Available online at www.sciencedirect.com
ScienceDirect
Artificial Intelligence 171 (2007) 453–490
Artificial Intelligence
www.elsevier.com/locate/artint

Partially observable Markov decision processes with imprecise parameters

Hideaki Itoh *, Kiyohiko Nakamura

Department of Computational Intelligence and Systems Science, Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, 4259-G3-46 Nagatsuta-cho, Midori-ku, Yokohama, Kanagawa 226-8502, Japan

Received 11 May 2005; received in revised form 7 March 2007; accepted 16 March 2007

Available online 24 March 2007

Abstract

This study extends the framework of *partially observable Markov decision processes* (POMDPs) to allow their parameters, i.e., the probability values in the state transition functions and the observation functions, to be imprecisely specified. It is shown that this extension can reduce the computational costs associated with the solution of these problems. First, the new framework, *POMDPs with imprecise parameters* (POMDPIPs), is formulated. We consider (1) the *interval case*, in which each parameter is imprecisely specified by an interval that indicates possible values of the parameter, and (2) the *point-set case*, in which each probability distribution is imprecisely specified by a set of possible distributions. Second, a new optimality criterion for POMDPIPs is introduced. As in POMDPs, the criterion is to regard a policy, i.e., an action-selection rule, as optimal if it maximizes the expected total reward. The expected total reward, however, cannot be calculated precisely in POMDPIPs, because of the parameter imprecision. Instead, we estimate the total reward by adopting arbitrary *second-order beliefs*, i.e., beliefs in the imprecisely specified state transition functions and observation functions. Although there are many possible choices for these second-order beliefs, we regard a policy as optimal as long as there is at least one of such choices with which the policy maximizes the total reward. Thus there can be multiple optimal policies for a POMDPIP. We regard these policies as equally optimal, and aim at obtaining one of them. By appropriately choosing which second-order beliefs to use in estimating the total reward, computational costs incurred in obtaining such an optimal policy can be reduced significantly. We provide an exact solution algorithm for POMDPIPs that does this efficiently. Third, the performance of such an optimal policy, as well as the computational complexity of the algorithm, are analyzed theoretically. Last, empirical studies show that our algorithm quickly obtains satisfactory policies to many POMDPIPs.

PACS: 07.05.Mh; 02.50.Le

Keywords: POMDP; Second-order beliefs; Parameter set; Probability interval

* Corresponding author.
E-mail addresses: hideaki@dis.titech.ac.jp (H. Itoh), nakamura@dis.titech.ac.jp (K. Nakamura).

0004-3702/$ – see front matter
doi:10.1016/j.artint.2007.03.004

1. Introduction

The theory of partially observable Markov decision processes (POMDPs) is normative for sequential decision making under uncertainty [2,15,35,51]. It provides a general framework for designing intelligent agents [8,30], and several real-world applications have been reported (e.g., [29,38]). Let us consider a toy example called the tiger problem:

> Imagine an agent standing in front of two closed doors. Behind one of the doors is a tiger and behind the other is a large reward. If the agent opens the door with the tiger, then a large penalty is received (presumably in the form of some amount of bodily injury). Instead of opening one of the two doors, the agent can listen, in order to gain some information about the location of the tiger. Unfortunately, listening is not free; in addition, it is also not entirely accurate. There is a chance that the agent will hear a tiger behind the left-hand door when the tiger is really behind the right-hand door, and vice versa [30].

What is the best action-selection rule for this agent? By a standard POMDP, this problem can be modeled as follows: The possible *states* of the environment are s_l (meaning that the tiger is behind the left-hand door) and s_r (behind the right-hand door). Assume that the agent's initial *belief* in the states is that s_l and s_r are equally probable (i.e., $\Pr(s_l) = \Pr(s_r) = 0.5$). The agent can choose its action among LEFT (open the left door), RIGHT (open the right door), and LISTEN (listen to the tiger). The action LEFT results in $+10$ *reward* when the state is s_r, but -100 when it is s_l, while these rewards are reversed for the action RIGHT. The action LISTEN costs as much as -1 reward, but the agent obtains a noisy *observation* TL (meaning that the tiger is likely to be behind the left-hand door) or TR (the tiger is likely to be behind the right-hand door). If the state is s_l, TL is observed with probability 0.85 and TR is observed with probability 0.15 (i.e., $\Pr(\mathrm{TL}|s_l) = 0.85$ and $\Pr(\mathrm{TR}|s_l) = 0.15$), while we similarly have $\Pr(\mathrm{TR}|s_r) = 0.85$ and $\Pr(\mathrm{TL}|s_r) = 0.15$.

An action-selection rule for the agent is called a *policy*. A policy is called optimal when it maximizes the expected total reward. In this example, the optimal policy is (1) choose the action LISTEN several times until the agent's belief that the tiger is on the right (or left) side becomes sufficiently strong and then (2) choose the action LEFT (or RIGHT) accordingly. POMDP theory tells us how the optimal policy can be derived.

POMDPs, however, cannot be used when their parameters (e.g., $\Pr(\mathrm{TL}|s_l) = 0.85$) are not specified precisely. The parameters remain imprecise for various reasons [5,13,32] including limited data, insufficient inference time, disagreement among experts [34,49], and model abstraction [12,20,23]. Here we mention four examples. First, suppose that $\Pr(\mathrm{TL}|s_l)$ is estimated by an experiment that examines the frequency with which TL is observed when the state is s_l. Such an estimate is subject to a statistical error. For this case, intervals (e.g., 95% confidence intervals) may be used to express the uncertainty. Second, suppose that a human expert can be consulted to determine the value of $\Pr(\mathrm{TL}|s_l)$. The expert would determine the value by his or her own subjective belief; he or she might say that $\Pr(\mathrm{TL}|s_l)$ would be equal to 0.85. However, he or she might find it hard to explain why $\Pr(\mathrm{TL}|s_l)$ should be precisely 0.85 and not 0.849 for example. Thus in this situation intervals can be used to express the expert's belief more faithfully. Third, suppose that there are multiple experts consulted. Even if each expert could specify a precise value for each parameter, the values might differ from each other. In this case, a set of distributions may be adopted to express the disagreed uncertainty. For example, we have $(\Pr(\mathrm{TL}|s_l), \Pr(\mathrm{TR}|s_l)) \in \{(0.84, 0.16), (0.85, 0.15)\}$ if one expert specified the distribution $(\Pr(\mathrm{TL}|s_l), \Pr(\mathrm{TR}|s_l))$ as $(0.84, 0.16)$, while another expert specified it as $(0.85, 0.15)$. Last, suppose that the tiger problem is an abstracted version of a more complex problem. For example, the probability distribution of observing TL or TR, given s_l, might actually also depend on the temperature of the sonic sensor. Suppose that if the temperature is high (denoted by t_{high}), the distribution is $(\Pr(\mathrm{TL}|s_l, t_{\mathrm{high}}), \Pr(\mathrm{TR}|s_l, t_{\mathrm{high}})) = (0.84, 0.16)$. Similarly, if the temperature is low, the distribution is $(\Pr(\mathrm{TL}|s_l, t_{\mathrm{low}}), \Pr(\mathrm{TR}|s_l, t_{\mathrm{low}})) = (0.85, 0.15)$. The agent, however, may want to neglect such a small dependence on the temperature. In this case, again a set of distributions may be adopted to express the abstracted uncertainty as $(\Pr(\mathrm{TL}|s_l), \Pr(\mathrm{TR}|s_l)) \in \{(0.84, 0.16), (0.85, 0.15)\}$.

Motivated by these examples, in this paper, we introduce *POMDPs with imprecise parameters* (POMDPIPs). We consider two cases. One is the *interval* case, in which each parameter is imprecisely specified by an interval, e.g., $\Pr(\mathrm{TL}|s_l) \in [0.84, 0.86]$. The other is the *point-set* case, in which each distribution is specified by a set of distributions, e.g., $(\Pr(\mathrm{TL}|s_l), \Pr(\mathrm{TR}|s_l)) \in \{(0.84, 0.16), (0.85, 0.15)\}$. In this paper, we will consider the parameter imprecision in the state transition functions (i.e., the probability functions that model how a state is changed by each action) and the

observation functions (i.e., the probability functions that model which observation is obtained). Other imprecisions, such as the imprecision in the reward function, remain to be considered in the future.

Another motivation for the introduction of POMDPIPs arises from the fact that the POMDPs are computationally expensive to solve [36,42]. By *solving* a POMDP, we mean obtaining its optimal policy. Although several algorithms that can solve POMDPs in finite time [11,24,51,57,58] have been developed, within a given non-prohibitive time period, only relatively small-sized POMDPs can be solved by these algorithms. This high computational cost can be due to the fact that the algorithms seek the optimal policy that strictly maximizes the expected reward. In many problems, however, such a strict optimization is meaningless, since the expected reward cannot be precisely evaluated because of the parameter imprecision. For such problems, it may often be sufficient to maximize the expected reward that is roughly estimated by using the imprecise parameters. Such rough optimization may require a lower computational cost.

Thus motivated, in this paper, we will formulate an optimality criterion for POMDPIPs in the following manner.

We will begin by considering a hypothetical situation in which strict optimization can be performed for POMDPIPs. To perform strict optimization, we need more information than POMDPIPs. Let us assume hypothetically that the agent can specify correct *second-order beliefs* (e.g., [19,41]), i.e., the beliefs in the *models*, where we define a model as a pair of the state transition function and the observation function. For instance, take the example of the abstracted uncertainty above, in which the probability distribution of observing TL or TR was either $(\Pr(\mathrm{TL}|s_l, t_{\mathrm{high}}), \Pr(\mathrm{TR}|s_l, t_{\mathrm{high}})) = (0.84, 0.16)$ or $(\Pr(\mathrm{TL}|s_l, t_{\mathrm{low}}), \Pr(\mathrm{TR}|s_l, t_{\mathrm{low}})) = (0.85, 0.15)$, depending on the temperature of the sonic sensor. Suppose, for simplicity, that the other probability distributions in the model (i.e., the state transition function and the observation function) are specified precisely (see Section 3 for the general case). Thus, we have two models, one for the high temperature and the other for the low temperature, which we denote by m_{high} and m_{low}, respectively. Suppose hypothetically that the agent has performed a detailed experiment and found that the temperature is high or low with the same probability 0.5. Then its second-order belief is that m_{high} and m_{low} are equally probable. If such additional beliefs are given, the total reward can be defined exactly, and hence strict optimization can be performed.

Then we will introduce a relaxed optimality criterion. Recall that the second-order beliefs are not given in POMDPIPs. None of the second-order beliefs are considered to be less reliable than the others. Thus, we adopt arbitrary second-order beliefs to estimate the total reward. We consider a policy optimal as long as it maximizes the estimated total reward. We refer to the policies that satisfy this optimality criterion as "quasi-optimal" policies. By "quasi" we mean that the policies are optimized by using the second-order beliefs that are not necessarily correct.

There are two possible approaches for adopting arbitrarily-selected second-order beliefs instead of the correct (but unknown) beliefs. When we estimate the total reward, we use a second-order belief for multiple purposes, i.e., not only to estimate the reward obtained immediately after each action but also to estimate the future rewards after the action and the subsequent observation. One approach is to adopt a *single* arbitrarily-selected second-order belief instead of the correct belief, and use it for all of these purposes. The other approach is to use *multiple* arbitrarily-selected second-order beliefs instead of the correct belief, and use different beliefs for different purposes. Although further research is necessary for detailed comparison, we will argue in Section 3.3 that the latter is at least not always disadvantageous in terms of the performance of the optimal policy obtained and that it is advantageous in terms of computational costs. Thus, we will choose the latter approach in this study.

Next, we will provide an exact algorithm to obtain the quasi-optimal policies. The algorithm exploits the fact that we have allowed the second-order beliefs to be adopted arbitrarily. To avoid confusion with the second-order beliefs, let us refer to the beliefs in the states as the *first-order beliefs*. A first-order belief is said to be *reachable* when it becomes the agent's first-order belief after some actions and observations. In POMDPs, it is usual that a large number of first-order beliefs are reachable, and this makes it hard for us to calculate the optimal policy. Our algorithm reduces the number of reachable first-order beliefs by picking the second-order beliefs with which same first-order beliefs are reached repeatedly. We will show that such second-order beliefs can be picked efficiently by solving linear programming problems.

We will also provide a theoretical bound on the amount of reward loss that can occur by using a quasi-optimal policy, when compared with the optimal policy in the hypothetical situation in which the correct second-order beliefs are given and the strict optimization is conducted. Furthermore, we study empirically the conditions whereby the quasi-optimal policies have satisfactory performance and are easy to obtain by our algorithm.

Our optimality criterion is closely related to E-admissibility [33,34]. If we take the former approach above, i.e., adopt a single arbitrarily-selected belief instead of the correct but unknown belief, our criterion is exactly the same as E-admissibility because the quasi-optimal policy is optimal with regard to the adopted single belief. However, since we select the latter approach, i.e., adopt multiple arbitrarily-selected beliefs, the quasi-optimal policy is not always optimal with regard to a single belief. Note that we do not argue that our criterion is always better than E-admissibility. Detailed comparisons await future research.

In POMDP literature, Cozman et al. [14] has already considered (using the E-admissibility criterion) parameter imprecision with the motivation of reducing computational costs. However, their study was limited to the case where the agent's actions do not affect the state of the environment, and the observation probability functions were limited to Gaussian distributions. The present study considers a more general case.

There are several other studies of decision-making under parameter imprecision with various definitions of optimality [21,34,50,54]. For (fully observable) Markov decision processes (MDPs) with parameter imprecision, several algorithms have been provided to obtain *max-min* policies [20,48,56], *max-max* policies [20,48], all *E-admissible* policies [55], and *maximal* policies [27]. For influence diagrams (IDs) with parameter imprecision, *admissible* policies have been studied [10,17,18]. The ideas in these studies may, in principle, be applied to POMDPs. However, their computational complexity is at least that of solving the corresponding problem whose parameters are precise. Thus none of these approaches offers a means of reducing computational complexity when applied to POMDPs.

Other studies have been devoted to solving large-sized POMDPs. One approach is to develop approximate algorithms [28], e.g., grid-based algorithms [7,28,59], α-vector-based algorithms [44,46,47,52,53], and policy-gradient methods [1]. Another approach is to exploit the structure of each POMDP problem: e.g., factored state representation [9,25] and hierarchies [26]. Their combinations, e.g., the approximate algorithms for factored state representation [16, 37], have also been studied. However, as yet no study has explored the possibility of exploiting parameter imprecision to reduce computational costs.

This paper is organized as follows; Section 2 reviews POMDPs and their two transformations (history-state MDPs and belief-state MDPs), as a background for the following sections. Section 3 formulates POMDPIPs and their quasi-optimal policies. In Section 4 we provide an algorithm to obtain the quasi-optimal policies. The performances of the quasi-optimal policies and the provided algorithm are analyzed theoretically in Section 5 and empirically in Section 6. Section 7 concludes this study.

2. POMDPs

In this section, we review POMDPs and their two transformations (see, e.g., [6,28] for introductory explanations).

2.1. Definition

We assume discrete time steps. We define a POMDP by a tuple $(S, A, \Theta, \{T_h \mid h \in H\}, \{O_h \mid h \in H\}, R, p_0)$ where

- S is a finite set of the states.
- A is a finite set of actions.
- Θ is a finite set of observations.
- $\{T_h \mid h \in H\}$ is a set of state transition functions. $T_h : S \times A \times S \to [0, 1]$ is the state transition function for a given $h \in H$, where H is a set of histories that is defined as follows. Let $s_t \in S$, $a_t \in A$, and $o_t \in \Theta$ denote the state, action, and observation at time $t (= 0, 1, 2, \ldots)$, respectively. Let $h_t := (a_0, o_1, a_1, o_2, \ldots, a_{t-1}, o_t)$ denote the history up to time t.[1] We say that the length of such a history is t. We further introduce $H_t := \{(a_0, o_1, a_1, o_2, \ldots, a_{t-1}, o_t) \mid a_0, a_1, \ldots, a_{t-1} \in A, o_1, o_2, \ldots, o_t \in \Theta\}$, which is the set of all the possible histories whose length is t. Finally, let $H := \bigcup_{t=0}^{\infty} H_t$ be the set of all the histories of any length. For every $h \in H$, $s, s' \in S$ and $a \in A$, let $T_h(s, a, s')$ be the probability that state s' is reached from state s on action a after history h. Note that the state transition function is history-dependent, i.e., T_{h_1} and T_{h_2} may be different functions if h_1 and h_2 are different histories. It holds that $\sum_{s' \in S} T_h(s, a, s') = 1$ for all $h \in H$, $s \in S$ and $a \in A$;

[1] For notational simplicity, we assume that the agent does not observe o_0 at time $t = 0$. The modification of this assumption to include o_0 is straightforward.

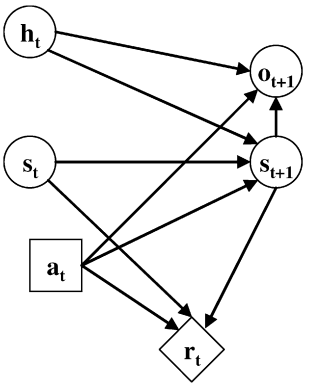

Fig. 1. An influence diagram of a single time step of a POMDP.

- $\{O_h \mid h \in H\}$ is a set of observation functions. $O_h : S \times A \times \Theta \to [0, 1]$ is the observation function for a given $h \in H$. Note that these functions are also history-dependent. $O_h(s', a, o)$ is the probability that observation o is given to the agent when state s' is reached on action a after history h. It holds that $\sum_{o \in \Theta} O_h(s', a, o) = 1$ for all $h \in H, s' \in S$ and $a \in A$;
- $R : S \times A \times S \to \mathbb{R}$ is the reward function, where $R(s, a, s')$ denotes the reward that the agent gains when state s changes into s' on action a, for all $s, s' \in S$ and $a \in A$;
- $p_0 : S \to [0, 1]$ is the *prior* probability function of the initial state $s \in S$. It holds that $\sum_{s \in S} p_0(s) = 1$.

