Automatica 155 (2023) 111140

Contents lists available at ScienceDirect

Automatica

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

Simultaneous perception–action design via invariant finite belief sets[☆]

Michael Hibbard [*], Takashi Tanaka, Ufuk Topcu

Department of Aerospace Engineering and Engineering Mechanics University of Texas at Austin, United States of America

ARTICLE INFO

Article history:
Received 18 August 2021
Received in revised form 5 July 2022
Accepted 17 May 2023
Available online 3 July 2023

Keywords:
Perception and sensing
Planning
Optimization under uncertainties

ABSTRACT

Although perception is an increasingly dominant portion of the overall computational cost for autonomous systems, only a fraction of the information perceived is likely to be relevant to the current task. To alleviate these perception costs, we develop a novel simultaneous perception–action design framework wherein an agent senses only the task-relevant information. This formulation differs from that of a partially observable Markov decision process, since the agent is free to synthesize not only its policy for action selection but also its belief-dependent observation function. The method enables the agent to balance its perception costs with those incurred by operating in its environment. To obtain a computationally tractable solution, we approximate the value function using a novel *method of invariant finite belief sets*, wherein the agent acts exclusively on a finite subset of the continuous belief space. We solve the approximate problem through value iteration in which a linear program is solved individually for each belief state in the set, in each iteration. Finally, we prove that the value functions, under an assumption on their structure, converge to their continuous state-space values as the sample density increases.

1. Introduction

Evolution has driven biological organisms to strike a balance between the conflicting desire of utilizing all available information in order to make a decision and the desire of minimizing the cost of perceiving that information from the environment. Such perception costs are not negligible: studies of the human brain, for instance, have shown that it constitutes 20% of our resting energy consumption rate, of which 50% is associated with signaling (Attwell & Laughlin, 2001). In order to optimally utilize only a portion of the available information, it is believed that organisms have evolved to strategically perceive only the *task-relevant* information from their environment (Berry et al., 1999; Egner & Hirsch, 2005). The mammalian visual cortex, for example, is sensitive to only particular features over a small region of the visual field (Hubel & Wiesel, 1968).

Perception costs have likewise become a bottleneck in many engineering applications. For example, experimental results have shown that over 94% of the computational time in autonomous driving is allocated to perception (Lin et al., 2018). Although state-of-the-art accelerator platforms like GPUs are effective for latency reduction, their power consumption is significant enough to degrade a vehicle's driving range. To alleviate these issues, Censi et al. (2015) argues that sensor hardware should extract only task-relevant information. Although intuitively appealing, such an idea is difficult to implement, as what constitutes "task-relevant" information is difficult to define.

To provide a methodological foundation for task-relevant sensing, we propose a simultaneous perception–action design (SPADE) framework based on the standard Markov decision process (MDP) formulation (Puterman, 2014) with a novel information-theoretic perception cost. The perception cost penalizes information flowing from the sensor to the down-stream decision-making unit. This framework allows for the synthesis of a sensing mechanism that extracts the minimum amount of task-relevant information from the underlying state of the controlled Markov chain.

The use of information theory to model perception costs has previously been studied. Viewing an agent as a communication channel, Sims (2003) proposes a model penalizing the mutual information between the state of the environment and the agent's action. Likewise, Sims (2016) proposes using rate–distortion theory to characterize perception costs. The analogy between the perception–action cycle and a communication channel was also studied in Tishby and Polani (2011), where algorithms to synthesize the optimal trade-off between the cost-to-go and the

[☆] This work was supported by DARPA, United States of America Grant D19AP00078, AFOSR, United States of America Grant FA9550-20-1-0101, AFRL, United States of America Grant FA9550-19-1-0169, NSF Awards 1944318 and 1652113, and NSF CNS Award 1836900. The material in this paper was not presented at any conference. This paper was recommended for publication in revised form by Associate Editor Subhrakanti Dey under the direction of Editor Ian R. Petersen.

[*] Corresponding author.
E-mail addresses: mwhibbard@utexas.edu (M. Hibbard), ttanaka@utexas.edu (T. Tanaka), utopcu@utexas.edu (U. Topcu).

https://doi.org/10.1016/j.automatica.2023.111140
0005-1098/© 2023 Elsevier Ltd. All rights reserved.

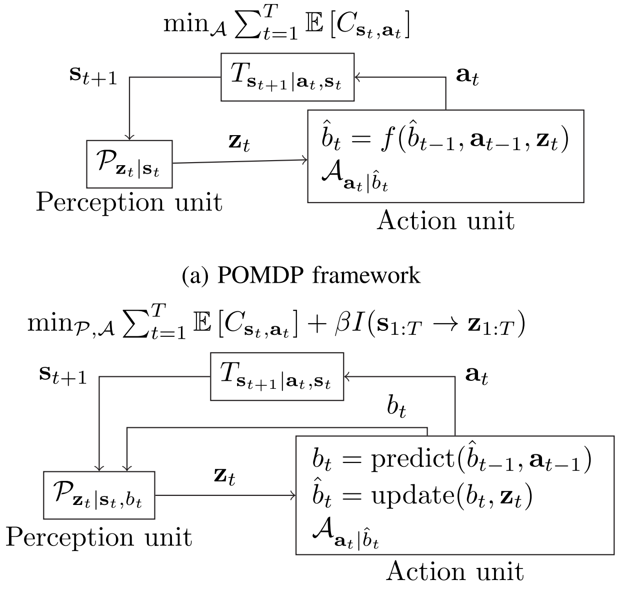
information-to-go are proposed. Alternatively, Ortega and Braun (2013) studied the problem of rational inattention through the lens of thermodynamics, where information processing costs are characterized through differences in free energy. Rational inattention was also studied in Shafieepoorfard et al. (2016), which provides theoretical results connecting controllers subject to information constraints to rate–distortion theory. Following Massey (1990), we use the information-theoretic concept of *directed information* to model the agent's perception costs. The directed information measures the information flow between a pair of random processes, such as from the environment to the observations made about the environment. To our knowledge, this paper is the first to apply directed information to the study of optimal perception.

The proposed SPADE framework is reminiscent of *active perception* (Aloimonos, 2013; Bajcsy et al., 2018). In active perception problems, an agent (or group of agents, as in Spaan, 2008) seeks to take actions that lead to desired observations, such as those that reduce the uncertainty in its underlying state (e.g., Araya et al., 2010; Ghasemi & Topcu, 2019). As opposed to existing works on active perception, the SPADE framework allows an agent to instead synthesize its own sensing mechanism, rather than acting in such a way as to exploit a prespecified sensing mechanism. In Section 5, we use this theoretical, synthesized sensing mechanism as the basis for a dynamic sensor selection strategy, a common problem in the field of active perception (e.g., Hashemi et al., 2018; Hibbard et al., 2020; Spaan & Lima, 2009).

The proposed SPADE framework allows more flexibility for an agent compared to a partially observable MDP (POMDP) (Kaelbling et al., 1998), with Fig. 1 highlighting three key distinctions. First and foremost, our formulation includes the perception strategy $\mathcal{P}$ as a decision variable, rather than using a sensing mechanism fixed a priori. In effect, the agent is capable of *choosing* what to observe. Furthermore, we allow for a belief-dependent perception mechanism, as shown in the lower left block in Fig. 1(b). This generalization is biologically plausible, e.g., eye movement can be controlled, and is crucial in developing a computationally-tractable synthesis of optimal perception and action strategies. Finally, in order for such a generalization to be meaningful, the SPADE framework incorporates a perception cost for the agent, denoted $I(\mathbf{s}_{1:T} \to \mathbf{z}_{1:T})$ in Fig. 1(b). As we discuss in Section 2.3, this quantity is the directed information, a statistical measure of information flow, between the state sequence $\{\mathbf{s}_1, \dots \mathbf{s}_t\}$ and the observation sequence $\{\mathbf{z}_1, \dots, \mathbf{z}_t\}$. The inclusion of this term is critical, since full-state measurement is always optimal in the absence of a perception cost.

Mathematically, our problem is closest to stochastic optimal control with directed information constraints, as studied in Tanaka et al. (2017, 2021). In Tanaka et al. (2017), the problem is studied in the linear-quadratic-Gaussian (LQG) regime, where it is shown that an optimal policy consists of linear perception and action units, whose combined structure is similar to Fig. 1(b). In the LQG case, the simultaneous perception–action synthesis problem can be reformulated as a computationally-efficient semidefinite programming problem. Likewise, the finite-state counterpart of the problem was studied in Tanaka et al. (2021), where an alternative solution method based on the so-called forward–backward Arimoto–Blahut algorithm is proposed. However, the synthesized strategy does not admit the perception–action separation structure, and the algorithm suffers due to the nonconvexity of the cost function. The SPADE framework is in part motivated to overcome these difficulties.

We first show the SPADE problem is solvable, in principle, through a backward-in-time dynamic programming algorithm over the belief space. In each iteration, each belief state value function is updated by solving a nonconvex optimization problem. However, such an approach is impractical, as the value functions must be evaluated over the entirety of the continuous belief space. To circumvent this difficulty, we propose a novel *method of invariant finite belief sets* for approximating the value functions, wherein we enforce that the agent operates exclusively on an invariant finite subset of the belief simplex. We do so by restricting the space of admissible perception strategies in such a way that the subsequent belief state always belongs to the invariant finite subset. We show this restriction is equivalent to imposing a set of linear constraints on the set of admissible perception strategies, and that each value function update on the invariant finite belief set can be obtained through the solution of a linear program, allowing the synthesis of an optimal simultaneous perception–action strategy through a computationally tractable and parallelizable value iteration. The idea of value function updates over sampled belief states is reminiscent of point-based value iteration for POMDPs (Pineau et al., 2003). Our approach is fundamentally different since, by construction, the synthesized perception strategy renders a user-specified set of sampled belief states invariant, allowing for *exact* value iteration over this set. For general POMDPs with fixed observation functions, exact value iteration is not possible. Furthermore, we show that as the cardinality of the invariant finite belief set increases, the linear constraints on the sensing strategies become less binding, yielding better approximations of the continuous belief space solution. Finally, as the sample density of the invariant finite belief set increases, we prove that the value functions of the sampled belief states converge to their continuous counterparts, under an assumption on the structure of these value functions.


$\min_{\mathcal{A}} \sum_{t=1}^{T} \mathbb{E}\left[C_{\mathbf{s}_t,\mathbf{a}_t}\right]$
$\mathbf{s}_{t+1}$
$T_{\mathbf{s}_{t+1}|\mathbf{a}_t,\mathbf{s}_t}$
$\mathbf{a}_t$
$\mathbf{z}_t$
$\mathcal{P}_{\mathbf{z}_t|\mathbf{s}_t}$
$\hat{b}_t = f(\hat{b}_{t-1}, \mathbf{a}_{t-1}, \mathbf{z}_t)$
$\mathcal{A}_{\mathbf{a}_t|\hat{b}_t}$
Perception unit
Action unit
$\min_{\mathcal{P},\mathcal{A}} \sum_{t=1}^{T} \mathbb{E}\left[C_{\mathbf{s}_t,\mathbf{a}_t}\right] + \beta I(\mathbf{s}_{1:T} \to \mathbf{z}_{1:T})$
$\mathbf{s}_{t+1}$
$T_{\mathbf{s}_{t+1}|\mathbf{a}_t,\mathbf{s}_t}$
$\mathbf{a}_t$
b_t
$\mathbf{z}_t$
$\mathcal{P}_{\mathbf{z}_t|\mathbf{s}_t,b_t}$
$b_t = \text{predict}(\hat{b}_{t-1}, \mathbf{a}_{t-1})$
$\hat{b}_t = \text{update}(b_t, \mathbf{z}_t)$
$\mathcal{A}_{\mathbf{a}_t|\hat{b}_t}$
Perception unit
Action unit

(a) POMDP framework

(b) Simultaneous perception-action (SPADE) framework

Fig. 1. Visualization of the differences between the standard POMDP framework and the SPADE framework. Note the inclusion of feedback between the action and the perception unit.

Notation. We use uppercase symbols to denote sets and lowercase, bold symbols to denote random variables. For a set X, $\Delta(X)$ denotes the set of all probability distributions over X. The probability a random variable $\mathbf{x}$ takes a value $x \in X$ is denoted by $\Pr(\mathbf{x} = x)$. We denote a sequence $\{x_1, \dots, x_k\}$ of length k by $x_{1:k}$. For a vector $v \in \mathbb{R}^n$, let $\text{diag}(v) \in \mathbb{R}^{n \times n}$ be a diagonal matrix of the elements of v. We compactly express the summation over a set $\bar{X}$ by $\sum_{\bar{x}}$, where the corresponding set is clear by the notation used.

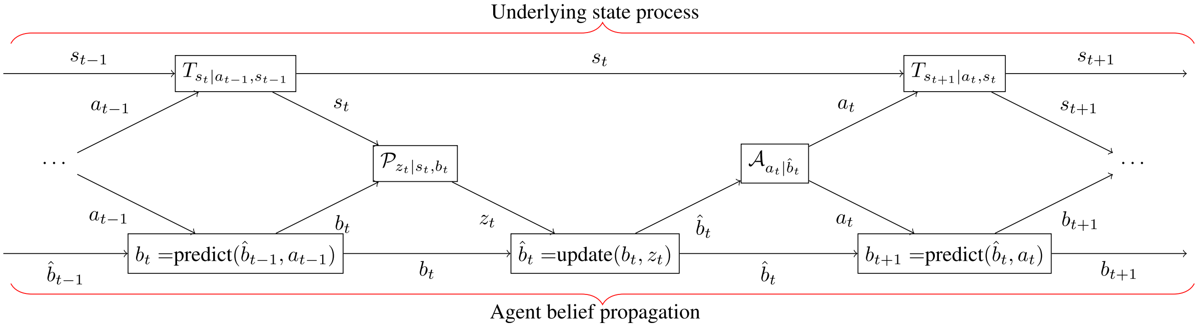


Fig. 2. Interplay between the underlying state process and the prior and posterior belief states of the agent.

2. Simultaneous perception–action design

We now formulate the simultaneous perception–action design (SPADE) problem of the agent.

2.1. Perception model

We use a *perception MDP* $\mathcal{M} = \langle S, A, T, C, Z, \gamma \rangle$ to model the environment of the agent, where $s \in S$ is a finite set of states, $a \in A$ is a finite set of actions, $T : S \times A \rightarrow \Delta(S)$ is a transition function mapping state–action pairs to probability distributions over successor states, $C : S \times A \rightarrow \mathbb{R} \geq 0$ is a cost function, $z \in Z$ is a set of observations, and $0 \leq \gamma < 1$ is a discount factor. Following the formulation of Shafieepoorfard et al. (2016), we assume the agent has access to an observation alphabet Z with infinite cardinality. The rationale behind this assumption will be discussed at the end of this subsection. We refer to the probability of transitioning to state s' after taking action a in state s by $T_{s'|a,s}$. Likewise, we refer to the cost of taking action a in state s by $C_{s,a}$ and the entire cost matrix by $C \in \mathbb{R}^{|S|\times|A|}$.

