrm{MDP}}(o_0) \mid \pi_0]$, $\mathbb{E}[\widetilde{V}_0^{\mathrm{Q}}(o_0) \mid \pi_0]$ etc. All of the numerical results in this section and the next were obtained using MATLAB release 2016b on a MacOS Sierra with a 1.3 GHz Intel Core i5 processor and 4 GB of RAM.

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.

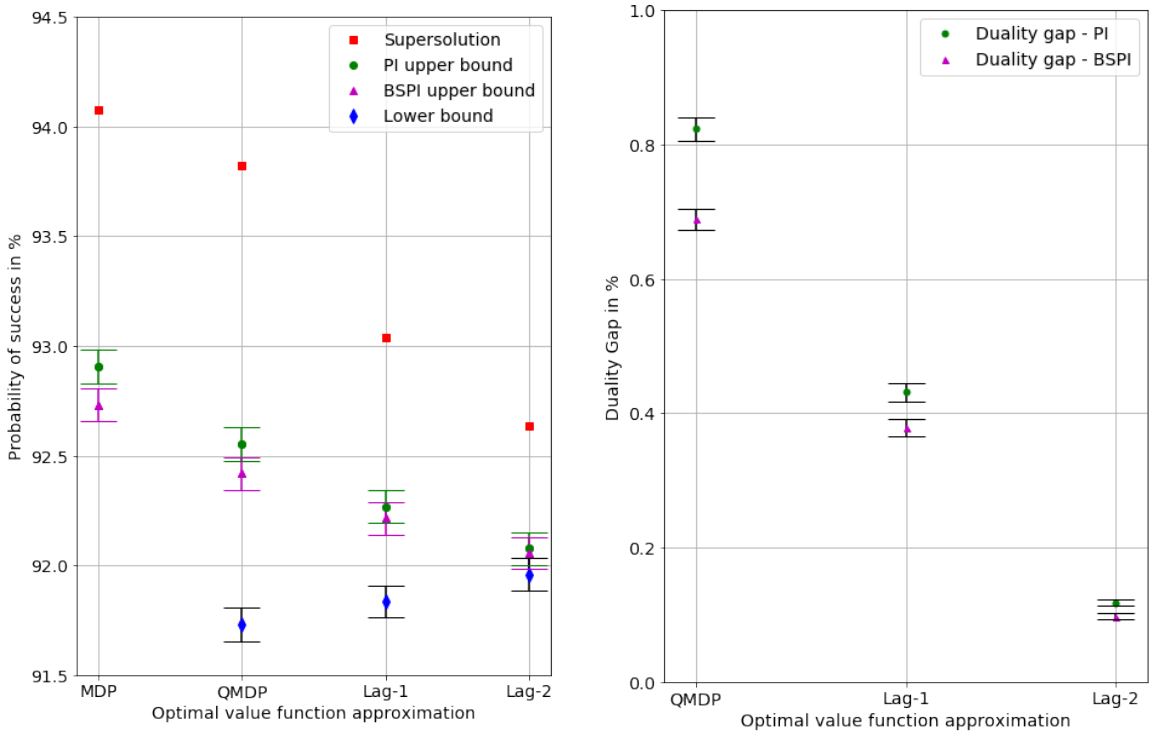


Figure 3: (a) Lower and upper bounds corresponding to each of the four AVFs. The supersolution upper bound is plotted together with the corresponding dual upper bounds obtained from the perfect information (PI) and belief state perfect information (BSPI) relaxations. Approximate 95% confidence intervals are also provided via error bars. The model parameters were $\alpha = 0.10$ and $T = 10$. (b) Duality gap estimates and confidence intervals for the value function approximations from Figure 3a. Details on how the duality gap can be estimated directly are provided in Appendix D.

Figure 3a displays lower and upper bounds corresponding to each of the four AVFs with $\alpha = 0.10$ and $T = 10$ while Figure 3b focuses directly on the corresponding duality gaps. Approximate 95% confidence intervals are also provided and so we see 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 observe that regardless of the AVF (that we used to construct the penalties and resulting change-of-measure), the BSPI bound is always superior to the corresponding PI bound.

We also note that the best duality gap (approximately 92.06% – 91.96% = 0.10%) is approximately an 85% 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

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 dual upper bounds and duality gaps (DG). All numbers are expressed as percentages. Run-times for control variate / penalty calculations were allocated to the lower bound run times.

Approx.	MDP		QMDP		Lag-1		Lag-2	
	LB*	DG	LB	DG	LB	DG	LB	DG
PI results								
Mean	-	1.15	91.73	0.82	91.84	0.43	91.96	0.12
Std. dev.	-	0.016	0.039	0.009	0.038	0.007	0.038	0.002
Run time (in minutes)	-	0.26	6.56	0.49	6.97	0.55	233	1.04
Supersolution UB	94.08		93.82		93.04		92.64	
DG reduction	51%		61%		64%		82%	
BSPI results								
Mean	-	0.97	91.73	0.69	91.84	0.38	91.96	0.10
Std. dev.	-	0.015	0.039	0.008	0.038	0.007	0.038	0.002
Run time (in minutes)	-	0.41	6.58	0.77	6.89	0.81	235	1.59
Supersolution UB	94.08		93.82		93.04		92.64	
DG reduction	59%		67%		68%		85%	

*There is no greedy policy w.r.t. the MDP AVF.

appear very significant[13] 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 be very significant.

The number of simulated paths that we used to generate the various PI and BSPI 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 BSPI Lag-2 duality gap is a mere 0.10%. The most obvious feature of the tables is how little time was required to compute the dual bounds in comparison to the lower bounds. This comparison is a somewhat 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. It is therefore fairer to add the run-times for the LB and DG columns and interpret that as the overall time required to compute the lower bounds and duality gap. We do note, however, that the reported standard errors are very small and so we could have used significantly fewer sample paths to still obtain sufficiently accurate estimates of the lower bounds and duality gaps.

8. An Application to Multiaccess Communication

Our second application is a well-known[14] multiaccess communication problem in which multiple remote users share a common channel. Users with information packets wish to transmit them through the channel and

[13] In addition, these improvements in the upper bound are conservative because the duality gaps include the difference between the best lower bound and the unknown optimal value function.

[14] See, for example, Chapter 4 of [7] for an overview of the problem.

this can only be done at integer times. Users only submit at most one packet per time slot. If only one user submits a packet through the channel in a given time slot then the packet will be successfully transmitted in that slot. If more than one user submits a packet, however, then the packets will collide, transmission fails and the packets are returned to their respective users to be sent at a later time slot. If no packet was sent during a time slot, then the system is said to be idle in that slot. Users cannot communicate with each other and therefore do not know the action histories of other users.

The total number of packets waiting to be delivered at time t is called the *backlog* and is denoted by h_t. While the backlog is not directly observed by the users, they do know the history of the channel activity via observations of collisions ($o_t = 2$), successful transmissions ($o_t = 1$) and idle time slots ($o_t = 0$). In addition, new packets arrive randomly to the backlog at the end of period t. The number of arrivals, denoted by $z_t \geq 0$, are assumed to follow some discrete probability distribution independent of prior arrivals, and they can be first scheduled for transmission beginning in period $t+1$. The backlog therefore evolves according to

$$h_{t+1} = \begin{cases} h_t + z_t - 1, & \text{if } o_t = 1 \\ h_t + z_t, & \text{otherwise.} \end{cases} \tag{49}$$

The *slotted Aloha* scheduling strategy prescribes each packet in the backlog to be scheduled for transmission with probability $a_t \in \mathcal{A} := [0,1]$. This probability is common to all waiting packets and transmission attempts are independent across packages. It is therefore easy to see that the probability of a transmission ($o_t = 1$) during slot t is $h_t a_t (1-a_t)^{h_t - 1}$. We assume a reward of $r_t(h_t)$ is obtained at time t where $r_t(\cdot)$ is a monotonically decreasing function of the backlog. The objective is to choose a transmission probability a_t to maintain a small backlog or equivalently, to maximize the probability of a transmission. In the fully-observable case where h_t is observed by the DM, it is straightforward to see that the maximum transmission probability is attained at $a_t = 1/h_t$ when $h_t \geq 1$. However, in the POMDP setting where h_t is not directly observable computing an optimal policy is generally intractable.

In order to adapt this problem to our finite state and action framework, we restrict the maximum number of packets in the backlog to be $M_h = 30$, so that $h_t \in \mathcal{H} = \{0, 1, \ldots, M_h\}$. We assume that arrivals z_t follow a Poisson distribution with mean λ, but truncate this distribution so that, if the current backlog is h_t, then the maximum number of arrivals is limited to $M_h - h_t$. This is easily accomplished by taking

$$P_z(k \mid h_t) := P(z_t = k \mid h_t) = \frac{f(k;\lambda)}{F(M_h - h_t;\lambda)}, \text{ for } k = 0, \ldots, M_h - h_t \tag{50}$$

where $f(\cdot;\lambda)$ and $F(\cdot;\lambda)$ denote the PMF and CDF, respectively, of the Poisson distribution with parameter λ. To deal with the continuous action space, we must discretize $[0,1]$. Following [12], and recalling that $a_t = 1/h_t$ maximizes the transmission probability for a given known state h_t, we set the discrete action set to be

$$\mathcal{A} := \left\{ \frac{1}{m} : m = 1, \ldots, M_h \right\} \tag{51}$$

As stated earlier, observations o_t of the channel history satisfy $o_t \in \mathcal{O} = \{0, 1, 2\}$. The observation

probabilities depend on the current backlog h_t and decision a_t, and satisfy

$$B_{ho}(a) := \begin{cases} (1-a)^h, & \text{if } o = 0 \\ ha(1-a)^{h-1}, & \text{if } o = 1 \\ 1-(1-a)^h - ha(1-a)^{h-1}, & \text{if } o = 2 \end{cases} \tag{52}$$

where $B_{ho}(a) := \mathbb{P}(o_t = o \mid h_t = h, a_t = a)$. The state transmission probabilities implied by (49) satisfy for $h, h' \in \{0, 1, \ldots, M_h\}$

$$P_{hh'}(o) = \begin{cases} 0, & \text{if } h' < h-1, \\ P_z(h'-h+1 \mid h) & \text{if } o = 1 \text{ and } h' \geq h-1, \\ P_z(h'-h \mid h) & \text{if } o \in \{0,2\} \text{ and } h' \geq h \end{cases} \tag{53}$$

where $P_{hh'}(o) := \mathbb{P}(h_{t+1} = h' \mid h_t = h, o_t = o)$ and where $P_z(k \mid h)$ corresponds to the probability mass function of the truncated Poisson arrivals given in (50).

A couple of observations are in order. First, we note that in contrast to our earlier description of the POMDP framework, we assume here that the observation o_t is a function of the *current action* a_t rather than the previous action a_{t-1}. This results in a slightly different but equally straightforward filtering algorithm to compute the belief-state any point in time. It also means that conditional on the observation o_t, the hidden-state dynamics are action-independent. This means that in defining an action-independent change-of-measure it will only be necessary to change the observation probabilities $B_{ho}(a)$.

8.1. Value Function Approximations

To simplify matters we only consider the MDP and QMDP AVFs in this application. They satisfy

$$\widetilde{V}_t^{\text{MDP}}(\pi_t) := \sum_{h \in \mathcal{H}} \pi_t(h) \max_{a_t \in \mathcal{A}} V_t^{\text{Q}}(h, a_t) \tag{54}$$

$$\widetilde{V}_t^{\text{Q}}(\pi_t) := \max_{a_t} \sum_{h \in \mathcal{H}} \pi_t(h) V_t^{\text{Q}}(h, a_t) \tag{55}$$

where

$$V_t^{\text{Q}}(h,a) := r_t(h) + \sum_{h' \in \mathcal{H}} \sum_{o \in \mathcal{O}} P_{hh'}(o) B_{ho}(a) V_{t+1}^{\text{MDP}}(h')$$

$$V_t^{\text{MDP}}(h) := \max_{a_t \in \mathcal{A}} V_t^{\text{Q}}(h, a_t)$$

for $t \in \{0, \ldots, T\}$ with terminal condition $V_{T+1}^{\text{MDP}} := 0$. Note that because the time t observation o_t is now a function of a_t, the belief state π_t is a function of the observation and action histories $o_{0:t-1}$ and $a_{0:t-1}$, respectively, rather than $o_{1:t}$ and $a_{0:t-1}$.