The process proceeds as follows; at time $t = 0$, the process has an initial state $s_0 \in S$ with probability $p_0(s_0)$, and the history is empty (i.e., $h_0 = \emptyset$). At time t $(= 0, 1, 2, \ldots)$, the state and the history is denoted by s_t and $h_t = (a_0, o_1, a_1, o_2, \ldots, a_{t-1}, o_t) \in H_t$, respectively. The agent selects an action $a_t \in A$ by which it changes the current state s_t into the subsequent state s_{t+1} with probability $T_{h_t}(s_t, a_t, s_{t+1})$. The agent then observes $o_{t+1} \in \Theta$ with probability $O_{h_t}(s_{t+1}, a_t, o_{t+1})$, and gains reward $r_t := R(s_t, a_t, s_{t+1})$, whereupon the history is changed into $h_{t+1} = (a_0, o_1, a_1, o_2, \ldots, a_{t-1}, o_t, a_t, o_{t+1}) \in H_{t+1}$. The relations between these variables within a single time step are shown in Fig. 1.

For the agent to select its action a_t at time t, we assume that the POMDP tuple $(S, A, \Theta, \{T_h \mid h \in H\}, \{O_h \mid h \in H\}, R, p_0)$ and the past history $h_t = (a_0, o_1, a_1, o_2, \ldots, a_{t-1}, o_t)$ are available. The action selection rule of the agent, which is a mapping from the available information to an action, is called a policy.

Solving a POMDP is to find the optimal policy that maximizes an objective function. We adopt the 'infinite-horizon discounted sum of the expected rewards' criterion, in which the objective function is defined as

$$E\left\{\sum_{t=1}^{\infty} \gamma^{t-1} r_t\right\}, \tag{1}$$

where $0 \leqslant \gamma < 1$ is a discount factor, and the expectation $E\{\cdot\}$ is taken over all the possible process paths.[2]

2.2. History-state MDP

Solving a POMDP problem is equivalent to solving a (fully observable) Markov decision process (MDP), which is called the history-state MDP.

In history-state MDPs, we take an alternative view of the POMDP process (Fig. 2). At time $t = 0$, it starts with the initial history that is the empty history $\emptyset$. At time t $(= 0, 1, 2, \ldots)$, let h be the current history. The agent selects an action $a \in A$, observes $o \in \Theta$, and gains some reward. The history h is then changed into a new history.

[2] The algorithm that we will provide later can be modified for the finite-horizon cases, in which the objective function is $E\{\sum_{t=1}^{\tilde{T}} r_t\}$ for a given termination time $\tilde{T}$.

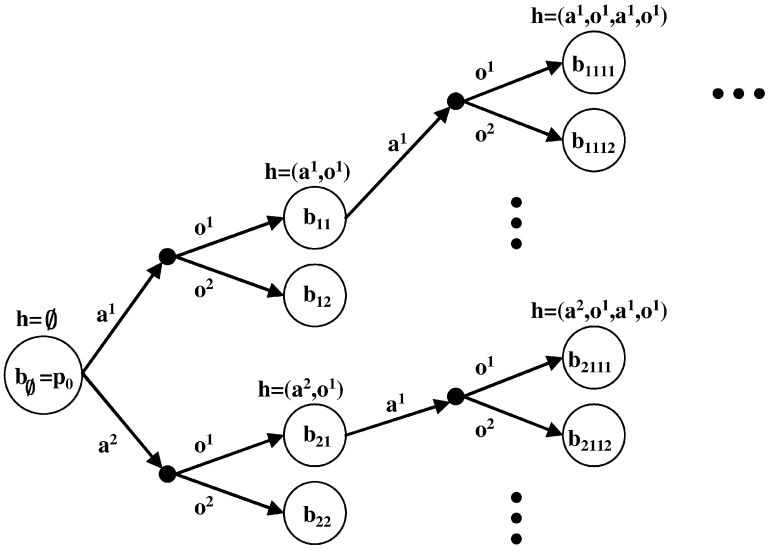

Fig. 2. History tree for a POMDP with two actions (a^1 and a^2) and two observations (o^1 and o^2). Each open circle corresponds to a history, which is treated as a state in the history-state MDP. We associate each history, h, with the corresponding belief b_h. The subscripts of b indicate the history with which the belief is associated; e.g., b_{11} is the belief associated with the history $h = (a^1, o^1)$.

Taking this view, we can consider the Bellman equation to derive the optimal policy. We require some further functions for this purpose. First, let b_h, which is termed the *belief*, be a probability function where $b_h(s)$ is the probability that the current state is s, given that the past history is h. It can be calculated recursively as follow; To begin with, let the initial belief for the empty history $h_0 = \emptyset$ be

$$b_\emptyset = p_0. \tag{2}$$

Note that '=' here means that $b_\emptyset$ and p_0 are identical functions, i.e., '$b_\emptyset(s) = p_0(s)$ for all $s \in S$.' Next, for every $h \in H$, $a \in A$ and $o \in \Theta$, let '$h; \langle a, o\rangle$' denote the history in which a and o have followed h. The belief $b_{h;\langle a,o\rangle}$ is calculated from b_h by Bayes' rule as

$$b_{h;\langle a,o\rangle}(s') = \frac{O_h(s', a, o)\sum_{s\in S} T_h(s, a, s')b_h(s)}{\sum_{s''\in S} O_h(s'', a, o)\sum_{s\in S} T_h(s, a, s'')b_h(s)} \tag{3}$$

for every $s' \in S$.

Second, let $P(o|h, a)$ be the probability that o is observed, given that action a is taken after history h. It can be calculated as

$$P(o|h, a) := \sum_{s'\in S} O_h(s', a, o)\sum_{s\in S} T_h(s, a, s')b_h(s). \tag{4}$$

Last, let $\rho(h, a)$ be the average reward that the agent will gain by taking action a at history h. We have

$$\rho(h, a) := \sum_{s\in S}\sum_{s'\in S} R(s, a, s')T_h(s, a, s')b_h(s). \tag{5}$$

Now, by the principle of optimality [4], we can write the Bellman equation as

$$V^*(h) = \max_{a\in A}\left\{\rho(h, a) + \gamma\sum_{o\in\Theta} P(o|h, a)V^*\big(h; \langle a, o\rangle\big)\right\} \tag{6}$$

for every $h \in H$, where $V^* : H \to \mathbb{R}$ is called the optimal value function. The optimal policy is defined as the mapping $\mu^* : H \to A$ that satisfies

$$\mu^*(h) = \arg\max_{a\in A}\left\{\rho(h, a) + \gamma\sum_{o\in\Theta} P(o|h, a)V^*\big(h; \langle a, o\rangle\big)\right\} \tag{7}$$

for every $h \in H$.

2.3. Belief-state MDP

Solving the history-state MDP in the previous section is equivalent to solving another MDP problem called the belief-state MDP [2] if we restrict the T's and O's as follows.

First, we restrict T's, so that they satisfy $T_{h_1} = T_{h_2}$ whenever $b_{h_1} = b_{h_2}$ holds for any h_1 and $h_2 \in H$. Note again that the '=' symbols here denote equality of functions. In other words, we impose the condition that $T_{h_1}(s, a, s') = T_{h_2}(s, a, s')$ for all $s, s' \in S$ and $a \in A$, whenever the beliefs after h_1 and h_2 are the same (i.e., $b_{h_1}(s) = b_{h_2}(s)$ for all $s \in S$). With this restriction, whenever $b_{h_1} = b_{h_2}$ we treat T_{h_1} and T_{h_2} as identical. Consequently, we re-define the set of T's as $\{T_b \mid b \in B\}$ instead of $\{T_h \mid h \in H\}$, where B is the set of all the possible beliefs, i.e., $B := \{b \mid b = b_h, h \in H\}$. Second, we similarly restrict O's, so that they satisfy $O_{h_1} = O_{h_2}$ whenever $b_{h_1} = b_{h_2}$ holds for any h_1 and $h_2 \in H$. Consequently, we re-define O's from $\{O_h \mid h \in H\}$ to $\{O_b \mid b \in B\}$.

With the T's and O's restricted, the belief $b \in B$ becomes a sufficient statistic. That is, after any history, the belief b (and also the tuple $(S,\ A,\ \Theta,\ \{T_b \mid b \in B\},\ \{O_b \mid b \in B\},\ R,\ p_0)$) summarizes all the information available at that time for the agent to predict what will happen (together with its probability) in the future. Thus, the Bellman equation (Eq. (6)) can be re-written as

$$V^*(b) = \max_{a \in A} \left\{ \rho(b, a) + \gamma \sum_{o \in \Theta} P(o|b, a) V^*\big(\tau(b, a, o)\big) \right\} \tag{8}$$

for every $b \in B$, where $\rho(b, a)$ and $P(o|b, a)$ are defined (re-defined from Eqs. (5) and (4)) as

$$\rho(b, a) := \sum_{s \in S} \sum_{s' \in S} R(s, a, s') T_b(s, a, s') b(s), \qquad P(o|b, a) := \sum_{s' \in S} O_b(s', a, o) \sum_{s \in S} T_b(s, a, s') b(s),$$

and τ, which we call the belief-update function, is defined as

$$\tau(b, a, o)(s') := \frac{O_b(s', a, o) \sum_{s \in S} T_b(s, a, s') b(s)}{\sum_{s'' \in S} O_b(s'', a, o) \sum_{s \in S} T_b(s, a, s'') b(s)} \tag{9}$$

for every $s' \in S$.

The optimal policy can be re-written as the mapping $\mu^* : B \to A$ that satisfies

$$\mu^*(b) = \arg\max_{a \in A} \left\{ \rho(b, a) + \gamma \sum_{o \in \Theta} P(o|b, a) V^*\big(\tau(b, a, o)\big) \right\} \tag{10}$$

for every $b \in B$.

3. POMDPIPs

3.1. Definition

Here, we describe the formulation of POMDPs with imprecise parameters (POMDPIPs). In POMDPIPs, the process proceeds in exactly the same way as POMDP in Section 2.1, except that the agent knows T_h and O_h at each history $h \in H$ only imprecisely. A POMDPIP is defined by a tuple $(S, A, \Theta, T^M, O^M, R, p_0)$, where

- S, A, Θ, R and p_0 are the same as those defined in POMDPs. Furthermore, let H be the set of all the histories.[3]
- T^M is what we call the *model-set function* for the state transition functions $\{T_h | h \in H\}$. We consider two cases (Fig. 3).[4] One is what we call the *interval case*, in which $T^M : S \times A \times S \to I$ indicates the range of the possible values of $T_h(s, a, s')$ for each $s, s' \in S$, $a \in A$, and $h \in H$, where I is the set of all the intervals within $[0, 1]$. For example, suppose that we have $T^M(s, a, s') = [0.8, 0.9]$ for certain s, a, and s'. This means that, for each $h \in H$, we have $T_h(s, a, s') \in [0.8, 0.9]$ for the s, a, and s'. The other case we consider is what we call the *point-set*

[3] Here we consider the imprecision only on T_h and O_h. In case p_0 is imprecise, an equivalent POMDPIP in which only T_h and O_h are imprecise can be constructed by introducing an auxiliary state. Further research is required to handle imprecision on R.

[4] For simplicity we consider these two cases separately. It is straightforward to consider a combination of both of these cases.

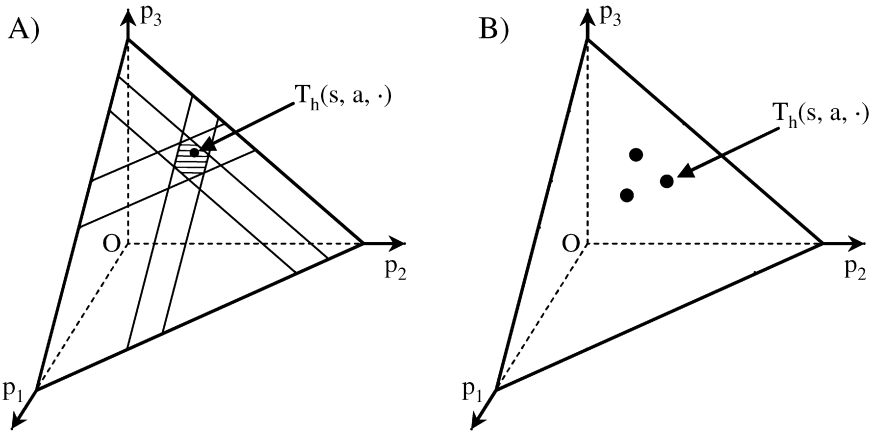

Fig. 3. Example of $T_h(s,a,\cdot)$ in POMDPIPs. The figure shows a probability simplex $\Delta_{|S|}$ for $|S|=3$, in which $T_h(s,a,\cdot)=(p_1,p_2,p_3)$, for certain $h\in H$, $s\in S$, and $a\in A$, is located. (A) the interval case. For each $s'\in S$, $T_h(s,a,s')$ should be within the interval specified by $T^M(s,a,s')$. Consequently, $T_h(s,a,\cdot)$ should lie within a convex area (striped in the figure). (B) the point-set case. $T_h(s,a,\cdot)$ should be one of the probability functions that $T^M(s,a)$ specifies (the three dots in the figure).

case, in which $T^M: S\times A\to \Delta^*_{|S|}$ indicates the probability functions that $T_h(s,a,\cdot)$ can possibly be identical to, where $\Delta^*_{|S|}$ is defined as follows. First, for any $n\geqslant 0$, let Δ_n be the n-dimensional probability simplex, i.e., $\Delta_n=\{(p_1,p_2,\ldots,p_n)\mid p_i\geqslant 0 \text{ for all } i,\ \sum_{i=1}^n p_i=1\}$. Next, let Δ^*_n be the set of all the finite sets of different points in Δ_n. Thus, for each $s\in S$ and $a\in A$, $T^M(s,a)$ is a finite set of probability functions. For example, suppose that there are three states, i.e., $S=\{s^1,s^2,s^3\}$, and that $T^M(s,a)=\{(0.8,0.1,0.1),(0.85,0.1,0.05)\}$ for a certain s and a. This means that, for each $h\in H$, we have $T_h(s,a,\cdot)=(0.8,0.1,0.1)$ or $T_h(s,a,\cdot)=(0.85,0.1,0.05)$, where by the relation $T_h(s,a,\cdot)=(p_1,p_2,\ldots)$ we mean that $T_h(s,a,s^1)=p_1$, $T_h(s,a,s^2)=p_2$, and so on.

- O^M is the model-set function for the observation functions. O^M is defined in the same way as T^M, that is, for the interval case, $O^M: S\times A\times\Theta\to I$ indicates the range of possible values of each parameter. Thus, we have $O_h(s',a,o)\in O^M(s',a,o)$ for every $s'\in S$, $a\in A$, $o\in\Theta$, and $h\in H$. For the point-set case, $O^M: S\times A\to\Delta^*_{|\Theta|}$ indicates the possible probability functions. Thus we have $O_h(s',a,\cdot)\in O^M(s',a)$ for each $s'\in S$, $a\in A$, and $h\in H$.[5]

The information available to the agent for the selection of its action a_t at time t is the tuple $(S, A, \Theta, T^M, O^M, R, p_0)$ and the past history h_t.

Let us define some basic notions for later use. We call a pair of state-transition function T_h and observation function O_h for each $h\in H$ a *model*. Further, let M_0 denote the set of all the possible models that is defined without regard to the model-set functions. That is, we define M_0 as

$$M_0:=\Bigg\{(T,O)\mid T: S\times A\times S\to[0,1], \sum_{s'\in S} T(s,a,s')=1 \text{ for all } s\in S \text{ and } a\in A,\\ O: S\times A\times\Theta\to[0,1], \sum_{o\in\Theta} O(s',a,o)=1 \text{ for all } s'\in S \text{ and } a\in A\Bigg\}.$$

Finally, let M denote the set of all the possible models that is defined with regard to the model-set functions. For the interval case,

$$M:=\big\{(T,O)\mid (T,O)\in M_0, T(s,a,s')\in T^M(s,a,s') \text{ for all } s,s'\in S, \text{ and } a\in A,\\ O(s',a,o)\in O^M(s',a,o) \text{ for all } s'\in S, a\in A, \text{ and } o\in\Theta\big\},$$

[5] Note that T_h and O_h are assumed to be history-dependent. This assumption should be natural for many problems, where the values of the parameters in T_h and O_h may fluctuate due to unknown or neglected dynamics (e.g., the dynamics of the sonic sensor's temperature in Section 1).

i.e., M is the set of all models such that every parameter in its state transition function T and observation function O is within the interval specified by the corresponding model-set function T^M or O^M.

For the point-set case,

$$M := \big\{(T, O) \mid (T, O) \in M_0, T(s, a, \cdot) \in T^M(s, a) \text{ for all } s \in S \text{ and } a \in A, \\ O(s', a, \cdot) \in O^M(s', a) \text{ for all } s' \in S \text{ and } a \in A\big\},$$

i.e., M is the set of all models such that every probability distribution in its state transition function T and observation function O is identical to one of the possible distributions specified by the corresponding model-set function T^M or O^M.

3.2. A "truly" optimal policy

Before we formulate the optimality criterion for POMDPIPs in the next section, let us consider a hypothetical situation where the optimal policies can be defined in a normative way as in POMDPs. This provides a basis for us to formulate a relaxed optimality criterion for POMDPIPs in the next section.