Due to imperfect information, an agent must estimate its current state $s_t \in S$ through its history of observations. Specifically, the agent maintains a *belief state* $b_t = [b_{t,s} : s \in S]^\top \in \Delta(S)$ at each time step t, where, for all $s \in S$, $b_{t,s} = \Pr(\mathbf{s}_t = s|z_{1:t-1})$ denotes the probability the agent believes it resides in state s at time t given the sequence of observations $z_{1:t-1}$. In our formulation, the agent maintains a parallel set of belief states. We refer to the first of these belief states as the *prior* belief state b_t, as defined previously. Particularly, the prior belief state characterizes the agent's belief *prior* to making an observation at time step t. The second belief state the agent maintains is referred to as the *posterior* belief state $\hat{b}_t = [\hat{b}_{t,s} : s \in S]^\top \in \Delta(S)$, where each $\hat{b}_{t,s}$ is defined according to $\hat{b}_{t,s} = \Pr(\mathbf{s}_t = s|z_{1:t})$, which is the belief state of the agent *after* making an observation about its underlying state s_t (but, we stress, before taking an action at time step t). Thus, the prior belief state captures the available information when choosing its perception strategy (i.e., its belief-dependent observation function), while the posterior belief state captures the available information when choosing an action. The differences between these parallel belief states are formalized in Section 2.2.

In our formulation, the agent is free to design *both* its action-selection strategy and a belief-dependent observation function. We refer to this joint process as a *simultaneous perception–action strategy*, which consists of both its action strategy $\mathcal{A}$ and its perception strategy $\mathcal{P}$. The agent's action strategy is a sequence $\mathcal{A} = \{\mathcal{A}_1, \ldots, \mathcal{A}_t, \ldots\}$, where $\mathcal{A}_t : \Delta(S) \rightarrow \Delta(A)$. The action strategy maps a posterior belief $\hat{b}$ at time t to a probability distribution over action selection. We denote the probability of taking action a in the posterior belief state $\hat{b}_t$ at time t as $\mathcal{A}_{a|\hat{b}_t}$. Likewise, the perception strategy is a sequence $\mathcal{P} = \{\mathcal{P}_1, \ldots, \mathcal{P}_t, \ldots\}$, where $\mathcal{P}_t : \Delta(S) \times S \rightarrow \Delta(Z)$. The perception strategy prescribes a belief-dependent observation function for the agent at each time step. We denote the probability of making observation z about state s while in the prior belief b_t at time t as $\mathcal{P}_{z|s,b_t}$.

Remark 1. Since the agent is able to synthesize its own perception strategy $\mathcal{P}$, our assumption that Z has infinite cardinality provides the agent the greatest freedom in this synthesis problem. However, as shown in Shafieepoorfard et al. (2016), there is no advantage to using a set Z with cardinality greater than that of $\Delta(S)$. In the main problem (4) that we formulate, we assume $Z = \Delta(S)$ without loss of generality, and assign each observation to a unique belief state.

2.2. Relation between prior and posterior belief states

Fig. 2 details the relation between the prior and posterior beliefs. Given a posterior belief $\hat{b}_t$, the agent first selects an action strategy $\mathcal{A}_{a|\hat{b}_t}$. Once selected, the agent then *predicts* the unique prior belief that it transitions to by

$$b_{t+1,s} = \sum_{s'} \sum_{a} T_{s|a,s'} \mathcal{A}_{a|\hat{b}_t} \hat{b}_{t,s'}. \tag{1}$$

When the action strategy is deterministic, i.e., there exists an action $a \in A$ such that $\mathcal{A}_{a|\hat{b}_t} = 1$, we can write $a = \mathcal{A}_{\hat{b}_t}$ and more succinctly express (1) as

$$b_{t+1,s} = \sum_{s'} T_{s|\mathcal{A}_{\hat{b}_t},s'} \hat{b}_{t,s'}. \tag{2}$$

Once the agent has transitioned to the prior belief state b_t, it then chooses its perception strategy $\mathcal{P}_{z|s,b_t}$. With probability $\Pr(\mathbf{z}_t = z) = \sum_s \mathcal{P}_{z|s,b_t} b_{t,s}$, the agent then makes an observation $\mathbf{z}_t = z$ and *updates* to the posterior belief state $\hat{b}_t$ according to the set of Bayesian updates

$$\hat{b}_{t,s} = (\mathcal{P}_{z|s,b_t} b_{t,s})/(\sum_{s'} \mathcal{P}_{z|s',b_t} b_{t,s'}) \tag{3}$$

for each $s \in S$. Note the transition to the posterior belief $\hat{b}_t$ occurs with probability $\Pr(\mathbf{z}_t = z)$. Furthermore, when Z is a countable set, the structure of (3) implies that, for any $b_t \in \Delta(S)$, $\hat{b}_t$ is a collection of point masses of weight $\Pr(\mathbf{z}_t = z)$. We adopt $\hat{b}_t^z$ to denote the unique posterior belief obtained when observation z is made while in prior belief state b_t. Similarly, we denote $b_{t+1}^{z,a}$ as the unique prior belief obtained when action a is chosen in posterior belief state $\hat{b}_t^z$.

2.3. An information-theoretic perception cost

In the absence of perception costs, it is always optimal to select a noiseless, full-state measurement (i.e., $Z = S$ and $z_t = s_t$)

as the perception strategy, although such measurements are often unrealistic. To make the perception design problem meaningful, it is necessary to introduce a metric to model the cost of information acquisition. Although the SPADE framework is general enough to support a variety of perception costs, we focus on a model where these costs are quantified by the information-theoretic concept of *directed information*. Directed information (Massey, 1990), sometimes referred to as transfer entropy (Schreiber, 2000), is a nonnegative quantity characterizing the information flow between random processes. For a pair of random processes $\mathbf{s}_{1:T}$ and $\mathbf{z}_{1:T}$, the directed information from $\mathbf{s}_{1:T}$ to $\mathbf{z}_{1:T}$ is

$$I(\mathbf{s}_{1:T} \to \mathbf{z}_{1:T}) \triangleq \sum_{t=1}^{T} I(\mathbf{s}_{1:t}; \mathbf{z}_t | \mathbf{z}_{1:t-1}),$$

where $I(\mathbf{s}_{1:t}; \mathbf{z}_t | \mathbf{z}_{1:t-1})$ is the conditional mutual information (Cover & Thomas, 2012), explicitly written as

$$\sum_{s_{1:t}, z_{1:t}} \Pr(s_{1:t}, z_{1:t}) \log \frac{\Pr(s_{1:t}, z_t | z_{1:t-1})}{\Pr(s_{1:t} | z_{1:t-1}) \Pr(z_t | z_{1:t-1})}.$$

In the SPADE model, the random processes $\mathbf{s}_{1:T}$ and $\mathbf{z}_{1:T}$ represent the state and the observation sequences, respectively. The directed information is closely related to the information traffic from the perception unit to the action unit (the lower left and right blocks of Fig. 1(b), respectively), and is a suitable metric to capture perception costs as, for our model, it is equivalent to the summation of the stage-additive information gains. A formal analysis providing the directed information with a Shannon-theoretic operational meaning is provided in Appendix A. For mathematical convenience, we introduce the discounted directed information:

$$\begin{aligned} I_\gamma(\mathbf{s}_{1:T} \to \mathbf{z}_{1:T}) &\triangleq \sum_{t=1}^{\infty} \gamma^{t-1} I(\mathbf{s}_{1:t}; \mathbf{z}_t | \mathbf{z}_{1:t-1}) \\ &= \sum_{t=1}^{\infty} \gamma^{t-1} I(\mathbf{s}_t; \mathbf{z}_t | \mathbf{z}_{1:t-1}). \end{aligned}$$

2.4. Main problem

In the SPADE problem, the objective of the agent is to minimize the discounted sum of its perception and environmental costs. The agent accomplishes this objective by optimally choosing its perception and action strategies, which we express through the objective function

$$\min_{\mathcal{P}_t, \mathcal{A}_t} \sum_{t=1}^{\infty} \gamma^{t-1} (\beta I(\mathbf{s}_t; \mathbf{z}_t | \mathbf{z}_{1:t-1}) + \mathbb{E}\left[C_{\mathbf{s}_t, \mathbf{a}_t}\right]), \tag{4}$$

where β is a parameter weighting the relative cost of information. We seek to formulate (4) in terms of a dynamic programming problem over the belief simplex $\Delta(S)$. To this end, consider a prior belief state with $b_{t,s} = \Pr(\mathbf{s}_t = s | z_{1:t-1})$. The stage-wise mutual information for b_t is

$$\begin{aligned} &I(\mathbf{s}_t; \mathbf{z}_t | \mathbf{z}_{1:t-1}) \\ &= \sum_s \int_Z \mathcal{P}_{z|s,b_t} b_{t,s} \log \frac{\mathcal{P}_{z|s,b_t}}{\sum_{s'} \mathcal{P}_{z|s',b_t} b_{t,s'}} \triangleq R_{b_t, \mathcal{P}_t}. \end{aligned} \tag{5}$$

Note that for a fixed b_t, the stage-wise mutual information in (5) is convex with respect to the perception strategy $\mathcal{P}_t$, since it is the weighted sum of relative entropies.

From the definitions of $R_{b_t, \mathcal{P}_t}$ and C_{s_t, a_t}, we can decompose the objective function in (4) into a set of recursive Bellman equations, where the value function is given by

$$V(b_t) \triangleq \inf_{\mathcal{A}_t, \mathcal{P}_t} \{\beta R_{b_t, \mathcal{P}_t} + \mathbb{E}_{b_t}^{\mathcal{A}_t, \mathcal{P}_t}[C_{\mathbf{s}_t, \mathbf{a}_t} + \gamma V(b_{t+1}^{\mathbf{z},\mathbf{a}})]\} \tag{6}$$

for each prior belief state b_t at time t. The notation $\mathbb{E}_{b_t}^{\mathcal{A}_t, \mathcal{P}_t}$ emphasizes that these expectations are evaluated under the joint distribution defined by $b_t, \mathcal{P}_t$ and $\mathcal{A}_t$. From a prior belief state b_t, recall that the agent first transitions to a posterior belief state $\hat{b}_t$ through the update Eq. (3) with probability given by $\Pr(\mathbf{z}_t = z)$. Once in $\hat{b}_t$, the agent then transitions to a prior belief state b_{t+1} through the predict Eq. (1). Through this evolution, we can decompose (6) into the perception and action stages, respectively:

$$V(b_t) = \inf_{\mathcal{P}_t} [\beta R_{b_t, \mathcal{P}_t} + \sum_s \int_Z \mathcal{P}_{z|s,b_t} b_{t,s} \hat{V}(\hat{b}_t^z) dz] \tag{7a}$$

$$\hat{V}(\hat{b}_t^z) = \min_{a_t} \left[\sum_s \hat{b}_{t,s}^z C_{s_t, a_t} + \gamma V(b_{t+1}^{z, a_t}) \right], \tag{7b}$$

for each $z \in Z$. In (7a), we have explicitly written the expectation by noting that, for a given prior belief state b_t, the posterior belief state $\hat{b}_t^z$ is a random variable realized with a probability $\Pr(\mathbf{z}_t = z)$ and a state distribution given by the update Eq. (3). To obtain (7b), recall for a posterior belief state $\hat{b}_t^z$, each action a yields a unique transition to the prior belief state $b_{t+1}^{z,a}$ given by the predict Eq. (2). In (7b), $\min_{\mathcal{A}_t}$ is replaced with $\min_a$ since it is straightforward to show the optimal action strategies are deterministic.

The combined set of recursive Bellman equations given by (7a) and (7b) suggests the use of dynamic programming to solve our main problem (4). Namely, denote by $B(\Delta(S))$ the space of functions $V : \Delta(S) \to \mathbb{R}$ such that $\|V\|_{\sup} \triangleq \sup_{b \in \Delta S} |V(b)| < +\infty$. Now, define the operator T by

$$(TV)(b) \triangleq \inf_{\mathcal{A}, \mathcal{P}} \{\beta R_{b, \mathcal{P}} + \mathbb{E}_b^{\mathcal{A}, \mathcal{P}}[C_{\mathbf{s}, \mathbf{a}} + \gamma V(b^{\mathbf{z}, \mathbf{a}})]\}. \tag{8}$$

Using T, the Bellman Eq. (6) can be written as $V = TV$. The following theorem states that T is a contractive mapping from $B(\Delta(S))$ to itself and that the corresponding value iteration is convergent.

Theorem 1. *The following results hold for the operator T:*

(a) For any $V \in B(\Delta(S))$ and $V' \in B(\Delta(S))$,

$$\|TV - TV'\|_{sup} \le \gamma \|V - V'\|_{sup}.$$

(b) For an arbitrary $V_0 \in B(\Delta(S))$, define a sequence of functions $\{V_k\}_{k=1,2,\ldots}$, $V_k \in B(\Delta(S))$, by $V_k = T^k V_0$, $k = 1, 2, \ldots$. Then, we have

$$\lim_{k \to \infty} \|V_k - V^*\|_{sup} = 0,$$

where $V^* \in B(\Delta(S))$ is the unique solution to $V^* = TV^*$.

Proof. (a) Let $q \triangleq \|V - V'\|_{\sup}$. Then,

$$V(b) - q \le V'(b) \le V(b) + q$$

for every $b \in \Delta(S)$. Applying the operator T to each side of the inequality, we have that, for each $b \in \Delta(S)$,

$$TV(b) - \gamma q \le TV'(b) \le TV(b) + \gamma q,$$

where we have made use of the fact that

$$\gamma \mathbb{E}_b^{\mathcal{A}, \mathcal{P}}[V(b^{\mathbf{z}, \mathbf{a}}) + q] = \gamma \mathbb{E}_b^{\mathcal{A}, \mathcal{P}}[V(b^{\mathbf{z}, \mathbf{a}})] + \gamma q$$

in (8). The result then follows.
(b) The space $B(\Delta(S))$ equipped with the sup norm $\| \cdot \|_{\sup}$ is a complete metric space. Since T is a contractive mapping from $B(\Delta(S))$ to itself, we apply the Banach fixed-point theorem (Khamsi & Kirk, 2011) to obtain the result. □

Unfortunately, the value iteration $V_k = T^k V_0$ is computationally intractable as $V(\cdot)$ must be evaluated everywhere on the continuous belief simplex $\Delta(S)$.

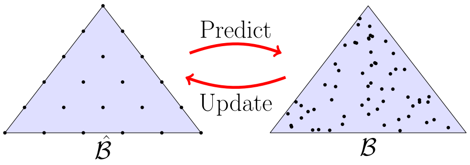


Fig. 3. As a visualization, consider the set $\hat{\mathcal{B}}$, which consists of the uniformly sampled belief states (black dots) drawn from the belief simplex shown at left. In turn, the set $\mathcal{B}$ of belief states is obtained by applying (1) to each belief state in $\hat{\mathcal{B}}$ for each action $a \in \mathcal{A}$. Note that the representative belief states in both $\hat{\mathcal{B}}$ and $\mathcal{B}$ shown above remain invariant over repeated predict and update steps.