8.2. The Uncontrolled Formulation

Since the MDP and QMDP AVFs are[15] supersolutions, we can ignore the absolute continuity requirement and define an uncontrolled emission probability matrix according to

$$E_{ij}^t \equiv B_{ij}\big(\operatorname*{argmax}_{a\in\mathcal{A}} V_t^{\mathrm{Q}}(i,a)\big), \tag{56}$$

That is, we use the emission probability matrix induced by following a policy that is greedy w.r.t the QMDP value function approximation. Because the hidden-state transitions are already action-independent (given the current observation) we leave those dynamics unchanged under $\tilde{\mathbb{P}}$. As previously mentioned, the POMDP dynamics here are different to the baseline case as defined in Section 2 because of the timing of observations and actions whereby the the observation o_t is a function of a_t rather than a_{t-1}. This results in slightly different filtering updates and RN derivative calculations and we give them explicitly in Appendix E.

8.3. Numerical Results

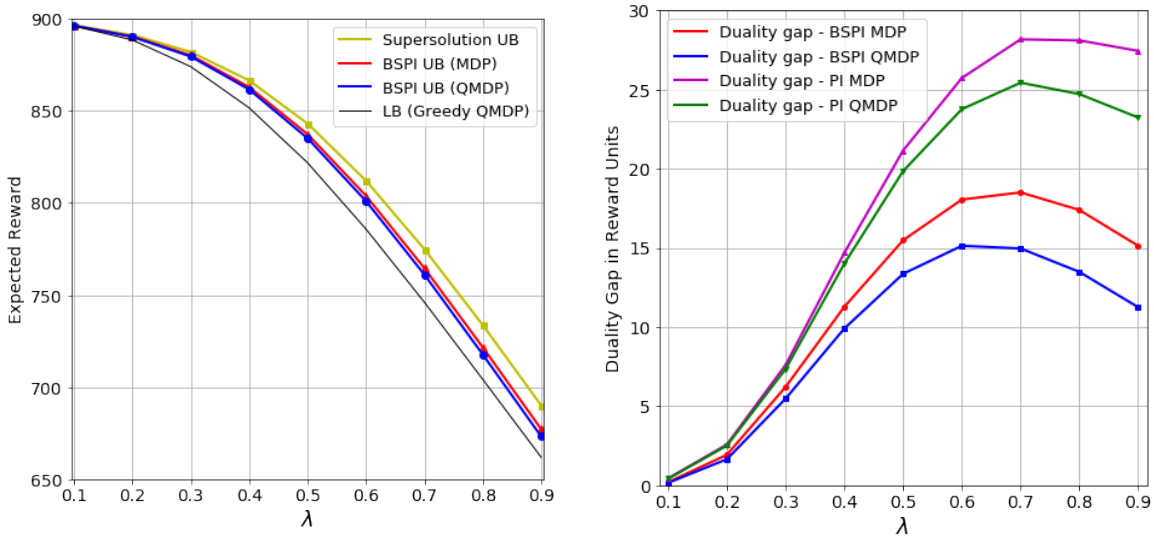
We consider a system with $T = 30$ periods and initial belief-state $\pi_0 = [1, 0, \ldots, 0]$ so that the system is initially empty w.p. 1. We assume a linear function $r_t(h_t) \coloneqq M_h - h_t$ so that the reward is maximal (and equal to M_h) when the backlog is zero and minimal (and equal to zero) when the backlog is at its maximum. We used 1,000 sample paths to estimate the dual upper bounds and duality gaps for the PI and BSPI relaxations.

Figure 4a displays the lower and upper bounds corresponding to each of the two AVFs used for various values of λ. We display the dual bounds in that figure for the BSPI relaxation but we remark that the PI dual bounds lie between the supersolution upper bound (the yellow curve) and the BSPI upper bound with penalties constructed using the MDP AVF (the red curve). We also note that the MDP and QMDP supersolution upper bounds are equal because by assumption the system is empty initially so that the left-hand-sides of (54) and (55) are equal at time $t = 0$. Figure 4b illustrates the duality gaps that we estimated directly for both value function approximations and for both relaxations.

A few additional observations are in order. First, we note the dual bounds for the QMDP approximation outperform the corresponding dual bounds for the MDP approximation. This is not surprising since the QMDP AVF should be a better approximation to the unknown optimal value function than the MDP approximation. Second, we observe from Figure 4a that both dual bounds obtained from the MDP and QMDP approximations improve upon the supersolution upper bound. (This was also true for the PI relaxation dual bounds.) Finally, we observe that the dual gaps increase in λ up to values of $\lambda \approx 0.7$, and decrease in λ thereafter. This non-monotonicity in λ can be explained by the fact that as $\lambda \nearrow 1$ the system becomes rapidly saturated in which case the DM can infer with a higher degree of confidence (than he would be able to at intermediate values of λ) that the time t backlog is likely to be close to the system cap M_h. As a result we expect the duality gap to decrease as $\lambda \nearrow 1$. Likewise when $\lambda \searrow 0$, we expect the best duality gap to also converge to 0 since the system will generally be empty and the DM will be able to infer this with increasing confidence as fewer and fewer collisions ($o_t = 2$) occur.

When we used the MDP AVF to construct the penalties, the total running time (to calculate the lower bound and duality gap for each value of λ) was 45.9 seconds and 52.3 seconds for the PI and BSPI relax-

[15] It is easy to adapt the proofs of Appendix C (to handle the fact that the observation o_t is a function of a_t rather than a_{t-1}) to show that the MDP and QMDP AVFs are supersolutions.



(a) (b)

Figure 4: (a) Upper bounds for the slotted Aloha system as a function of the arrival parameter λ. The lower bound is obtained by simulating the policy that is greedy w.r.t. the QMDP AVF. The dual bounds are generated using the BSPI relaxation. (b) Duality gap estimates for the BSPI and PI relaxations as a function of the arrival parameter λ. The widths of the (non-displayed) 95% confidence intervals varied between approximately 0.2 for lower values of λ, to 1 for higher values of λ. The supersolution bound is the supersolution given by the MDP and Q-value functions which are coincide at time $t = 0$.

ations, respectively. Using the QMDP approximation, the corresponding times were 53.6 and 58.9 seconds, respectively.

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. We have exploited the structure of POMDPs to construct various value function approximations and show that they are supersolutions. Numerical applications to robotic control and multiaccess communications have demonstrated that significant bound improvements can be obtained using information relaxations when the penalties are constructed from supersolutions. We also used the supersolution property to estimate the duality gap directly and take advantage of the significant variance reduction that follows from this approach.

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 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* [23] framework which is popular in the decision science literature.

A third 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

[1] N. Aleksandrov and B.M. Hambly. A dual approach to multiple exercise option problems under constraints. *Mathematical Methods of Operations Research*, 71(3):503–533, 2010.

[2] L. Andersen and M. Broadie. Primal-dual simulation algorithm for pricing multidimensional american options. *Management Science*, 50(9):1222–1234, 2004.

[3] S. R. Balseiro, D. B. Brown, and C. Chen. Static routing in stochastic scheduling: performance guarantees and asymptotic optimality. Working paper, Duke University, 2016.

[4] S.R. Balseiro and D.B. Brown. Approximations to stochastic dynamic programs via information relaxation duality. Working paper, Duke University, 2016.

[5] C. Bender. Primal and dual pricing of multiple exercise options in continuous time. *SIAM Journal of Financial Mathematics*, 2:562–586, 2011.

[6] C. Bender, C. Gartner, and N. Schweizer. Pathwise dynamic programming. Preprint, 2016.

[7] Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control*, volume 1. Athena Scientific, 4th edition, 2017.

[8] D. B. Brown and M. B. Haugh. Information relaxation bounds for infinite horizon markov decision processes. *Operations Research*, 65(5):1355–1379, 2017.

[9] D. B. Brown and J. E. Smith. Information relaxations, duality, and convex stochastic dynamic programs. *Operations Research*, 62(6):1394–1415, 2014.

[10] D. B. Brown, J. E. Smith, and P. Sun. Information relaxations and duality in stochastic dynamic programs. *Operations Research*, 58(4-part-1):785–801, 2010.

[11] D.B. Brown and J.E. Smith. Dynamic portfolio optimization with transaction costs: heuristics and dual bounds. *Management Science*, 57(10):1752–1770, 2011.

[12] Anthony R. Cassandra. *Exact and Approximate Algorithms for Partially Observed Markov Decision Processes*. PhD thesis, Brown University, 1998.

[13] S. Chandramouli and M.B. Haugh. A unified approach to multiple stopping and duality. *Operations Research Letters*, 2012.

[14] N. Chen and P. Glasserman. Additive and multiplicative duals for american option pricing. *Finance and Stochastics*, 11:153–179, 2007.

[15] V.V. Desai, V.F. Farias, and C.C. Moallemi. Bounds for markov decision processes. In *Reinforcement Learning and Approximate Dynamic Programming for Feedback Control, (F. L. Lewis, D. Liu, eds.)*. IEEE Press, 2011.

[16] A. Federgruen, D. Guetta, and G. Iyengar. Information relaxation-based lower bounds for the stochastic lot sizing problem with advanced demand information. Working paper, Columbia University, 2015.

[17] M. B. Haugh, G. Iyengar, and C. Wang. Tax-aware dynamic asset allocation. *Operations Research*, 64(4):849–866, 2016.

[18] M. B. Haugh and L. Kogan. Pricing american options: A duality approach. *Operations Research*, 52(2):258–270, 2004.

[19] M. B. Haugh and C. Wang. Dynamic portfolio execution and information relaxations. *SIAM J. Financial Math.*, 5:316–359, 2014.

[20] M. B. Haugh and C. Wang. Information relaxations and dynamic zero-sum games. Working paper, Columbia University, 2014.

[21] M. Hauskrecht. Value-function approximations for partially observable markov decision processes. *Journal of Artificial Intelligence Research*, 13:33–94, 2000.

[22] M. C. Horsch and D. Poole. An anytime algorithm for decision making under uncertainty. In *Proceedings of the 14th Annual Conference on Uncertainty in Artificial Intelligence (UAI-98)*, pages 246–55, 1998.

[23] F.V. Jensen and T.D. Nielsen. *Bayesian Networks and Decision Graphs*. Springer Verlag, 2007.

[24] L. Kogan and I. Mitra. Accuracy verification for numerical solutions of equilibrium models. Working paper, Massachusetts Insitute of Technology, 2013.

[25] D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. MIT Press, 2009.

[26] V. Krishnamurthy. *Partially Observed Markov Decision Processes: From Filtering to Controlled Sensing*. Cambride University Press, 2016.

[27] G. Lai, F. Margot, and N. Secomandi. An approximate dynamic programming approach to benchmark practice-based heuristics for natural gas storage valuation. *Operations Research*, 58(3):564–582, 2010.

[28] M.L. Littman, A.R. Cassandra, and L.P. Kaelbling. Learning policies for partially observable environments. In *Proceedings of the 12th International Conference on Machine Learning*, pages 362–70, 1995.

[29] N. Meinshausen and B.M. Hambly. Monte carlo methods for the valuation of multiple-exercise options. *Mathematical Finance*, 14(4):557–583, 2004.

[30] D. Nilsson and M. Hohle. Computing bounds on expected utilities for optimal policies based on limited information. Technical report, Dina Research, 2001.

[31] L.C.G. Rogers. Monte carlo valuation of american options. *Mathematical Finance*, 12:271–286, 2002.

[32] L.C.G. Rogers. Pathwise stochastic optimal control. *SIAM Journal on Control and Optimization*, 46:1116–1132, 2007.

[33] J. Schoenmakers. A pure martingale dual for multiple stopping. *Finance and Stochastics*, pages 1–16, 2010.

[34] F. Ye and E. Zhou. Information relaxation and dual formulation of controlled markov diffusions. *IEEE Transactions on Automatic Control*, 2014.

[35] F. Ye, H. Zhu, and E. Zhou. Weakly coupled dynamic program: Information and lagrangian relaxations. Working paper, Georgia Institute of Technology, 2014.

[36] H. Zhu, F. Ye, and E. Zhou. Solving the dual problems of dynamic programs via regression. arXiv preprint arXiv:1610.07726, 2016.

A. Change-of-Measures and RN Derivative Calculations

We now present the RN derivative calculations for the non-belief-state and belief-state formulations and we also provide explicit calculations for the robotic navigation application. The details for the multiaccess communication application are deferred to Appendix E since that model had a slightly different dependence structure for which the filtering equations must also be updated.