In POMDPIPs, we have assumed that the agent knows only that, for each $h \in H$, the model $m_h (:= (T_h, O_h))$ is a member of M. One way to deal with this uncertainty about the model is to use the *second-order belief*, i.e., the belief in the models. Let us suppose hypothetically that the agent had more information (in any form) to specify its second-order belief by a probability density function $b_h^M : M_0 \to [0, 1]$, where we let $b_h^M(m_h)$ be the probability that, given a history h, the model m_h will govern the process immediately after h.

For instance, let us again consider the example of the abstracted uncertainty in Section 1. Suppose that the model $m_h = (T_h, O_h)$ depends on the temperature of the sonic sensor, and that it is m_{high} and m_{low} for high and low temperatures, respectively. If the agent performs a detailed experiment and finds that the temperature is high or low with the same probability, then the agent sets the second-order belief as $b_h^M(m_h) = 1/2(\delta(m_h - m_{\text{high}}) + \delta(m_h - m_{\text{low}}))$, where δ is Dirac's delta function.

The second-order belief b_h^M should be naturally assumed to satisfy

$$b_h^M(m_h) = 0 \quad \text{for } m_h \notin M, \tag{11}$$

and

$$\int_{m_h \in M} b_h^M(m_h)\, \mathrm{d}m_h = 1 \tag{12}$$

for every $h \in H$.[6] We call Eqs. (11) and (12) the *permissibility condition* for any second-order belief. That is, we say that a function $f : M_0 \to [0, 1]$ satisfies the permissibility condition, if and only if it satisfies $f(m) = 0$ for $m \notin M$ and $\int_{m \in M} f(m)\, \mathrm{d}m = 1$.

Now we can consider a modified version of POMDPs, in which the process proceeds in exactly the same way as the original POMDPs in Section 2.1, except that the model m_h for each $h \in H$ is determined stochastically with probability $b_h^M(m_h)$. We refer to this modified POMDP with given second-order beliefs as a *hypothetical POMDP*.

In hypothetical POMDPs, there are no imprecise parameters. Thus we can define the optimal policy as that which maximizes the discounted sum of the expected rewards (Eq. (1)). We call these optimal policies the *truly optimal* policies.

For later use, let us derive Bellman equations that the truly optimal policy satisfies. Let us consider an equivalent history-state MDP as in Section 2.2 (Fig. 2).

First, the belief in the states, $b_h : S \to [0, 1]$ for each $h \in H$, is calculated recursively as follows. To avoid confusion, we refer to this belief as the *first-order* belief. Let the initial first-order belief be

$$b_\emptyset = p_0. \tag{13}$$

Next, for every $h \in H$, $a \in A$ and $o \in \Theta$, the belief $b_{h;\langle a,o\rangle}$ is calculated from b_h by Bayes' rule as

[6] We implicitly assume that the integral in Eq. (12) and the others throughout this paper exist.

$$b_{h;\langle a,o\rangle}(s') = \frac{\int_{m_h=(T_h,O_h)\in M} O_h(s',a,o)\sum_{s\in S} T_h(s,a,s')b_h(s)b_h^M(m_h)\,\mathrm{d}m_h}{\int_{m_h=(T_h,O_h)\in M}\sum_{s''\in S} O_h(s'',a,o)\sum_{s\in S} T_h(s,a,s'')b_h(s)b_h^M(m_h)\,\mathrm{d}m_h} =: \tau\big(b_h,a,o,b_h^M\big)(s'), \tag{14}$$

for all $s'\in S$, where we re-defined the belief-update function τ of Eq. (9).

Then, by the principle of optimality [4], the truly optimal policy can be defined as the mapping $\mu^*: H\to A$ that satisfies the Bellman equations (exactly the same as Eqs. (6) and (7))

$$V^*(h) = \max_{a\in A}\left\{\rho(h,a) + \gamma\sum_{o\in\Theta} P(o|h,a)V^*\big(h;\langle a,o\rangle\big)\right\}, \tag{15}$$

$$\mu^*(h) = \arg\max_{a\in A}\left\{\rho(h,a) + \gamma\sum_{o\in\Theta} P(o|h,a)V^*\big(h;\langle a,o\rangle\big)\right\}, \tag{16}$$

for each $h\in H$, where we re-define $\rho(h,a)$ and $P(o|h,a)$ (from Eqs. (5) and (4), respectively) as

$$\rho(h,a) := \int_{m_h=(T_h,O_h)\in M}\sum_{s\in S}\sum_{s'\in S} R(s,a,s')T_h(s,a,s')b_h(s)b_h^M(m_h)\,\mathrm{d}m_h, \tag{17}$$

$$P(o|h,a) := \int_{m_h=(T_h,O_h)\in M}\sum_{s'\in S} O_h(s',a,o)\sum_{s\in S} T_h(s,a,s')b_h(s)b_h^M(m_h)\,\mathrm{d}m_h. \tag{18}$$

Below are some notes regarding the second-order beliefs introduced in this section.

First, although we have introduced the *beliefs in the models* (i.e., the beliefs in m_h's), there are other types of beliefs that could be considered instead. For example, we could introduce the *beliefs in the conditional distributions*, i.e., the beliefs in $T_h(s,a,\cdot)$'s, for each $s\in S$, $a\in A$, and $h\in H$, and the beliefs in $O_h(s',a,\cdot)$'s, for each $s'\in S$, $a\in A$, and $h\in H$. Another example is the *beliefs in the sets of all the models*, i.e., the beliefs in $\{m_h|h\in H\}$'s. Comparisons among them remain to be studied in the future (but see the following paragraph).

Second, we will actually use the second-order beliefs in the conditional distributions when we provide a solution algorithm in Section 4. A belief in the conditional distributions is equivalent to a belief in the models if the latter is decomposable into a product of the beliefs in the conditional distributions (see Section 4.2). We will use only these decomposable beliefs when we provide a solution algorithm in Section 4. Thus, we actually consider the beliefs in the conditional distributions as far as the solution algorithm is concerned. However, we began with the non-decomposed second-order beliefs in this section because they are a wider class of the beliefs that could be utilized in a wider range of problems, and the theoretical result in Section 5.1 is applicable to these beliefs.

Last, the parameter imprecision handled with the second-order beliefs in this paper can be handled equivalently with the convex hulls of possible probability measures (or the credal sets) [13,33]. Obviously, a second-order belief $b_h^M(m_h)$ that satisfies the permissibility condition (Eqs. (11) and (12)) can be referred to also as a member of the convex hull of the possible measures over the models, i.e., the convex hull of $f_i(m_h)=\delta(m_h-m_h^i)$, $i=1,2,\ldots,$ where $f_i: M_0\to[0,1]$, each m_h^i is a member of M, and δ is Dirac's delta function. Similarly, as will become obvious in Section 4.2, using a second-order belief in the conditional distributions is equivalent to using a convex hull of possible conditional distributions.

3.3. Quasi-optimal policies

In the previous section, we defined the truly optimal policies with a strict optimal criterion, assuming that the agent could specify the second-order beliefs. In POMDPIPs, however, the agent has no precise idea how to specify the second-order beliefs. In this section, we formulate a relaxed optimality criterion, allowing the agent to employ arbitrary functions for the second-order beliefs.

First, recall that, to define the truly optimal policy, the second-order belief b_h^M has to be determined in order to Bayes-update the first-order belief by Eq. (14) after every $h\in H$, $a\in A$, and $o\in\Theta$. Since b_h^M is unknown in POMDPIPs, we allow the agent to employ any function that satisfies the permissibility condition (Section 3.2) for this purpose. Let $\hat{b}_{h,a,o}^M$ denote the employed function.

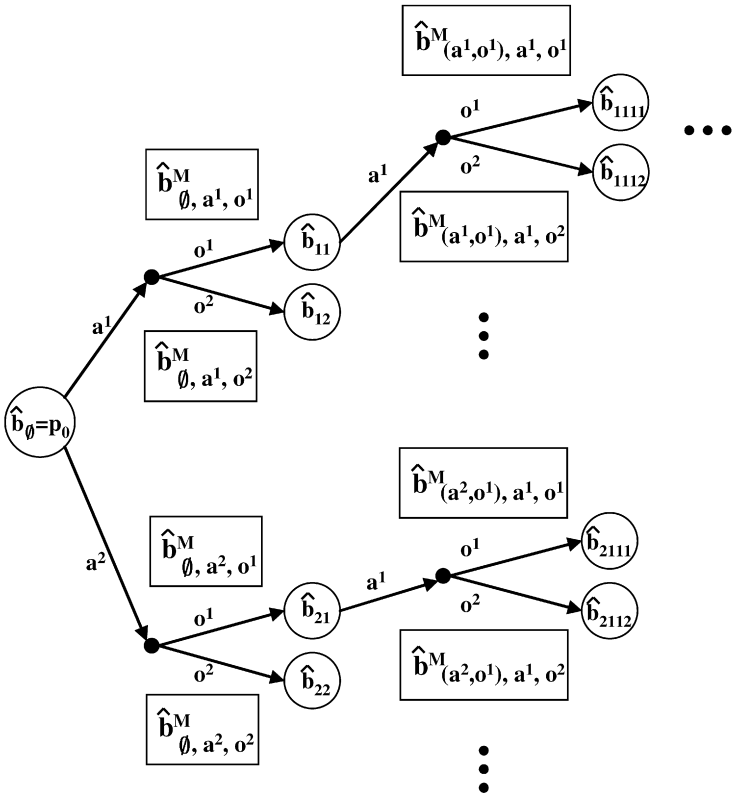

Fig. 4. History tree for a POMDPIP with two actions and two observations. The possibly-correct first-order beliefs are calculated recursively using the possibly-correct second-order beliefs; $\hat{b}_{11} = \tau(\hat{b}_\emptyset, a^1, o^1, \hat{b}^M_{\emptyset,a^1,o^1})$, $\hat{b}_{12} = \tau(\hat{b}_\emptyset, a^1, o^2, \hat{b}^M_{\emptyset,a^1,o^2})$, and so on.

In the following, we call b^M_h in Section 3.2 the "correct" second-order belief. By "correct" we stress that b^M_h has been defined as the belief that the agent would employ if sufficient information is available. On the other hand, we call $\hat{b}^M_{h,a,o}$ the "possibly-correct" second-order belief. By "possibly correct" we mean that it could possibly be identical to the correct belief.

Let $\hat{b}_h$ denote the first-order beliefs that are calculated by these possibly-correct second-order beliefs. That is, $\hat{b}_h$ for each $h \in H$ is calculated recursively (Fig. 4) as

$$\hat{b}_\emptyset = p_0 \tag{19}$$

and

$$\hat{b}_{h;\langle a,o\rangle} = \tau\big(\hat{b}_h, a, o, \hat{b}^M_{h,a,o}\big) \tag{20}$$

for each $h \in H$, $a \in A$, and $o \in \Theta$. We call b_h in Section 3.2 the *correct* first-order beliefs. We call $\hat{b}_h$ the *possibly-correct* first-order beliefs.

Second, recall that, for each $h \in H$, the second-order belief b^M_h is also used to estimate the expected one-step reward, $\rho(h,a)$, and the probability of the next observation, $P(o|h,a)$, in Eqs. (17) and (18), respectively. Again, we allow the agent to employ any function that satisfies the permissibility condition for these purposes. Let $\hat{b}^M_h$ denote the employed function, which we shall also label as the possibly-correct second-order beliefs.

Last, we define the quasi-optimal policy as the solution of Eqs. (13)–(18) in which b^M_h is replaced with arbitrary, but possibly-correct, functions $\hat{b}^M_{h,a,o}$ and $\hat{b}^M_h$. That is, the quasi-optimal policy is the policy $\hat{\mu}^* : H \to A$ that satisfies

$$\hat{V}^*(h) = \max_{a\in A}\left\{\hat{\rho}(h,a) + \gamma \sum_{o\in\Theta} \hat{P}(o|h,a)\hat{V}^*\big(h;\langle a,o\rangle\big)\right\}, \tag{21}$$

$$\hat{\mu}^*(h) = \arg\max_{a\in A}\left\{\hat{\rho}(h,a) + \gamma \sum_{o\in\Theta} \hat{P}(o|h,a)\hat{V}^*\big(h;\langle a,o\rangle\big)\right\}, \tag{22}$$

for every $h \in H$, where

$$\hat{\rho}(h,a) := \int_{m_h=(T_h,O_h)\in M} \sum_{s\in S}\sum_{s'\in S} R(s,a,s')T_h(s,a,s')\hat{b}_h(s)\hat{b}_h^M(m_h)\,\mathrm{d}m_h, \tag{23}$$

$$\hat{P}(o|h,a) := \int_{m_h=(T_h,O_h)\in M} \sum_{s'\in S} O_h(s',a,o)\sum_{s\in S} T_h(s,a,s')\hat{b}_h(s)\hat{b}_h^M(m_h)\,\mathrm{d}m_h, \tag{24}$$

in which $\hat{b}_h$ is defined by Eqs. (19) and (20). The hat mark ($\hat{}$) on V^*, μ^*, $\rho(h,a)$, and $P(o|h,a)$, has been introduced in order to stress that these functions are not calculated with the "correct" (but unknown) second-order beliefs.

In summary, a quasi-optimal policy is the policy that satisfies the same Bellman equations as the hypothetical POMDPs, except that the second-order beliefs are replaced with arbitrary, but possibly-correct, ones. We regard quasi-optimal policies as the solution of POMDPIPs. We now proceed to make some additional comments regarding these policies.

We first note that, for each $h \in H$, we have replaced a correct second-order belief b_h^M with *multiple* possibly-correct second-order beliefs ($\hat{b}_{h,a,o}^M$ for each $a \in A$ and $o \in \Theta$ and $\hat{b}_h^M$). We could have instead used a single possibly-correct second-order belief. Doing this may sometimes have its own merit; if we use a single belief, then the quasi-optimal policy is guaranteed to be the truly optimal policy for at least one possible hypothetical POMDP, and hence can be termed as E-admissible [33,34].

However, in this paper, we study the use of multiple beliefs for the following reasons. First, at least for some POMDPIPs, using multiple beliefs can lead to a more robust policy than using a single belief; it can be risky to rely on a single belief. We provide a simple example in Appendix A. Second, since by using multiple beliefs we have a higher degree of freedom, the quasi-optimal policy should be easier to obtain. This is a useful property because one of the motivations in this paper is to solve the problems with low computational costs.

Note that we do not argue that the use of multiple beliefs is always better than the use of a single belief. Further research is necessary for detailed comparisons (however, see also Section 6.1 for an empirical study). Note also that the algorithm in the next section can be modified for the use of a single belief, although the modification will increase the computational costs.

We further note that, in our definition, the possibly-correct first-order belief $\hat{b}_h$ for each history $h \in H$ is a single probability function. Another possibility is to use a set of probability functions (e.g., [13,18]). This possibility may be worth investigating in the future. The manner of changing the set of probability functions, given a new action and observation, is currently a topic of debate [3,22]; there can be multiple future research directions. In this paper, we do not use a set of probability functions. We regard the quasi-optimal policy as an optimal policy (in a broader sense) for an agent who is allowed to use only a single probability function for expressing its belief in the states.

Note also that although we have introduced the second-order beliefs, we do not use them to make any inference about the models. In a future study, it may be interesting to use the second-order beliefs to make some inference about the uncertain models, e.g., Bayesian inference to identify the true model from the action-observation history. Although strict inference may be impossible because of prohibitive computational costs, we would be able to focus on identifying the true model to the extent that the remaining uncertainty does not significantly affect the total reward. Making such an inference in POMDPIPs can be an interesting future research topic.

We finally note that, since $\hat{b}_{h,a,o}^M$ and $\hat{b}_h^M$ are arbitrary, there can be multiple quasi-optimal policies for a single POMDPIP. We regard any of those policies as the solution of the POMDPIP. In the next section, we will provide an efficient algorithm to obtain one of these quasi-optimal policies.

4. Algorithm for solving POMDPIPs

4.1. Determining $\hat{b}_{h,a,o}^M$

The first step in our algorithm is to determine $\hat{b}_{h,a,o}^M$ for each $h \in H$, $a \in A$, and $o \in \Theta$ under the permissibility condition. Recall that if we determine all the $\hat{b}_{h,a,o}^M$'s, then all the possibly-correct first-order beliefs ($\hat{b}_h$ for every

$h \in H$) are given by Eqs. (19) and (20). Let $\hat{B}$ be the set of all the different first-order beliefs so derived; i.e., $\hat{B} := \{\hat{b} \mid \hat{b} = \hat{b}_h, h \in H\}$. We try to keep $|\hat{B}|$, i.e., the number of the different first-order beliefs, as small as possible in order to reduce computational costs. This is achieved using the following procedure, which we term the FIND-A-SMALL-BELIEF-SET procedure.

We determine these $\hat{b}^M_{h,a,o}$'s in a breadth-first manner, and calculate the $\hat{b}_h$'s whenever it becomes possible; that is, in Fig. 4 for example, we first calculate $\hat{b}_\emptyset$ (i.e., set it equal to p_0), then determine $\hat{b}^M_{\emptyset,a^1,o^1}$, calculate $\hat{b}_{11}$, determine $\hat{b}^M_{\emptyset,a^1,o^2}$, calculate $\hat{b}_{12}, \ldots$, determine $\hat{b}^M_{(a^1,o^1),a^1,o^1}$, calculate $\hat{b}_{1111}, \ldots$, and so forth. In determining each $\hat{b}^M_{h,a,o}$, we try to make $\hat{b}_{h;\langle a,o\rangle}$ $(= \tau(\hat{b}_h, a, o, \hat{b}^M_{h,a,o}))$ identical to one of the other first-order beliefs that have already been calculated. For example, in Fig. 4, suppose that we have already determined $\hat{b}^M_{\emptyset,a^1,o^1}$ and $\hat{b}^M_{\emptyset,a^1,o^2}$, and that we have calculated $\hat{b}_\emptyset$, $\hat{b}_{11}$, and $\hat{b}_{12}$. Now, we search for the $\hat{b}^M_{\emptyset,a^2,o^1}$ for which $\hat{b}_{21}$ $(= \tau(\hat{b}_\emptyset, a^2, o^1, \hat{b}^M_{\emptyset,a^2,o^1}))$ is identical to one of $\hat{b}_\emptyset$, $\hat{b}_{11}$, or $\hat{b}_{12}$. For this search, we use the IS-FEASIBLE procedure in Section 4.2. If such a $\hat{b}^M_{\emptyset,a^2,o^1}$ is found, it is adopted, otherwise we employ an arbitrary $\hat{b}^M_{\emptyset,a^2,o^1}$ under the permissibility condition.