Remark 2. It is possible to express our main objective (4) in terms of the standard POMDP paradigm described in Fig. 1(a) by augmenting the space of actions with the space of perception strategies; i.e., $\mathcal{A}_{\text{aug}} = (\mathcal{P}, \mathcal{A})$. Suppose that we then discretize the space of perception strategies such that $\mathcal{A}_{\text{aug, disc}} = (\mathcal{P}_{\text{disc}}, \mathcal{A})$ contains only a finite number of actions. By doing so, we have converted the problem into the standard form of a POMDP, for which we can use off-the-shelf POMDP solvers. However, this approach introduces two layers of approximation. One must first approximate the continuous space $\mathcal{P}$ of observation functions by a finite set $\mathcal{P}_{\text{disc}}$. The second approximation is due to the inherent hardness of POMDPs; standard PBVI provides a universal scheme to approximately solve a POMDP by discretizing the belief space. It is not clear how to cleverly perform both discretizations to achieve the best computational performance.

3. Method of invariant finite belief sets

Due to the continuity of both $\Delta(S)$ and Z, exactly solving for an optimal simultaneous perception–action strategy is computationally intractable. We now focus on approximating (7a) and (7b) by developing a novel method in which the agent operates on an invariant, finite subset of the continuous belief simplex. We refer to this set as an *invariant finite belief set* (IFBS). We then show that, as the number of sampled belief states approaches infinity, the value functions converge to their continuous state-space counterparts, under an appropriate assumption. For notational clarity, we omit the time index t of all variables.

3.1. Method of invariant finite belief sets

To construct a model approximating (7a) and (7b), we first pick a representative sample $\hat{\mathcal{B}} \subset \Delta(S)$ of posterior belief states. The set $\hat{\mathcal{B}}$ consists of a finite number of elements that must be chosen such that the following assumption is met.

Assumption 1. The set $\hat{\mathcal{B}}$ contains all extreme points of the belief simplex.

Recall from Remark 1 that we set $Z = \Delta(S)$ without loss of generality for the continuous problem. Similarly, for the method of invariant finite belief sets, we set $Z = \hat{\mathcal{B}}$ without loss of generality. Thus, we have that $|Z| = |\hat{\mathcal{B}}| \triangleq M$. Based on the selection of $\hat{\mathcal{B}}$, we can apply the predict Eq. (2) for each $\hat{b} \in \hat{\mathcal{B}}$ and each $a \in A$ to obtain the corresponding set of prior belief states, which we refer to as $\mathcal{B} \subset \Delta(S)$. Since the subsequent prior belief state b given a posterior belief state $\hat{b}$ and action a is unique, we have $|\mathcal{B}| = M \cdot |A|$.

Given the sets $\mathcal{B}$ and $\hat{\mathcal{B}}$, we now seek conditions under which they are *invariant*; i.e., over repeated predict and update steps using (2) and (3), the agent remains in $\mathcal{B}$ and $\hat{\mathcal{B}}$, as illustrated in Fig. 3. By the construction of $\mathcal{B}$, the predict Eq. (2) trivially yields prior beliefs exclusively in $\mathcal{B}$ for any $\hat{b} \in \hat{\mathcal{B}}$ and any action $a \in A$. It remains to show we can restrict the update equation such that the resulting posterior belief state remains in $\hat{\mathcal{B}}$. This condition requires that, for each prior belief $b \in \mathcal{B}$ and any observation $z \in Z$, the resulting posterior belief state is guaranteed to exist in the set $\hat{\mathcal{B}}$; i.e., $\hat{b}^z \in \hat{\mathcal{B}}$, where $\hat{b}^z$ is the posterior belief state that results from making observation z while in prior belief state b. Recalling the update Eq. (3), the individual probabilities $\hat{b}^z_s$ are

$$\hat{b}^z_s = \mathcal{P}_{z|s,b} b_s / (\sum_{s'} \mathcal{P}_{z|s',b} b_{s'}) \quad \forall s \in S. \tag{9}$$

To ensure that the posterior belief state $\hat{b}^z$ obtained by (9) remains in our invariant set $\hat{\mathcal{B}}$, we seek to impose restrictions on the set of admissible perception strategies $\mathcal{P}$ such that $\hat{b}^z \in \hat{\mathcal{B}}$ is guaranteed for all possible observations $z \in Z$. Recalling Remark 1, such restrictions can readily be imposed since the sets $\hat{\mathcal{B}}$ and Z have equal cardinalities, as well as the fact that the perception strategy is belief-dependent.

We now show these restrictions are linear constraints on $\mathcal{P}$ and are algorithmically straightforward to incorporate. To start with, recall the prior and posterior belief states are the vectors $b = [b_s : s \in S]^\top$ and $\hat{b} = [\hat{b}_s : s \in S]^\top$, respectively. We now introduce $\mathcal{P}_{z|:,b} \in \mathbb{R}^{|S|}$ for each observation $z \in Z$ and prior belief state $b \in \mathcal{B}$ to denote the vector $[\mathcal{P}_{z|s,b} : s \in S]^\top \in \mathbb{R}^{|S|}$. Using this notation, we introduce

$$\alpha^z_b \triangleq \sum_s \mathcal{P}_{z|s,b} b_s = b^\top \mathcal{P}_{z|:,b}$$

to encode the probability of observing z when in the prior belief state b and the perception strategy $\mathcal{P}$ is applied. Eq. (9) implies α^z_b and $\mathcal{P}_{z|:,b}$ have the linear relation

$$\alpha^z_b \hat{b}^z = \text{diag}(b) \mathcal{P}_{z|:,b}. \tag{10}$$

We use (10) to ensure that the updated posterior belief state remains in the set $\hat{\mathcal{B}}$ as follows. For a fixed prior belief state b, since the cardinalities of both the set $\hat{\mathcal{B}}$ and Z are equal, we can assign each z in (9) to a unique posterior belief state $\hat{b} \in \hat{\mathcal{B}}$. Explicitly writing $\hat{\mathcal{B}}$ as $\hat{\mathcal{B}} = \{\hat{b}^1, \dots, \hat{b}^M\}$, and Z as $Z = \{1, \dots, M\}$, we can alternatively express (10) as

$$\alpha^m_b \hat{b}^m = \text{diag}(b) \mathcal{P}_{m|:,b} \qquad \forall m = 1 \dots M, \tag{11}$$

where we have assigned, without loss of generality, the mth observation to update the prior belief state b to the mth posterior belief state $\hat{b}^m$ through (9). Rearranging terms in (11) and writing it for each $m = 1, \dots, M$, we obtain

$$\left[\begin{array}{ccc:ccc} \text{diag}(b) & & 0 & -\hat{b}^1 & & 0 \\ & \ddots & & & \ddots & \\ 0 & & \text{diag}(b) & 0 & & -\hat{b}^M \\ \hdashline I_{|S|} & \cdots & I_{|S|} & & 0_{|S| \times M} & \end{array}\right] \left[\begin{array}{c} \mathcal{P}_{:|:,b} \\ \hdashline \alpha_b \end{array}\right] = \left[\begin{array}{c} 0_{|S|M \times 1} \\ \hdashline 1_{|S| \times 1} \end{array}\right], \tag{12}$$

where $\mathcal{P}_{:|:,b} = [\mathcal{P}^\top_{1|:,b}, \dots, \mathcal{P}^\top_{m|:,b}]^\top$, $\alpha_b = [\alpha^1_b, \dots, \alpha^M_b]^\top$, and the lower set of constraints encodes that the perception strategy must be a valid probability distribution; i.e., $\sum_z \mathcal{P}_{z|s,b} = 1$ for each $s \in S$. Since valid probability distributions must have nonnegative elements, we impose that

$$\left[\mathcal{P}^m_b, \alpha^m_b\right]^\top \geq 0, \qquad \forall m = 1, \dots, M, \tag{13}$$

entry-wise. Assuming the number of observations is greater than the number of states; i.e., $M > |S|$, (12) is an underdetermined linear system with $(M \cdot |S| + M)$ variables and $(M \cdot |S| + |S|)$ constraints, so there exist infinitely many solutions to (12)–(13). We note that in realistic scenarios, we typically have $M \gg |S|$. We now state the following lemma.

Lemma 1. *Under Assumption 1, (12)–(13) admit a feasible solution.*

Proof. See Appendix B. □

In what follows, we denote $\mathcal{P}_{b\to\hat{\mathcal{B}}}$ as the subset of $\mathcal{P}$ satisfying the linear constraints (12)–(13).

3.2. Dynamic programming revisited

We now propose a method to approximate the dynamic programming formulas (7a) and (7b) using our invariant finite belief set. For the user-defined set of posterior belief states $\hat{\mathcal{B}}$ satisfying Assumption 1, denote its associated set of prior belief states as $\mathcal{B}$. For each $b \in \mathcal{B}$, we modify (7a) to

$$V(b) = \min_{\mathcal{P}\in\mathcal{P}_{b\to\hat{\mathcal{B}}}} [\beta R_{b,\mathcal{P}} + \sum_{s,z} \mathcal{P}_{z|s,b} b_s \hat{V}(\hat{b}^z)]. \tag{14}$$

We have modified (7a) by including the additional constraint $\mathcal{P} \in \mathcal{P}_{b\to\hat{\mathcal{B}}}$ which ensures that the perception strategy causes the agent to remain on the IFBS. Furthermore, recalling that we have set $|Z| = |\hat{\mathcal{B}}| = M$, we have replaced the integral operator in (7a) with a summation in (14). Notably, (14) is a convex optimization problem and can be further simplified. In the following lemma, we show that (14) can be reduced to an equivalent linear program (LP).

Lemma 2. *For a given prior belief state b, define*

$$\bar{S}(b) \triangleq \{s \in S \mid b_s \neq 0\}$$

$$\overline{M}(b) \triangleq \{m \mid \mathrm{supp}(\hat{b}^m) \subseteq \mathrm{supp}(b)\}.$$

Then, introducing the notation $b[\bar{S}(b)] \triangleq col_{s\in\bar{S}(b)}\{b(s)\}$ *and* $\hat{b}[\bar{S}(b)] \triangleq col_{s\in\bar{S}(b)}\{\hat{b}(s)\}$, *(14) is equivalent to the LP:*

$$\min_{\alpha_b\geq 0} \sum_{\bar{m}\in\overline{M}(b)} F_{\bar{m}}\alpha_b^{\bar{m}} \tag{15a}$$

$$\text{s.t.} \sum_{\bar{m}\in\overline{M}(b)} \alpha_b^{\bar{m}}\hat{b}^{\bar{m}}[\bar{S}(b)] = b[\bar{S}(b)] \tag{15b}$$

$$\alpha_b^m = 0 \quad \forall m \notin \overline{M}(b), \tag{15c}$$

where

$$F_{\bar{m}} = \beta D(\hat{b}^{\bar{m}} \,\|\, b) + \hat{V}(\hat{b}^{\bar{m}}),$$
$$D(\hat{b}^{\bar{m}} \,\|\, b) = \sum_{\bar{s}} \hat{b}^{\bar{m}}_{\bar{s}} \log(\hat{b}^{\bar{m}}_{\bar{s}}/b_{\bar{s}}). \tag{16}$$

Proof. See Appendix B. □

We introduce $C_{:,a} \in \mathbb{R}^{|S|}$ to denote the column of the cost matrix C corresponding to a given action $a \in A$. Then, for each $\hat{b}^m \in \hat{\mathcal{B}}$, we can express (7b) as

$$\hat{V}(\hat{b}^m) = \min_{a\in A}[C_{:,a}^\top \hat{b}^m + \gamma V(b^{m,a})]. \tag{17}$$

For a given IFBS, both (15) and (17) are computationally tractable. Furthermore, (15) is parallelizable for each $b \in \mathcal{B}$. Based on this discussion, the following backward dynamic programming problem is suggested: for each $b \in \mathcal{B}$, solve (15), while for each $\hat{b} \in \hat{\mathcal{B}}$, solve (17). Applying (17) to V followed by (15) is equivalent to applying the operator $\tilde{T} : \mathbb{R}^M \to \mathbb{R}^M$ defined according to

$$(\tilde{T}\tilde{V})(b) = \min_{\mathcal{A},\mathcal{P}\in\mathcal{P}_{b\to\hat{\mathcal{B}}}} \{\beta R_{b,\mathcal{P}} + \mathbb{E}_b^{\mathcal{A},\mathcal{P}}[C_{\mathbf{s},\mathbf{a}}] + \gamma\mathbb{E}_b^{\mathcal{A},\mathcal{P}}[\tilde{V}(b^{\mathbf{m},\mathbf{a}})]\}. \tag{18}$$

For an initial bounded vector $\tilde{V}_0 \in \mathbb{R}^M$, the value iteration procedure $\tilde{V}_k = \tilde{T}^k\tilde{V}_0$ can be viewed as an approximation of the original value iteration $V_k = T^kV_0$ in Theorem 1. Similar to Theorem 1, the convergence of the modified value iteration procedure is formalized as follows.

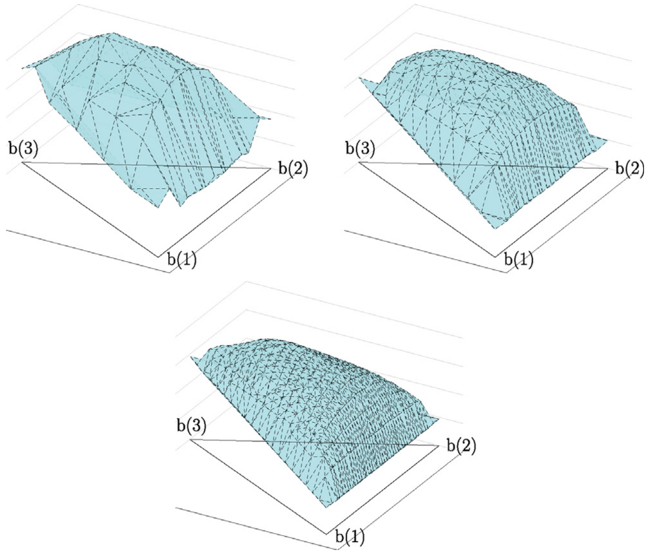


Fig. 4. Prior belief state value functions plotted on the 3D belief simplex for an increasing number of sampled belief states.

Theorem 2. *The following results hold for the operator* $\tilde{T}$*:*

(a) *For any bounded vectors* $\tilde{V} \in \mathbb{R}^M$ *and* $\tilde{V}' \in \mathbb{R}^M$,

$$\|\tilde{T}\tilde{V} - \tilde{T}\tilde{V}'\|_\infty \leq \gamma\|\tilde{V} - \tilde{V}'\|_\infty.$$

(b) *For an arbitrary bounded vector* $\tilde{V}_0 \in \mathbb{R}^M$, *define a sequence of bounded vectors* $\{\tilde{V}_k\}_{k=1,2,\ldots}$ *by* $\tilde{V}_k = \tilde{T}^k\tilde{V}_0$, $k = 1, 2, \ldots$. *Then, we have*

$$\lim_{k\to\infty} \tilde{V}_k = \tilde{V}^*,$$

where $\tilde{V}^* \in \mathbb{R}^M$ *is the unique solution to* $\tilde{V}^* = \tilde{T}\tilde{V}^*$.