A.1. The Uncontrolled Non-Belief-State POMDP Formulation

While the belief-state formulation and corresponding BSPI bound are the main focus of the paper, it is convenient to begin with the non-belief-state formulation. We first recall the RN derivatives of (23) and (24) which we repeat here for the sake of convenience:

$$\phi(i,j,k,a) := \frac{P_{ij}(a)}{Q_{ij}} \cdot \frac{B_{jk}(a)}{E_{jk}} \tag{57}$$

$$\Phi_t(h_{0:t}, o_{1:t}, a_{0:t-1}) := \prod_{s=0}^{t-1} \phi(h_s, h_{s+1}, o_{s+1}, a_s). \tag{58}$$

To show that the general RN derivatives in (57) and (58) are correct under the PI relaxation framework, it suffices to prove that

$$\mathbb{E}\Big[r_t(h_t,a_t)\,\Big|\,\mathcal{F}_0\Big] = \widetilde{\mathbb{E}}\Big[\Phi_t r_t(h_t,a_t)\,\Big|\,\mathcal{F}_0\Big] \tag{59}$$

for all $t \in \{0,\dots,T\}$, where we recall that $\mathbb{E}$ and $\widetilde{\mathbb{E}}$ correspond to expectations under $\mathbb{P}$ and $\tilde{\mathbb{P}}$, respectively. We first write the expectation on the r.h.s. of (59) explicitly to obtain

$$\sum_{o_{1:t},h_{0:t}} \Phi_t(h_{0:t}, o_{1:t}, a_{0:t-1}) r_t(h_t,a_t) \tilde{\mathbb{P}}(o_{1:t}, h_{0:t} \mid \pi_0) \tag{60}$$

where π_0 is the initial hidden state distribution. From (57) and (58) we have

$$\begin{aligned}
\Phi_t(\cdot) = \prod_{s=0}^{t-1} \frac{P_{h_s h_{s+1}}(a_s) B_{h_{s+1} o_{s+1}}(a_s)}{Q_{h_s h_{s+1}} E_{h_{s+1} o_{s+1}}} &\equiv \prod_{s=0}^{t-1} \frac{\mathbb{P}_{a_s}(h_{s+1} \mid h_s)\mathbb{P}_{a_s}(o_{s+1} \mid h_{s+1})}{\tilde{\mathbb{P}}(h_{s+1} \mid h_s)\tilde{\mathbb{P}}(o_{s+1} \mid h_{s+1})} \\
&= \frac{\mathbb{P}_{a_{0:t-1}}(o_{1:t}, h_{1:t} \mid h_0)\pi_0(h_0)}{\tilde{\mathbb{P}}(o_{1:t}, h_{1:t} \mid h_0)\pi_0(h_0)} \\
&= \frac{\mathbb{P}_{a_{0:t-1}}(o_{1:t}, h_{0:t} \mid \pi_0)}{\tilde{\mathbb{P}}(o_{1:t}, h_{0:t} \mid \pi_0)}
\end{aligned} \tag{61}$$

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 (61) into (60) we obtain

$$\sum_{o_{1:t},h_{0:t}} r_t(h_t,a_t) \mathbb{P}_{a_{0:t-1}}(o_{1:t}, h_{0:t} \mid \pi_0) = \mathbb{E}\Big[r_t(h_t,a_t)\,\Big|\,\mathcal{F}_0\Big] \tag{62}$$

which establishes the correctness of (57) and (58).

The Robotic Navigation Application

In the numerical experiments of Sections 7 and 8 we constructed the penalties using supersolutions and, as explained in Appendix D, the absolute continuity of $\mathbb{P}$ w.r.t. $\tilde{\mathbb{P}}$ is not required to construct dual bounds. We therefore defined $\tilde{\mathbb{P}}$ to be the measure induced by following the policy that was greedy with respect to the AVF under consideration, i.e. the QMDP, Lag-1 or Lag-2 AVF. In the case of the robotic navigation application, the action-independent transition probabilities induced by following the policy that is greedy with respect to the QMDP AVF were defined according to

$$Q_{ij}^t \equiv P_{ij}\big(\operatorname*{argmax}_{a\in\mathcal{A}} V_t^Q(i,a)\big) \tag{63}$$

for $t \in \{0,\dots,T-1\}$. Similarly, the action-independent transition probabilities induced by following the policy that is greedy with respect to the Lag-1 AVF (41) were defined according to

$$Q_{ij}^t := P_{ij}\left(\operatorname*{argmax}_{a\in\mathcal{A}} \mathbb{E}\left[r_t(h_t,a) + V_{t+1}^{\mathrm{L}_1}(h_t,a,o_{t+1}) \mid h_t = i\right]\right) \tag{64}$$

and for the Lag-2 AVF (43) we defined

$$Q_{ij}^t \equiv P_{ij}\left(\operatorname*{argmax}_{a\in\mathcal{A}} \mathbb{E}\left[\max_{a_{t+1}} \mathbb{E}\left[r_t(h_t,a) + r_{t+1}(h_{t+1},a_{t+1}) + V_{t+2}^{L_2}(h_t,a_{t:t+1},o_{t+1:t+2}) \mid h_t = i, o_{t+1}\right] \mid h_t = i\right]\right). \tag{65}$$

Regardless of the AVF, we defined $E_{jk} := B_{jk}$ since the emission matrix B under $\mathbb{P}$ was already action-independent. The RN derivatives for the uncontrolled PI formulation are then given by

$$\begin{aligned}\Phi_t(h_{0:t},a_{0:t-1}) &:= \prod_{s=0}^{t-1}\phi_s(h_s,h_{s+1},a_s)\\ \phi_s(i,j,a) &:= \frac{P_{ij}(a)}{Q_{ij}^s}.\end{aligned} \tag{66}$$

where Q_{ij}^s is given by (63), (64) or (65), depending on the AVF under consideration.

A.2. The Uncontrolled Belief-State POMDP Formulation

We only consider uncontrolled measure changes $\tilde{\mathbb{P}}$ that are Markovian in that $\tilde{\mathbb{P}}(\pi_{1:t}) = \prod_{s=0}^{t-1}\tilde{\mathbb{P}}(\pi_{s+1} \mid \pi_s)$ for some $\tilde{\mathbb{P}}(\pi_{s+1} \mid \pi_s)$ that we must define. The corresponding RN derivatives then take the form

$$\frac{d\mathbb{P}}{d\tilde{\mathbb{P}}} =: \Phi_T^\pi(\pi_{0:T},a_{0:T-1}) := \prod_{s=0}^{T-1}\phi(\pi_s,\pi_{s+1},a_s) \tag{67}$$

$$\phi(\pi_s,\pi_{s+1},a_s) := \frac{\sum_{i,j,k}\pi(i)P_{ij}(a_s)B_{jk}(a_s)\mathbf{1}_{\{\pi_{s+1}=f(\pi_s,a_s,k)\}}}{\tilde{\mathbb{P}}(\pi_{s+1} \mid \pi_s)} \tag{68}$$

where $\pi_{0:T} := \{\pi_0,\pi_1,\dots,\pi_T\}$ and $f(\pi,a,k)$ is as defined in (18) and is the new filtered belief-state that results under $\mathbb{P}$ from taking action a and observing k when the current belief-state is π. In order to justify (67) and (68) we must show

$$\mathbb{E}\left[r_t(\pi_t,a_t)\,\middle|\,\mathcal{F}_0^\pi\right] = \tilde{\mathbb{E}}\left[\Phi_T^\pi r_t(\pi_t,a_t)\,\middle|\,\mathcal{F}_0^\pi\right] = \tilde{\mathbb{E}}\left[\Phi_t^\pi r_t(\pi_t,a_t)\,\middle|\,\mathcal{F}_0^\pi\right]$$

where the second equality follows from a standard conditioning argument. Writing the expectations explicitly, we must have

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

where $\mathbb{P}_{a_{0:t-1}}$ explicitly recognizes the dependence of the given probabilities on $a_{0:t-1}$. It is clear then that the RN derivatives must satisfy

$$\Phi_t^\pi := \frac{\mathbb{P}_{a_{0:t-1}}(\pi_{1:t})}{\tilde{\mathbb{P}}(\pi_{1:t})}. \tag{69}$$

We can compute the numerator of (69) as

$$\begin{aligned}
\mathbb{P}_{a_{0:t-1}}(\pi_{1:t}) = \prod_{s=0}^{t-1} \mathbb{P}_{a_s}(\pi_{s+1} \mid \pi_s) &= \prod_{s=0}^{t-1} \sum_{o_{s+1}} \mathbb{P}_{a_s}(o_{s+1} \mid \pi_s)\mathbb{P}_{a_s}(\pi_{s+1} \mid o_{s+1}, \pi_s) \\
&= \prod_{s=0}^{t-1} \sum_{h,h',o_{s+1}} \pi_s(h)\mathbb{P}_{a_s}(h' \mid h)\mathbb{P}_{a_s}(o_{s+1} \mid h')\mathbf{1}_{\{\pi_{s+1}=f(\pi_s,a_s,o_{s+1})\}} \\
&= \prod_{s=0}^{t-1} \sum_{h,h',o} \pi_s(h)P_{hh'}(a_s)B_{h'o}(a_s)\mathbf{1}_{\{\pi_{s+1}=f(\pi_s,a_s,o)\}}.
\end{aligned} \tag{70}$$

Substituting $\tilde{\mathbb{P}}(\pi_{1:t}) = \prod_{s=0}^{t-1} \tilde{\mathbb{P}}(\pi_{s+1} \mid \pi_s)$ and (70) into (69) then establishes that (67) and (68) are correct.

The Robotic Navigation Application

In the specific applications considered in this paper, we don't need to concern ourselves with the absolute continuity of the measure change because of the use of supersolutions to construct penalties. We can therefore utilize the same measure-changes described in Appendix A.1 for the non-belief-state formulation but with a layer of filtering included. Specifically, we define

$$\tilde{\mathbb{P}}(\pi_{s+1} \mid \pi_s) := \sum_{h,h',o} \pi_s(h)Q^s_{hh'}B_{h'o}\mathbf{1}_{\{\pi_{s+1}=\tilde{f}_s(\pi_s,o)\}} \tag{71}$$

where $Q^s_{hh'}$ is given by either (63), (64) or (65), depending on the AVF under consideration, and where the j^{th} component of $\tilde{f}_s(\pi_s; o)$ in the $|\mathcal{H}|$-dimensional simplex is defined according to

$$\tilde{f}_s(j; \pi_s; k) = \frac{\sum_i \pi_s(i)Q^s_{ij}B_{jk}}{\sum_{i,l} \pi_s(i)Q^s_{il}B_{lk}}. \tag{72}$$

The RN derivatives for the BSPI relaxation are then given by (67) with the denominator in (68) given by (71).

A.3. Relating the RN Derivatives for the Uncontrolled PI and BSPI Formulations

It is interesting to compare the one-step RN derivatives given by (57) and (68) for the PI and BSPI uncontrolled formulations, respectively. As discussed at the end of Section 5.2, the PI and BSPI change-of-measures are closely related when both are constructed from the same action-independent transition and emission matrices Q^t and E^t. In particular, we argued that the RN derivative term in (68) (or equivalently (15)) could

be loosely viewed as a filtered version of the RN derivative term in (57) (or equivalently (23)). To understand this relationship more clearly, we compute the $\tilde{\mathbb{P}}$-expectation of the PI one-step RN derivative conditional on π_s and π_{s+1}. Omitting the dependence of $\tilde{f}_s$, ϕ_s, Q^s, E^s on s, we obtain

$$\begin{aligned}\tilde{\mathbb{E}}[\phi(h_s, h_{s+1}, o_{s+1}, a_s) \mid \pi_{s+1}, \pi_s] &= \sum_{h_s, h_{s+1}, o_{s+1}} \phi(h_s, h_{s+1}, o_{s+1}, a_s) \tilde{\mathbb{P}}(h_s, h_{s+1}, o_{s+1} \mid \pi_{s+1}, \pi_s) \\ &= \sum_{h_s, h_{s+1}, o_{s+1}} \phi(h_s, h_{s+1}, o_{s+1}, a_s) \frac{\tilde{\mathbb{P}}(h_s, h_{s+1}, o_{s+1}, \pi_{s+1} \mid \pi_s)}{\tilde{\mathbb{P}}(\pi_{s+1} \mid \pi_s)} \end{aligned} \tag{73}$$

where the second equality was obtained by a standard conditioning argument. The probability in the numerator in the r.h.s. of (73) can be written as