If it is possible to make one first-order belief identical to another, we will then also make the descendant first- and second-order beliefs identical. For example, suppose that $\hat{b}_{21}$ was made identical to $\hat{b}_{11}$. Since we determine $\hat{b}^M_{(a^2,o^1),a^1,o^1}$, $\hat{b}^M_{(a^2,o^1),a^1,o^2}, \ldots$, to be the same as $\hat{b}^M_{(a^1,o^1),a^1,o^1}$, $\hat{b}^M_{(a^1,o^1),a^1,o^2}, \ldots$, respectively, we have $\hat{b}_{2111} = \hat{b}_{1111}$, $\hat{b}_{2112} = \hat{b}_{1112}$, and so on. By following this rule, we can skip the determination and calculation of the beliefs following $\hat{b}_{21}$; we need only consider the beliefs following $\hat{b}_{11}$.

We continue determining and calculating the beliefs, skipping them whenever possible. Eventually, we see that we can skip all the remaining ones (proof in Section 5.2). This means that all the $\hat{b}^M_{h,a,o}$'s and $\hat{b}_h$'s have been determined and calculated, respectively. We may then proceed to the next step described in Section 4.3.

4.2. IS-FEASIBLE procedure

The IS-FEASIBLE procedure searches, under the permissibility condition, for a function $\hat{b}^M : M_0 \to [0, 1]$ for which $\tau(\hat{b}, a, o, \hat{b}^M) = \hat{b}'$ holds, given two first-order beliefs $\hat{b}$ and $\hat{b}'$, an action a, an observation o, and a part of the POMDPIP tuple (S, A, Θ, T^M, O^M). The search for the desired second-order belief in the FIND-A-SMALL-BELIEF-SET procedure (in Section 4.1) can be performed by using this procedure several times. Each time the IS-FEASIBLE procedure is used, we let $\hat{b}'$ be one of the other first-order beliefs that have already been calculated, and we let $\hat{b}$ be $\hat{b}_h$. For example, in Fig. 4, suppose again that we have already determined $\hat{b}^M_{\emptyset,a^1,o^1}$ and $\hat{b}^M_{\emptyset,a^1,o^2}$, and that we have calculated $\hat{b}_\emptyset$, $\hat{b}_{11}$, and $\hat{b}_{12}$. We now have to search for the $\hat{b}^M_{\emptyset,a^2,o^1}$ for which $\tau(\hat{b}_\emptyset, a^2, o^1, \hat{b}^M_{\emptyset,a^2,o^1})$ is identical to $\hat{b}_\emptyset$, $\hat{b}_{11}$, or $\hat{b}_{12}$. This search can be performed by using the IS-FEASIBLE procedure three times with $\hat{b}' = \hat{b}_\emptyset$, $\hat{b}_{11}$, or $\hat{b}_{12}$ and with $\hat{b} = \hat{b}_\emptyset$.

We begin with replacing the required task with an easier one. Recall that our task is to search, under the permissibility condition, for $\hat{b}^M$ for which

$$\hat{b}'(s') = \frac{\int_{m=(T,O)\in M} O(s', a, o) \sum_{s\in S} T(s, a, s')\hat{b}(s)\hat{b}^M(m)\,\mathrm{d}m}{\int_{m=(T,O)\in M} \sum_{s''\in S} O(s'', a, o) \sum_{s\in S} T(s, a, s'')\hat{b}(s)\hat{b}^M(m)\,\mathrm{d}m} \quad \text{for every } s' \in S \tag{25}$$

holds, where $\hat{b}$, $\hat{b}'$, a, and o are given. We restrict our search to those $\hat{b}^M$ that can be decomposed according to

$$\hat{b}^M(m) = \prod_{s\in S,\ \tilde{a}\in A} F_{s,\tilde{a}}\big(T(s, \tilde{a}, \cdot)\big) \prod_{s'\in S,\ \tilde{a}\in A} G_{s',\tilde{a}}\big(O(s', \tilde{a}, \cdot)\big), \tag{26}$$

where $m = (T, O) \in M_0$, and $F_{s,\tilde{a}}$ (for every $s \in S$ and $\tilde{a} \in A$) and $G_{s',\tilde{a}}$ (for every $s' \in S$ and $\tilde{a} \in A$) are probability density functions. Here we denote an action by $\tilde{a}$ to distinguish it from the action, a, that is a constant (specified by the FIND-A-SMALL-BELIEF-SET procedure) in this procedure. Note that such decomposable second-order beliefs always exist. Note also that, by this restriction of the search, we might fail to find a desired $\hat{b}^M$ even when it exists. Hence the number of resultant first-order beliefs, $|\hat{B}|$, might increase. However, the restricted search can be performed quickly, and the total computational time may be reduced. Note that we can still find a quasi-optimal policy despite the restriction.

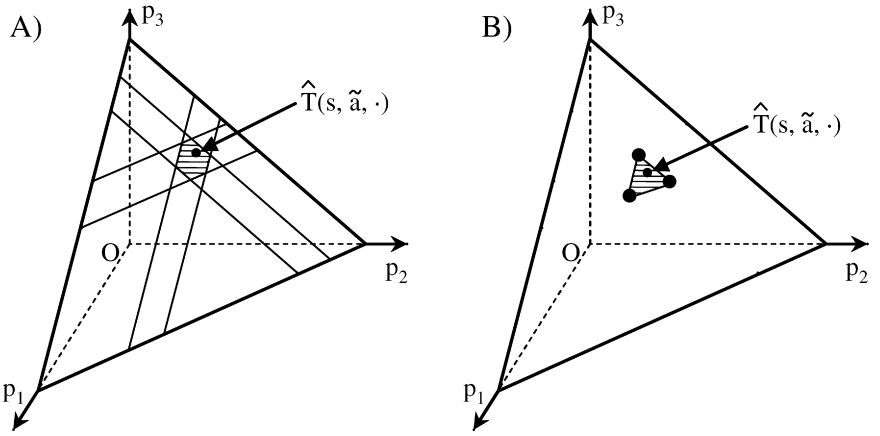

Fig. 5. Example of the averaged model function $\hat{T}(s,\tilde{a},\cdot)$. The striped areas indicate where $\hat{T}(s,\tilde{a},\cdot)$ can be located. Compare this figure with Fig. 3, letting $a=\tilde{a}$. (A) the interval case. $\hat{T}(s,\tilde{a},s')$ should be in the same convex area that constrained $T_h(s,\tilde{a},s')$ in Fig. 3. (B) the point-set case. $\hat{T}(s,\tilde{a},\cdot)$ should be a convex combination of the possible probability functions that $T_h(s,\tilde{a},\cdot)$ in Fig. 3 could be identical to.

Let us define the *averaged model functions* as

$$\hat{T}(s,\tilde{a},s') := \int\limits_{T(s,\tilde{a},\cdot)\in\Delta_{|S|}} T(s,\tilde{a},s')F_{s,\tilde{a}}\big(T(s,\tilde{a},\cdot)\big)\,\mathrm{d}T(s,\tilde{a},\cdot) \tag{27}$$

for each $s,s'\in S$, and $\tilde{a}\in A$, and

$$\hat{O}(s',\tilde{a},\tilde{o}) := \int\limits_{O(s',\tilde{a},\cdot)\in\Delta_{|\Theta|}} O(s',\tilde{a},\tilde{o})G_{s',\tilde{a}}\big(O(s',\tilde{a},\cdot)\big)\,\mathrm{d}O(s',\tilde{a},\cdot) \tag{28}$$

for each $s'\in S,\tilde{a}\in A$, and $\tilde{o}\in\Theta$. Again, we denote an observation by $\tilde{o}$ to distinguish it from the specified observation o. From Eqs. (26)–(28), we can re-write Eq. (25) as

$$\hat{b}'(s') = \frac{\hat{O}(s',a,o)\sum_{s\in S}\hat{T}(s,a,s')\hat{b}(s)}{\sum_{s''\in S}\hat{O}(s'',a,o)\sum_{s\in S}\hat{T}(s,a,s'')\hat{b}(s)} \quad \text{for every } s'\in S. \tag{29}$$

From the permissibility condition together with Eqs. (26) and (27), we have that, for each $s\in S$ and $\tilde{a}\in A$, $\hat{T}(s,\tilde{a},\cdot)$ should be a convex combination of the possible probability functions that the model-set function T^M specifies. In the interval case (Fig. 3A), the possible probability functions are those within a convex area that are constrained by the intervals for each parameter. Thus, their convex combination, $\hat{T}(s,\tilde{a},\cdot)$, is also constrained by the same intervals (Fig. 5A). In the point-set case (Fig. 3B), the possible probability functions are those that are directly specified by the model-set functions. Thus $\hat{T}(s,\tilde{a},\cdot)$ is a convex combination of these probability functions (Fig. 5B). We call these conditions that $\hat{T}(s,\tilde{a},\cdot)$ for each $s\in S$ and $\tilde{a}\in A$ should satisfy the *convexity conditions* for $\hat{T}$. Similarly we have conditions for $\hat{O}(s',\tilde{a},\cdot)$ for each $s'\in S$ and $\tilde{a}\in A$, which we call the *convexity conditions* for $\hat{O}$.

In summary, the search for a $\hat{b}^M$ that satisfies both Eq. (25) and the permissibility condition has been replaced with the search for a pair of the averaged model functions, $(\hat{T},\hat{O})$, that satisfies Eq. (29) and the convexity conditions for $\hat{T}$ and for $\hat{O}$.

Whether such a pair of the averaged model functions exists is easily determined by the feasibility test of Fig. 6. A proof of the validity of the test is detailed in Appendix B. Henceforth, we shall label the implementation of this test the IS-FEASIBLE procedure. The averaged model functions $(\hat{T},\hat{O})$, if they exist, can be derived from the solution of this test.[7]

In the procedure and thereafter, we let $\underline{T}$, $\overline{T}$, $\underline{O}$, and $\overline{O}$ denote the lower and upper bounds of each parameter; in the interval case, the lower bounds are defined by

[7] Let $\hat{O}(s',a,o)=Z/q(s')$ for every $s'\in S$. Employ arbitrary values (under the convexity conditions) for the undetermined parameters.

(A)
Procedure IS-FEASIBLE (for the interval case)
input: $(\hat{b}, a, o, \hat{b}', S, A, \Theta, T^M, O^M)$
output: True or False
Test if there is a solution that satisfies the following constraints:
variables:
$\hat{T}(s, a, s') \in \mathbb{R}$ for all $s, s' \in S$,
$q(s') \in \mathbb{R}$ for all $s' \in S$,
$Z \in \mathbb{R}$,
constraints:
$\underline{T}(s, a, s') \leqslant \hat{T}(s, a, s') \leqslant \overline{T}(s, a, s')$ for all $s, s' \in S$,
$\sum_{s' \in S} \hat{T}(s, a, s') = 1$ for all $s \in S$,
$\sum_{s \in S} \hat{b}(s)\hat{T}(s, a, s') = \hat{b}'(s')q(s')$ for all $s' \in S$,
$\frac{Z}{\overline{O}(s', a, o)} \leqslant q(s') \leqslant \frac{Z}{\underline{O}(s', a, o)}$ for all $s' \in S$,
$Z \geqslant 0$.
Return True if a solution is found. Return False otherwise.
end procedure

(B)
Procedure IS-FEASIBLE (for the point-set case)
input: $(\hat{b}, a, o, \hat{b}', S, A, \Theta, T^M, O^M)$
output: True or False
Test if there is a solution that satisfies the following constraints:
variables:
$\hat{T}(s, a, s') \in \mathbb{R}$ for all $s, s' \in S$,
$\lambda_s^i \in [0, 1]$ for all $s \in S$ and $i = 1, \ldots, |T^M(s, a)|$
$q(s') \in \mathbb{R}$ for all $s' \in S$,
$Z \in \mathbb{R}$,
constraints:
$\hat{T}(s, a, s') = \sum_i \lambda_s^i T_i^M(s, a, s')$ for all $s, s' \in S$,
$\sum_i \lambda_s^i = 1$ for all $s \in S$,
$\sum_{s \in S} \hat{b}(s)\hat{T}(s, a, s') = \hat{b}'(s')q(s')$ for all $s' \in S$,
$\frac{Z}{\overline{O}(s', a, o)} \leqslant q(s') \leqslant \frac{Z}{\underline{O}(s', a, o)}$ for all $s' \in S$,
$Z \geqslant 0$.
Return True if a solution is found. Return False otherwise.
end procedure

Fig. 6. Sub-routine that checks if the averaged model functions $(\hat{T}, \hat{O})$ that change $\hat{b}$ into $\hat{b}'$ after $a \in A$ and $o \in \Theta$ can exist. (A) is for the interval case; (B) is for the point-set case. In (B), we define $|T^M(s, a)|$ as the number of possible probability functions that $T^M(s, a)$ specifies. Each probability function is indexed as $T_i^M(s, a)$, where $i = 1, 2, \ldots, |T^M(s, a)|$.

$$\underline{T}(s, \tilde{a}, s') := \min T^M(s, \tilde{a}, s'), \tag{30}$$

$$\underline{O}(s', \tilde{a}, \tilde{o}) := \min O^M(s', \tilde{a}, \tilde{o}), \tag{31}$$

for all $s, s' \in S$, $\tilde{a} \in A$, and $\tilde{o} \in \Theta$. In the point-set case, they are defined by

$$\underline{T}(s, \tilde{a}, s') := \min_{T(s, \tilde{a}, \cdot) \in T^M(s, \tilde{a})} T(s, \tilde{a}, s'), \tag{32}$$

$$\underline{O}(s', \tilde{a}, \tilde{o}) := \min_{O(s', \tilde{a}, \cdot) \in O^M(s', \tilde{a})} O(s', \tilde{a}, \tilde{o}), \tag{33}$$

for all $s, s' \in S$, $\tilde{a} \in A$, and $\tilde{o} \in \Theta$. The upper bounds are defined by substituting max for min in the above definitions.

The feasibility test in Fig. 6 can be conducted efficiently; all the constraints are linear, and, in the interval case, there are $|S|(|S| + 1) + 1$ variables and $2|S|(|S| + 2) + 1$ (in)equalities, which are almost minimum in comparison

with the number of the parameters and the bounds within the problem at hand.[8] It is also efficient for the point-set case. The feasibility test can be solved by sub-procedures of the linear programming routines that are implemented in many programming languages. The worst-case complexity is a polynomial order of $|S|$ [31]. See also Section 5.2 for the computational complexity of the overall algorithm.

For later use, we define some notations here. Recall that we use this procedure when we determine $\hat{b}^M_{h,a,o}$ for each $h \in H$, $a \in A$, and $o \in \Theta$ in the FIND-A-SMALL-BELIEF-SET procedure. We denote the desired averaged model functions, if found, by $\hat{T}_{h,a,o}$ and $\hat{O}_{h,a,o}$. If they are not found, we employ arbitrary $\hat{T}$ and $\hat{O}$ that satisfy the convexity conditions, and denote them by $\hat{T}_{h,a,o}$ and $\hat{O}_{h,a,o}$, also. Thus, $\hat{T}_{h,a,o}$ and $\hat{O}_{h,a,o}$ always indicate (implicitly) the $\hat{b}^M_{h,a,o}$ to be employed.

4.3. Determining $\hat{b}^M_h$

The second step in our algorithm is to determine, for each $h \in H$, the $\hat{b}^M_h$ that satisfies the permissibility condition. Although there are several possible ways to do this, here we determine them as

$$\hat{b}^M_h(m_h) = \frac{1}{|A\|\Theta|} \sum_{a\in A, o\in\Theta} \hat{b}^M_{h,a,o}(m_h) \quad \text{for all } m_h \in M_0, \tag{34}$$

for every $h \in H$. That is, we take the average of the $\hat{b}_{h,a,o}$'s. Clearly, if the $\hat{b}^M_{h,a,o}$'s satisfy the permissibility condition, then so will the $\hat{b}^M_h$'s.

Recall that, in the previous section, we have replaced the determination of $\hat{b}^M_{h,a,o}$ with the determination of $\hat{T}_{h,a,o}$ and $\hat{O}_{h,a,o}$; we have determined $\hat{b}^M_{h,a,o}$ only implicitly. Thus we cannot determine $\hat{b}^M_h$ explicitly from Eq. (34). Recall, however, that we need $\hat{b}^M_h$ only for the evaluation of $\hat{\rho}(h,a)$ and $\hat{P}(o|h,a)$ from Eqs. (23) and (24). Substituting $\hat{b}^M_h$ in Eqs. (23) and (24) with the right-hand side of Eq. (34), and using Eqs. (26)–(28), we have

$$\hat{\rho}(h,a) = \frac{1}{|A\|\Theta|} \sum_{\tilde{a}\in A,\ \tilde{o}\in\Theta} \sum_{s\in S}\sum_{s'\in S} R(s,a,s')\hat{T}_{h,\tilde{a},\tilde{o}}(s,a,s')\hat{b}_h(s), \tag{35}$$

$$\hat{P}(o|h,a) = \frac{1}{|A\|\Theta|} \sum_{\tilde{a}\in A,\ \tilde{o}\in\Theta} \sum_{s'\in S} \hat{O}_{h,\tilde{a},\tilde{o}}(s',a,o) \sum_{s\in S} \hat{T}_{h,\tilde{a},\tilde{o}}(s,a,s')\hat{b}_h(s). \tag{36}$$

Thus $\hat{\rho}(h,a)$ and $\hat{P}(o|h,a)$ are given in terms of the $\hat{T}_{h,a,o}$'s, $\hat{O}_{h,a,o}$'s, and $\hat{b}_h$'s, all of which have already been determined.