Proof. The proof follows that of Theorem 1 and is omitted. □

To limit computation time, it is advantageous to construct $\hat{\mathcal{B}}$ using as few belief states as possible. However, doing so may yield value functions in (15) and (17) that poorly approximate the true value functions in (7a) and (7b). Thus, it is desirable to possess some method for improving the approximation. To this end, we present the following lemma.

Lemma 3. *Let* $\hat{\mathcal{B}}$ *be a set of posterior belief states with the corresponding set of prior belief states* $\mathcal{B}$ *found through the solution of* (2) *for each* $\hat{b} \in \hat{\mathcal{B}}$ *and* $a \in A$. *Denote their respective value functions by* $\hat{V}(\cdot)$ *and* $V(\cdot)$. *Consider a new set* $\hat{\mathcal{B}}' = \hat{\mathcal{B}} \cup \hat{b}$ *where* $\hat{b} \notin \hat{\mathcal{B}}$; *i.e.,* $\hat{\mathcal{B}}'$ *is formed by adding a sample belief state* $\hat{b}$ *to* $\hat{\mathcal{B}}$. *For* $\hat{\mathcal{B}}'$, *denote its associated set of prior belief states by* $\mathcal{B}'$ *and their value functions by* $\hat{V}'(\cdot)$ *and* $V'(\cdot)$, *respectively. Then,* $V(b) \geq V'(b)$ *for all* $b \in \mathcal{B}$ *and* $\hat{V}(\hat{b}) \geq \hat{V}'(\hat{b})$ *for all* $\hat{b} \in \hat{\mathcal{B}}$; *i.e., the value functions are monotonically non-increasing as* $|\hat{\mathcal{B}}|$ *increases.*

Proof. See Appendix B. □

Following from Lemma 3, if it is believed that (15)–(17) yield poor approximations, then one can add sample beliefs to $\hat{\mathcal{B}}$, and recompute. Recall, by Assumption 1, that increasing the cardinality of $\hat{\mathcal{B}}$ likewise requires increasing the cardinality of the observation alphabet such that $|\hat{\mathcal{B}}| = M$. In this sense, Lemma 3 implies an agent with a larger observation alphabet can do no worse than an agent with a smaller one when synthesizing its perception strategy. Fig. 4 illustrates that the fixed point $V(b)$ of the value iteration is monotonically non-increasing in M for a simple three-state example. Additional information on this example is provided in Appendix D.

3.3. Approximation of value functions

Theorem 1 implies the value function V^* for the main problem (4) can be obtained by the value iteration $V_k = T^k V_0$. Although such a procedure is computationally intractable, we showed an approximation $\tilde{V}^*$ of V^* can be obtained through the modified value iteration $\tilde{V}_k = \tilde{T}^k \tilde{V}_0$, which is performed exclusively on the IFBS. We now study how the gap between V^* and $\tilde{V}^*$ depends on the IFBS sample density, characterized as follows: for each $\hat{b} \in \Delta(S)$, denote by $\pi_{\hat{b}} \in \hat{\mathcal{B}}$ the nearest element to $\hat{b}$ in $\hat{\mathcal{B}}$ whose support is contained in the support of $\hat{b}$. That is,

$$\begin{aligned} \pi_{\hat{b}} \triangleq \operatorname{argmin}_{\hat{b}^m \in \hat{\mathcal{B}}} \quad & \|\hat{b} - \hat{b}^m\|_\infty \\ \text{s.t.} \quad & \operatorname{supp}(\hat{b}^m) \subseteq \operatorname{supp}(\hat{b}). \end{aligned} \tag{19}$$

The constraint that $\operatorname{supp}(\hat{b}^m) \subseteq \operatorname{supp}(\hat{b})$ will be necessary for the proof of Lemma 4 below. Using π, define

$$\hat{\epsilon} \triangleq \max_{\hat{b} \in \Delta(S)} \|\hat{b} - \pi_{\hat{b}}\|_\infty \tag{20}$$

as a sampling density parameter. How well the function V^* can be approximated also depends on the "regularity" of V^*. We define the regularity parameter $\hat{\delta} \geq 0$ such that

$$|V^*(b) - V^*(b')| \leq \hat{\delta} \tag{21}$$

holds for all $b, b' \in \Delta(S)$ such that $\|b - b'\|_\infty \leq \hat{\epsilon}|S|$. The main result of this subsection critically relies on the following lemma, which provides an upper bound on the difference between the operators T and $\tilde{T}$ applied to the same function V. For a given function $V \in B(\Delta(S))$, denote by $V|_{\mathcal{B}} \in \mathbb{R}^M$ the restriction to the set $\mathcal{B}$; i.e., $V|_{\mathcal{B}}$ is the function V evaluated only at points in the set $\mathcal{B}$.

Lemma 4. *Suppose that a function $V \in B(\Delta(S))$ satisfies*

$$|V(b) - V(b')| \leq \hat{\delta} \tag{22}$$

for all $b, b' \in \Delta(S)$ such that $\|b - b'\|_\infty \leq \hat{\epsilon}|S|$. Then,

$$\|(TV)|_{\mathcal{B}} - \tilde{T}(V|_{\mathcal{B}})\|_\infty \leq \epsilon, \tag{23}$$

$$\begin{aligned} \epsilon \triangleq\ & \gamma\hat{\delta} + \hat{\epsilon}\beta|\log\hat{\epsilon}\,\|S| + \hat{\epsilon}\cdot \\ & (\beta\sum_{\bar{s}} |\log b_{\bar{s}}| + \sum_{s,a} |C_{s,a}|). \end{aligned} \tag{24}$$

Proof. See Appendix C. □

The main result of this subsection is summarized as follows:

Theorem 3. *Let $\mathcal{B}$ and $\hat{\mathcal{B}}$ be fixed, and define $\hat{\epsilon}$ by (20). Let $V^* \in B(\Delta(S))$ be the unique function satisfying $V^* = TV^*$, and define $\hat{\delta}$ by (21). Define the sequence $\tilde{V}_k \in \mathbb{R}^M$, $k = 1, 2, \ldots$ by $\tilde{V}_k = \tilde{T}^k \tilde{V}_0$, where $\tilde{V}_0 \in \mathbb{R}^M$ is an arbitrary bounded vector. Then, for the ϵ defined by (24),*

$$\limsup_{k\to\infty} \|V^*|_{\mathcal{B}} - \tilde{V}_k\|_\infty \leq \epsilon / (1 - \gamma)\,. \tag{25}$$

Proof. Notice that

$$\begin{aligned} & \|V^*|_{\mathcal{B}} - \tilde{V}_{k+1}\|_\infty \\ & = \|(TV^*)|_{\mathcal{B}} - \tilde{T}\tilde{V}_k\|_\infty \\ & = \|(TV^*)|_{\mathcal{B}} - \tilde{T}(V^*|_{\mathcal{B}}) + \tilde{T}(V^*|_{\mathcal{B}}) - \tilde{T}\tilde{V}_k\|_\infty \\ & \leq \|\tilde{T}(V^*|_{\mathcal{B}}) - \tilde{T}\tilde{V}_k\|_\infty + \|(TV^*)|_{\mathcal{B}} - \tilde{T}(V^*|_{\mathcal{B}})\|_\infty \\ & \leq \gamma\|V^*|_{\mathcal{B}} - \tilde{V}_k\|_\infty + \epsilon. \end{aligned} \tag{26}$$

Algorithm 1 Method of invariant finite belief sets.

```
Require: $S, A, C, T, \gamma, \beta, \hat{\mathcal{B}}$, tol $> 0$
  $\mathcal{B} \leftarrow \emptyset$
  for $\hat{b} \in \hat{\mathcal{B}}, a \in A$ do
      $b \leftarrow$ (1) for fixed $\hat{b}, a$;    $\mathcal{B} \leftarrow \mathcal{B} \cup \{b\}$
  $\hat{V}(\hat{b}), V(b) \leftarrow 0$ for all $\hat{b} \in \hat{\mathcal{B}}, b \in \mathcal{B}$
  while err $>$ tol do
      $V'(b), \mathcal{P}_b \leftarrow$ (15) $\forall b \in \mathcal{B}$;    $\hat{V}'(\hat{b}), \mathcal{A}(\hat{b}) \leftarrow$ (17) $\forall \hat{b} \in \hat{\mathcal{B}}$
      err $\leftarrow \max\{\max_{b\in\mathcal{B}} |V(b) - V'(b)|, \max_{\hat{b}\in\hat{\mathcal{B}}} |\hat{V}(\hat{b}) - \hat{V}'(\hat{b})|\}$
      $V(b) \leftarrow V'(b)\ \forall b \in \mathcal{B}$;    $\hat{V}(b) \leftarrow \hat{V}'(b)\ \forall \hat{b} \in \hat{\mathcal{B}}$
```

The first equality is obtained by invoking $V^* = TV^*$ and $\tilde{V}_{k+1} = \tilde{T}\tilde{V}_k$. In the last step, we used the fact that $\tilde{T}$ is contractive (Theorem 2) and the result of Lemma 4. Define a sequence e_k of positive numbers by

$$e_{k+1} = \gamma e_k + \epsilon, \tag{27}$$

taking $e_0 = \|V^*|_{\mathcal{B}} - \tilde{V}_0\|_\infty$. Then,

$$\lim_{k\to\infty} e_k = \epsilon / (1 - \gamma)\,. \tag{28}$$

Combining the results of (26), (27), and (28), it is straightforward to show by induction that

$$\|V^*|_{\mathcal{B}} - \tilde{V}_k\|_\infty \leq e_k \quad \forall k = 0, 1, \ldots,$$

from which (25) follows. □

Notice the constant ϵ in (25) depends on $\hat{\epsilon}$ and $\hat{\delta}$. Since the optimal value function V^* is not known in advance, it is in general not possible to compute $\hat{\delta}$. However, in circumstances where V^* is known to be uniformly continuous, for each $\hat{\delta} > 0$, the condition (21) can always be guaranteed by choosing a sufficiently small $\hat{\epsilon} > 0$, i.e., by making the set $\hat{\mathcal{B}}$ sufficiently dense in $\Delta(S)$. In such cases, $\tilde{V}^*$ can approximate $V^*|_{\mathcal{B}}$ arbitrarily well by increasing the sample density of $\hat{\mathcal{B}}$ in $\Delta(S)$. It is currently not known under what conditions the uniform continuity of V^* is guaranteed. Obtaining these conditions remains the subject of future work.

3.4. Algorithmic implementation

The pseudocode for solving the SPADE problem via the method of invariant finite belief sets is shown in Algorithm 1. Given an initial sample $\hat{\mathcal{B}}$ of posterior belief states, the first step is to compute the corresponding set of prior belief states $\mathcal{B}$ by propagating each $\hat{b} \in \hat{\mathcal{B}}$ through the transition function T for each $a \in A$, as given in the predict equation (2). Once both sets are constructed, the value functions are initialized to zero. These value functions are then updated until the maximum error between subsequent iterations is below a threshold value. In each iteration, the value functions of all states in $\mathcal{B}$ and $\hat{\mathcal{B}}$ are updated. Note that each of the M posterior belief state value function updates has a time complexity of $O(|S| \cdot |A|)$, yielding a total time complexity of $O(M \cdot |S| \cdot |A|)$. Similarly, each of the $M \cdot |A|$ prior belief state value function updates involves solving an LP with a time complexity of $O(M^3)$ (Gonzaga, 1989), yielding a total time complexity of $O(M^4 \cdot |A|)$. Combining these results, the time complexity of each iteration is $O(M \cdot |S| \cdot |A| + M^4 \cdot |A|)$.

4. Numerical example: Mars rover

Consider a Mars rover that must maneuver from its initial position to a target position as shown in Fig. 5, where the blue state is the initial position, the green states are the target position,

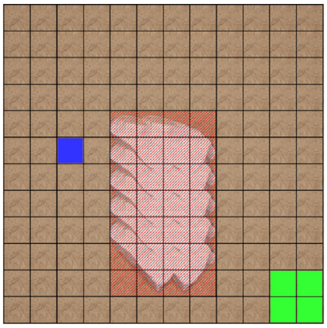


Fig. 5. Mars gridworld environment considered. The blue state is the initial position, the green states are the target position, and the red, dashed states are the dangerous, rocky region to be avoided.

and the red, dashed states are the dangerous, rocky region to be avoided. To reach its target position, the rover must avoid a dangerous section in the center of its environment. To accomplish this objective, the rover can take one of two possible paths: a more direct but dangerous path that travels along the bottom of the environment, or a longer but safer path around the top of the rocky area. In this example, we study the relation between the relative cost of perception and the resulting trajectory of the rover. In this context, the notion of "task-relevant" information pertains to observations identifying the rover's underlying state.

In each state, the rover can choose between one of 4 available actions: move one step left, right, up, or down. Due to stochasticity, however, the rover may either remain in its current state or slip into one of the other surrounding states, each with probability $^{0.05}/_{8}$. If the rover were to transition to a state outside the environment, it instead transitions to the closest state still within the environment. In Fig. 5, the green and red-dashed states are absorbing; i.e., if the rover reaches any of its target states or one of the rocky states, it remains there. Furthermore, for all non-target states, the rover incurs a cost of 1 for taking any action.

We sample six posterior belief states for each state $s \in S$. We start by sampling the extreme points of the belief simplex, satisfying Assumption 1. For each $s \in S$, we then set $\hat{b}_s = 0.5$ and $\hat{b}_{s'} = {}^{0.5}/_{8}$ for all s', $s' \neq s$ in the 3×3 square centered around s. For the third posterior belief state, we repeat the previous using $\hat{b}_s = 0.75$. We then follow a similar process considering the 5×5 square centered around s. First, we set $\hat{b}_s = 0.5$, $\hat{b}_{s'} = {}^{0.5}/_{16}$ for the eight states s' in the 3×3 square around s, and $\hat{b}_{s''} = {}^{0.5}/_{32}$ for the remaining sixteen states in the 5×5 square. We repeat this process using $\hat{b}_s = 0.35$ and $\hat{b}_s = 0.20$. If this procedure allocates non-zero probability mass to a state outside the environment, it is instead allocated to the closest state within the environment. This procedure yields a set of 864 posterior belief states, which subsequently yields a set of 3456 prior belief states using (2). We use a discount factor of $\gamma = 0.95$ and consider values of $\beta = 0$ and $\beta = 20$, where $\beta = 0$ corresponds to the case where the rover incurs no perception cost. Considering an infinite time horizon, we perform value iteration until convergence for each value of β, wherein we solve (15) for each prior belief state and (17) for each posterior belief state at each iteration. To solve each LP, we use the Gurobi LP solver (Gurobi Optimization, 2022).

To discuss the qualitative differences between the synthesized joint perception–action strategies for the values of β considered, we examine the sample trajectories they generate. Fig. 6 plots the expected state residence averaged over 1000 trials for each value of β. In the case $\beta = 0$, we see that the rover takes the shorter path underneath the rocky obstacles to reach its target, since, in the absence of perception costs, it is strictly focused on minimizing environmental costs. The rover is able to follow this path as it has perfect state knowledge. Because it incurs no perception costs, the synthesized perception strategy will always uniquely indicate the true underlying state of the rover. The rover is thus able to leverage this perfect state information towards taking a more direct path to reach the target states.