$$\begin{aligned}\tilde{\mathbb{P}}(h_s, h_{s+1}, o_{s+1}, \pi_{s+1} \mid \pi_s) &\overset{(a)}{=} \tilde{\mathbb{P}}(\pi_{s+1} \mid h_s, h_{s+1}, o_{s+1}, \pi_s) \tilde{\mathbb{P}}(o_{s+1} \mid h_s, h_{s+1}, \pi_s) \tilde{\mathbb{P}}(h_{s+1} \mid h_s, \pi_s) \pi_s(h_s) \\ &\overset{(b)}{=} \tilde{\mathbb{P}}(\pi_{s+1} \mid o_{s+1}, \pi_s) \tilde{\mathbb{P}}(o_{s+1} \mid h_{s+1}) \tilde{\mathbb{P}}(h_{s+1} \mid h_s) \pi_s(h_s) \\ &\overset{(c)}{=} \mathbf{1}_{\{\pi_{s+1} = \tilde{f}(\pi_s, o_{s+1})\}} \tilde{\mathbb{P}}(o_{s+1} \mid h_{s+1}) \tilde{\mathbb{P}}(h_{s+1} \mid h_s) \pi_s(h_s) \end{aligned} \tag{74}$$

where (a) is obtained by a sequence of conditioning arguments and noting that $\tilde{\mathbb{P}}(h_s \mid \pi_s) = \pi_s(h_s)$ (b) comes from noting that π_{s+1} is independent of h_s and h_{s+1} given o_{s+1} and π_s, that o_{s+1} is independent of h_s and π_s conditional on h_{s+1}, and that h_{s+1} is independent of π_s conditional on h_s. Finally, (c) follows by noting that $\tilde{\mathbb{P}}(\pi_{s+1} \mid o_{s+1}, \pi_s) = \mathbf{1}_{\{\pi_{s+1} = \tilde{f}(\pi_s, o_{s+1})\}}$ where the components of $\tilde{f}$ are given by the filtering expression (72). Substituting (74) and the expression for $\phi(\cdot)$ given in (57) into (73), we obtain

$$\begin{aligned}&\tilde{\mathbb{E}}[\phi(h_s, h_{s+1}, o_{s+1}, a_s) \mid \pi_{s+1}, \pi_s] \\ &= \sum_{h_s, h_{s+1}, o_{s+1}} \frac{P_{h_s, h_{s+1}}(a_s)}{Q_{h_s, h_{s+1}}} \frac{B_{h_{s+1}, o_{s+1}}(a_s)}{E_{h_{s+1}, o_{s+1}}} \frac{\mathbf{1}_{\{\pi_{s+1} = \tilde{f}(\pi_s, o_{s+1})\}} \tilde{\mathbb{P}}(o_{s+1} \mid h_{s+1}) \tilde{\mathbb{P}}(h_{s+1} \mid h_s) \pi_s(h_s)}{\tilde{\mathbb{P}}(\pi_{s+1} \mid \pi_s)} \\ &= \sum_{h_s, h_{s+1}, o_{s+1}} \frac{\pi_s(h_s) P_{h_s, h_{s+1}}(a_s) B_{h_{s+1}, o_{s+1}}(a_s) \mathbf{1}_{\{\pi_{s+1} = \tilde{f}(\pi_s, o_{s+1})\}}}{\tilde{\mathbb{P}}(\pi_{s+1} \mid \pi_s)} \end{aligned} \tag{75}$$

where the second equality follows because $Q_{h_s, h_{s+1}} \equiv \tilde{\mathbb{P}}(h_{s+1} \mid h_s)$ and $E_{h_{s+1}, o_{s+1}} \equiv \tilde{\mathbb{P}}(o_{s+1} \mid h_{s+1})$.

We can now compare the definition of the RN derivative given in (68) for the BSPI relaxation with the r.h.s. of (75). The only difference between the two expressions are the indicator functions $\mathbf{1}_{\{\pi' = f(\pi, a, k)\}}$ and $\mathbf{1}_{\{\pi_{s+1} = \tilde{f}(\pi_s, o_{s+1})\}}$ appearing in the numerators of (68) and (75), respectively. We recall that $f(\pi, a, k)$ is the $\mathbb{P}$-filtering update function whereas $\tilde{f}(\pi_s, o_{s+1})$ is the action-independent $\tilde{\mathbb{P}}$-filtering update function and so it's not surprising that $\mathbf{1}_{\{\pi_{s+1} = \tilde{f}(\pi_s, o_{s+1})\}}$ appears in the numerator of (75) as it's a $\tilde{\mathbb{P}}$ conditional expectation that led to it. If $\tilde{\mathbb{P}}$ is obtained by taking $Q = P(a')$ and $E = B(a')$ for some fixed action a', however, then it's not hard to see that both (68) and (75) collapse to the value 1 when $a_s = a'$. This is to be expected of course since if $\tilde{\mathbb{P}}$ is the action-independent measure induced by following some feasible policy, then the RN derivatives (for both uncontrolled formulations) should collapse to 1 when the action a coincides with the action chosen by the policy.

A.4. Defining a Belief-State Measure Change Where $\mathbb{P}$ is Absolutely Continuous W.R.T. $\tilde{\mathbb{P}}$

In some circumstances it may be desirable to use penalties constructed from AVFs that are *not* supersolutions. In that case, it will be necessary to ensure that the measure change $\tilde{\mathbb{P}}$ will be such that $\mathbb{P}$ will be absolutely continuous w.r.t $\tilde{\mathbb{P}}$. An obvious approach to defining uncontrolled belief-state dynamics for π_t would be to use (16) and (17) to generate uncontrolled hidden state / observation sequences and then simply use the generated observations to update the belief state appropriately, beginning with π_0. This is precisely what we did in (14) and (15). 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 (16) and (17) do satisfy their absolute continuity conditions as discussed at the beginning of Section 4.3. To see this note that the belief state updates under $\mathbb{P}$ are computed according to

$$\pi_{t+1}(j;a,k) = \frac{\sum_i \pi_t(i)P_{ij}(a)B_{jk}(a)}{\sum_{i,l} \pi_t(i)P_{il}(a)B_{lk}(a)} \tag{76}$$

where we explicitly recognize the $\mathbb{P}$-dependence of π_{t+1} on $a_t = a$ and $o_{t+1} = k$. In contrast, the belief state updates under $\tilde{\mathbb{P}}$ are computed according to

$$\tilde{\pi}_{t+1}(j;k) = \frac{\sum_i \pi_t(i)Q_{ij}E_{jk}}{\sum_{i,l} \pi_t(i)Q_{il}E_{lk}}. \tag{77}$$

Even if Q and E satisfy their absolute continuity conditions, there will in general be $\pi_{t+1}(\cdot\,;a,k)$'s that satisfy $\mathbb{P}(\pi_{t+1}(\cdot\,;a,k) \mid \pi_t) > 0$ and $\tilde{\mathbb{P}}(\pi_{t+1}(\cdot\,;a,k) \mid \pi_t) = 0$. As such, $\mathbb{P}$ will not be absolutely continuous w.r.t. $\tilde{\mathbb{P}}$.

There are many ways to resolve this issue and we now describe one such approach. Specifically, we assume that under $\tilde{\mathbb{P}}$ the current belief state π transitions with strictly positive probability to any[16] belief state π' which is feasible for some available action $a \in \mathcal{A}$ given π. We then define the belief-state transition probability

$$\tilde{\mathbb{P}}(\pi' \mid \pi) := \frac{1}{|\mathcal{A}| \times |\mathcal{O}|} \sum_{(a,o)\in\mathcal{A}\times\mathcal{O}} \mathbf{1}_{\{\pi' = f(\pi;a,o)\}} \tag{78}$$

where $f(\pi;a,o)$ lies in the $|\mathcal{H}|$-dimensional simplex and is defined in (18). In this case all π''s which have strictly positive probability (conditional on π) under $\mathbb{P}$ become equally likely under $\tilde{\mathbb{P}}$ (conditional on π).

It is of course possible to define other $\tilde{\mathbb{P}}$'s and still guarantee that $\mathbb{P}$ is absolutely continuous w.r.t. $\tilde{\mathbb{P}}$. For example, we could define $\tilde{\mathbb{P}}$ so that at each time t every feasible action $a \in \mathcal{A}$ is chosen with strictly positive probability. Then any π' that has strictly positive probability under $\mathbb{P}$ will also have strictly positive probability under $\tilde{\mathbb{P}}$.

B. The Lag-1 and Lag-2 Approximate Value Functions

B.1. Computing the Optimal Value Function for the Lag-1 MDP

The Lag-1 formulation corresponds to the relaxed problem in which the time t DM knows the true state h_{t-1} that prevailed at time $t-1$, the observation history $o_{0:t}$ and the action history $a_{0:t-1}$. Given the dependence structure of the hidden states and observations in the POMDP, it follows that the Lag-1 optimal value function V_t^{L1} only depends on (h_{t-1}, a_{t-1}, o_t). The terminal value function satisfies $V_T^{L_1}(h_{T-1}, a_{T-1}, o_T) :=$

[16] Note that while there are infinitely many points in the $|\mathcal{H}|$-dimensional simplex only a finite number of these points will have a strictly positive probability under $\mathbb{P}$ conditional on π_0. These points with strictly positive $\mathbb{P}$-probability arise from the various possible combinations of action / observation sequences which are finite in number by assumption.

$r_T(o_T) = r_T(h_T)$ with

$$V_t^{\mathrm{L}_1}(h_{t-1}, a_{t-1}, o_t) := \max_{a_t} \mathbb{E}[r_t(h_t, a_t) + V_{t+1}^{\mathrm{L}_1}(h_t, a_t, o_{t+1}) \mid h_{t-1}, o_t]$$
$$= \max_{a_t} \sum_{h_t, o_{t+1}} \mathbb{P}_{a_{t-1:t}}(h_t, o_{t+1} \mid h_{t-1}, o_t)[r_t(h_t, a_t) + V_{t+1}^{\mathrm{L}_1}(h_t, a_t, o_{t+1})]$$

for $t \in \{1, \ldots, T-1\}$ and where $\mathbb{P}_{a_{t-1:t}}$ recognizes the dependence of the conditional PMF on the actions $a_{t-1:t}$. These probabilities can be calculated explicitly using standard manipulations. In particular, we have

$$\begin{aligned}\mathbb{P}_{a_{t-1:t}}(h_t, o_{t+1} \mid h_{t-1}, o_t) &= \frac{\mathbb{P}_{a_{t-1:t}}(h_t, o_t, o_{t+1} \mid h_{t-1})}{\mathbb{P}_{a_{t-1:t}}(o_t \mid h_{t-1})} \\ &= \frac{\sum_{h_{t+1}} \mathbb{P}_{a_{t-1:t}}(h_t, o_t, h_{t+1}, o_{t+1} \mid h_{t-1})}{\sum_{h_t} \mathbb{P}_{a_{t-1:t}}(h_t, o_t \mid h_{t-1})} \\ &= \frac{P_{h_{t-1}h_t} B_{h_t o_t} \sum_{h_{t+1}} P_{h_t h_{t+1}} B_{h_{t+1} o_{t+1}}}{\sum_{h_t} P_{h_{t-1}h_t} B_{h_t o_t}} \end{aligned} \tag{79}$$

where for ease of exposition we have suppressed[17] the dependence of the various quantities in (79) on the various actions. We can calculate $V_0^{L_1}$ in a similar fashion by noting that

$$V_0^{\mathrm{L}_1}(o_0) := \max_{a_0} \mathbb{E}[r_0(h_0, a_0) + V_1^{L_1}(h_0, a_0, o_1) \mid o_0]$$
$$= \max_{a_0} \sum_{h_0, o_1} \mathbb{P}_{a_0}(h_0, o_1 \mid o_0)[r_0(h_0, a_0) + V_1^{\mathrm{L}_1}(h_0, a_0, o_1)]$$

where $\mathbb{P}_{a_0}(h_0, o_1 \mid o_0)$ can be calculated as in (79) but with $P_{h_{t-1}h_t}(a_{t-1})$ replaced by $P(h_0)$.

B.2. The Lag-2 Approximate Value Function

We must first show how the optimal value function for the Lag-2 MDP can be calculated.