4.4. Dynamic programming over the possibly-correct first-order beliefs

The last step in our algorithm is to solve the Bellman equations, i.e., Eqs. (21), (22), (35), and (36). As in Section 2.3, solving these equations is equivalent to solving a belief-state MDP:

$$\hat{V}^*(\hat{b}) = \max_{a\in A}\left\{\hat{\rho}(\hat{b},a) + \gamma \sum_{o\in\Theta} \hat{P}(o|\hat{b},a)\hat{V}^*\big(\tau\big(\hat{b},a,o,\hat{b}^M_{\hat{b},a,o}\big)\big)\right\}, \tag{37}$$

$$\hat{\mu}^*(\hat{b}) = \arg\max_{a\in A}\left\{\hat{\rho}(\hat{b},a) + \gamma \sum_{o\in\Theta} \hat{P}(o|\hat{b},a)\hat{V}^*\big(\tau\big(\hat{b},a,o,\hat{b}^M_{\hat{b},a,o}\big)\big)\right\}, \tag{38}$$

$$\hat{\rho}(\hat{b},a) = \frac{1}{|A\|\Theta|} \sum_{\tilde{a}\in A,\ \tilde{o}\in\Theta} \sum_{s\in S}\sum_{s'\in S} R(s,a,s')\hat{T}_{\hat{b},\tilde{a},\tilde{o}}(s,a,s')\hat{b}(s), \tag{39}$$

$$\hat{P}(o|\hat{b},a) = \frac{1}{|A\|\Theta|} \sum_{\tilde{a}\in A,\ \tilde{o}\in\Theta} \sum_{s'\in S} \hat{O}_{\hat{b},\tilde{a},\tilde{o}}(s',a,o) \sum_{s\in S} \hat{T}_{\hat{b},\tilde{a},\tilde{o}}(s,a,s')\hat{b}(s), \tag{40}$$

[8] To be precise, the procedure in Fig. 6 is correct only under the condition that $\hat{b}'(s') \neq 0$ and $(\underline{O}(s',a,o), \overline{O}(s',a,o)) \neq (0,0)$ for all $s' \in S$. For the general case, we require a little more complicated procedure that is described in Appendix B.

where $\hat{V}^*$, $\hat{\mu}$, $\hat{\rho}$, $\hat{P}$, $\hat{T}$, and $\hat{O}$ are all re-defined on $\hat{B}$, instead of H: for example, we substitute $\hat{\rho}(\hat{b}, a)$ for all the $\hat{\rho}(h, a)$'s for which $\hat{b}_h$ is the same. To see that solving these equations is equivalent to solving Eqs. (21), (22), (35), and (36), note that we have, by construction, $\hat{T}_{h_1,a,o} = \hat{T}_{h_2,a,o}$, $\hat{O}_{h_1,a,o} = \hat{O}_{h_2,a,o}$, and $\tau(\hat{b}_{h_1}, a, o, \hat{b}^M_{h_1,a,o}) = \tau(\hat{b}_{h_2}, a, o, \hat{b}^M_{h_2,a,o})$ whenever $\hat{b}_{h_1} = \hat{b}_{h_2}$ holds, for any $h_1, h_2 \in H$, $a \in A$, and $o \in \Theta$.

Finally, if $|\hat{B}|$ is finite, the quasi-optimal policy can be obtained by solving these equations (Eqs. (37)–(40)) numerically, for example by the value iteration methods [6].

5. Theoretical analyses

5.1. A bound on the reward losses of the quasi-optimal policies

We provide a bound on the amount of reward loss that may occur by using a quasi-optimal policy instead of specifying the correct second-order beliefs and performing the strict optimization. Note that this bound is applicable not only to the quasi-optimal policies obtained by the algorithm in Section 4 but also to quasi-optimal policies obtained by any other method.

In the following, let $\epsilon^T_{\max}$ and $\epsilon^O_{\max}$ denote the maximum imprecision of the parameters in T_h and O_h, i.e.,

$$\epsilon^T_{\max} := \max_{s,s' \in S,\ a \in A} \left| \overline{T}(s, a, s') - \underline{T}(s, a, s') \right| \tag{41}$$

and

$$\epsilon^O_{\max} := \max_{s' \in S,\ a \in A,\ o \in \Theta} \left| \overline{O}(s', a, o) - \underline{O}(s', a, o) \right|. \tag{42}$$

Let $V^{\hat{\mu}^*}$ be the value (i.e., the infinite-horizon discounted sum of the expected rewards) of a given quasi-optimal policy $\hat{\mu}^*$ that is evaluated in the hypothetical POMDP. Let V^{μ^*} be the value of the truly optimal policy μ^* that is also evaluated in the hypothetical POMDP. Note that, by definition, $V^{\mu^*} \geqslant V^{\hat{\mu}^*}$.

Theorem 1. *The reward loss, i.e., the difference between $V^{\hat{\mu}^*}$ and V^{μ^*}, is bounded by*

$$V^{\mu^*} - V^{\hat{\mu}^*} \leqslant \frac{((1-\gamma)|S|\epsilon^T_{\max} + 16\gamma d)R_{\max} + (1+15\gamma)\gamma d(\hat{V}^{\hat{\mu}^*}_{\max} + \hat{V}^{\mu^*}_{\max})}{(1-\gamma)^2},$$

where

$$d := |S|\epsilon^T_{\max} + |\Theta|\epsilon^O_{\max} + 2|\Theta||S|\epsilon^T_{\max}\epsilon^O_{\max}, \tag{43}$$

$$R_{\max} := \max_{s,s' \in S,\ a \in A} \left| R(s, a, s') \right|, \tag{44}$$

and

$$\hat{V}^{\mu}_{\max} := \max_{h \in H,\ a \in A,\ o \in \Theta} \left| \hat{V}^{\mu}\big(h; \langle a, o \rangle\big) \right| \tag{45}$$

for any policy $\mu : H \to A$, in which $\hat{V}^{\mu} : H \to \mathbb{R}$ is the solution of

$$\hat{V}^{\mu}(h) = \hat{\rho}\big(h, \mu(h)\big) + \gamma \sum_{o \in \Theta} \hat{P}\big(o|h, \mu(h)\big)\hat{V}^{\mu}\big(h; \langle \mu(h), o \rangle\big) \tag{46}$$

for every $h \in H$, where $\hat{\rho}$ and $\hat{P}$ are the functions in Eqs. (21) *and* (22) *that the quasi-optimal policy $\hat{\mu}^*$ satisfies.*

Proof. See Appendix C. □

Here are some notes on how this bound can be calculated (or upper-bounded):

First, both $\hat{V}^{\hat{\mu}^*}_{\max}$ and $\hat{V}^{\mu^*}_{\max}$ are upper-bounded by $V_{\max} := R_{\max} + \gamma R_{\max} + \gamma^2 R_{\max} + \cdots = R_{\max}/(1-\gamma)$, which is easy to calculate.

Second, $\hat{V}^{\hat{\mu}^*}_{\max}$ is readily available if our algorithm (Section 4) has been used to derive the quasi-optimal policy $\hat{\mu}^*$. Note that $\hat{V}^{\mu}$ in Eq. (46) is an extension of $\hat{V}^*$ in Eq. (21) in the sense that $\hat{V}^{\mu}$ is the value (that is estimated by

using possibly-correct beliefs) for any policy, whereas $\hat{V}^*$ is the value for the quasi-optimal policy $\hat{\mu}^*$. Thus, $\hat{V}^{\hat{\mu}^*}$ is identical to $\hat{V}^*$. Since our algorithm calculates $\hat{V}^*$, it is easy to obtain $\hat{V}^{\hat{\mu}^*}_{\max}$ by Eq. (45).

Last, for $\hat{V}^{\mu^*}_{\max}$, there is a tighter bound than $V_{\max}$. If $R(s,a,s')$ is non-negative for all s, a, and s', then $\hat{V}^{\mu^*}_{\max}$ is upper-bounded by $\hat{V}^{\hat{\mu}^*}_{\max}$. Otherwise, $\hat{V}^{\mu}_{\max}$ is upper-bounded by the larger value of $\hat{V}^{\hat{\mu}^*}_{\max}$ and $\hat{V}^{\hat{\mu}^*_-,-}_{\max}$, where $\hat{V}^{\hat{\mu}^*_-,-}_{\max}$ is calculated in exactly the same manner as $\hat{V}^{\hat{\mu}^*}_{\max}$, except that the reward function R is replaced by $-R$ and the quasi-optimal policy $\hat{\mu}^*_-$ is optimized for this modified problem (i.e., $\hat{\mu}^*_-$ maximizes the estimated total negative reward).[9]

Thus, this theoretical bound can be easily calculated. Unfortunately, this is not a tight bound. Unless γ, $\epsilon^T_{\max}$, and $\epsilon^O_{\max}$ are sufficiently small, this bound is looser than the default bound $2R_{\max}/(1-\gamma)$ that can be applied to any policy.[10] Given the presence of the $(1-\gamma)^{-2}$ factor, our bound is of limited practical use for problems in which γ is close to 1. Tighter bounds remain to be derived in future studies. We will study the reward loss empirically in Section 6.

5.2. Computational complexity of the provided algorithm

Here we summarize the computational costs our algorithm (Section 4) incurs in solving for the quasi-optimal policy.

First, we show that the FIND-A-SMALL-BELIEF-SET procedure of Section 4.1, which uses the IS-FEASIBLE procedure discussed in Section 4.2, terminates with finite $|\hat{B}|$ under moderate conditions. Suppose, for example, that every parameter has non-zero imprecision (i.e., for every parameter, the lower bound (Eqs. (30)–(33)) does not equal the upper bound). Let us consider a set $\Delta'_{\hat{b}}$ into which a first-order belief $\hat{b}$ can be Bayes-updated (Eq. (29)) by choosing a set of averaged model functions ($\hat{T}$ and $\hat{O}$) arbitrarily under the convexity conditions. Note that $\Delta'_{\hat{b}}$ is a subset of the probability simplex $\Delta_{|S|}$. Let b_c be the first-order belief that is added when the desired averaged model functions are not found in the IS-FEASIBLE procedure (Section 4.2). Since every parameter has non-zero imprecision, we can always place b_c inside $\Delta'_{\hat{b}}$ at a positive distance from its boundary; i.e., we can guarantee that the set $\Delta'_{\hat{b}}$ includes a non-empty ball

$$K = \left\{ b \mid \|b - b_c\| \leqslant \epsilon_r,\ b \in \Delta_{|S|} \right\} \subset \Delta'_{\hat{b}}, \tag{47}$$

where $\epsilon_r > 0$ is the radius of the ball, and $\|\cdot\|$ denotes the L_1 norm. Thus, when the FIND-A-SMALL-BELIEF-SET procedure adds a new first-order belief, this belief should be a distance of more than ϵ_r from the other first-order beliefs. Since every first-order belief is in the simplex $\Delta_{|S|}$, the procedure cannot continue adding new first-order beliefs an infinite number of times. It has therefore been proved that the procedure terminates with a finite number of the first-order beliefs. The condition that every parameter has non-zero imprecision can be relaxed, as long as the non-empty ball is guaranteed to exist.

As is evident from this discussion, the number of first-order beliefs $|\hat{B}|$ can, in the worst case, increase exponentially with $|S|$. As described below, $|\hat{B}|$ is an important factor in the computational complexity of the provided algorithm. However, currently, we have no other theoretical results concerning $|\hat{B}|$. We will instead provide some empirical results in the next section; for example, we will show that $|\hat{B}|$ does not always increase exponentially with $|S|$.

Next, we quantify the computational complexity of the present algorithm. We use the notation $O(\cdot)$ to indicate the order of the complexity. First, for the FIND-A-SMALL-BELIEF-SET procedure, we need, in the worst case, $O(|\hat{B}|^2|A||\Theta|)$ iterations of the IS-FEASIBLE procedure. Second, as noted in Section 4.2, the IS-FEASIBLE procedure can be conducted efficiently; in the interval case, the complexity is $O(poly(|S|))$, that is the polynomial order of $|S|$. In the point-set case, it is $O(poly(|S|N_T))$, where $N_T := \max_{s\in S, a\in A} |T^M(s,a)|$. Third, to construct the belief-state MDP (Eqs. (39) and (40)), we require $O(|\hat{B}||A|^2|\Theta|^2|S|^2)$ time. Last, to solve the belief-state MDP (Eqs. (37) and (38)) by the value iteration method, we need $O(|\hat{B}||A||\Theta|)$ time per iteration.

---

[9] The former bound is derived by noting that $\hat{V}^{\mu^*}(h;\langle a,o\rangle) \leqslant \hat{V}^{\hat{\mu}^*}(h;\langle a,o\rangle)$ holds for any h, a, and o. The latter bound is derived by noting additionally that $|\hat{V}^{\mu^*}(h;\langle a,o\rangle)| \leqslant \max(\hat{V}^{\mu^*}(h;\langle a,o\rangle), -\hat{V}^{\mu^*}(h;\langle a,o\rangle))$ and $-\hat{V}^{\mu^*}(h;\langle a,o\rangle) \leqslant \hat{V}^{\hat{\mu}^*_-,-}(h;\langle a,o\rangle)$ hold for any h, a, and o, where we define $\hat{V}^{\mu,-}$ for any μ in the same manner as $\hat{V}^{\mu}$, except that we replace the reward function R with $-R$.

[10] This default bound is obtained by observing that any policy may miss at most $2R_{\max}$ reward at each time step. Calculating the discounted sum over an infinite horizon gives this bound; i.e., $2R_{\max} + 2R_{\max}\gamma + 2R_{\max}\gamma^2 + \cdots = 2R_{\max}/(1-\gamma)$.

We can make a small modification to our algorithm, which often reduces the computational cost. Combining the complexity of the FIND-A-SMALL-BELIEF-SET procedure and that of the IS-FEASIBLE procedure, we see that the time that the IS-FEASIBLE procedure requires is $O(|\hat{B}|^2|A||\Theta|poly(|S|))$ in the interval case, and $O(|\hat{B}|^2|A||\Theta|poly(|S|N_T))$ in the point-set case. These can be changed into $O(|\hat{B}||A||\Theta|(poly(|S|) + |\hat{B}||S|))$ and $O(|\hat{B}||A||\Theta|(poly(|S|N_T) + |\hat{B}||S|))$, respectively, by the following modification. Recall that, in the FIND-A-SMALL-BELIEF-SET procedure, given each $\hat{b}_h$ and $a \in A$ and $o \in \Theta$, we search through all the candidates (i.e., all the first-order beliefs that have already been calculated), seeking one that $\tau(\hat{b}_h, a, o, \hat{b}^M_{h,a,o})$ can be identical to. However, in many cases, searching through all the candidates can be avoided; we can search through only a small number of the candidates that are likely to include the required belief. To do this, we first calculate a tentatively Bayes-updated belief $\hat{b}'' = \tau(\hat{b}_h, a, o, \hat{b}^M_{h,a,o})$, using an arbitrary $\hat{b}^M_{h,a,o}$ that satisfies the permissibility condition. Then we pick k-nearest neighbors of $\hat{b}''$ among all the candidate beliefs, and we search only within them. Note that, with this modification, we can still find the quasi-optimal policy. Although this modification may increase the resultant size of $\hat{B}$, empirically, it was found to significantly reduce the total computational time. In the next section we adopt this technique with $k = 5$, using the L_1 norm for measuring the distance between the first-order beliefs. Further, note that with this modification, the FIND-A-SMALL-BELIEF-SET procedure of Section 4.1 still terminates with finite $|\hat{B}|$ under moderate conditions. This is proved in exactly the same manner as in the beginning of this section if we let $\hat{b}''$ be b_c, i.e., if we use $\hat{b}''$ as the first-order belief to be added when the search fails.

There is another modification that we can make to our algorithm in order to reduce the computational cost. When the number of states, $|S|$, is large, it takes long time to solve the linear programming problem in the IS-FEASIBLE procedure. Recall that it takes $O(poly(|S|))$ time in the interval case and $O(poly(|S|N_T))$ time in the point-set case. We can modify the procedure so that we search only for the values for $\hat{O}$. The transition function $\hat{T}$ is set arbitrarily (under the convexity conditions). With $\hat{T}$ fixed, the procedure (Fig. 6) requires $O(|S|^2)$ time for both the interval case and the point-set case. This modification often reduced the solution time when $|S|$ is large. We use this modification in Sections 6.2 and 6.3. Note that since this modification restricts the search space of the IS-FEASIBLE procedure, the number of required first-order beliefs might increase. However, this modification often reduced significantly the time spent in each search, and consequently the total solution time was reduced. Note also that with this modification, the FIND-A-SMALL-BELIEF-SET procedure of Section 4.1 still terminates with finite $|\hat{B}|$ under moderate conditions. For example, if every parameter in $\hat{O}$ has non-zero imprecision, the non-empty ball K in Eq. (47) can be guaranteed to exist, and hence the procedure terminates with finite $|\hat{B}|$.