For $\beta = 20$, perception costs drive the rover to exhibit different behavior. In most simulations, the rover takes the longer path around the top of the rocky area. Intuitively, the synthesized perception strategy drives the rover to maintain a diffuse belief state, as such belief states correspond to lower perception costs. Thus, to balance perception and environmental costs, the rover remains in belief states that are diffuse yet have a low probability of residing in a rocky state. In some cases, the rover initially slips several states in the opposite direction of the safe path. Once in such a state, the environmental costs associated with following an even longer trajectory dominate the perception costs. In these cases, the rover follows trajectories similar to those of $\beta = 0$. Through this example, we see how the relative costs associated with perception and the agent's environment can yield significantly different behavior.

5. Numerical example: Dynamic sensor selection

A common problem in the field of active perception is that of dynamic sensor selection (Spaan & Lima, 2009). Often, due to power or computational constraints, it is infeasible for an agent to simultaneously use all available sensors to make observations, and the agent must instead synthesize a strategy for selecting a subset of the sensors to use at each time step. Here, we use the output of (15)–(17), namely, the perception strategy $\mathcal{P}$, to synthesize a dynamic sensor selection strategy. We consider the ring environment in Fig. 7, where the agent must estimate the state of an uncontrolled target. As shown in Fig. 7 for state s_1, the uncontrolled dynamics are described by a Markov chain in which the target transitions to the same state with probability $^{1}/_{2}$, to a neighboring state with probability $^{1}/_{6}$, and to a state two steps away with probability $^{1}/_{12}$. The goal for the agent is to minimize its total maximum a posteriori (MAP) estimation error of the true target state, which we encode using the set of actions $A = \{a_1, \ldots, a_8\}$ and the cost function $C = \mathbf{1}_{8\times 8} - \mathbf{I}_8$, where $\mathbf{1}_{8\times 8}$ is an 8×8 matrix of ones. The action a_i corresponds to picking state s_i to be the true state of the target.

The agent can use a set of 8 sensors, with one centered on each state, labeled $\mathcal{S}^i$ for $i = 1, \ldots, N$, as shown in Fig. 7 for the sensor centered on state s_4. Associated with each sensor $\mathcal{S}^i$ is an observation function $\mathcal{P}^i$ that maps each state to a probability distribution over 6 possible observations. We provide the observation function for the sensor centered on state s_4 in Table 1, where the observations correspond to observing the target in state s_2, s_3, s_4, s_5, s_6, or simply making "no observation", respectively. Note the probability that the sensor makes the correct observation increases as the true target state moves closer to s_4. The remaining observation functions are constructed similarly.

The dynamic sensor selection strategy is constructed as follows. To start, let $\hat{b}^{ri}_{t-1} = [\hat{b}^{ri}_{t-1,s_1}, \ldots, \hat{b}^{ri}_{t-1,s_8}]^\top$ be the belief state of the agent at time step $t-1$, and let b^{ri}_t be the belief state obtained after propagating $\hat{b}^{ri}_{t-1}$ through the transition dynamics (1). Using the propagated belief state b^{ri}_t, define $b^* \triangleq \min_{b\in\mathcal{B}} \|b - b^{ri}_t\|_2$, i.e., b^* is the closest sampled prior belief state to the agent's current belief state. We now select a subset of sensors, denoted $\mathcal{S}^t \subseteq \{\mathcal{S}^1, \ldots, \mathcal{S}^8\}$, such that their combined observation function, denoted $\mathcal{P}^t$, is closest to the synthesized observation function $\mathcal{P}^{b^*} \triangleq \mathcal{P}_{z|s,b^*}$. For a given belief state b, we use the

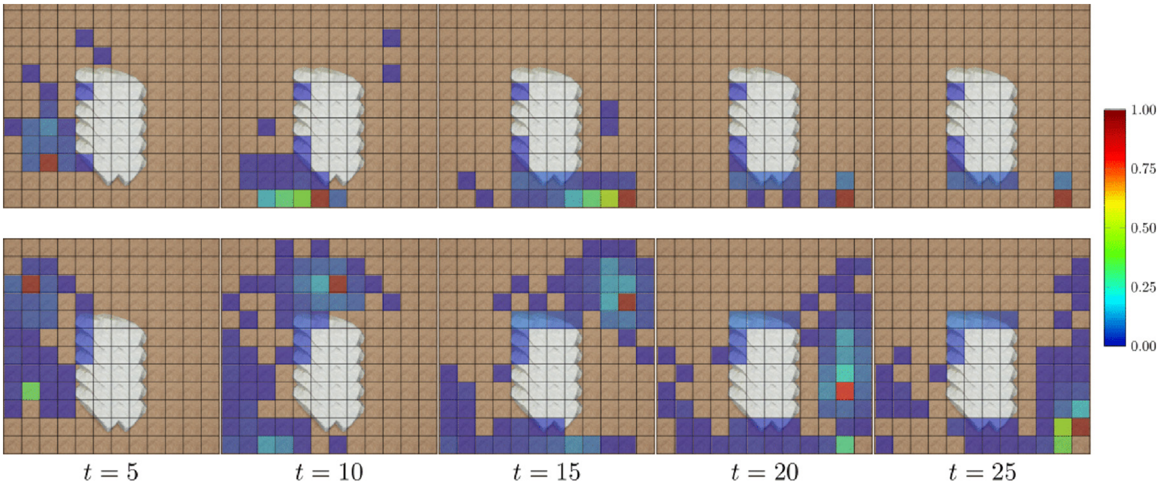


Fig. 6. Expected rover state residences averaged over 1000 trials for $\beta = 0$ (upper row) and $\beta = 20$ (lower row) plotted at various time steps. The shading of each state corresponds to the fraction of the trials that occupied that state at the given time step. When $\beta = 0$, the rover incurs no perception costs and is able to take a more direct path to the goal state. For $\beta = 20$, the rover faces prohibitive perception-related costs and instead favors a longer path around the rocky obstacle.

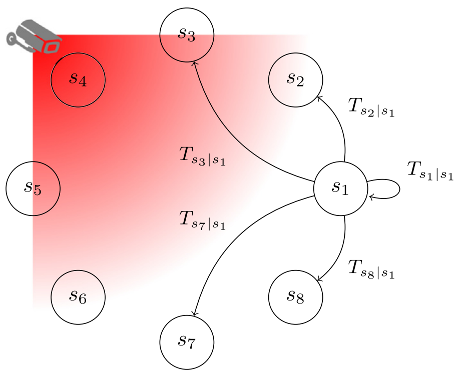


Fig. 7. Visualization of the ring environment considered.

Table 1
Probabilities $\mathcal{P}^4_{z|s}$ for the sensor centered on state s_4.

	z_1	z_2	z_3	z_4	z_5	z_6
s_1	0.068	0.034	0	0	0	0.898
s_2	0.384	0.085	0.043	0	0	0.488
s_3	0.107	0.480	0.107	0.053	0	0.253
s_4	0.067	0.133	0.600	0.133	0.068	0
s_5	0	0.053	0.107	0.480	0.107	0.253
s_6	0	0	0.043	0.085	0.384	0.488
s_7	0	0	0	0.034	0.068	0.898
s_8	0.027	0	0	0	0.027	0.945

following metric, denoted $d(\mathcal{P}^1, \mathcal{P}^2|b)$, to define how "close" a pair of observation functions $\mathcal{P}^1$ and $\mathcal{P}^2$ are:

$$d(\mathcal{P}^1, \mathcal{P}^2|b) \triangleq \mathbb{E}[\,\|\,\mathbb{E}[\text{bayes}(b,z)|\mathcal{P}^1] - \mathbb{E}[\text{bayes}(b,z)|\mathcal{P}^2]\,\|_1 \mid b\,], \qquad (29)$$

where, for a given $\mathcal{P}$, bayes(b, z) denotes the standard Bayesian update given by $b'_s = (\mathcal{P}_{z|s} b_s)/(\sum_{s'} \mathcal{P}_{z|s'} b_{s'})$ for each $s \in S$. Intuitively, $d(\mathcal{P}^1, \mathcal{P}^2|b)$ measures the distance between the expected belief states obtained by each of the observation functions given a fixed ground-truth state, weighted by the probability that the state is the ground-truth state. Note that using this metric requires checking 2^8 sensor combinations at each time step. For systems with a larger number of sensors, one could alternatively implement a greedy approach to minimize $d(\mathcal{P}^t, \mathcal{P}^{b^*}|b^{ri}_t)$.

We sample a total of 193 posterior belief states, beginning with the extreme points of the belief simplex and iteratively add increasingly diffuse belief states centered on each of the 8 states. Furthermore, since the dynamics of the target are a Markov chain, the set $\mathcal{B}$ likewise consists of 193 belief states. We use a discount factor of $\gamma = 0.95$ and vary the values of β from 0.3 to 0.8, solving the value iteration procedure (15)–(17) for each. After completing this value iteration for a given value of β, we then simulate the proposed dynamic sensor selection strategy, starting with a uniform belief state $b^{ri}_1 = [{}^1\!/_8, \ldots, {}^1\!/_8]^\top$. The resulting sensor selections over a horizon of 50 times steps are shown in Fig. 8 for each value of β. For values of β below 0.3, our proposed algorithm always picks all available sensors, whereas if the value of β is greater than 0.8, it almost always picks no sensor. As expected, as the value of β increases, so too does the number of sensors picked at each time step. In Fig. 9, we further plot the cumulative MAP estimation error for each value of β considered. As β increases, the algorithm trades off an increase in the estimation error with the use of fewer sensors. Note that, due to the construction of each $\mathcal{P}^i$, even selecting all sensors yields nonzero estimation error.

As a comparison, we implement the work of Spaan and Lima (2009), which uses a PBVI-based approach in which the set of POMDP actions consists of the actions $\{a_1, \ldots, a_8\}$ and the set of all possible sensor selections $2^{\{\mathcal{S}_1,\ldots,\mathcal{S}_8\}}$. In this method, the sensing constraint is directly imposed as a cardinality constraint on the number of sensors that can be chosen at each time step. We refer the reader to Pineau et al. (2003) for additional information on PBVI. For each cardinality constraint considered, we begin with a sample of 73 belief states and perform 5 rounds of belief set expansion, backing up the α-vectors 25 times per iteration. Note that, although the PBVI approach starts with fewer belief states than our algorithm, the belief expansion step nearly doubles the cardinality after each round.

Table 2 displays the comparison between this approach and our proposed algorithm. The simulations are run on a 1.8 GHz Intel Core i7-8550 CPU with 16 GB RAM (note we have not fully optimized either implementation). When the average cardinalities of the chosen sensor subsets are similar, (e.g., in the case $k = 2$ and $\beta = 0.7$), the PBVI-based approach outperforms our algorithm in both the average time required per online iteration and in minimizing the cumulative MAP estimation error. Since at runtime the PBVI output is simply finding the minimizing

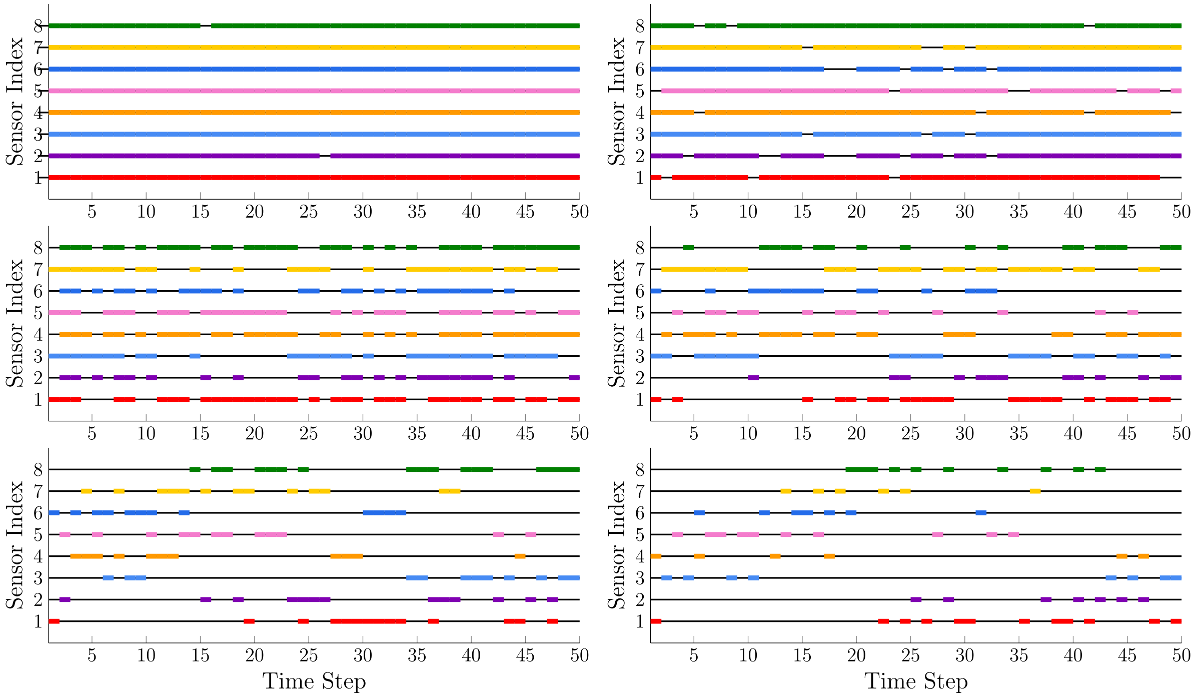


Fig. 8. Comparison of the sensor selection at each time step for various values of β. Clockwise from top left are the plots for $\beta = 0.3$, $\beta = 0.4$, $\beta = 0.5$, $\beta = 0.6$, $\beta = 0.7$, and $\beta = 0.8$, respectively.

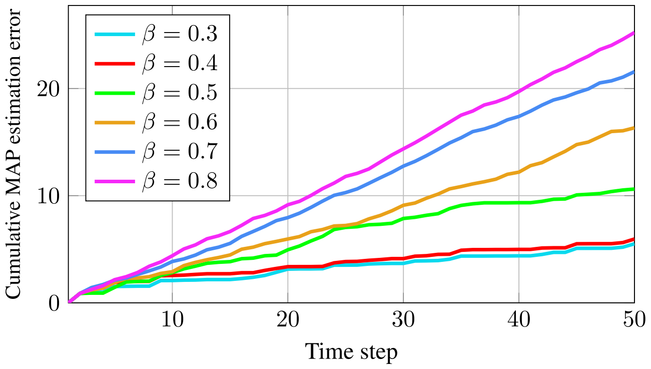


Fig. 9. MAP estimation error as a function of β.

hyperplane, it is significantly more efficient than the exhaustive search our algorithm performs at each iteration. Furthermore, since the offline portion of the PBVI-based approach explicitly incorporates the sensors $\{\mathcal{S}^1, \ldots, \mathcal{S}^8\}$, it is unsurprising that the resulting cumulative MAP error is less than that of our algorithm. However, the offline portion of PBVI quickly becomes intractable as the cardinality constraint increases. Although each iteration of PBVI can be done in polynomial time, it requires constructing a prohibitively large set of α-vectors at each backup step, as one vector is required for each of the exponential number of observations for a combinatorial number of sensors. On the other hand, the offline portion of our algorithm is *sensor agnostic* and only sees a slight change in the offline runtime due to varying β. Another benefit of this agnosticism is that changing the number or the specific observation functions of the sensors at runtime does not require resolving the offline portion, whereas doing so is required for PBVI. Thus, our algorithm shows promise for use in large sensor networks for which the offline solution of PBVI-based methods is prohibitively expensive, and sensors may be added or updated at runtime.