Computing the Optimal Value Function for the Lag-2 MDP

The Lag-2 formulation corresponds to the relaxed problem in which the time t DM knows the true state h_{t-2} that prevailed at time $t-2$, the observation history $o_{0:t}$ and the action history $a_{0:t-1}$. The terminal value function satisfies $V_T^{\mathrm{L}_2}(h_{T-2}, a_{T-2:T-1}, o_{T-1:T}) := r_T(o_T) = r_T(h_T)$ with

$$V_t^{\mathrm{L}_2}(h_{t-2}, a_{t-2:t-1}, o_{t-1:t}) := \max_{a_t} \mathbb{E}[r_t(h_t, a_t) + V_{t+1}^{\mathrm{L}_2}(h_{t-1}, a_{t-1:t}, o_{t:t+1}) \mid h_{t-2}, o_{t-1:t}] \tag{80}$$
$$= \max_{a_t} \sum_{h_{t-1:t}, o_{t+1}} \mathbb{P}_{a_{t-2:t}}(h_{t-1:t}, o_{t+1} \mid h_{t-2}, o_{t-1:t})[r_t(h_t, a_t) + V_{t+1}^{\mathrm{L}_2}(h_{t-1}, a_{t-1:t}, o_{t:t+1})]$$

[17] In these appendices we will often suppress the dependence of the various transition and observation probabilities on the chosen actions. For example, it should be clear in (79) that $P_{h_{t-1}h_t}$ depends on a_{t-1} while $B_{h_{t+1}o_{t+1}}$ depend on a_t.

for $t \in \{2, \ldots, T-1\}$ and where we use $\mathbb{P}_{a_{t-2:t}}$ to denote a probability that depends on $a_{t-2:t}$. We note it is straightforward to calculate $\mathbb{P}_{a_{t-2:t}}(\cdot \mid \cdot)$ using standard arguments. Specifically, we have

$$\begin{aligned}\mathbb{P}_{a_{t-2:t}}(h_{t-1:t}, o_{t+1} \mid h_{t-2}, o_{t-1:t}) &= \frac{\mathbb{P}_{a_{t-2:t}}(h_{t-1:t}, o_{t-1:t+1} \mid h_{t-2})}{\mathbb{P}_{a_{t-2:t}}(o_{t-1:t} \mid h_{t-2})} \\ &= \frac{\sum_{h_{t+1}} \mathbb{P}_{a_{t-2:t}}(h_{t-1:t+1}, o_{t-1:t+1} \mid h_{t-2})}{\sum_{h_{t-1:t}} \mathbb{P}_{a_{t-2:t}}(h_{t-1:t}, o_{t-1:t} \mid h_{t-2})} \\ &= \frac{PB_{t-2} PB_{t-1} \sum_{h_{t+1}} PB_t}{\sum_{h_{t-1}, h_t} PB_{t-2} PB_{t-1}} \end{aligned} \tag{81}$$

where we use PB_t to denote $P_{h_t h_{t+1}} B_{h_{t+1} o_{t+1}}$ and again we have suppressed the action dependence of the various terms. A slightly different calculation is required for each of V_0^{L2} and V_1^{L2} as there is no hidden state information available at times $t = 0$ and $t = 1$. For $t = 1$ we have

$$\begin{aligned} V_1^{\text{L2}}(o_{0:1}, a_0) &\coloneqq \max_{a_1} \mathbb{E}[r_1(h_1, a_1) + V_2^{\text{L2}}(h_0, a_{0:1}, o_{1:2}) \mid o_{0:1}] \\ &= \max_{a_1} \sum_{h_0, o_2} \mathbb{P}_{a_{0:1}}(h_{0:1}, o_1 \mid o_{0:1})[r_1(h_1, a_1) + V_2^{\text{L2}}(h_0, a_{0:1}, o_{1:2})] \end{aligned}$$

where $\mathbb{P}_{a_{0:1}}(h_{0:1}, o_1 \mid o_{0:1})$ is calculated as in (81) 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

$$\begin{aligned} V_0^{\text{L2}}(o_0) &\coloneqq \max_{a_0} \mathbb{E}[r_0(h_0, a_0) + V_1^{\text{L2}}(o_{0:1}, a_0) \mid o_0] \\ &= \max_{a_0} \sum_{o_1} \mathbb{P}_{a_0}(h_0, o_1 \mid o_0)[r_0(h_0, a_0) + V_1^{\text{L2}}(o_{0:1}, a_0)] \end{aligned}$$

and where

$$\mathbb{P}_{a_0}(h_0, o_1 | o_0) = \frac{P(h_0) B_{h_0 o_0} \sum_{h_1} PB_0}{\sum_{h_0} P(h_0) B_{h_0 o_0}}.$$

Computing the Lag-2 Approximate Value Function for the POMDP

Following (43) we can write the Lag-2 AVF as

$$\widetilde{V}_t^{\text{L2}}(\pi_t) = \max_{a_t} \mathbb{E}[r_t(h_t, a_t) + \max_{a_{t+1}} \mathbb{E}[r_{t+1}(h_{t+1}, a_{t+1}) + V_{t+2}^{\text{L2}}(h_t, a_{t:t+1}, o_{t+1:t+2}) \mid \mathcal{F}_t^{\pi}, o_{t+1}] \mid \mathcal{F}_t^{\pi}]. \tag{82}$$

The inner expectation in (82) can be calculated according to

$$\sum_{h_{t:t+1}, o_{t+2}} \mathbb{P}_{a_{t:t+1}}(h_{t:t+1}, o_{t+2} \mid \pi_t, o_{t+1})[r_{t+1}(h_{t+1}, a_{t+1}) + V_{t+2}^{\text{L2}}(h_t, a_{t:t+1}, o_{t+1:t+2})]. \tag{83}$$

The probability in (83) can then be computed using standard arguments. In particular, we have

$$\begin{aligned}\mathbb{P}_{a_{t:t+1}}(h_{t:t+1}, o_{t+2} \mid \pi_t, o_{t+1}) &= \frac{\mathbb{P}_{a_{t:t+1}}(h_{t:t+1}, o_{t+1:t+2} \mid \pi_t)}{\mathbb{P}_{a_t}(o_{t+1} \mid \pi_t)} \\ &= \frac{\sum_{h_{t+2}} \mathbb{P}_{a_{t:t+1}}(h_{t:t+2}, o_{t+1:t+2} \mid \pi_t)}{\sum_{h_{t:t+1}} \mathbb{P}_{a_t}(h_{t:t+1}, o_{t+1} \mid \pi_t)} \\ &= \frac{\pi_t(h_t) PB_t \sum_{h_{t+2}} PB_{t+1}}{\sum_{h_t, h_{t+1}} \pi_t(h_t) PB_t} \end{aligned} \tag{84}$$

where we once again denote by $PB_t \equiv P_{h_t h_{t+1}}(a_t) B_{h_{t+1} o_{t+1}}(a_t)$.

Remark B.1. *We note that if $T = 2$, then we recover the optimal value $V_0^*(\pi_0)$ of the POMDP. In particular,*

$$\begin{aligned}\widetilde{V}_0^{L_2}(\pi_0) &= \max_{a_0} \mathbb{E}[\max_{a_1} \mathbb{E}[r_0(h_0, a_0) + r_1(h_1, a_1) + V_2^{L_2}(h_0, a_{0:1}, o_{1:2}) \mid \mathcal{F}_0^\pi, o_1] \mid \mathcal{F}_0^\pi] \\ &= \max_{a_0} \mathbb{E}[r_0(h_0, a_0) + \max_{a_1} \mathbb{E}[r_1(h_1, a_1) + r_2(h_2) \mid \mathcal{F}_0^\pi, o_1] \mid \mathcal{F}_0^\pi] \;=\; V_0^*(\pi_0)\end{aligned}$$

where the second equality follows from the tower property of conditional expectations.

B.3. Comparing the Lag-1 and Lag-2 Approximate Value Functions

We begin by proving the unsurprising result that the Lag-2 AVF is tighter than the Lag-1 AVF.

Proposition B.1. *For all t we have $V_t^*(\pi_t) \le \widetilde{V}_t^{L_2}(\pi_t) \le \widetilde{V}_t^{L_1}(\pi_t)$.*

Proof. We show in Appendix C that $\widetilde{V}_t^{\mathrm{L}_2}(\pi_t)$ is a supersolution and so it follows that $V_t^*(\pi_t) \le \widetilde{V}_t^{\mathrm{L}_2}(\pi_t)$. To prove the second inequality we begin with the definition of $\widetilde{V}_t^{\mathrm{L}_1}(\pi_t)$ in (41) for $t = 0, \ldots, T-2$. (We recall that at $t = T-1$ and $t = T$ we have that $\widetilde{V}_t^{\mathrm{L}_2}(\pi_t) = \widetilde{V}_t^{\mathrm{L}_1}(\pi_t)$ for all π_t.) We obtain

$$\begin{aligned}\widetilde{V}_t^{\mathrm{L}_1}(\pi_t) &:= \max_{a_t} \mathbb{E}\left[r_t(h_t, a_t) + V_{t+1}^{\mathrm{L}_1}(h_t, a_t, o_{t+1}) \mid \mathcal{F}_t^\pi\right] \\ &\overset{(a)}{=} \max_{a_t} \mathbb{E}_{h_t, o_{t+1}}\left[r_t(h_t, a_t) + \max_{a_{t+1}} \mathbb{E}_{h_{t+1}, o_{t+2}}\left[r_{t+1}(h_{t+1}, a_{t+1}) + V_{t+2}^{\mathrm{L}_1}(h_{t+1}, a_{t+1}, o_{t+2}) \mid h_t, o_{t+1}\right] \mid \mathcal{F}_t^\pi\right] \\ &\overset{(b)}{=} \max_{a_t} \mathbb{E}_{h_t, o_{t+1}}\left[\max_{a_{t+1}} r_t(h_t, a_t) + \mathbb{E}_{h_{t+1}, o_{t+2}}\left[r_{t+1}(h_{t+1}, a_{t+1}) + V_{t+2}^{\mathrm{L}_1}(h_{t+1}, a_{t+1}, o_{t+2}) \mid h_t, o_{t+1}\right] \mid \mathcal{F}_t^\pi\right] \\ &\overset{(c)}{\ge} \max_{a_t} \mathbb{E}_{o_{t+1}}\left[\max_{a_{t+1}} \mathbb{E}_{h_t}\left[r_t(h_t, a_t) + \mathbb{E}_{h_{t+1}, o_{t+2}}\left[r_{t+1}(h_{t+1}, a_{t+1}) + V_{t+2}^{\mathrm{L}_1}(h_{t+1}, a_{t+1}, o_{t+2}) \mid h_t, o_{t+1}\right] \mid \mathcal{F}_t^\pi, o_{t+1}\right] \mid \mathcal{F}_t^\pi\right] \\ &\overset{(d)}{=} \max_{a_t} \mathbb{E}_{o_{t+1}}\left[\max_{a_{t+1}} \mathbb{E}_{h_{t:t+1}, o_{t+2}}\left[r_t(h_t, a_t) + r_{t+1}(h_{t+1}, a_{t+1}) + V_{t+2}^{\mathrm{L}_1}(h_{t+1}, a_{t+1}, o_{t+2}) \mid \mathcal{F}_t^\pi, o_{t+1}\right] \mid \mathcal{F}_t^\pi\right] \\ &\overset{(e)}{=} \max_{a_t} \mathbb{E}_{o_{t+1}}\left[\max_{a_{t+1}} \mathbb{E}_{h_{t:t+1}, o_{t+2}}\left[r_t(h_t, a_t) + r_{t+1}(h_{t+1}, a_{t+1}) + V_{t+2}^{\mathrm{L}_2}(h_{t+1}, a_{t:t+1}, o_{t+1:t+2}) \mid \mathcal{F}_t^\pi, o_{t+1}\right] \mid \mathcal{F}_t^\pi\right] \\ &= \widetilde{V}_t^{\mathrm{L}_2}(\pi_t)\end{aligned}$$

where (a) results from using the definition of $V_{t+1}^{\mathrm{L}_1}$ and (b) follows by simply moving $r_t(h_t, a_t)$ inside the maximization of a_{t+1}. Inequality (c) results from applying Jensen's inequality when exchanging the order of the expectation w.r.t. h_t and the maximization of a_{t+1}. Equality (d) results from the tower property and noting that the argument inside the inner expectation, conditional on h_t, is independent of $\mathcal{F}_t^\pi$. Inequality (e) follows by replacing $V_{t+2}^{\mathrm{L}_1}$ with $V_{t+2}^{\mathrm{L}_2}$ and then using Lemma B.1 below. Finally, the last equality follows from the definition of $\widetilde{V}_t^{\mathrm{L}_2}(\pi_t)$ in (43). □

Lemma B.1. $\mathbb{E}[V_{t+2}^{L_1}(h_{t+1}, a_{t+1}, o_{t+2}) \mid \mathcal{F}_t^\pi, o_{t+1}] \ge \mathbb{E}[V_{t+2}^{L_2}(h_t, a_{t:t+1}, o_{t+1:t+2}) \mid \mathcal{F}_t^\pi, o_{t+1}]$ *for all $t = 0, \ldots, T-2$.*