6. Experiments

We applied our algorithm to several POMDPIPs. We report only results for the interval case. Those for the point-set case are expected to be similar. All results were obtained by Matlab codes on a Pentium4 PC.[11]

6.1. Small-sized POMDPIPs

Here, we used nine small-sized POMDPs for which the optimal policies are known [11,24,57].

First, for each POMDP (Table 1) and a parameter $\epsilon \in \{0.0125, 0.025, 0.05, 0.1, 0.2, 0.4, 0.6\}$ that controls the imprecision, we constructed a POMDPIP as follows. We set T^M as $T^M(s, a, s') := [T(s, a, s') - \epsilon, T(s, a, s') + \epsilon]$ for all $s, s' \in S$ and $a \in A$, where $T(s, a, s')$ is the parameter value of the original POMDP. Similarly we set $O^M(s', a, o) := [O(s', a, o) - \epsilon, O(s', a, o) + \epsilon]$ for all $s' \in S, a \in A$ and $o \in \Theta$, where $O(s', a, o)$ is the original parameter value. We made these bounds saturated as 0 or 1, when they exceeded the range of [0, 1]. We then calculated the maximum imprecision $\epsilon^T_{\max}$ and $\epsilon^O_{\max}$ (Eqs. (41) and (42)), and let $\epsilon_{\max}$ be the maximum of these.

Second, we applied our algorithm to each of these POMDPIPs, and obtained the quasi-optimal policies. Since the algorithm sometimes requires arbitrary averaged model functions ($\hat{T}$ and $\hat{O}$), we calculated "typical" model functions and used them whenever necessary. To make the typical model functions, we first calculated the middle points of the upper and lower bounds. Since such points sometimes break the sum-to-1 constraints (e.g., $\sum_{s'} \hat{T}(s, a, s') = 1$), we

[11] The codes are freely available at http://www.brn.dis.titech.ac.jp/~hideaki/pomdpips/index.htm.

Table 1
Small-sized POMDPs with known optimal policies

Test problem	$\|S\|$	$\|A\|$	$\|\Theta\|$
Tiger	2	3	2
1D maze	4	2	2
4× 3	11	4	6
4 × 3 CO	11	4	11
Cheese	11	4	7
Part painting	4	4	2
Network	7	4	2
Shuttle	8	3	5
Aircraft ID	12	6	5

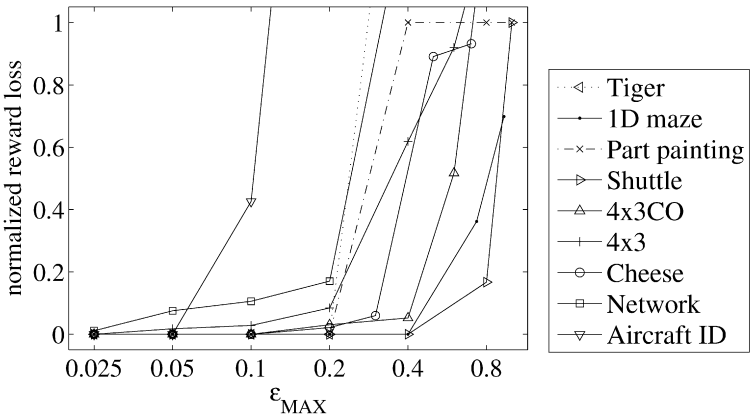

Fig. 7. Small-sized POMDPIPs; the reward losses of the quasi-optimal policies evaluated in the original POMDPs.

then picked the nearest points (measured by Euclidean norm) that do not violate all the constraints (both the sum-to-1 constraints and the convexity conditions).

Then we studied whether the quasi-optimal policies obtained were nontrivial (Fig. 7). For each POMDPIP problem, we assumed that the original POMDP was the true environment, i.e., the hypothetical POMDP in Section 3.2. We evaluated $V^{\hat{\mu}^*}$, i.e., the value of the obtained quasi-optimal policy in the original POMDP. We also evaluated V^{μ^*}, i.e., the value of the truly optimal policy of the original POMDP. The difference, $V^{\mu^*} - V^{\hat{\mu}^*}$, indicates the reward loss that occurred by using the quasi-optimal policy, instead of identifying the true environment and performing the strict optimization for it. Fig. 7 shows the reward losses, each of which was normalized by V^{μ^*}. For all of the problems, the reward losses were small when $\epsilon_{\max}$ was small. This indicates that the quasi-optimal policies found were nontrivial. These policies are admissible as solutions of the POMDPIPs in the sense that they are almost optimal for at least one possibly true environment.

Furthermore, we studied robustness of the obtained policies (Fig. 8). In this study, we regard a policy to be robust if the reward loss, $V^{\mu^*} - V^{\hat{\mu}^*}$, is small for various true environments. For each POMDPIP problem, we randomly generated twenty hypothetical POMDPs. For each of the hypothetical POMDPs, we calculated the reward loss in the same way as in Fig. 7. Then we calculated the largest reward loss in these twenty environments. The results (Fig. 8) suggest that, in many problems, the reward loss is kept small as long as $\epsilon_{\max}$ is within a range. However, in Aircraft ID and Tiger, relatively large reward losses were observed. In these problems, the agent receives a huge negative reward by taking some action in some state. To avoid the huge negative reward, the first-order belief (i.e., the belief in the state) needs to be inferred precisely. For such problems, quasi-optimal policies may not be a good solution; we may need precise probabilities and strict optimization, or we may need other policies such as maximin.

We compared the robustness of the quasi-optimal policies with that of the E-admissible policies. Note that the quasi-optimal policies could be more robust than the E-admissible policies (Section 3.3 and Appendix A). For each POMPDIP problem, we obtained three kinds of E-admissible policies: (1) the optimal policy for the original POMDP problem, (2) the optimal policy for the POMDP that consists of the "typical" model functions, and (3) the optimal

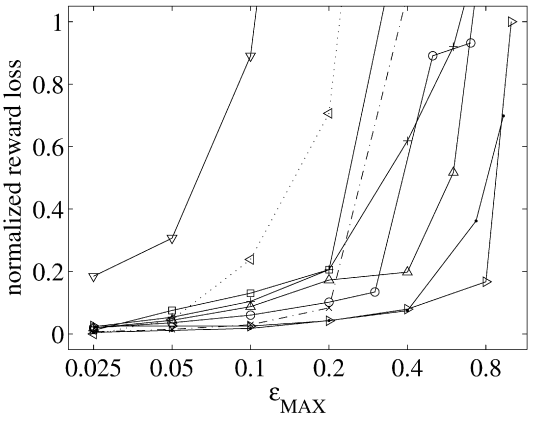

Fig. 8. Small-sized POMDPIPs; the largest reward losses of the quasi-optimal policies evaluated in twenty randomly-generated hypothetical POMDPs. See Fig. 7 for the legend.

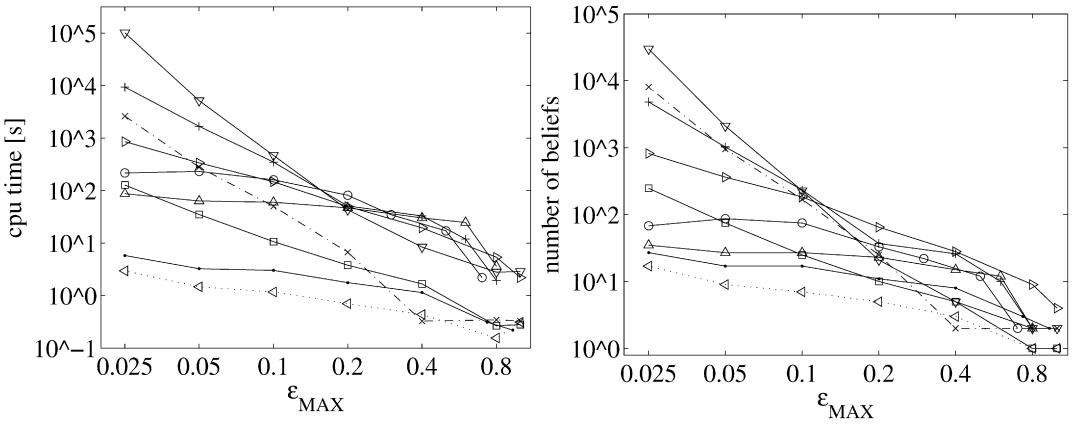

Fig. 9. Small-sized POMDPIPs; solution time (left) and $|\hat{B}|$ (right). See Fig. 7 for the legend.

policy for a randomly-generated hypothetical POMDP. For each of these E-admissible policies, we calculated the largest reward loss as in Fig. 8. Unfortunately, the results (not shown) were not significantly different from those of the quasi-optimal policies in Fig. 8. Among the three kinds of the E-admissible policies tested, the first ones (i.e., the optimal policies for the original POMDPs) were the most robust (i.e., the largest reward loss was the smallest) in most of the problems. Thus, we compared these E-admissible policies with the quasi-optimal policies. Out of the 63 POMDPIP problems (9 original POMDP problems times 7 values of $\epsilon_{\max}$) in Fig. 8, these E-admissible policies were more robust than the quasi-optimal policies in 29 problems. In the other 34 problems, the quasi-optimal policies were more robust. Thus, the quasi-optimal policies tended to be more robust. However, the difference was subtle, and further research is required for detailed comparisons.

Fig. 9 (left) shows the cpu time required to obtain the quasi-optimal policies. Our algorithm terminated in a reasonably short time in many cases. For example, for the "Aircraft ID" problem, the shortest solution time ever reported is 27,676 seconds [57]. Although direct comparison is impossible, our algorithm required 5,152 seconds when $\epsilon_{\max}$ was around 0.05, and only 467 seconds when it was around 0.1 (with a degraded value, though). For the "Network" problem, the shortest solution time ever was 140 seconds [57], whereas our algorithm required 35 seconds when $\epsilon_{\max}$ was around 0.05 and 11 seconds when it was around 0.1. In some of the other problems, however, our algorithm recorded longer solution times than those reported for other algorithms.

The number of beliefs, $|\hat{B}|$, is plotted in Fig. 9 (right). It may be noted that the plots of the $|\hat{B}|$ appear similar to those of the cpu time. This result is in agreement with expectations due to the algorithm's complexity (Section 5.2). We further note that $|\hat{B}|$ increases as $\epsilon_{\max}$ becomes smaller. All the plots in the figure appear to be well approximated

by a straight line. Since the figure is the log-log scale plot, we may conclude that $|\hat{B}|$ is approximately proportional to $poly(1/\epsilon_{\max})$.

The slopes of the plots in Fig. 9 clearly depend on the problem being solved; The slopes of the plots of the Aircraft ID and Part painting problems are large, whereas they are smaller for the Cheese and 4×3 CO problems. This indicates that $|\hat{B}|$ does not depend directly on the size of the problem (i.e., $|S|$, $|A|$ and $|\Theta|$). Rather, it appears that the slopes tended to be smaller for problems for which $|\Theta|$ is large compared with $|S|$. This result is in agreement with expectations; if $|\Theta|$ is large, and if by each $o \in \Theta$ the agent tends to be able to obtain a relatively large amount of information about the state s, then the first-order beliefs $\hat{b}_h \in \hat{B}$ will be mostly located around the rims of the probability simplex (i.e., $\hat{b}_h(s) \ll 1$ for most $s \in S$), and hence only a small number of the beliefs would be required in the FIND-A-SMALL-BELIEF-SET procedure of Section 4.1.

6.2. *POMDPIPs with nearly full observations*

Next we study larger-sized problems. As suggested in the previous section, our algorithm can be expected to quickly solve POMDPIPs in which the agent's first-order beliefs are mostly located around the rims of the probability simplex. As an example, we consider here some problems in which the agent can perform low-noise observations of the states.

First, we constructed POMDPs. They are maze-type environments with n states and four actions, and the agent's state can be observed by n kinds of observations (i.e., $|S| = |\Theta| = n, |A| = 4$); we constructed three mazes with $n = 320, 640$, and 1280. These POMDPs are simple models of a navigation problem in which a robot does not have a complete capability of moving and sensing, but has a capability close to this condition. The agent's action changes, with a probability of 0.9, its current state to another one in agreement with its intention. However, with a probability of 0.1, the agent stays in the same state or moves to an unintended state that is selected uniformly randomly. Each observation corresponds to each state by a one-to-one mapping, and usually (with probability 0.9) the agent observes the current state correctly. However, with a probability of 0.1, it observes a false state that is selected uniformly randomly. The initial state is distributed uniformly over the states. There is a single goal state. A reward of 1 is given for reaching the goal state, there is no reward otherwise. We set $\gamma = 0.95$.

Having constructed the POMDPs, we made POMDPIPs from these POMDPs, in the manner described in Section 6.1.

Then we obtained their solutions. We used a modified algorithm in which the IS-FEASIBLE procedure searches only for the values in $\hat{O}$ (see Section 5.2 for details). We set the transition function $\hat{T}$ as the "typical" model function defined in Section 6.1, instead of setting it by linear programming.

The results are shown in Figs. 10–12. The formats are the same as those of Figs. 7–9, except that, in Figs. 10 and 11, the value of the truly optimal policy, V^{μ^*}, of each problem was calculated approximately by the Perseus algorithm [52,53] because we could not solve these large problems exactly.

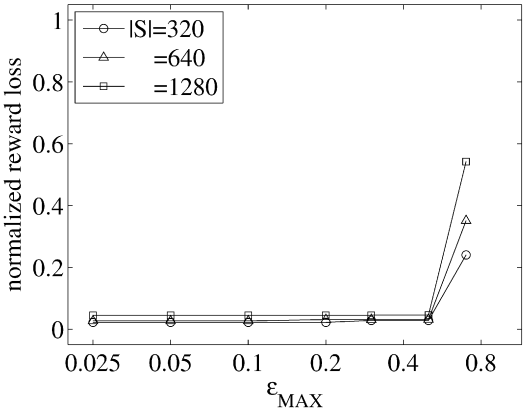

Fig. 10. POMDPIPs with nearly full observations; the reward losses of the quasi-optimal policies evaluated in the original POMDPs.

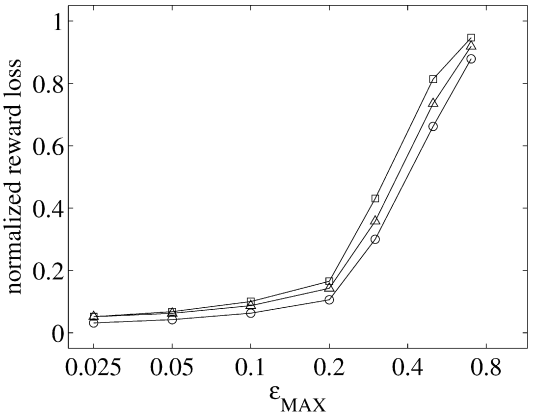

Fig. 11. POMDPIPs with nearly full observations; the largest reward losses of the quasi-optimal policies evaluated in twenty randomly-generated hypothetical POMDPs. See Fig. 10 for the legend.

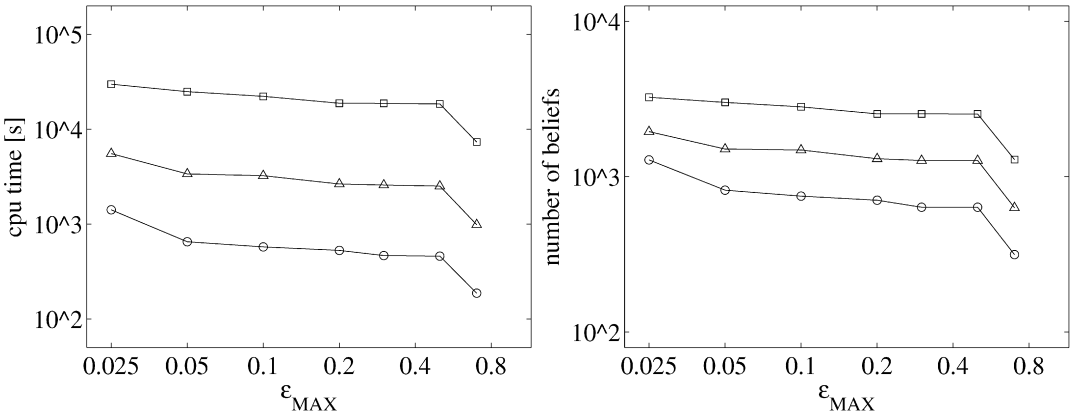

Fig. 12. POMDPIPs with nearly full observations; solution time (left) and $|\hat{B}|$ (right). See Fig. 10 for the legend.

The quasi-optimal policies were obtained successfully for these POMDPIPs, even for the cases in which $|S| = |\Theta| = 1240$, without a significant reduction in the performance. The policies obtained were suggested to be nontrivial (Fig. 10) and robust (Fig. 11) when the parameter imprecision was small. As expected, the algorithm terminated in a reasonable time (Fig. 12 left) with a small number of first-order beliefs (Fig. 12 right). In all of the problems, the number of required first-order beliefs was less than 4 times larger than $|S|$.

6.3. Problems that have not been solved

Although our algorithm was able to solve the large-sized problems up to $|S| = 1240$ in Section 6.2, it failed to solve some other problems in a reasonable amount of time. We report on these problems in this section. Further research is required to solve these problems.

We tried to solve the Tag-Avoid problem [44], Cycle10 problem, and 3leg10 problem [45,47] (Table 2). For each of these POMDP problems, we created POMDPIPs in the manner described in Section 6.1, with $\epsilon = 0.0125$, 0.1, 0.4, 0.6, or 0.8. In each POMDPIP problem, we applied the algorithm used in Section 6.2, which searches only for the values of the parameters in $\hat{O}$. The algorithm used in Section 6.1, which searches for the values of the parameters in $\hat{T}$ and $\hat{O}$, was not applicable due to running out of memory.