Table 2
Comparison of the PBVI-based approach and our algorithm's offline solution times, online solution times, cumulative MAP estimation error, and average number of chosen sensors.

Method	Offline [s]	Online [s]	Error	$\lvert\mathcal{S}^t\rvert$
PBVI, $k = 1$	24.95	8.1×10^{-4}	23.6191	1.00
PBVI, $k = 2$	8.6×10^2	6.0×10^{-4}	15.84	2.00
PBVI, $k = 3$	1.5×10^4	7.1×10^{-4}	11.4284	3.00
PBVI, $k = 4$	TO	–	–	–
$\beta = 0.3$	39.75	0.98	5.51	7.96
$\beta = 0.4$	42.73	0.95	5.96	7.19
$\beta = 0.5$	46.44	1.05	10.62	4.91
$\beta = 0.6$	50.43	1.15	16.33	3.18
$\beta = 0.7$	50.37	1.14	21.57	2.12
$\beta = 0.8$	54.13	1.13	25.23	1.35

6. Conclusion and future work

We consider a simultaneous perception–action design problem wherein the perception costs are modeled using the directed information. The agent's objective function is decomposed into two coupled sets of recursive Bellman equations, allowing a tractable, approximate solution through a novel *method of invariant finite belief sets* to be obtained. The proposed method restricts the agent to operate exclusively on a finite subset of the continuous belief space. An optimal simultaneous perception–action strategy is then obtained using a dynamic programming approach wherein a linear program is solved for each prior belief state at each iteration.

Several natural extensions of the SPADE framework are as follows. To start with, once the optimal perception strategy is obtained, the next step is to select, or develop, a sensor that "physically realizes" the perception strategy. The types of additional constraints that must be imposed on the perception strategy to allow for such a sensor remains the subject of future work. Furthermore, for the tabular algorithm we propose in Eqs. (15) and (17), naïvely constructing the posterior belief set $\hat{\mathcal{B}}$ yields an impractical cardinality for realistic, large-scale

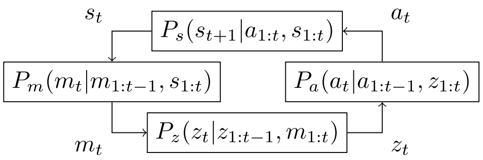

s_t
$P_s(s_{t+1}|a_{1:t}, s_{1:t})$
a_t
$P_m(m_t|m_{1:t-1}, s_{1:t})$
$P_a(a_t|a_{1:t-1}, z_{1:t})$
m_t
$P_z(z_t|z_{1:t-1}, m_{1:t})$
z_t

Fig. A.1. Feedback system considered.

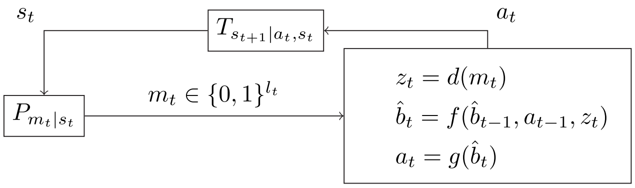


Fig. A.2. Perception through bitpipe.

problems. Developing methods to cleverly construct $\hat{\mathcal{B}}$ remains an important direction for future study. An alternative direction for mitigating computational costs is to generalize the state space in terms of a set of features, as in the case of tile or Kanerva coding in the reinforcement learning literature (Sutton & Barto, 2018) and to solve the SPADE problem on this feature space.

Appendix A. Details on information-theoretic perception cost

We now provide a rationale for using the directed information to model perception costs. To this end, notice that

$$I(\mathbf{s}_{1:T} \to \mathbf{z}_{1:T}) = \sum_{t=1}^{T} I(\mathbf{s}_{1:t}; z_t|\mathbf{z}_{1:t-1}) \tag{A.1a}$$

$$= \sum_{t=1}^{T} I(\mathbf{s}_t; \mathbf{z}_t|\mathbf{z}_{1:t-1}) - I(\mathbf{s}_{1:t-1}; \mathbf{z}_t|\mathbf{s}_t, \mathbf{z}_{1:t-1}) \tag{A.1b}$$

$$= \sum_{t=1}^{T} I(\mathbf{s}_t; \mathbf{z}_t|\mathbf{z}_{1:t-1}) \tag{A.1c}$$

$$= \sum_{t=1}^{T} H(\mathbf{s}_t|\mathbf{z}_{1:t-1}) - H(\mathbf{s}_t|\mathbf{z}_{1:t}). \tag{A.1d}$$

The second term in (A.1b) is zero as our model assumes $\mathbf{z}_t$ is independent of $\mathbf{s}_{1:t-1}$ given $\mathbf{s}_t$, as in Fig. 1(b). Therefore, the directed information is equivalent to the summation of the stage-additive information gains; i.e., the difference in entropy of the state variable before and after incorporating $\mathbf{z}_t$. We interpret this information gain as the minimum number of information *bits* that must be delivered from the perception unit to the action unit in each time step. The set of SPADE parameters that minimize $I(\mathbf{s}_{1:T} \to \mathbf{z}_{1:T})$ are advantageous, since an optimal source coding (i.e., data compression) scheme can potentially reduce the data traffic from the perception unit to the action unit to $I(\mathbf{s}_{1:T} \to \mathbf{z}_{1:T})$ bits. We now present a formal analysis providing the directed information a Shannon-theoretic operational meaning wherein we model the communication channel from the perception unit to the action unit as a noiseless bitpipe, through which the message $\mathbf{z}_t$ is delivered in the form of a variable-length, uniquely-decodable binary code. Specifically, let P_s, P_m, P_z and P_a be conditional probability distributions with the structures shown in Fig. A.1. Given an initial distribution $P_s(\mathbf{s}_1)$, let $\mathbf{s}_{1:T}, \mathbf{m}_{1:T}, \mathbf{z}_{1:T}$ and $\mathbf{a}_{1:T}$ be random processes defined by the feedback diagram in Fig. A.1. Then, we obtain the following lemma.

Lemma 5. *Given the Feedback system shown in Fig. A.1, the following inequalities hold:*

$$I(\mathbf{s}_{1:T} \to \mathbf{a}_{1:T}) \leq I(\mathbf{s}_{1:T} \to \mathbf{z}_{1:T}) \leq I(\mathbf{m}_{1:T} \to \mathbf{z}_{1:T}),$$
$$I(\mathbf{s}_{1:T} \to \mathbf{a}_{1:T}) \leq I(\mathbf{m}_{1:T} \to \mathbf{a}_{1:T}) \leq I(\mathbf{m}_{1:T} \to \mathbf{z}_{1:T}).$$

Proof. See, e.g., Derpich and Østergaard (2021). □

To provide the directed information a Shannon-theoretic operational meaning, assume messages from the perception unit to the action unit are communicated through a noiseless bitpipe, as shown in Fig. A.2, and that delivering an individual bit incurs a unit cost. This formulation provides a meaningful model to estimate the cost of communication in applications where sensor data is transmitted over a digital communication channel. In time step t, the perception unit produces a uniquely decodable variable-length binary code $m_t \in \{0, 1\}^{\ell_t}$, where ℓ_t is the length of the code. The message m_t is decoded in the action unit to reproduce the observation signal $z_t \in Z$. Since communication is costly, the design goal of the simultaneous perception–action system, including message encoder and decoder, is to minimize

$$\sum_{t=1}^{T} \mathbb{E}[C_{\mathbf{s}_t,\mathbf{a}_t}] + \mathbb{E}[\boldsymbol{\ell}_t], \tag{A.2}$$

where $\mathbb{E}[\boldsymbol{\ell}_t]$ is the expected codeword length. Although it is difficult to evaluate $\mathbb{E}[\boldsymbol{\ell}_t]$ directly, it can be approximated using directed information, as shown in the following lemma.

Lemma 6. *For the communication system in Fig. A.2*

$$I(\mathbf{s}_{1:T} \to \mathbf{z}_{1:T}) \leq \sum_{t=1}^{T} \mathbb{E}[\boldsymbol{\ell}_t]. \tag{A.3}$$

Proof. The following inequalities establish the claim:

$$I(\mathbf{s}_{1:t} \to \mathbf{z}_{1:t}) \leq I(\mathbf{m}_{1:t} \to \mathbf{z}_{1:t}) \tag{A.4a}$$

$$= \sum_{t=1}^{T} I(\mathbf{m}_{1:t}; \mathbf{z}_t|\mathbf{z}_{1:t-1})$$

$$= \sum_{t=1}^{T} I(\mathbf{m}_t; \mathbf{z}_t|\mathbf{z}_{1:t-1}) - \underbrace{I(\mathbf{m}_{t-1}; \mathbf{z}_t|\mathbf{m}_t, \mathbf{z}_{1:t-1})}_{=0}$$

$$= \sum_{t=1}^{T} H(\mathbf{m}_t|\mathbf{z}_{1:t-1}) - \underbrace{H(\mathbf{m}_t|\mathbf{z}_{1:t})}_{\geq 0}$$

$$\leq \sum_{t=1}^{T} H(\mathbf{m}_t) \tag{A.4b}$$

$$\leq \sum_{t=1}^{T} \mathbb{E}[\boldsymbol{\ell}_t]. \tag{A.4c}$$

Lemma 5 is applied to obtain (A.4a). The inequality in (A.4b) holds since conditioning can only reduce the entropy. Finally, (A.4c) follows from the fact any uniquely-decodable code is a uniquely-decodable code of itself, and thus its expected codeword length is lower-bounded by its entropy, see, e.g., Theorem 5.3.1 of Cover and Thomas (2012). □

Evaluating the tightness of the lower bound (A.3) is more challenging. However, it is reported (Kostina & Hassibi, 2019; Tanaka et al., 2017) that the construction of source coders operating at a rate close to this lower bound is possible under some special circumstances, such as LQG settings.

Appendix B. Proofs of Lemmas 1–3

Proof of Lemma 1. Starting with the constraint given in (11), we sum over all $m = 1 \dots M$ to obtain

$$\sum_m \alpha_b^m \hat{b}^m = \sum_m \mathrm{diag}(b)\mathcal{P}_{m|:,b}. \tag{B.1}$$

Since $\mathrm{diag}(b)$ is present in each term, we can move it outside the summation. Then, noting $\mathcal{P}$ must be a valid probability distribution for each $s \in S$; i.e., $\sum_m \mathcal{P}_{m|s,b} = 1$, we must have that $\sum_m \mathcal{P}_{m|:,b} = 1_{|S|\times 1}$. Substituting into (B.1) yields

$$\sum_m \alpha_b^m \hat{b}^m = \mathrm{diag}(b)\mathbf{1}_{|S|\times 1}, \tag{B.2}$$

which we can write more intuitively in matrix notation as

$$\begin{bmatrix}\hat{b}^1 & \cdots & \hat{b}^M\end{bmatrix}\begin{bmatrix}\alpha_b^1 & \cdots & \alpha_b^M\end{bmatrix}^\top = b. \tag{B.3}$$

Since we chose $\hat{\mathcal{B}}$ such that Assumption 1 is satisfied, we can express any b as a convex combination of extreme points of $\Delta(S)$. Thus, there exist nonnegative coefficients α_b^m, $m = 1 \dots M$, satisfying (B.3). Furthermore, in realistic applications, we will often have that $M \gg |S|$. By this condition, there exist infinitely many solutions to (12)–(13). □

Proof of Lemma 2. Considering only states $s \in \bar{S}(b)$, we use (10) to parameterize the perception strategy variables as

$$\mathcal{P}_{m|\bar{s},b} = \alpha_b^m(\hat{b}_{\bar{s}}^m / b_{\bar{s}}) \tag{B.4}$$

for all $\bar{s} \in \bar{S}(b)$ and all $m = 1, \dots, M$. Since b and $\hat{b}^m$, $m = 1, \dots, M$, are nonnegative vectors, and α_b^m is constrained to be nonnegative, the parameterized variables are likewise guaranteed to be nonnegative. Substituting this parameterization into the first set of linear constraints in (12) and multiplying each side by $\mathrm{diag}(b[\bar{S}(b)])$, we obtain

$$\sum_{\bar{m}} \alpha_b^{\bar{m}} \hat{b}^{\bar{m}}[\bar{S}(b)] = b[\bar{S}(b)] \tag{B.5}$$

Now, substituting the parameterization for the perception strategy variables into (14), we see that

$$\sum_{\bar{m}}\sum_{\bar{s}} b_{\bar{s}}\mathcal{P}_{\bar{m}|\bar{s},b}(\beta \log \frac{\mathcal{P}_{\bar{m}|\bar{s},b}}{\alpha_b^{\bar{m}}} + \hat{V}(\hat{b}^{\bar{m}})) \tag{B.6a}$$

$$= \sum_{\bar{m}}\sum_{\bar{s}} \alpha_b^{\bar{m}} \hat{b}_{\bar{s}}^{\bar{m}}(\beta \log(\hat{b}_{\bar{s}}^{\bar{m}}/b_{\bar{s}}) + \hat{V}(\hat{b}^{\bar{m}})) \tag{B.6b}$$

$$= \sum_{\bar{m}} \alpha_b^{\bar{m}} \sum_{\bar{s}} \hat{b}_{\bar{s}}^{\bar{m}}(\beta \log(\hat{b}_{\bar{s}}^{\bar{m}}/b_{\bar{s}}) + \hat{V}(\hat{b}^{\bar{m}})) \tag{B.6c}$$

$$= \sum_{\bar{m}} \alpha_b^{\bar{m}}(\beta \sum_{\bar{s}} \hat{b}_{\bar{s}}^{\bar{m}} \log(\hat{b}_{\bar{s}}^{\bar{m}}/b_{\bar{s}}) + \hat{V}(\hat{b}^{\bar{m}})). \tag{B.6d}$$

By defining F_m and $D(\hat{b}^m \parallel b)$ as in (16), we can rewrite the objective function in (B.6d) together with the constraints in(B.5) to obtain the desired LP, completing the proof. □