Proof. To begin we note that it suffices to prove that

$$\mathbb{E}[V_{t+2}^{\mathrm{L}_1}(h_{t+1}, a_{t+1}, o_{t+2}) \mid h_t, o_{t+1:t+2}] \ge V_{t+2}^{\mathrm{L}_2}(h_t, a_{t:t+1}, o_{t+1:t+2}) \tag{85}$$

since taking expectation $\mathbb{E}[\cdot \mid \mathcal{F}_t^\pi, o_{t+1}]$ on both sides of (85) and applying the tower property yields[18] the

[18] Note that the term inside the expectation on the left-hand-side of (85), conditional on h_t, is independent of $\mathcal{F}_t^\pi$, and so the tower property indeed yields the result.

desired result. We now prove (85) by induction for $t = 0, \ldots, T-2$. The base case follows immediately since $V_T^{\mathrm{L}_1} = V_T^{\mathrm{L}_2} = r_T(h_T)$ and so $\mathbb{E}[V_T^{\mathrm{L}_1} \mid h_{T-2}, o_{T-1:T}] = r_T(h_T) = V_T^{\mathrm{L}_2}$ where we recall that $o_T \equiv h_T$. We now assume the result is true for time $t+3$ so that $\mathbb{E}[V_{t+3}^{\mathrm{L}_1} \mid h_{t+1}, o_{t+2:t+3}] \geq V_{t+3}^{\mathrm{L}_2}$. An application of the tower property then implies

$$\mathbb{E}[V_{t+3}^{\mathrm{L}_1}(h_{t+2}, a_{t+2}, o_{t+3}) \mid h_{t+1}, o_{t+2}] \geq \mathbb{E}[V_{t+3}^{\mathrm{L}_2}(h_{t+1}, a_{t+1:t+2}, o_{t+2:t+3}) \mid h_{t+1}, o_{t+2}]. \tag{86}$$

It then follows that

$$\begin{aligned}
&\mathbb{E}\left[V_{t+2}^{\mathrm{L}_1}(h_{t+1}, a_{t+1}, o_{t+2}) \mid h_t, o_{t+1:t+2}\right] \\
&\overset{(a)}{=} \mathbb{E}\left[\max_{a_{t+2}} \mathbb{E}\left[r_{t+2}(h_{t+2}, a_{t+2}) + V_{t+3}^{\mathrm{L}_1}(h_{t+2}, a_{t+2}, o_{t+3}) \mid h_{t+1}, o_{t+2}\right] \mid h_t, o_{t+1:t+2}\right] \\
&\overset{(b)}{\geq} \max_{a_{t+2}} \mathbb{E}\left[\mathbb{E}\left[r_{t+2}(h_{t+2}, a_{t+2}) + V_{t+3}^{\mathrm{L}_1}(h_{t+2}, a_{t+2}, o_{t+3}) \mid h_{t+1}, o_{t+2}\right] \mid h_t, o_{t+1:t+2}\right] \\
&\overset{(c)}{\geq} \max_{a_{t+2}} \mathbb{E}\left[\mathbb{E}\left[r_{t+2}(h_{t+2}, a_{t+2}) + V_{t+3}^{\mathrm{L}_2}(h_{t+1}, a_{t+1:t+2}, o_{t+2:t+3}) \mid h_{t+1}, o_{t+2}\right] \mid h_t, o_{t+1:t+2}\right] \\
&\overset{(d)}{=} \max_{a_{t+2}} \mathbb{E}\left[\mathbb{E}\left[r_{t+2}(h_{t+2}, a_{t+2}) + V_{t+3}^{\mathrm{L}_2}(h_{t+1}, a_{t+1:t+2}, o_{t+2:t+3}) \mid h_{t:t+1}, o_{t+1:t+2}\right] \mid h_t, o_{t+1:t+2}\right] \\
&\overset{(e)}{=} \max_{a_{t+2}} \mathbb{E}\left[r_{t+2}(h_{t+2}, a_{t+2}) + V_{t+3}^{\mathrm{L}_2}(h_{t+1}, a_{t+1:t+2}, o_{t+2:t+3}) \mid h_t, o_{t+1:t+2}\right] \\
&= V_{t+2}^{\mathrm{L}_2}(h_t, a_{t:t+1}, o_{t+1:t+2})
\end{aligned}$$

where we use the definition of $V_{t+2}^{\mathrm{L}_1}$ in (a). Inequality (b) follows from Jensen's inequality after exchanging the outer expectation with $\max_{a_{t+2}}$. We obtain (c) from the induction hypothesis and inequality (86). Equality (d) follows by noting that the argument inside the inner expectation, conditional on h_{t+1}, is independent of h_t and o_{t+1}. Equality (e) then follows from the tower property and the final equality results from the definition of $V_{t+2}^{\mathrm{L}_2}$. We have therefore shown the desired result for time $t+2$ and so the proof is complete. □

C. Proving that the Approximate Value Functions Are Supersolutions

We now prove Proposition 6.2 which states that all of our AVFs are supersolutions. These results are not surprising and the proof for each AVF is quite straightforward but we include them for completeness. Recall that a supersolution is an AVF ϑ that for all possible time t belief states π_t satisfies

$$\vartheta_t(\pi_t) \geq \max_{a_t \in \mathcal{A}} \left\{ r_t(\pi_t, a_t) + \mathbb{E}\left[\vartheta_{t+1}(\pi_t) \mid \mathcal{F}_t^{\pi}\right] \right\}. \tag{87}$$

Before proceeding we note that given the current belief state π_t and the next observation o_{t+1}, the belief state π_{t+1} can be computed according to

$$\pi_{t+1}(h') = \frac{\sum_h \pi_t(h) P_{hh'} B_{h'o}}{\sigma(o, \pi_t)} \tag{88}$$

where[19] $\sigma(o, \pi_t) := P(o_{t+1} \mid \pi_t) = \sum_{h,h'} \pi_t(h) P_{hh'} B_{h'o}$ for $t \in \{0, \ldots, T-1\}$.

[19] As before, we will often suppress the dependence of the various transmission and emission probabilities on the actions.

Proof that the MDP Approximation is a Supersolution

Following (36) and (37) we have

$$\begin{aligned}
\widetilde{V}_t^{\mathrm{MDP}}(\pi_t) &= \sum_h \pi_t(h) \max_{a_t} \left\{ r_t(h, a_t) + \sum_{h'} P_{hh'}(a_t) V_{t+1}^{\mathrm{MDP}}(h') \right\} \\
&\overset{(a)}{\geq} \max_{a_t} \sum_h \pi_t(h) \left\{ r_t(h, a_t) + \sum_{h'} P_{hh'}(a_t) V_{t+1}^{\mathrm{MDP}}(h') \right\} \\
&\overset{(b)}{=} \max_{a_t} \left\{ r_t(\pi_t, a_t) + \sum_{h', o_{t+1}} \left[\sum_h \pi_t(h) P_{hh'}(a_t) B_{h' o_{t+1}}(a_t) \right] V_{t+1}^{\mathrm{MDP}}(h') \right\} \\
&\overset{(c)}{=} \max_{a_t} \left\{ r_t(\pi_t, a_t) + \sum_{h', o_{t+1}} P(o_{t+1} \mid \pi_t) \pi_{t+1}(h') V_{t+1}^{\mathrm{MDP}}(h') \right\} \\
&\overset{(d)}{=} \max_{a_t} \left\{ r_t(\pi_t, a_t) + \sum_{o_{t+1}} P(o_{t+1} \mid \pi_t) \widetilde{V}_{t+1}^{\mathrm{MDP}}(\pi_{t+1}) \right\} \\
&\equiv \max_{a_t} \left\{ r_t(\pi_t, a_t) + \mathbb{E}\left[\widetilde{V}_{t+1}^{\mathrm{MDP}}(\pi_{t+1}) \mid \mathcal{F}_t^{\pi} \right] \right\}
\end{aligned}$$

where (a) results from Jensen's inequality and (b) follows from including the factor $\sum_{o_{t+1}} B_{h' o_{t+1}} = 1$ and then a simple re-ordering of the terms. Equality (c) follows from (88) while we have used the definition of $\widetilde{V}_{t+1}^{\mathrm{MDP}}(\pi_{t+1})$ to obtain (d).

Proof that the QMDP Approximation is a Supersolution

The proof for the QMDP approximation follows a similar argument. From (38) and (39) we have

$$\begin{aligned}
\widetilde{V}_t^{\mathrm{Q}}(\pi_t) &= \max_{a_t} \sum_h \pi_t(h) \left\{ r_t(h, a_t) + \sum_{h'} P_{hh'}(a_t) V_{t+1}^{\mathrm{MDP}}(h') \right\} \\
&\overset{(a)}{=} \max_{a_t} \left\{ r_t(\pi_t, a_t) + \sum_{h', o_{t+1}} P(o_{t+1} \mid \pi_t) \pi_{t+1}(h') V_{t+1}^{\mathrm{MDP}}(h') \right\} \\
&\overset{(b)}{=} \max_{a_t} \left\{ r_t(\pi_t, a_t) + \sum_{h', o_{t+1}} P(o_{t+1} \mid \pi_t) \pi_{t+1}(h') \max_{a'} V_{t+1}^{Q}(h', a') \right\} \\
&\overset{(c)}{\geq} \max_{a_t} \left\{ r_t(\pi_t, a_t) + \sum_{o_{t+1}} P(o_{t+1} \mid \pi_t) \max_{a'} \sum_{h'} \pi_{t+1}(h') V_{t+1}^{\mathrm{Q}}(h', a') \right\} \\
&\overset{(d)}{=} \max_{a_t} \left\{ r_t(\pi_t, a_t) + \sum_{o_{t+1}} P(o_{t+1} \mid \pi_t) \widetilde{V}_{t+1}^{\mathrm{Q}}(\pi_{t+1}) \right\} \\
&\equiv \max_{a_t} \left\{ r_t(\pi_t, a_t) + \mathbb{E}\left[\widetilde{V}_{t+1}^{\mathrm{Q}}(\pi_{t+1}) \mid \mathcal{F}_t^{\pi} \right] \right\}
\end{aligned}$$

where (a) follows from following steps (b) to (d) of the MDP proof above and (b) then follows from the definition of both V_{t+1}^{MDP} and V_{t+1}^{Q}. Inequality (c) follows from Jensen's inequality after changing the order of $\max_{a'}$ and the marginalization of h'. Finally (d) follows from the definition of $\widetilde{V}_{t+1}^{\mathrm{Q}}(\pi_{t+1})$.

Proof that the Lag-1 Approximation is a Supersolution

Because of the many terms involved, throughout the proof we will write the relevant quantities as expectations and we will use $\mathbb{E}_X$ to denote an expectation taken over the random variable X. Following its definition in

(41), the Lag-1 AVF satisfies

$$
\begin{aligned}
\widetilde{V}_t^{\mathrm{L}1}(\pi_t) &\overset{(a)}{=} \max_{a_t} \mathbb{E}_{h_t,o_{t+1}}\left[r_t(h_t,a_t) + \max_{a_{t+1}} \mathbb{E}_{h_{t+1},o_{t+2}}\left[r_{t+1}(h_{t+1},a_{t+1}) + V_{t+2}^{\mathrm{L}1}(h_{t+1},a_{t+1},o_{t+2}) \mid h_t, o_{t+1}\right] \mid \mathcal{F}_t^\pi\right] \\
&\overset{(b)}{=} \max_{a_t}\left\{r_t(\pi_t,a_t) + \mathbb{E}_{h_t,o_{t+1}}\left[\max_{a_{t+1}} \mathbb{E}_{h_{t+1},o_{t+2}}\left[r_{t+1}(h_{t+1},a_{t+1}) + V_{t+2}^{\mathrm{L}1}(h_{t+1},a_{t+1},o_{t+2}) \mid h_t, o_{t+1}\right] \mid \mathcal{F}_t^\pi\right]\right\} \\
&\overset{(c)}{=} \max_{a_t}\left\{r_t(\pi_t,a_t) + \mathbb{E}_{o_{t+1}}\left[\mathbb{E}_{h_t}\left[\max_{a_{t+1}} \mathbb{E}_{h_{t+1},o_{t+2}}\left[r_{t+1}(h_{t+1},a_{t+1}) + V_{t+2}^{\mathrm{L}1}(h_{t+1},a_{t+1},o_{t+2}) \mid h_t, o_{t+1}\right] \mid o_{t+1}, \mathcal{F}_t^\pi\right] \mid \mathcal{F}_t^\pi\right]\right\} \\
&\overset{(d)}{\geq} \max_{a_t}\left\{r_t(\pi_t,a_t) + \mathbb{E}_{o_{t+1}}\left[\max_{a_{t+1}} \mathbb{E}_{h_t}\left[\mathbb{E}_{h_{t+1},o_{t+2}}\left[r_{t+1}(h_{t+1},a_{t+1}) + V_{t+2}^{\mathrm{L}1}(h_{t+1},a_{t+1},o_{t+2}) \mid h_t, o_{t+1}\right] \mid o_{t+1}, \mathcal{F}_t^\pi\right] \mid \mathcal{F}_t^\pi\right]\right\} \\
&\overset{(e)}{=} \max_{a_t}\left\{r_t(\pi_t,a_t) + \mathbb{E}_{o_{t+1}}\left[\max_{a_{t+1}} \mathbb{E}_{h_t}\left[\mathbb{E}_{h_{t+1},o_{t+2}}\left[r_{t+1}(h_{t+1},a_{t+1}) + V_{t+2}^{\mathrm{L}1}(h_{t+1},a_{t+1},o_{t+2}) \mid h_t, o_{t+1}, \mathcal{F}_t^\pi\right] \mid o_{t+1}, \mathcal{F}_t^\pi\right] \mid \mathcal{F}_t^\pi\right]\right\} \\
&\overset{(f)}{=} \max_{a_t}\left\{r_t(\pi_t,a_t) + \mathbb{E}_{o_{t+1}}\left[\max_{a_{t+1}} \mathbb{E}_{h_{t+1},o_{t+2}}\left[r_{t+1}(h_{t+1},a_{t+1}) + V_{t+2}^{\mathrm{L}1}(h_{t+1},a_{t+1},o_{t+2}) \mid o_{t+1}, \mathcal{F}_t^\pi\right] \mid \mathcal{F}_t^\pi\right]\right\} \\
&\overset{(g)}{=} \max_{a_t}\left\{r_t(\pi_t,a_t) + \mathbb{E}\left[\widetilde{V}_{t+1}^{\mathrm{L}1}(\pi_{t+1}) \mid \mathcal{F}_t^\pi\right]\right\}
\end{aligned}
$$