For each POMDPIP problem, we allowed the algorithm to run for up to 48 hours. For $\epsilon \in \{0.0125, 0.1, 0.4, 0.6\}$, none of the problems were solved within the time limit. Many first-order beliefs had been generated (Table 2), and more beliefs were being generated. For ϵ as large as 0.8, the algorithm terminated after generating a tractable number

Table 2
Problems for which the algorithm did not terminate or terminated but failed to find a satisfactory solution

Test problem	$\|S\|$	$\|A\|$	$\|\Theta\|$	# beliefs ($\times 10^4$) generated within 48 h				# beliefs ($\times 10^0$) when terminated[a]
				$\epsilon = 0.0125$	0.1	0.4	0.6	0.8
Tag-Avoid	870	5	30	3.2	3.1	2.0	1.4	439 (759, 2.1)
Cycle10	1280	21	2	2.9	2.8	2.8	1.4	33 (23, 0.41)
3leg10	1280	21	2	2.9	2.8	2.8	1.6	127 (76, 0.42)

[a] Shown with (time elapsed [s], normalized reward loss of the policy found).

of beliefs (Table 2) for every problem. However, the policies found were unsatisfactory; the normalized reward loss, which we calculated in the same manner as in Fig. 10, was 0.41 or larger.

7. Conclusion

In this paper, we formulated POMDPIPs and their quasi-optimal policies, and provided an efficient algorithm to obtain these policies. We also provided a theoretical bound on the reward losses of the quasi-optimal policies and the computational complexity of the algorithm. Empirical studies showed that the algorithm can find nontrivial policies in a reasonable time for many POMDPIPs.

There are several directions for future research.

First, characteristics of the quasi-optimal policy can be studied more deeply. For example, derivation of tighter theoretical bounds may be investigated. Also, detailed comparisons between the quasi-optimal policies and the E-admissible policies remain to be performed (Section 6.1).

Second, robustness of the policy can be pursued further. In this paper, we did not attempt to obtain the most robust policies possible. For some problems, the quasi-optimal policies were relatively less robust (Section 6.1). It would be desirable if we could obtain the robust policies, e.g., using the maximin approach, within a reasonable time. Recently, Nilim and El Ghaoui proposed handling imprecision by using likelihood-bounded sets (as opposed to intervals) in order to efficiently obtain robust policies in MDPs [39,40]. It would be interesting to investigate POMDPs as well.

Third, other types of the solution algorithm can be considered. Although our algorithm solved many POMDPIPs, it failed (Section 6.3) to solve the problems for which recent POMDP algorithms are able to find good (although approximate) policies [43,45,47]. Our algorithm directly constructs grid-based belief state MDPs. It can suffer from exponential growth of the number of first-order beliefs, as in the other grid-based approaches [7,28,59]. Other approaches, e.g., those based on α-vectors [24,44,46,47,51–53,57], may offer reduced solution times for the classes of POMDPIPs for which our algorithm is slow.

Last, we note that our algorithm can also be used as an approximate planning method for large-sized POMDPs. Even when parameters are given precisely, we can choose to introduce a parameter imprecision, balancing the gains in reduction in solution time against the reduction in accuracy. The empirical results (Section 6) are encouraging because they suggest that satisfactory policies will be obtained within a certain range of the parameter imprecision. Note that the idea that the parameter imprecision can be exploited to reduce the computational complexity of solving POMDPs is orthogonal to other approximation methods (reviewed in Section 1), i.e., every combination of this idea and other approximation methods can be pursued.

Acknowledgements

We thank Minseok Kim, Shigemi Sawa, and the reviewers for their valuable comments. We also thank Anthony R. Cassandra and Joelle Pineau for making the codes of the POMDP problems available. We thank Matthijs Spaan for making the code of the Perseus algorithm available. This work is supported in part by the Grant-in-Aid for Young Scientists (B-15700180) from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

Appendix A. An example in which using multiple second-order beliefs leads to a more robust policy than using a single second-order belief

Let us consider a very simple POMDPIP problem. Suppose that there are four states, $S = \{s^1, s^2, s^3, s^4\}$, two actions, $A = \{a^1, a^2\}$, and two observations, $\Theta = \{o^1, o^2\}$. Suppose $p_0 = (1, 0, 0, 0)$, i.e., the process starts with s^1. Let us consider the point-set case. Let T^M be $T^M(s, a) = \{(0, 0, 0, 1)\}$ for all s and a, except $T^M(s^1, a^1) = \{(0, 0.4, 0.6, 0), (0, 0.6, 0.4, 0)\}$. That is, the state always changes to s^4 no matter which action the agent takes, except that it changes to s^2 or s^3 with the imprecise probability if the agent takes action a^1 in state s^1. Let O^M be $O^M(s', a) = \{(0.5, 0.5)\}$ for all s' and a; that is, the state is completely unobservable. Let R be $R(s, a, s') = 0$ for all s, a, and s', except that $R(s, a, s') = 1$ for $s = s^2$, $a = a^1$, and $s' = s^4$, and that $R(s, a, s') = 1$ for $s = s^3$, $a = a^2$, and $s' = s^4$. That is, a reward is gained only when the agent takes action a^1 in state s^2 or when the agent takes action a^2 in state s^3.

For this POMDPIP problem, let us consider the history tree shown in Fig. 4. The possibly-correct first-order beliefs are (1) $\hat{b}_\emptyset = p_0 = (1, 0, 0, 0)$, (2) $\hat{b}_{11}$, which is the belief after taking action a^1 and observing o^1, equals a convex combination of $(0, 0.4, 0.6, 0)$ and $(0, 0.6, 0.4, 0)$, (3) $\hat{b}_{12}$ also equals a convex combination of $(0, 0.4, 0.6, 0)$ and $(0, 0.6, 0.4, 0)$, and (4) $(0, 0, 0, 1)$ otherwise. Note that $\hat{b}_{11}$ and $\hat{b}_{12}$ are calculated depending on the second-order beliefs $\hat{b}^M_{\emptyset,a^1,o^1}$ and $\hat{b}^M_{\emptyset,a^1,o^2}$, respectively. Thus, if only a single second-order belief is allowed, $\hat{b}_{11}$ and $\hat{b}_{12}$ should be identical. If multiple second-order beliefs are allowed, $\hat{b}_{11}$ and $\hat{b}_{12}$ can be different.

Now, let us consider which action is to be taken after each action-observation history. Since a reward can be gained only when the state is in s^2 or s^3, the initial action (i.e., the action after history $\emptyset$) should be a^1. For the same reason, the actions after two actions (i.e., actions after $\hat{b}_{1111}$, $\hat{b}_{1112}$, and so on) do not affect the total reward. Thus, the only actions to be optimized are those immediately after $\hat{b}_{11}$ and $\hat{b}_{12}$.

Let us consider the action after $\hat{b}_{11}$. Recall that a unit reward is gained for action a^1 in state s^2 or action a^2 in state s^3. The estimated rewards for actions a^1 and a^2 depend on the value of $\hat{b}_{11}$, as shown in Fig. A.1. For example, if $\hat{b}_{11}(s^2) = 0.4$ (which means $\hat{b}_{11} = (0, 0.4, 0.6, 0)$), then the estimated reward is 0.4 for action a^1 and 0.6 for action a^2; hence action a^2 should be taken.

The same holds for the action after $\hat{b}_{12}$. Thus, if only a single second-order belief is allowed, then $\hat{b}_{11}$ and $\hat{b}_{12}$ are identical, and hence the actions after $\hat{b}_{11}$ and $\hat{b}_{12}$ are identical. Let us suppose, for example, that $\hat{b}_{11}(s^2) = \hat{b}_{11}(s^3) = 0.4$. Then action a^2 is selected both after $\hat{b}_{11}$ and $\hat{b}_{12}$.

On the other hand, if multiple second-order beliefs are allowed, and if $\hat{b}_{11}$ and $\hat{b}_{12}$ are different, then the actions after $\hat{b}_{11}$ and $\hat{b}_{12}$ can be different. Let us suppose, for example, that the agent adopts $\hat{b}_{11}(s^2) = 0.4$ and $\hat{b}_{12}(s^2) = 0.6$. Then action a^2 is selected after $\hat{b}_{11}$ and action a^1 is selected after $\hat{b}_{12}$. Since both $\hat{b}_{11}$ and $\hat{b}_{12}$ are reached with probability 0.5, this agent selects action a^1 or a^2 with probability 0.5.

Now, suppose that the correct second-order beliefs are given (as in Section 3.2). Depending on their values, the expected total reward varies from 0.4γ to 0.6γ for the agent who always selects a^2 (where γ is the discount factor). On the other hand, the expected reward is always 0.5γ for the agent who selects a^1 or a^2 with probability 0.5.

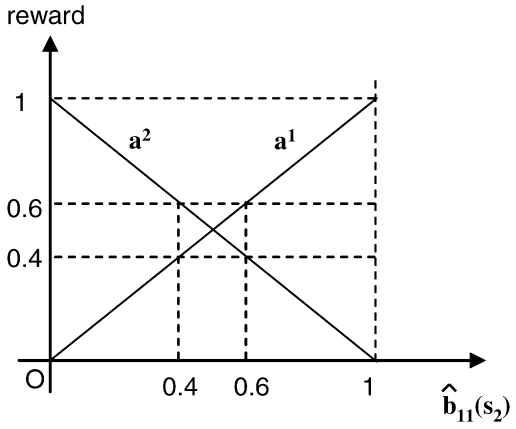

Fig. A.1. Reward expected by taking action a^1 or a^2 after $\hat{b}_{11}$.

Thus, although the agent who adopts a single second-order belief gains more reward than the agent with multiple beliefs if the belief adopted is correct, the worst case performance can be better in the agent with multiple beliefs. In this sense, the use of multiple beliefs can lead to a more robust policy.

Note that in other parts of this paper, we use the term "robust" with a slightly different meaning; i.e., we regard a policy to be robust if the reward loss incurred by using the policy, instead of using the optimal policy for each environment, is small for various environments (Section 6). The conclusion in Appendix A is not changed by adopting this meaning because the maximum reward loss is 0.2γ for the agent who always selects a^2 and 0.1γ for the agent who selects a^1 or a^2 with probability 0.5.

Note that we do not argue that the use of multiple beliefs is always better. Note also that, in our algorithm, we do not try to find a policy that is as robust as possible. The example in this section suggests the use of randomized policies (i.e., stochastic action-selection rules) for robustness. Finding such robust policies is beyond the scope of the present paper. The only purpose of introducing this simple example is to show that it is not always advantageous to use a single second-order belief compared to the use of multiple second-order beliefs.

Appendix B. Derivation of the IS-FEASIBLE procedure

First, we consider the interval case. In the IS-FEASIBLE procedure, the query that we want to answer as true or false is formalized as follows:

Q1: There exists $\hat{T}(s,a,s') \in \mathbb{R}$ for each s and $s' \in S$, $\hat{O}(s',a,o) \in \mathbb{R}$ for each $s' \in S$, and $Z \in \mathbb{R}$, s.t.

$$Z > 0, \tag{B.1}$$

$$Z = \sum_{s,s' \in S} \hat{b}(s)\hat{T}(s,a,s')\hat{O}(s',a,o), \tag{B.2}$$

$$\underline{T}(s,a,s') \leqslant \hat{T}(s,a,s') \leqslant \overline{T}(s,a,s') \quad \text{for all } s,s' \in S, \tag{B.3}$$

$$\sum_{s' \in S} \hat{T}(s,a,s') = 1 \quad \text{for all } s \in S, \tag{B.4}$$

$$\underline{O}(s',a,o) \leqslant \hat{O}(s',a,o) \leqslant \overline{O}(s',a,o) \quad \text{for all } s' \in S, \tag{B.5}$$

$$\sum_{s \in S} \hat{b}(s)\hat{T}(s,a,s')\hat{O}(s',a,o)/Z = \hat{b}'(s') \quad \text{for all } s' \in S \tag{B.6}$$

hold.

There are some non-linear equations in this query. In the following, we transform this query into an equivalent query that consists of linear equations only.

To begin with, in Q1, Eq. (B.2) is unnecessary because we already have Eqs. (B.1) and (B.6) and $\sum_{s' \in S} \hat{b}'(s') = 1$ holds by construction. Consequently, the query Q1 is equivalent to:

Q2: There exists $\hat{T}(s,a,s') \in \mathbb{R}$ for each s and $s' \in S$, $\hat{O}(s',a,o) \in \mathbb{R}$ for each $s' \in S$, and $Z \in \mathbb{R}$, s.t. Eqs. (B.1), (B.3)–(B.6) hold.

Q1 is equivalent to Q2, since Q1 is true (or false) when Q2 is true (or false), respectively.

Next, let us divide S into mutually exclusive sets S_1, S_2, S_3, and S_4, as:

$$S_1 := \left\{ s' \mid \hat{b}'(s') \neq 0, \underline{O}(s',a,o) = 0, \overline{O}(s',a,o) = 0, s' \in S \right\},$$

$$S_2 := \left\{ s' \mid \hat{b}'(s') \neq 0, s' \in S - S_1 \right\},$$

$$S_3 := \left\{ s' \mid \hat{b}'(s') = 0, \underline{O}(s',a,o) = 0, s' \in S \right\},$$

$$S_4 := \left\{ s' \mid \hat{b}'(s') = 0, \underline{O}(s',a,o) \neq 0, s' \in S \right\}.$$

Note that the sets $S1$, $S2$, $S3$, and $S4$ form a partition of S. It is straightforward to see that Q2 is equivalent to:

Q3: There exists $\hat{T}(s,a,s') \in \mathbb{R}$ for each s and $s' \in S$, $\hat{O}(s',a,o) \in \mathbb{R}$ for each $s' \in S$, and $Z \in \mathbb{R}$, s.t.

Eqs. (B.1), (B.3), (B.4),

$$\hat{O}(s', a, o) = 0 \quad \text{for all } s' \in S_1, \tag{B.7}$$

$$\sum_{s \in S} \hat{b}(s)\hat{T}(s, a, s')\hat{O}(s', a, o)/Z = \hat{b}'(s') \neq 0 \quad \text{for all } s' \in S_1, \tag{B.8}$$

$$\underline{O}(s', a, o) \leqslant \hat{O}(s', a, o) \leqslant \overline{O}(s', a, o) \neq 0 \quad \text{for all } s' \in S_2, \tag{B.9}$$

$$\sum_{s \in S} \hat{b}(s)\hat{T}(s, a, s')\hat{O}(s', a, o)/Z = \hat{b}'(s') \quad \text{for all } s' \in S_2, \tag{B.10}$$

$$0 \leqslant \hat{O}(s', a, o) \leqslant \overline{O}(s', a, o) \quad \text{for all } s' \in S_3, \tag{B.11}$$

$$\sum_{s \in S} \hat{b}(s)\hat{T}(s, a, s')\hat{O}(s', a, o)/Z = 0 \quad \text{for all } s' \in S_3, \tag{B.12}$$

$$0 < \underline{O}(s', a, o) \leqslant \hat{O}(s', a, o) \leqslant \overline{O}(s', a, o) \quad \text{for all } s' \in S_4, \tag{B.13}$$

$$\sum_{s \in S} \hat{b}(s)\hat{T}(s, a, s')\hat{O}(s', a, o)/Z = 0 \quad \text{for all } s' \in S_4 \tag{B.14}$$

hold.

Note that, in Eq. (B.9), we have $\overline{O}(s', a, o) \neq 0$, since, if $\overline{O}(s', a, o) = 0$, then, from $0 \leqslant \underline{O}(s', a, o) \leqslant \overline{O}(s', a, o)$, it is concluded that $\underline{O}(s', a, o) = \overline{O}(s', a, o) = 0$, which means that s' is in S_1, not in S_2.

Next, we prove that Q3 is equivalent to:

Q4: S_1 is an empty set, and there exists $\hat{T}(s, a, s') \in \mathbb{R}$ for each s and $s' \in S$, $\hat{O}(s', a, o) \in \mathbb{R}$ for each $s' \in S$, and $Z \in \mathbb{R}$, s.t.

Eqs. (B.1), (B.3), (B.4),

$$\underline{O}(s', a, o) \leqslant \hat{O}(s', a, o) \leqslant \overline{O}(s', a, o) \neq 0 \quad \text{for all } s' \in S_2,$$

$$\sum_{s \in S} \hat{b}(s)\hat{T}(s, a, s')\hat{O}(s', a, o)/Z = \hat{b}'(s') \quad \text{for all } s' \in S_2,$$

$$\hat{O}(s', a, o) > 0 \quad \text{for all } s' \in S_2,$$

$$\sum_{s \in S} \hat{b}(s)\hat{T}(s, a, s') = 0 \quad \text{for all } s' \in S_4$$

hold.

To prove that Q3 is equivalent to Q4, check that Q4 is true if Q3 is true, and that Q3 is true if Q4 is true. To check the former, note the followings: For $s' \in S_1$, Eqs. (B.7) and (B.8) can never be satisfied. For $s' \in S_2$, Eqs. (B.9) and (B.10) can be satisfied only if $\hat{O}(s', a, o) > 0$ holds. For $s' \in S_4$, Eqs. (B.13) and (B.14) are satisfied only if $\sum_{s \in S} \hat{b}(s)\hat{T}(s, a, s') = 0$.

To check the latter, note the followings: For $s' \in S_3$, Eqs. (B.11) and (B.12) can always be satisfied by setting

$$\hat{O}(s', a, o) = 0, \tag{B.15}$$

which is always possible by the definition of S_3. For $s' \in S_4$, Eq. (B.13) can always be satisfied by setting

$$\hat{O}(s', a, o) \text{ arbitrarily s.t. } \underline{O}(s', a, o) \leqslant \hat{O}(s', a, o) \leqslant \overline{O}(s', a, o), \tag{B.16}$$

which is always possible by definition.