Proof of Lemma 3. Consider a perception strategy $\mathcal{P} \in \mathcal{P}_{b\to\hat{\mathcal{B}}}$ with observation probabilities $\mathcal{P}_{z^m|s,b}$, where the observation alphabet has cardinality M. Now, consider a perception strategy $\mathcal{P}'$ with observation alphabet cardinality $|Z'| = M + 1$ constructed as follows. For each z^m, $m = 1, \dots, M$, let $\mathcal{P}'_{z^m|s,b} = \mathcal{P}_{z^m|s,b}$ for all $S \times \mathcal{B}$, and let all remaining $\mathcal{P}'_{z^m|s,b}$ be arbitrarily chosen such that $\mathcal{P}' \in \mathcal{P}'_{b\to\hat{\mathcal{B}}'}$. Then, there is a one-to-one correspondence between the expectations over successor states for both the prior and posterior belief states in the sets $\mathcal{B}$ and $\hat{\mathcal{B}}$; thus, their respective value functions are equal. Since we chose $\mathcal{P}$ arbitrarily, the value functions for each belief state when synthesizing a perception strategy with an observation alphabet containing $M + 1$ elements cannot be greater than the case of synthesizing a perception strategy containing M elements; i.e.,

$$V(b) \geq V'(b) \;\; \forall b \in \mathcal{B}, \quad \hat{V}(\hat{b}) \geq \hat{V}'(\hat{b}), \quad \forall \hat{b} \in \hat{\mathcal{B}}. \quad □$$

Appendix C. Proof of Lemma 4

We will use the following basic lemma:

Lemma 7. *Let $p, q \in \Delta(S)$ be two probability distributions such that $\|p - q\|_\infty \leq \epsilon \leq 1/2$. Then, it holds that*

$$|H(p) - H(q)| \leq \epsilon|\log \epsilon||S|.$$

Proof. The proof follows that of Theorem 17.3.3 of Cover and Thomas (2012). Consider the concave function $f(t) = -t \log t$. Since $f(0) = f(1) = 0$, it follows that $f(t) \geq 0$ for all $t \in [0, 1]$. The maximum absolute slope of the chord of $f(t)$ from t to $t + \epsilon$ is obtained at either end, where either $t = 0$ or $t = 1 - \epsilon$. Thus, for $0 \leq t \leq 1 - \epsilon$, it follows that

$$|f(t) - f(t + \epsilon)| \leq \max\{f(\epsilon), f(1 - \epsilon)\} = -\epsilon \log \epsilon \tag{C.1}$$

since $\epsilon \leq 1/2$. Then,

$$|H(p) - H(q)| = |\sum_s(-p(s) \log p(s)$$

$$+ q(s) \log q(s))| \tag{C.2a}$$

$$\leq \sum_s | - p(s) \log p(s) + q(s) \log q(s)| \tag{C.2b}$$

$$\leq \sum_s -\epsilon \log \epsilon = -\epsilon \log \epsilon|S|, \tag{C.2c}$$

where (C.2c) follows from (C.1). □

To show (23), we need to prove the gap between $(TV)(b)$ and $(\tilde{T}V|_{\mathcal{B}})(b)$, given by (8) and (18), respectively, is bounded by ϵ for each $b \in \mathcal{B}$. Let $(\mathcal{A}^*, \mathcal{P}^*)$ be a minimizer for (8). If a minimizer does not exist, one can instead consider an ϵ_0-suboptimal solution for a sufficiently small $\epsilon_0 > 0$. In this case, the following proof can be adapted with only minor adjustments. Since $\mathcal{P}^*_{z|s,b}$ is unconstrained, the posterior belief $\hat{b}^z$ can take general values in $\Delta(S)$. Particularly, $\mathcal{P}^*_{z|s,b}$ drives the prior belief b to a posterior belief $\hat{b}^z \in \Delta(S)$ with probability $\alpha_b^z = \sum_s \mathcal{P}^*_{z|s,b} b_s$. For each $s \in S$ such that $b_s \neq 0$, it follows from Bayes' rule (9) that

$$\mathcal{P}^*_{z|s,b} = \alpha_b^z(\hat{b}_s^z / b_s). \tag{C.3}$$

Note $(\mathcal{A}^*, \mathcal{P}^*)$ may not be admissible for (18). Instead, we construct an admissible $(\tilde{\mathcal{A}}^*, \tilde{\mathcal{P}}^*)$ for (18) (i.e., $\tilde{\mathcal{P}}^* \in \mathcal{P}_{b\to\hat{\mathcal{B}}}$) such that $(\tilde{\mathcal{A}}^*, \tilde{\mathcal{P}}^*)$ is "close" to $(\mathcal{A}^*, \mathcal{P}^*)$. We construct this perception policy $\tilde{\mathcal{P}}^* \in \mathcal{P}_{b\to\hat{\mathcal{B}}}$ from $\mathcal{P}^*$ as follows: for each $\hat{b}^m \in \hat{\mathcal{B}}$, define the neighborhood in Z as $N(\hat{b}^m) \triangleq \{z \in Z : \pi_{\hat{b}^z} = \hat{b}^m\}$. Clearly, we have $Z = \bigcup_{m=1}^M N(\hat{b}^m)$. Set

$$\alpha_b^m = \int_{N(\hat{b}^m)} \alpha_b^z dz, \tag{C.4}$$

for each $m = 1, \dots, M$. Now, define $\tilde{\mathcal{P}}^*$ as

$$\tilde{\mathcal{P}}^*_{m|s,b} = \begin{cases}\alpha_b^m(\hat{b}_s^m / b_s) & \text{for } s \text{ such that } b(s) \neq 0\\ \text{arbitrary} & \text{for } s \text{ such that } b(s) = 0\end{cases} \tag{C.5}$$

for each $m = 1, \dots, M$ and $s \in S$. Note $\tilde{\mathcal{P}}^*_{m|s,b}$ is well-defined by the constraint $\mathrm{supp}(\hat{b}^m) \subseteq \mathrm{supp}(\hat{b})$ in the definition of $\pi_{\hat{b}}$ in (19). Specifically, if $b_s = 0$, then by (C.3), it must hold that

$\hat{b}^z_s = 0$ as well. Then, under the constraint in (19), it follows that $\hat{b}^m_s = \pi_{\hat{b}^z,s} = 0$. Likewise, define $\tilde{\mathcal{A}}^*$ as

$$\tilde{\mathcal{A}}^*_{a|\hat{b}^m} = \begin{cases} \frac{1}{\alpha^m_b} \int_{N(\hat{b}^m)} \alpha^z_b \mathcal{A}^*_{a|\hat{b}^z} dz & \text{if } \alpha^m_b \neq 0 \\ \text{arbitrary} & \text{if } \alpha^m_b = 0 \end{cases} \tag{C.6}$$

for each $a \in A$ and $\hat{b}^m \in \hat{\mathcal{B}}$. Now, notice that under the perception strategy $\tilde{\mathcal{P}}^*$, an observation m is observed with probability $\sum_s \tilde{\mathcal{P}}^*_{m|s,b} b_s = \sum_s \alpha^m_b \hat{b}^m_s = \alpha^m_b$. Then, by Bayes' rule (9), the posterior belief becomes $\hat{b}^m$ when m is observed since, by the construction of $\tilde{\mathcal{P}}^*$,

$$\frac{\tilde{\mathcal{P}}^*_{m|s,b} b_s}{\sum_{s'} \tilde{\mathcal{P}}^*_{m|s',b} bs'} = \frac{(\alpha^m_b \hat{b}^m_s / b_s) b_s}{\alpha^m_b} = \hat{b}^m_s .$$

Therefore, $\tilde{\mathcal{P}}^* \in \mathcal{P}_{b \to \hat{\mathcal{B}}}$ as desired. We now compare the right hand sides of (8) and (18) evaluated under $(\mathcal{A}^*, \mathcal{P}^*)$ and $(\tilde{\mathcal{A}}^*, \tilde{\mathcal{P}}^*)$, respectively, term-by-term.

C.1. Comparison of $R^*_{b,\mathcal{P}}$ and $R^*_{b,\tilde{\mathcal{P}}}$

Using (C.3), the perception cost $R^*_{b,\mathcal{P}}$ is

$$\begin{aligned} R^*_{b,\mathcal{P}} &= \sum_s \int_Z \mathcal{P}^*_{z|s,b} b_s \log(\mathcal{P}^*_{z|s,b}/\alpha^z_b) dz \\ &= \sum_s \int_Z \alpha^z_b \hat{b}^z_s \log(\hat{b}^z_s / b_s) dz = \int_Z \alpha^z_b D(\hat{b}^z \,\|\, b) dz \qquad \text{(C.7)} \end{aligned}$$

On the other hand, the perception cost $R^*_{b,\tilde{\mathcal{P}}}$ is

$$\begin{aligned} R^*_{b,\tilde{\mathcal{P}}} &= \sum_{s,m} \tilde{\mathcal{P}}^*_{m|s,b} b_s \log(\tilde{\mathcal{P}}^*_{m|s,b}/\alpha^m_b) \\ &= \sum_{s,m} \alpha^m_b \hat{b}^m_s \log(\hat{b}^m_s / b_s) && \text{(C.8a)} \\ &= \sum_m \alpha^m_b D(\hat{b}^m \,\|\, b) \\ &= \sum_m \int_{N(\hat{b}^m)} \alpha^z_b D(\hat{b}^m \,\|\, b) dz && \text{(C.8b)} \\ &= \sum_m \int_{N(\hat{b}^m)} \alpha^z_b D(\pi_{\hat{b}^z} \,\|\, b) dz && \text{(C.8c)} \\ &= \int_Z \alpha^z_b D(\pi_{\hat{b}^z} \,\|\, b) dz && \text{(C.8d)} \end{aligned}$$

Step (C.8a) follows by the construction of $\tilde{\mathcal{P}}^*$ in (C.5). The definition of α^m_b in (C.4) is used to obtain (C.8b). To obtain (C.8c), recall that $\pi_{\hat{b}^z} = \hat{b}^m$ for each $z \in N(\hat{b}^m)$. Finally, applying the fact that $\bigcup_{m=1}^M N(\hat{b}^m) = Z$ yields (C.8d). Comparing (C.7) and (C.8d), we obtain

$$\begin{aligned} |R^*_{b,\mathcal{P}} - R^*_{b,\tilde{\mathcal{P}}}| &= |\int_Z (D(\hat{b}^z \| b) - D(\pi_{\hat{b}^z} \| b)) dz| \\ &\leq \int_Z |D(\hat{b}^z \| b) - D(\pi_{\hat{b}^z} \| b)| dz \\ &\leq \max_{\hat{b} \in \Delta(S)} |D(\hat{b} \| b) - D(\pi_{\hat{b}} \| b)|. \end{aligned}$$

The last expression can be upper bounded as follows:

$$\begin{aligned} |D(\hat{b} \| b) - D(\pi_{\hat{b}} \| b)| &= |\sum_{\bar{s}} -\hat{b}_{\bar{s}} \log(\hat{b}_{\bar{s}}/b_{\bar{s}}) \\ &\qquad + \sum_{\bar{s}} \pi_{\hat{b},\bar{s}} \log(\pi_{\hat{b},\bar{s}}/b_{\bar{s}})| && \text{(C.9a)} \\ &= |\sum_{\bar{s}} -\hat{b}_{\bar{s}} \log \hat{b}_{\bar{s}} + \sum_{\bar{s}} \pi_{\hat{b},\bar{s}} \log \pi_{\hat{b},\bar{s}} \\ &\qquad + \sum_{\bar{s}} (\pi_{\hat{b},\bar{s}} - \hat{b}_{\bar{s}}) \log b_{\bar{s}}| && \text{(C.9b)} \\ &= |H(\hat{b}) - H(\pi_{\hat{b}}) + \sum_{\bar{s}} (\pi_{\hat{b},\bar{s}} - \hat{b}_{\bar{s}}) \log b_{\bar{s}}| && \text{(C.9c)} \\ &\leq |H(\hat{b}) - H(\pi_{\hat{b}})| + |\sum_{\bar{s}} (\pi_{\hat{b},\bar{s}} - \hat{b}_{\bar{s}}) \log b_{\bar{s}}| && \text{(C.9d)} \\ &\leq |H(\hat{b}) - H(\pi_{\hat{b}})| + \hat{\epsilon} |\sum_{\bar{s}} \log b_{\bar{s}}| && \text{(C.9e)} \\ &\leq \hat{\epsilon} |\log \hat{\epsilon}||S| + \hat{\epsilon} |\sum_{\bar{s}} \log b_{\bar{s}}|, && \text{(C.9f)} \end{aligned}$$

where (C.9b) follows from standard properties of logarithms, (C.9c) by the definition of the entropy, (C.9d) by the triangle inequality, (C.9e) by the fact $\|\hat{b} - \pi_{\hat{b}}\|_\infty \leq \hat{\epsilon}$, and (C.9f) from Lemma 7. Recalling the relative entropy is summed over only $\bar{S}(b)$, $\log b(\bar{s})$ is finite for all $b \in \mathcal{B}$ and $\bar{s} \in \bar{S}(b)$.

C.2. Comparison of $\mathbb{E}^{\mathcal{A}^*,\mathcal{P}^*}_b[C_{\mathbf{s},\mathbf{a}}]$ and $\mathbb{E}^{\tilde{\mathcal{A}}^*,\tilde{\mathcal{P}}^*}_b[C_{\mathbf{s},\mathbf{a}}]$

We now compare $\mathbb{E}^{\mathcal{A},\mathcal{P}}_b[C_{\mathbf{s},\mathbf{a}}]$ in (8) and $\mathbb{E}^{\tilde{\mathcal{A}}^*,\tilde{\mathcal{P}}^*}_b[C_{\mathbf{s},\mathbf{a}}]$ in (18). Notice that

$$\mathbb{E}^{\mathcal{A}^*,\mathcal{P}^*}_b[C_{\mathbf{s},\mathbf{a}}] = \sum_{s,a} \left(\int_Z \alpha^z_b \mathcal{A}^*_{a|\hat{b}^z} \hat{b}^z_s dz \right) C_{s,a}. \tag{C.10}$$

On the other hand,

$$\begin{aligned} \mathbb{E}^{\tilde{\mathcal{A}}^*,\tilde{\mathcal{P}}^*}_b[C_{\mathbf{s},\mathbf{a}}] &= \sum_{s,a,m} \alpha^m_b \tilde{\mathcal{A}}^*_{a|\hat{b}^m} \hat{b}^m_s C_{s,a} \\ &= \sum_{s,a,m} (\int_{N(\hat{b}^m)} \alpha^z_b \mathcal{A}^*_{a|\hat{b}^z} dz) \hat{b}^m_s C_{s,a} && \text{(C.11a)} \\ &= \sum_{s,a,m} (\int_{N(\hat{b}^m)} \alpha^z_b \mathcal{A}^*_{a|\hat{b}^z} \pi_{\hat{b}^z,s} dz) C_{s,a} && \text{(C.11b)} \\ &= \sum_{s,a} (\int_Z \alpha^z_b \mathcal{A}^*_{a|\hat{b}^z} \pi_{\hat{b}^z,s} dz) C_{s,a} && \text{(C.11c)} \end{aligned}$$