where (a) follows from the definition of $V_{t+1}^{\mathrm{L}1}$ in (40) and (b) follows from noting that the expectation of $r_t(h_t,a_t)$ conditional on $\mathcal{F}_t^\pi$ is $r_t(\pi_t,a_t)$. Equality (c) follows from the tower property while (d) follows from Jensen's inequality after changing the order of $\max_{a_{t+1}}$ and the expectation over h_t. Equality (e) follows since the function inside the expectation $\mathbb{E}[\cdot \mid h_t, o_{t+1}]$ is independent of $\mathcal{F}_t^\pi$ after conditioning on h_t. Equality (f) follows from applying the tower property to the nested expectations. Finally (g) follows from the definition of $\widetilde{V}_{t+1}^{\mathrm{L}1}(\pi_{t+1})$ and where we note that π_{t+1} is completely determined given π_t, o_{t+1} and a_t.

Proof that the Lag-2 Approximation is a Supersolution

Proving that the Lag-2 AVF is a supersolution is similar to proving that the Lag-1 AVF is a supersolution but the details are a little more involved. From (43) we have

$$
\begin{aligned}
\widetilde{V}_t^{\mathrm{L}2}(\pi_t) &:= \max_{a_t} \mathbb{E}_{o_{t+1}}\left[\max_{a_{t+1}} \mathbb{E}_{h_{t:t+1},o_{t+2}}\left[r_t(h_t,a_t) + r_{t+1}(h_{t+1},a_{t+1}) + V_{t+2}^{\mathrm{L}2}(h_t,a_{t:t+1},o_{t+1:t+2}) \mid \mathcal{F}_t^\pi, o_{t+1}\right] \mid \mathcal{F}_t^\pi\right] \\
&\overset{(a)}{=} \max_{a_t} \mathbb{E}_{o_{t+1}}\Bigg[\max_{a_{t+1}} \mathbb{E}_{h_{t:t+1},o_{t+2}}\Big[r_t(h_t,a_t) + r_{t+1}(h_{t+1},a_{t+1}) + \\
&\qquad \max_{a_{t+2}} \mathbb{E}_{h_{t+1:t+2},o_{t+3}}\left[r_{t+2}(h_{t+2},a_{t+2}) + V_{t+3}^{\mathrm{L}2}(h_{t+1},a_{t+1:t+2},o_{t+2:t+3}) \mid h_t, o_{t+1:t+2}\right] \mid \mathcal{F}_t^\pi, o_{t+1}\Big] \mid \mathcal{F}_t^\pi\Bigg] \\
&\overset{(b)}{=} \max_{a_t}\Bigg\{r_t(\pi_t,a_t) + \mathbb{E}_{o_{t+1}}\Bigg[\max_{a_{t+1}} \mathbb{E}_{h_{t:t+1},o_{t+2}}\Big[r_{t+1}(h_{t+1},a_{t+1}) + \\
&\qquad \max_{a_{t+2}} \mathbb{E}_{h_{t+1:t+2},o_{t+3}}\left[r_{t+2}(h_{t+2},a_{t+2}) + V_{t+3}^{\mathrm{L}2}(h_{t+1},a_{t+1:t+2},o_{t+2:t+3}) \mid h_t, o_{t+1:t+2}\right] \mid \mathcal{F}_t^\pi, o_{t+1}\Big] \mid \mathcal{F}_t^\pi\Bigg]\Bigg\} \\
&\overset{(c)}{=} \max_{a_t}\Bigg\{r_t(\pi_t,a_t) + \mathbb{E}_{o_{t+1}}\Bigg[\max_{a_{t+1}} \mathbb{E}_{h_{t:t+1},o_{t+2}}\Big[\max_{a_{t+2}}\Big\{r_{t+1}(h_{t+1},a_{t+1}) + \\
&\qquad \mathbb{E}_{h_{t+1:t+2},o_{t+3}}\left[r_{t+2}(h_{t+2},a_{t+2}) + V_{t+3}^{\mathrm{L}2}(h_{t+1},a_{t+1:t+2},o_{t+2:t+3}) \mid h_t, o_{t+1:t+2}\right]\Big\} \mid \mathcal{F}_t^\pi, o_{t+1}\Big] \mid \mathcal{F}_t^\pi\Bigg]\Bigg\}
\end{aligned}
\tag{89}
$$

where (a) follows from the definition of $V_{t+2}^{\mathrm{L}2}$ in (80). We obtain (b) by taking the expectation of $r_t(h_t,a_t)$ outside the maximization of a_{t+1} (which is fine since a_{t+1} has no bearing on $r_t(h_t,a_t)$) and then using the tower property with the outer expectation to obtain $r_t(\pi_t,a_t)$. Equality (c) follows from taking $r_{t+1}(h_{t+1},a_{t+1})$ inside the maximization of a_{t+2} which is again fine since a_{t+2} has no bearing on $r_{t+1}(h_{t+1},a_{t+1})$. We focus

now on the term inside the outermost expectation $\mathbb{E}_{o_{t+1}}[\cdot \mid \mathcal{F}_t^\pi]$ of (89). It satisfies

$$
\begin{aligned}
&\max_{a_{t+1}} \mathbb{E}_{h_{t:t+1},o_{t+2}}\Big[\max_{a_{t+2}}\Big\{r_{t+1}(h_{t+1},a_{t+1}) + \mathbb{E}_{h_{t+1:t+2},o_{t+3}}\big[r_{t+2}(h_{t+2},a_{t+2}) + V_{t+3}^{\mathrm{L}_2}(h_{t+1},a_{t+1:t+2},o_{t+2:t+3}) \mid h_t, o_{t+1:t+2}\big]\Big\} \mid \mathcal{F}_t^\pi, o_{t+1}\Big] \\
&\overset{(d)}{=} \max_{a_{t+1}} \mathbb{E}_{o_{t+2}}\Big[\mathbb{E}_{h_{t:t+1}}\Big[\max_{a_{t+2}}\Big\{r_{t+1}(h_{t+1},a_{t+1}) + \\
&\qquad \mathbb{E}_{h_{t+1:t+2},o_{t+3}}\big[r_{t+2}(h_{t+2},a_{t+2}) + V_{t+3}^{\mathrm{L}_2}(h_{t+1},a_{t+1:t+2},o_{t+2:t+3}) \mid h_t, o_{t+1:t+2}\big]\Big\} \mid \mathcal{F}_t^\pi, o_{t+1:t+2}\Big] \mid \mathcal{F}_t^\pi, o_{t+1}\Big] \\
&\overset{(e)}{\geq} \max_{a_{t+1}} \mathbb{E}_{o_{t+2}}\Big[\max_{a_{t+2}}\Big\{\mathbb{E}_{h_{t:t+1}}\big[r_{t+1}(h_{t+1},a_{t+1}) + \\
&\qquad \mathbb{E}_{h_{t+1:t+2},o_{t+3}}\big[r_{t+2}(h_{t+2},a_{t+2}) + V_{t+3}^{\mathrm{L}_2}(h_{t+1},a_{t+1:t+2},o_{t+2:t+3}) \mid h_t, o_{t+1:t+2}\big] \mid \mathcal{F}_t^\pi, o_{t+1:t+2}\big]\Big\} \mid \mathcal{F}_t^\pi, o_{t+1}\Big] \\
&\overset{(f)}{=} \max_{a_{t+1}} \mathbb{E}_{o_{t+2}}\Big[\max_{a_{t+2}}\Big\{\mathbb{E}_{h_{t+1}}\big[r_{t+1}(h_{t+1},a_{t+1}) \mid \mathcal{F}_t^\pi, o_{t+1:t+2}\big] + \\
&\qquad \mathbb{E}_{h_t}\Big[\mathbb{E}_{h_{t+1:t+2},o_{t+3}}\big[r_{t+2}(h_{t+2},a_{t+2}) + V_{t+3}^{\mathrm{L}_2}(h_{t+1},a_{t+1:t+2},o_{t+2:t+3}) \mid h_t, o_{t+1:t+2}\big] \mid \mathcal{F}_t^\pi, o_{t+1:t+2}\Big]\Big\} \mid \mathcal{F}_t^\pi, o_{t+1}\Big] \\
&\overset{(g)}{=} \max_{a_{t+1}} \mathbb{E}_{o_{t+2}}\Big[\max_{a_{t+2}}\Big\{\mathbb{E}_{h_{t+1}}\big[r_{t+1}(h_{t+1},a_{t+1}) \mid \mathcal{F}_t^\pi, o_{t+1:t+2}\big] + \\
&\qquad \mathbb{E}_{h_{t+1:t+2},o_{t+3}}\big[r_{t+2}(h_{t+2},a_{t+2}) + V_{t+3}^{\mathrm{L}_2}(h_{t+1},a_{t+1:t+2},o_{t+2:t+3}) \mid \mathcal{F}_t^\pi, o_{t+1:t+2}\big]\Big\} \mid \mathcal{F}_t^\pi, o_{t+1}\Big] \\
&\overset{(h)}{=} \max_{a_{t+1}} \mathbb{E}_{o_{t+2}}\Big[\max_{a_{t+2}} \mathbb{E}_{h_{t+1:t+2},o_{t+3}}\big[r_{t+1}(h_{t+1},a_{t+1}) + r_{t+2}(h_{t+2},a_{t+2}) + V_{t+3}^{\mathrm{L}_2}(h_{t+1},a_{t+1:t+2},o_{t+2:t+3}) \mid \mathcal{F}_t^\pi, o_{t+1:t+2}\big] \mid \mathcal{F}_t^\pi, o_{t+1}\Big] \\
&\overset{(i)}{=} \widetilde{V}_{t+1}^{\mathrm{L}_2}(\pi_{t+1})
\end{aligned}
$$

where (d) follows from the tower property so that $\mathbb{E}_{h_{t:t+1},o_{t+2}}\left[\cdot \mid \mathcal{F}_t^\pi, o_{t+1}\right] = \mathbb{E}_{o_{t+2}}\Big[\mathbb{E}_{h_{t:t+1}}\left[\cdot \mid \mathcal{F}_t^\pi, o_{t+1:t+2}\right] \mid \mathcal{F}_t^\pi, o_{t+1}\Big]$ and (e) follows from Jensen's inequality after changing the order of the $\max_{a_{t+2}}$ operator and the marginalization of h_t and h_{t+1}. We obtain (f) by simply writing the conditional expectation of a sum as the sum of conditional expectations. Equality (g) follows from applying the tower property to the nested expectations while (h) follows from grouping together the two conditional expectations $\mathbb{E}[\cdot \mid \mathcal{F}_t^\pi, o_{t+1:t+2}]$. Finally, (i) follows from the definition of the $\widetilde{V}_{t+1}^{\mathrm{L}_2}(\pi_{t+1})$ and where we note again that π_{t+1} is completely determined given π_t, o_{t+1} and a_t.

The overall result now follows by substituting $\widetilde{V}_{t+1}^{\mathrm{L}_2}(\pi_{t+1})$ in for the conditional expectation $\mathbb{E}_{o_{t+1}}[\cdot \mid \mathcal{F}_t^\pi]$ in (89) with the equality there replaced by a greater-than-or-equal to inequality.

D. Dropping the Requirement that $\mathbb{P} \ll \tilde{\mathbb{P}}$

We explain here why we do not require $\mathbb{P}$, the probability measure for the controlled formulation, to be absolutely continuous w.r.t $\tilde{\mathbb{P}}$ (the probability measure for the original uncontrolled formulation), when the penalties in (22) are constructed from supersolutions. This result was originally shown by BH 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 of belief-state POMDP formulation, i.e. the BSPI relaxation, but it should be clear that the result is general and holds for general information relaxations.

We therefore assume the penalty function, $c_t := \mathbb{E}[\vartheta_{t+1}(\pi_{t+1}) \mid \mathcal{F}_t^\pi] - \vartheta_{t+1}(\pi_{t+1})$, is such that ϑ_t is a supersolution satisfying[20] $\vartheta_{T+1} \equiv 0$. From Definition 6.1, it follows that for each $t \in \{0, \ldots, T\}$ and π_t we

[20]There is no difficulty in assuming $\vartheta_{T+1} \equiv 0$ since $\vartheta_t(\pi_t)$ represents an AVF and all of our AVFs naturally satisfy this assumption.

have

$$\vartheta_t(\pi_t) \geq r_t(\pi_t, a_t) + \mathbb{E}[\vartheta_{t+1}(\pi_{t+1}) \mid \mathcal{F}_t^{\pi}] \qquad \forall a_t \in \mathcal{A}. \tag{90}$$

Subtracting $\vartheta_t(\pi_t)$ from both sides of (90), summing over t and recalling that $\vartheta_{T+1} \equiv 0$, we obtain

$$\begin{aligned} 0 &\geq \sum_{t=0}^{T} \Big\{ r_t(\pi_t, a_t) + \mathbb{E}[\vartheta_{t+1}(\pi_{0:t+1}) \mid \mathcal{F}_t^{\pi}] - \vartheta_t(\pi_t) \Big\} \\ &= \sum_{t=0}^{T} \Big\{ r_t(\pi_t, a_t) + c_t \Big\} - \vartheta_0(\pi_0). \end{aligned} \tag{91}$$

We now obtain

$$\begin{aligned} V_0^* - \vartheta_0 &= \max_{\mu \in \mathcal{U}_{\mathbb{F}^\pi}} V_0^\mu - \vartheta_0 \\ &\overset{(a)}{=} \max_{\mu \in \mathcal{U}_{\mathbb{F}^\pi}} \mathbb{E}\Big[\sum_{t=0}^{T} (r_t + c_t) - \vartheta_0 \mid \mathcal{F}_0^{\pi} \Big] \qquad (92) \\ &\overset{(b)}{\leq} \max_{\mu \in \mathcal{U}_{\mathbb{F}^\pi}} \widetilde{\mathbb{E}}\Big[\sum_{t=0}^{T} \Phi_t\big(r_t + c_t\big) - \vartheta_0 \mid \mathcal{F}_0^{\pi} \Big] \\ &\overset{(c)}{\leq} \widetilde{\mathbb{E}}\Big[\max_{a_{0:T-1}} \sum_{t=0}^{T} \Phi_t\big(r_t + c_t\big) \mid \mathcal{F}_0^{\pi} \Big] - \vartheta_0. \qquad (93) \end{aligned}$$

where we have omitted the arguments of r_t and ϑ_0 for the sake of clarity. Equality (a) follows since $\mathbb{E}[\sum_{t=0}^{T} c_t \mid \mathcal{F}_0^{\pi}] = 0$ for any $\mathbb{F}^{\pi}$-adapted policy and since $\vartheta_0(\pi_0)$ is $\mathcal{F}_0^{\pi}$-adapted. In order to establish inequality (b), we first note that (91) implies the random quantity inside the expectation in (92) is non-positive w.p. 1. The inequality then follows[21] for any probability measure, $\tilde{\mathbb{P}}$, regardless of whether or not $\mathbb{P}$ is absolutely continuous w.r.t $\tilde{\mathbb{P}}$. 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 $\vartheta_0(\pi_0)$.

We can now add $\vartheta_0(\pi_0)$ across both sides of (93) 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 $\tilde{\mathbb{P}}$ 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 $\tilde{\mathbb{P}}$ is the measure induced by following an optimal policy for the primal problem since in that case $\mathbb{P}$ and $\tilde{\mathbb{P}}$ will coincide. Strong duality will then also continue to hold. In particular, (c) will then also be an equality if ϑ_t coincides with the optimal value function, V_t^*, which is itself a supersolution.

E. Further Details for the Multiaccess Communication Application

The main difference between the multiaccess communication application and the POMDP framework as defined in Section 2 is the timing of observations. Specifically, in the multiaccess communication application an observation occurs immediately after an action is taken and is therefore a function of the current hidden state and the *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. Therefore the filtering algorithm for the

[21] This result was stated as Lemma A.1 in [8] and we state it here for the sake of completeness. Consider a measurable space (Ω, Σ) and two probability measures P and Q. Let ϕ 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 $\mathbb{E}^P[Y] \leq \mathbb{E}^Q[\phi Y]$.

belief-state update is different than the standard update as given in (88) (where the action dependence was suppressed). The belief update for the slotted Aloha dynamics satisfies

$$\pi_{t+1}(h') = \frac{\sum_h \pi_t(h) B_{ho_t}(a_t) P_{hh'}(o_t)}{\sum_h \pi_t(h) B_{ho_t}(a_t)} \tag{94}$$

for $t \in \{0, \ldots, T-1\}$ and where we recognize the denominator in (94) as $\mathbb{P}_{a_t}(o_t \mid \pi_t)$. It is worth emphasizing that the belief state for time $t+1$ is a function of the time t action and observation. Moreover, the hidden-state transition probabilities under $\mathbb{P}$ are action-independent given the current observation. As a result we assume the hidden state transitions probabilities are unchanged when we go from $\mathbb{P}$ to $\tilde{\mathbb{P}}$.

RN Derivatives for the Belief-State Formulation

These alternative dynamics also impact the calculations of the RN derivatives. In the case of the belief-state formulation, the arguments in Appendix A.2 that led to (69) still apply. However, in the multiaccess communication application the numerator of (69) now satisfies

$$\begin{aligned}
\mathbb{P}_{a_{0:t-1}}(\pi_{1:t}) &= \prod_{s=0}^{t-1} \mathbb{P}_{a_s}(\pi_{s+1} \mid \pi_s) \\
&= \prod_{s=0}^{t-1} \sum_{o_s} \mathbb{P}_{a_s}(o_s \mid \pi_s) \mathbb{P}_{a_s}(\pi_{s+1} \mid o_s, \pi_s) \\
&= \prod_{s=0}^{t-1} \sum_{h,h',o_s} \pi_s(h) \mathbb{P}_{a_s}(o_s \mid h) \mathbb{P}(h' \mid h, o_s) \mathbf{1}_{\{\pi_{s+1} = f(\pi_s, a_s, o_s)\}} \\
&= \prod_{s=0}^{t-1} \sum_{h,o} \pi_s(h) B_{ho}(a_s) \mathbf{1}_{\{\pi_{s+1} = f(\pi_s, a_s, o)\}}
\end{aligned} \tag{95}$$

where $f(\pi_s, a_s, o_s)$ lies in the $|\mathcal{H}|$-dimensional simplex with each of its components defined according to

$$f(h'; \pi_s, a_s, o_s) := \frac{\sum_h \pi_s(h) B_{ho}(a_s) P_{hh'}(o)}{\sum_h \pi_s(h) B_{ho}(a_s)}.$$

Using similar arguments, we see that the denominator of (69) satisfies

$$\tilde{\mathbb{P}}(\pi_{1:t}) = \prod_{s=0}^{t-1} \sum_{h,o} \pi_s(h) E^s_{ho} \mathbf{1}_{\{\pi_{s+1} = \tilde{f}_s(\pi_s; o)\}} \tag{96}$$

where E^s_{ho} is the uncontrolled emission matrix defined in (56) and where $\tilde{f}_s(\pi_s; o)$ lies in the $|\mathcal{H}|$-dimensional simplex with each of its components defined according to

$$\tilde{f}_s(h'; \pi_s; o) := \frac{\sum_h \pi_s(h) E^s_{ho} P_{hh'}(o)}{\sum_h \pi_s(h) E^s_{ho}}.$$

The RN derivatives are now given by (67) but using (95) and (96) we see that each ϕ is now given by

$$\phi(\pi_s, \pi_{s+1}, a_s) := \frac{\sum_{h,o} \pi_s(h) B_{ho}(a_s) \mathbf{1}_{\{\pi_{s+1} = f(\pi_s, a_s, o)\}}}{\sum_{h,o} \pi_s(h) E^s_{ho} \mathbf{1}_{\{\pi_{s+1} = \tilde{f}_s(\pi_s; o)\}}}. \tag{97}$$

RN Derivatives for the Non-Belief-State Formulation

In the case of the non-belief-state formulation of the problem, the RN derivatives satisfy

$$\Phi_t := \frac{\mathbb{P}_{a_{0:t-1}}(o_{0:t-1}, h_{0:t})}{\tilde{\mathbb{P}}(o_{0:t-1}, h_{0:t})} \tag{98}$$

where

$$\mathbb{P}_{a_{0:t-1}}(o_{0:t-1}, h_{0:t}) = \pi_0(h_0) \prod_{s=0}^{t-1} \mathbb{P}_{a_s}(o_s \mid h_s) \mathbb{P}(h_{s+1} \mid h_s, o_s) \tag{99}$$

$$\tilde{\mathbb{P}}(o_{0:t-1}, h_{0:t}) = \pi_0(h_0) \prod_{s=0}^{t-1} \tilde{\mathbb{P}}(o_s \mid h_s) \mathbb{P}(h_{s+1} \mid h_s, o_s). \tag{100}$$

It immediately follows from (99) and (100) that the RN derivatives for the uncontrolled non-belief-state formulation satisfy

$$\phi_t(i, k, a) := \frac{B_{ik}(a)}{E^t_{ik}}$$

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

F. 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 [35]. 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

$$V_0^* := \max_{\mu \in \mathcal{U}_{\mathbb{F}^\pi}} \mathbb{E}\left[\sum_{t=0}^{\infty} \delta^t r(\pi_t, \mu_t) \,\middle|\, \mathcal{F}_0^\pi \right] \tag{101}$$

In order to solve the dual problem using a BSPI relaxation, we would have to simulate an infinite sequence of random variables $\{u_t\}_{t \geq 0}$, which is not possible in practice. An equivalent formulation, however, is to replace the discounting by a costless, absorbing state π^a which can be reached from every belief-state and feasible action with probability $1 - \delta$, at each t. The state transition distribution remains as in (6), conditional on not reaching the absorbing state. The equivalent absorbing state formulation is then given by

$$V_0^* := \max_{\mu \in \mathcal{U}_{\mathbb{F}^\pi}} \mathbb{E}\left[\sum_{t=0}^{\tau} r(\pi_t, \mu_t) \,\middle|\, \mathcal{F}_0^\pi \right] \tag{102}$$

where $\tau = \inf\{t : \pi_t = \pi^a\}$ is the absorption time, distributed as a geometric random variable with parameter $1 - \delta$. In (102) 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 BSPI relaxation, the dual upper bound can be expressed as

$$V_0^*(\pi_0) \leq \widetilde{\mathbb{E}}\left[\max_{a_{0:\tau-1}} \sum_{t=0}^{\tau} \Phi_t[r_t(\pi_t, a_t) + c_t] \,\middle|\, \mathcal{F}_0^\pi \right]. \tag{103}$$

An inner problem inside the expectation on the r.h.s of (103) can be generated by first simulating the absorption time $\tau \sim \text{Geom}(1-\delta)$, and then generating the belief states π_t using some action-independent change of measure. A lower bound can be obtained of course by simply simulating many paths of some feasible policy.

One concern with the bound of (103) is that the optimal objective of the inner problem in (103) 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 τ is now random and the presence of the RN derivative terms Φ_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[22] and other considerations allowed them to conclude that the variance of the upper bound estimator in (103) 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 $\bar{r} := \max_{\pi,a} r(\pi, a)$. 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.

[22] See also the discussion immediately following our Proposition 6.2.