Equality (C.11a) follows from the definition of $\tilde{\mathcal{A}}^*_{a|\hat{b}^m}$ in (C.6). Equality (C.11b) then holds by recalling $\pi_{\hat{b}^z} = \hat{b}^m$ for all $z \in N(\hat{b}^m)$. The fact $\bigcup_{m=1}^M N(\hat{b}^m) = Z$ is used to obtain (C.11c). Now, comparing (C.10) and (C.11c),

$$\begin{aligned} &|\mathbb{E}^{\mathcal{A}^*,\mathcal{P}^*}_b[C_{\mathbf{s},\mathbf{a}}] - \mathbb{E}^{\tilde{\mathcal{A}}^*,\tilde{\mathcal{P}}^*}_b[C_{\mathbf{s},\mathbf{a}}]| \\ &= |\sum_{s,a} (\int_Z \alpha^z_b \mathcal{A}^*_{a|\hat{b}^z} \hat{b}^z_s dz) C_{s,a} \\ &\qquad - \sum_{s,a} (\int_Z \alpha^z_b \mathcal{A}^*_{a|\hat{b}^z} \pi_{\hat{b}^z,s} dz) C_{s,a}| \\ &\leq \sum_{s,a} |C_{s,a}| (\int_Z \alpha^z_b \mathcal{A}^*_{a|\hat{b}^z} \cdot |\hat{b}^z_s - \pi_{\hat{b}^z,s}| dz) \\ &\leq \epsilon \sum_{s,a} |C_{s,a}| (\int_Z \alpha^z_b \mathcal{A}^*_{a|\hat{b}^z} dz) \\ &\leq \epsilon \sum_{s,a} |C_{s,a}| (\int_Z \alpha^z_b dz) = \epsilon \sum_{s,a} |C_{s,a}|. \end{aligned}$$

C.3. Comparison of $\mathbb{E}_b^{\mathcal{A}^*,\mathcal{P}^*}[V(b^{\mathbf{z},\mathbf{a}})]$ *and* $\mathbb{E}_b^{\tilde{\mathcal{A}},\tilde{\mathcal{P}}^*}[V(b^{\mathbf{z},\mathbf{a}})]$

We now compare $\mathbb{E}_b^{\mathcal{A},\mathcal{P}}[V(b^{\mathbf{z},\mathbf{a}})]$ in (8) and $\mathbb{E}_b^{\mathcal{A},\mathcal{P}}[V(b^{\mathbf{m},\mathbf{a}})]$ in (18). Under $(\mathcal{P}^*, \mathcal{A}^*)$, the random variables $(\mathbf{z}, \mathbf{a})$ are realized according to the probability distribution $\mathcal{A}^*_{a|\hat{b}^z}\alpha_b^z = \mathcal{A}^*_{a|\hat{b}^z}\sum_s \mathcal{P}^*_{z|s,b}b_s$, while under $(\tilde{\mathcal{P}}^*, \tilde{\mathcal{A}}^*)$, the random variables $(\mathbf{m}, \mathbf{a})$ are realized according to $\tilde{\mathcal{A}}^*_{a|\hat{b}^m}\alpha_b^m$. Thus,

$$
\begin{aligned}
&|\mathbb{E}_b^{\mathcal{A}^*,\mathcal{P}^*}[V_k(b^{\mathbf{z},\mathbf{a}})] - \mathbb{E}_b^{\tilde{\mathcal{A}}^*,\tilde{\mathcal{P}}^*}[V_k(b^{\mathbf{m},\mathbf{a}})]| \\
&= |\sum_a \int_Z \mathcal{A}^*_{a|\hat{b}^z}\alpha_b^z V_k(b^{z,a})dz \\
&\quad - \sum_{a,m} \tilde{\mathcal{A}}^*_{a|\hat{b}^m}\alpha_b^m V_k(b^{m,a})| \\
&= |\sum_a \int_Z \mathcal{A}^*_{a|\hat{b}^z}\alpha_b^z V_k(b^{z,a})dz \\
&\quad - \sum_{a,m} \int_{N(\hat{b}^m)} \mathcal{A}^*_{a|\hat{b}^z}\alpha_b^z V_k(b^{m,a})|. \qquad \text{(C.12)}
\end{aligned}
$$

To obtain (C.12), we again use the definition of $\tilde{\mathcal{A}}^*_{a|\hat{b}^m}$ in (C.6). Denote by $b^{m(z),a}$ the prior belief obtained when action a is selected in $\pi_{\hat{b}^z}$, i.e., $b_s^{m(z),a} = \sum_{s'} T_{s|a,s'}\pi_{\hat{b}^z,s'}$. It is noteworthy that for each $s \in S$ and $z \in Z$,

$$
\begin{aligned}
&\left|b^{z,a}(s) - b^{m(z),a}(s)\right| \\
&= \left|\sum_{s'} T_{s|a,s'}\hat{b}^z_{s'} - \sum_{s'} T_{s|a,s'}\pi_{\hat{b}^z,s'}\right| \\
&\leq \sum_{s'} \underbrace{T_{s|a,s'}}_{\leq 1} \underbrace{\left|\hat{b}^z(s) - \pi_{\hat{b}^z,s}\right|}_{\leq \hat{\epsilon}} \leq \hat{\epsilon}|S|.
\end{aligned}
$$

Since $b^{m(z),a} = b^{m,a}$ for each $z \in N(\hat{b}^m)$, (C.12) can be upper bounded as follows:

$$
\begin{aligned}
\text{(C.12)} &= \left|\sum_a \int_Z \mathcal{A}^*_{a|\hat{b}^z}\alpha_b^z V_k(b^{z,a})dz \right.\\
&\qquad \left. - \sum_a \int_Z \mathcal{A}^*_{a|\hat{b}^z}\alpha_b^z V_k(b^{m(z),a})\right| \\
&\leq \sum_a \int_Z \mathcal{A}^*_{a|\hat{b}^z}\alpha_b^z \underbrace{\left|V_k(b^{z,a}) - V_k(b^{m(z),a})\right|}_{\leq \hat{\delta}} dz \\
&\leq \hat{\delta} \sum_a \int_Z \mathcal{A}^*_{a|\hat{b}^z}\alpha_b^z \leq \hat{\delta}, \qquad \text{(C.13)}
\end{aligned}
$$

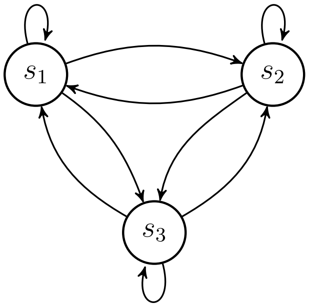
where the first inequality in (C.13) is obtained using (22).

C.4. Summary

Summarizing (i), (ii) and (iii) above, we obtain $|(TV)(b) - (\tilde{T}(V|_{\mathcal{B}}))(b)| \leq \epsilon$ for each $b \in \mathcal{B}$, completing the proof. □

Appendix D. Additional information on three-state example

Consider the MDP in Fig. D.1, where the three actions a_1, a_2, and a_3 have the given transition probabilities. The agent's objective is to avoid s_3, for which we set $C_{s_3,:} = 1$ and 0 otherwise. To construct $\hat{\mathcal{B}}$, we uniformly grid the three-dimensional belief simplex using three different spacings. For the uppermost plot in Fig. 4, a spacing of 0.2 is used, producing a total of 21 posterior belief states and 63 prior belief states. Similarly, for the center plot in Fig. 4, a spacing of 0.1 is used, yielding 62 posterior belief states and 186 prior belief states. Finally, the lowermost plot in Fig. 4 is produced using a spacing of 0.05, resulting in a total of 217 posterior belief states and a corresponding set of 651 prior belief states. Considering an infinite time horizon, the plotted values of prior belief states are obtained via value iteration using values of $\gamma = 0.95$ and $\beta = 1$, wherein we solve (15) for each prior belief and (17) for each posterior belief at each iteration. To solve each linear program, we use the Gurobi linear program solver (Gurobi Optimization, 2022).



$$
T_{:|a_1,:} = \begin{bmatrix} 0.1 & 0.9 & 0.0 \\ 0.0 & 0.1 & 0.9 \\ 0.5 & 0.5 & 0.0 \end{bmatrix}
$$

$$
T_{:|a_2,:} = \begin{bmatrix} 0.1 & 0.0 & 0.9 \\ 0.9 & 0.1 & 0.0 \\ 0.5 & 0.5 & 0.0 \end{bmatrix}
$$

$$
T_{:|a_3,:} = \begin{bmatrix} 0.998 & 0.001 & 0.001 \\ 0.001 & 0.998 & 0.001 \\ 0.001 & 0.001 & 0.998 \end{bmatrix}
$$

Fig. D.1. 3-state environment considered.

References

Aloimonos, Y. (2013). *Active perception*. Psychology Press.

Araya, M., Buffet, O., Thomas, V., & Charpillet, F. (2010). A POMDP extension with belief-dependent rewards. In *Advances in neural information processing systems* (pp. 64–72).

Attwell, D., & Laughlin, S. B. (2001). An energy budget for signaling in the grey matter of the brain. *Journal of Cerebral Blood Flow & Metabolism*, *21*(10), 1133–1145.

Bajcsy, R., Aloimonos, Y., & Tsotsos, J. K. (2018). Revisiting active perception. *Autonomous Robots*, *42*(2), 177–196.

Berry, M. J., Brivanlou, I. H., Jordan, T. A., & Meister, M. (1999). Anticipation of moving stimuli by the retina. *Nature*, *398*(6725), 334–338.

Censi, A., Mueller, E., Frazzoli, E., & Soatto, S. (2015). A power-performance approach to comparing sensor families, with application to comparing neuromorphic to traditional vision sensors. In *2015 IEEE international conference on robotics and automation* ICRA, (pp. 3319–3326). IEEE.

Cover, T. M., & Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons.

Derpich, M. S., & Østergaard, J. (2021). Directed data-processing inequalities for systems with feedback. *Entropy*, *23*(5), 533.

Egner, T., & Hirsch, J. (2005). Cognitive control mechanisms resolve conflict through cortical amplification of task-relevant information. *Nature Neuroscience*, *8*(12), 1784–1790.

Ghasemi, M., & Topcu, U. (2019). Online active perception for partially observable Markov decision processes with limited budget. In *2019 IEEE 58th conference on decision and control* CDC, (pp. 6169–6174). IEEE.

Gonzaga, C. C. (1989). An algorithm for solving linear programming problems in $O(n^3L)$ operations. In *Progress in mathematical programming* (pp. 1–28). Springer.

Gurobi Optimization, LLC (2022). Gurobi optimizer reference manual. URL https://www.gurobi.com.

Hashemi, A., Ghasemi, M., Vikalo, H., & Topcu, U. (2018). A randomized greedy algorithm for near-optimal sensor scheduling in large-scale sensor networks. In *2018 annual american control conference* ACC, (pp. 1027–1032). IEEE.

Hibbard, M., Tuggle, K., & Tanaka, T. (2020). Point-based value iteration and approximately optimal dynamic sensor selection for linear-Gaussian processes. *IEEE Control Systems Letters*, *5*(6), 2192–2197.

Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, *195*(1), 215–243.

Kaelbling, L. P., Littman, M. L., & Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, *101*(1–2), 99–134.

Khamsi, M. A., & Kirk, W. A. (2011). *An introduction to metric spaces and fixed point theory, volume 53*. John Wiley & Sons.

Kostina, V., & Hassibi, B. (2019). Rate-cost tradeoffs in control. *IEEE Transactions on Automatic Control*, *64*(11), 4525–4540.

Lin, S. C., Zhang, Y., Hsu, C. H., Skach, M., Haque, M. E., Tang, L., & Mars, J. (2018). The architectural implications of autonomous driving: Constraints and acceleration. In *Proceedings of the twenty-third international conference on architectural support for programming languages and operating systems* (pp. 751–766).

Massey, J. (1990). Causality, feedback and directed information. In *Proc. 1990 intl. symp. on info. th. and its applications*.

Ortega, P. A., & Braun, D. A. (2013). Thermodynamics as a theory of decision-making with information-processing costs. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *469*(2153), Article 20120683.

Pineau, J., Gordon, G. J., & Thrun, S. (2003). Point-based value iteration: An anytime algorithm for POMDPs. In *IJCAI-03, proceedings of the eighteenth international joint conference on artificial intelligence* (pp. 1025–1032).

Puterman, M. L. (2014). *Markov decision processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons.

Schreiber, T. (2000). Measuring information transfer. *Physical Review Letters*, *85*(2), 461.

Shafieepoorfard, E., Raginsky, M., & Meyn, S. P. (2016). Rationally inattentive control of Markov processes. *SIAM Journal on Control and Optimization*, *54*(2), 987–1016.

Sims, C. A. (2003). Implications of rational inattention. *Journal of Monetary Economics*, *50*(3), 665–690.

Sims, C. R. (2016). Rate–distortion theory and human perception. *Cognition*, *152*, 181–198.

Spaan, M. T. J. (2008). Cooperative active perception using POMDPs. In *AAAI 2008 workshop on advancements in POMDP Solvers*.

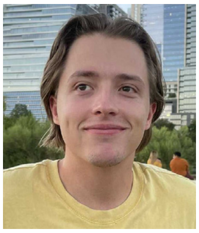
Spaan, M., & Lima, P. (2009). A decision-theoretic approach to dynamic sensor selection in camera networks. *Vol.19*, In *Proceedings of the international conference on automated planning and scheduling* (pp. 297–304).

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An Introduction*. MIT Press.

Tanaka, T., Esfahani, P. M., & Mitter, S. K. (2017). LQG control with minimum directed information: Semidefinite programming approach. *IEEE Transactions on Automatic Control*, *63*(1), 37–52.

Tanaka, T., Sandberg, H., & Skoglund, M. (2021). Transfer-entropy-regularized Markov decision processes. *IEEE Transactions on Automatic Control*.

Tishby, N., & Polani, D. (2011). Information theory of decisions and actions. In *Perception-action cycle* (pp. 601–636). Springer.



Michael Hibbard joined the Department of Aerospace Engineering at the University of Texas at Austin as a Ph.D. student in Fall 2018, where he received his M.S. degree in aerospace engineering in 2020. Prior to UT-Austin, he received his B.S. degree in engineering mechanics and astronautics from the University of Wisconsin-Madison in 2018. His research interests lie in the development of theory and algorithms providing formal guarantees for the mission success of autonomous agents with limited sensing capabilities, especially in the context of outer-space applications.

Takashi Tanaka received the B.S. degree from the University of Tokyo, Tokyo, Japan, in 2006, and the M.S. and Ph.D. degrees in aerospace engineering (automatic control) from the University of Illinois at Urbana-Champaign, Champaign, IL, USA, in 2009 and 2012, respectively. He was a Postdoctoral Associate with the Laboratory for Information and Decision Systems at the Massachusetts Institute of Technology, Cambridge, MA, USA, from 2012 to 2015, and a postdoctoral researcher at KTH Royal Institute of Technology, Stockholm, Sweden, from 2015 to 2017. Since 2017, he has been an Assistant Professor in the Department of Aerospace Engineering and Engineering Mechanics at the University of Texas at Austin. His research interests include control theory and its applications, most recently the information-theoretic perspectives of optimal control problems. He was the recipient of the DARPA Young Faculty Award, the AFOSR Young Investigator Program Award, and the NSF Career Award.

Ufuk Topcu is an Associate Professor in the Department of Aerospace Engineering and Engineering Mechanics at The University of Texas at Austin, where he holds the W. A. “Tex” Moncrief, Jr. Professorship in Computational Engineering and Sciences I. He received his Ph.D. from the University of California at Berkeley in 2008. Ufuk’s research focuses on the theoretical and algorithmic aspects of the design and verification of autonomous systems, typically in the intersection of formal methods, reinforcement learning, and control theory.