Next, let us define $q(s') = Z/\hat{O}(s', a, o)$ for all $s' \in S_2$. Now Q4 is equivalent to:

Q5: S_1 is an empty set, and there exists $\hat{T}(s, a, s') \in \mathbb{R}$ for each s and $s' \in S$, $q(s') \in \mathbb{R}$ for each $s' \in S_2$, and $Z \in \mathbb{R}$, s.t.

Eqs. (B.1), (B.3), (B.4),

$$\frac{Z}{\underline{O}(s', a, o)} \geqslant q(s') \geqslant \frac{Z}{\overline{O}(s', a, o)} \quad \text{for all } s' \in S_2, \tag{B.17}$$

$$\sum_{s\in S}\hat{b}(s)\hat{T}(s,a,s')=\hat{b}'(s')q(s') \quad \text{for all } s'\in S_2, \tag{B.18}$$

$$q(s')<\infty \quad \text{for all } s'\in S_2, \tag{B.19}$$

$$\sum_{s\in S}\hat{b}(s)\hat{T}(s,a,s')=0 \quad \text{for all } s'\in S_4 \tag{B.20}$$

hold, where we define $\frac{Z}{\underline{O}(s',a,o)}=\infty$ when $\underline{O}(s',a,o)=0$ in Eq. (B.17). Note that $\frac{Z}{\overline{O}(s',a,o)}\neq\infty$ since $\overline{O}(s',a,o)\neq 0$ for any $s'\in S_2$.

In Q5, Eq. (B.19) is unnecessary, since we have Eq. (B.18) in which the left-hand side is finite and $\hat{b}'(s')\neq 0$ by definition of S_2. Thus, finally, we have that Q5 is equivalent to:

Q6: S_1 is an empty set, and there exists $\hat{T}(s,a,s')\in\mathbb{R}$ for each s and $s'\in S$, $q(s')\in\mathbb{R}$ for each $s'\in S_2$, and $Z\in\mathbb{R}$, s.t. Eqs. (B.1), (B.3), (B.4), (B.17), (B.18), and (B.20) hold.

Therefore, if S_1, S_2, and S_4 are empty (which is often the case), we obtain the IS-FEASIBLE procedure in Fig. 6. Otherwise, we need a little more (but almost negligible) computational cost to answer the Q6 correctly; we need to test if S_1 is empty and to include Eq. (B.20) as a constraint. We also need to use Eqs. (B.15) and (B.16) to set $\hat{O}(s',a,o)$ for $s'\in S_3$ and S_4, respectively. Still, note that we can easily answer Q6, since all the constraints are linear.

We can consider the point-set case in a similar manner. The query that we want to answer as true or false is formalized as follows:

Q7: There exists $\hat{T}(s,a,s')\in\mathbb{R}$ for each s and $s'\in S$, $\hat{O}(s',a,o)\in\mathbb{R}$ for each $s'\in S$, $\lambda_s^i\in[0,1]$ for all $s\in S$ and $i=1,\ldots,|T^M(s,a)|$, $\nu_{s'}^i\in[0,1]$ for all $s'\in S$ and $i=1,\ldots,|O^M(s',a)|$, and $Z\in\mathbb{R}$, s.t.

$$Z>0,$$

$$Z=\sum_{s,s'\in S}\hat{b}(s)\hat{T}(s,a,s')\hat{O}(s',a,o),$$

$$T(s,a,s')=\sum_i\lambda_s^i T_i^M(s,a,s') \quad \text{for all } s,s'\in S,$$

$$\sum_i\lambda_s^i=1 \quad \text{for all } s\in S,$$

$$O(s',a,o)=\sum_i\nu_{s'}^i O_i^M(s',a,o) \quad \text{for all } s'\in S, \tag{B.21}$$

$$\sum_i\nu_{s'}^i=1 \quad \text{for all } s'\in S, \tag{B.22}$$

$$\sum_{s\in S}\hat{b}(s)\hat{T}(s,a,s')\hat{O}(s',a,o)/Z=\hat{b}'(s') \quad \text{for all } s'\in S$$

hold.

The constraints, Eqs. (B.21) and (B.22) with $\nu_{s'}^i\in[0,1]$, are equivalent to Eq. (B.5) (note that a and o are fixed here). After replacing them with Eq. (B.5), we can obtain the final form in Fig. 6 in the same way as the interval case described above.

Appendix C. Proof of Theorem 1

Here, we provide the proof of Theorem 1. This proof is based on McAllester et al. [37].

The proof strategy is as follows. Note that Theorem 1 gives a bound on the error of the optimal value that can occur by using possibly-correct second-order beliefs instead of the correct ones. To prove Theorem 1, we will bound various errors that can occur by using possibly-correct second-order beliefs instead of the correct ones. After proving some basic properties of the norm and the probability functions in Lemmas 1–4, we first bound the error of the first-order belief after an action and an observation in Lemma 5. Next, using Lemma 5, we bound the error of the first-order belief

after t steps in Lemma 6. Subsequently, using Lemma 6, we bound the error of the value of a policy in Lemma 7. Finally, using Lemma 7, we bound the error of the optimal value to prove Theorem 1.

We begin by providing basic lemmas. Let $\|\cdot\|$ denote the L_1 norm; that is, for any probability function $P(s)$, let $\|P(s)\| := \sum_s |P(s)|$. For this norm, we have the following lemmas:

Lemma 1. *Let P and Q be two probability functions on the same set. We have*

$$\big\|P(s) - Q(s)\big\| \leqslant 2.$$

Proof.

$$\big\|P(s) - Q(s)\big\| = \sum_s \big|P(s) - Q(s)\big| \leqslant \sum_s P(s) + \sum_s Q(s) = 2. \qquad \square$$

Lemma 2. *(Modified from Lemma 15 in [37].) Let X and O be two sets. Let P and Q be any two probability functions on $X \times O$. Let $P(o)$ denote the marginal probability function on O, i.e., $P(o) = \sum_x P(x, o)$, and similarly for $Q(o)$. Let $P(x|o)$ be the conditional probability function on X, i.e., $P(x|o) = P(x, o)/P(o)$ if $P(o) \neq 0$. Let $P(x|o)$ be an arbitrary probability function if $P(o) = 0$. Similarly for $Q(x|o)$. We then have the following.*

$$E_{o\sim P(o)}\big\|P(x|o) - Q(x|o)\big\| \leqslant \big\|P(x, o) - Q(x, o)\big\| + \big\|P(o) - Q(o)\big\|.$$

Proof. Let O^+ be the set of o for which $P(o) \neq 0$ holds. First, we prove that for any $o \in O^+$, we have

$$P(o)\sum_{x\in X}\left|\frac{Q(x, o)}{P(o)} - Q(x|o)\right| = \big|P(o) - Q(o)\big|. \tag{C.1}$$

When $Q(o) \neq 0$, Eq. (C.1) holds because

$$\begin{aligned}
P(o)\sum_{x\in X}\left|\frac{Q(x, o)}{P(o)} - Q(x|o)\right| &= P(o)\sum_{x\in X}\left|\frac{Q(x, o)}{P(o)} - \frac{Q(x, o)}{Q(o)}\right| \\
&= P(o)\left|\frac{1}{P(o)} - \frac{1}{Q(o)}\right|\sum_{x\in X} Q(x, o) = P(o)\left|\frac{1}{P(o)} - \frac{1}{Q(o)}\right|Q(o) \\
&= \big|P(o) - Q(o)\big|.
\end{aligned}$$

When $Q(o) = 0$, Eq. (C.1) holds because

$$\begin{aligned}
P(o)\sum_{x\in X}\left|\frac{Q(x, o)}{P(o)} - Q(x|o)\right| &= P(o)\sum_{x\in X}\big|Q(x|o)\big| \quad (\because Q(x, o) = 0 \text{ when } Q(o) = 0) \\
&= P(o) = \big|P(o) - Q(o)\big|. \quad (\because Q(o) = 0)
\end{aligned}$$

Therefore, Eq. (C.1) holds for any $o \in O^+$.

Next, we prove the lemma as follows:

$$\begin{aligned}
E_{o\sim P(o)}\big\|P(x|o) - Q(x|o)\big\| &= \sum_{o\in O^+} P(o)\sum_{x\in X}\big|P(x, o)/P(o) - Q(x|o)\big| \\
&\leqslant \sum_{o\in O^+} P(o)\sum_{x\in X}\left|\frac{P(x, o)}{P(o)} - \frac{Q(x, o)}{P(o)}\right| + \sum_{o\in O^+} P(o)\sum_{x\in X}\left|\frac{Q(x, o)}{P(o)} - Q(x|o)\right| \\
&= \sum_{o\in O^+, x\in X}\big|P(x, o) - Q(x, o)\big| + \sum_{o\in O^+}\big|P(o) - Q(o)\big| \quad (\because \text{Eq. (C.1)}) \\
&\leqslant \big\|P(x, o) - Q(x, o)\big\| + \big\|P(o) - Q(o)\big\|. \qquad \square
\end{aligned}$$

Next, let us consider the hypothetical POMDP in which the correct second-order beliefs are specified (Section 3.2). For later use, let $\Pr(a|h, \mu)$ denote the probability of taking action $a \in A$, given history $h \in H$ and policy μ. Define

$P(s', h'|h, s, t, \mu)$ as the probability that action-observation sequence h' is generated and the final state is s', given that time t has elapsed after history h had ended with state s and that policy μ is used. For a first-order belief b, define $P(s', h'|h, b, t, \mu) := \sum_{s\in S} b(s) P(s', h'|h, s, t, \mu)$. Note that h' is an action-observation sequence whose length is t; to denote this, we say $h' \in H_t$, with a little abuse of notation.

Let us define $P(h'|h, b, t, \mu)$ as $P(h'|h, b, t, \mu) := \sum_{s'\in S} P(s', h'|h, b, t, \mu)$. Also, define $P(s'|h', h, b, t, \mu)$ as $P(s'|h', h, b, t, \mu) := P(s', h'|h, b, t, \mu)/P(h'|h, b, t, \mu)$ if $P(h'|h, b, t, \mu) \neq 0$. If $P(h'|h, b, t, \mu) = 0$, let $P(s'|h', h, b, t, \mu)$ be an arbitrary probability function on S. We denote $P(s'|h', h, b, t, \mu)$ by $P(s'|h', h, b)$ for short, since t and μ are redundant, given h', h, and b. For these functions, we have the following bounds:

Lemma 3. *Let b and $\hat{b}$ be any first-order beliefs, t be any elapsed time, μ be any policy, and h be any history. Then we have*

$$\left\| P(h'|h, b, t, \mu) - P(h'|h, \hat{b}, t, \mu) \right\| \leqslant \|b - \hat{b}\|.$$

Proof.

$$\begin{aligned}\left\| P(h'|h, b, t, \mu) - P(h'|h, \hat{b}, t, \mu) \right\| &= \sum_{h'\in H_t} \left| \sum_{s\in S} P(h'|h, s, t, \mu) b(s) - \sum_{s\in S} P(h'|h, s, t, \mu) \hat{b}(s) \right| \\ &\leqslant \sum_{h'\in H_t} \sum_{s\in S} P(h'|h, s, t, \mu) \left| b(s) - \hat{b}(s) \right| = \|b - \hat{b}\|. \qquad \square\end{aligned}$$

Lemma 4. *Let b and $\hat{b}$ be any first-order beliefs, t be any elapsed time, μ be any policy, and h be any history. Then we have*

$$\sum_{h'\in H_t} P(h'|h, b, t, \mu) \left\| P(s'|h', h, b) - P(s'|h', h, \hat{b}) \right\| \leqslant 2\|b - \hat{b}\|.$$

Proof.

$$\begin{aligned}&\sum_{h'\in H_t} P(h'|h, b, t, \mu) \left\| P(s'|h', h, b) - P(s'|h', h, \hat{b}) \right\| \\ &\quad\leqslant \left\| P(s', h'|h, b, t, \mu) - P(s', h'|h, \hat{b}, t, \mu) \right\| \\ &\qquad + \left\| P(h'|h, b, t, \mu) - P(h'|h, \hat{b}, t, \mu) \right\| \quad (\because \text{Lemma 2}) \\ &\quad\leqslant \|b - \hat{b}\| \quad (\because \text{proof is similar to Lemma 3}) \\ &\qquad + \|b - \hat{b}\| \quad (\because \text{Lemma 3}) \\ &\quad= 2\|b - \hat{b}\|. \qquad \square\end{aligned}$$

Next, we prove the bounds on the errors of the possibly-correct first-order beliefs in Section 3.3, compared to the correct beliefs of the hypothetical POMDPs. In the following, suppose that the possibly-correct second-order beliefs $\hat{b}^M_{h,a,o}$ and $\hat{b}^M_h$ have been determined for every $h \in H$, $a \in A$, and $o \in \Theta$. First, we provide a bound on the error that may be caused by a one-step update of a first-order belief.

Lemma 5. *Let b be any first-order belief. Let h be any history. Let b' and $\hat{b}'$ be the correct and possibly-correct beliefs which are Bayes-updated from b, i.e., $b' := \tau(b, a, o, b^M_h)$ and $\hat{b}' := \tau(b, a, o, \hat{b}^M_{h,a,o})$, respectively. Let b' be an arbitrary probability function on S if $\tau(b, a, o, b^M_h)$ cannot be calculated (i.e., if the denominator is zero in Eq.* (14)*). Similarly for $\hat{b}'$. Define d as Eq.* (43)*. Then we have*

$$\sum_{a\in A, o\in\Theta} P\big(\langle a, o\rangle|h, b, 1, \mu\big) \|b' - \hat{b}'\| \leqslant 4d.$$

Proof. First, define $Q(s', o|h, b, a)$ and $Q(o|h, b, a)$ as

$$Q(s', o|h, b, a) := \int_{m_h=(T_h,O_h)\in M} O_h(s', a, o) \sum_{s\in S} T_h(s, a, s')b(s)b_h^M(m_h)\,\mathrm{d}m_h$$

and

$$Q(o|h, b, a) := \sum_{s'\in S} Q(s', o|h, b, a).$$

Then we have

$$b'(s) = \frac{Q(s', o|h, b, a)}{Q(o|h, b, a)} \quad \text{for all } s \in S \tag{C.2}$$

when $Q(o|h, b, a) \neq 0$. Similarly, define $\hat{Q}(s', o|h, b, a)$ and $\hat{Q}(o|h, b, a)$ as

$$\hat{Q}(s', o|h, b, a) := \int_{m_h=(T_h,O_h)\in M} O_h(s', a, o) \sum_{s\in S} T_h(s, a, s')b(s)\hat{b}_{h,a,o}^M(m_h)\,\mathrm{d}m_h$$

and

$$\hat{Q}(o|h, b, a) := \sum_{s'\in S} \hat{Q}(s', o|h, b, a).$$

Then we have

$$\hat{b}' = \frac{\hat{Q}(s', o|h, b, a)}{\hat{Q}(o|h, b, a)} \quad \text{for all } s \in S \tag{C.3}$$

when $\hat{Q}(o|h, b, a) \neq 0$. Next, define their differences as

$$\epsilon(s', o|h, b, a) := Q(s', o|h, b, a) - \hat{Q}(s', o|h, b, a)$$

and

$$\epsilon(o|h, b, a) := Q(o|h, b, a) - \hat{Q}(o|h, b, a).$$

Note that

$$\sum_{a\in A,\ o\in\Theta} P\big(\langle a, o\rangle|h, b, 1, \mu\big)\|b' - \hat{b}'\| = \sum_{a\in A} \Pr(a|h, \mu) \sum_{o\in\Theta} Q(o|h, b, a)\|b' - \hat{b}'\| \tag{C.4}$$

holds. For any $h \in H$, first-order belief b, and $a \in A$, if $o \in \Theta$ satisfies $Q(o|h, b, a) \leqslant 2|\epsilon(o|h, b, a)|$, then we have

$$\begin{aligned} Q(o|h, b, a)\|b' - \hat{b}'\| &\leqslant 4\big|\epsilon(o|h, b, a)\big| \quad (\because \text{Lemma 1 and assumption}) \\ &\leqslant 4\sum_{s'\in S}\big|\epsilon(s', o|h, b, a)\big|. \end{aligned} \tag{C.5}$$

If $o \in \Theta$ satisfies $Q(o|h, b, a) > 2|\epsilon(o|h, b, a)|$, then $Q(o|h, b, a) \neq 0$ and $\hat{Q}(o|h, b, a) \neq 0$ should hold (proof by contradiction), and hence we have

$$\begin{aligned} Q(o|h, b, a)\|b' - \hat{b}'\| &= Q(o|h, b, a)\sum_{s'\in S}\left|\frac{Q(s', o|h, b, a)}{Q(o|h, b, a)} - \frac{\hat{Q}(s', o|h, b, a)}{\hat{Q}(o|h, b, a)}\right| \quad (\because \text{Eqs. (C.2) and (C.3)}) \\ &= Q(o|h, b, a)\sum_{s'\in S}\left|\frac{Q(s', o|h, b, a)}{Q(o|h, b, a)} - \frac{Q(s', o|h, b, a) + \epsilon(s', o|h, b, a)}{Q(o|h, b, a) + \epsilon(o|h, b, a)}\right| \\ &= \sum_{s'\in S}\left|\frac{Q(s', o|h, b, a)\epsilon(o|h, b, a) - Q(o|h, b, a)\epsilon(s', o|h, b, a)}{Q(o|h, b, a) + \epsilon(o|h, b, a)}\right| \end{aligned}$$

$$
\begin{aligned}
&< 2\sum_{s'\in S}\left|\frac{Q(s',o|h,b,a)\epsilon(o|h,b,a)-Q(o|h,b,a)\epsilon(s',o|h,b,a)}{Q(o|h,b,a)}\right| \quad (\because \text{assumption})\\
&= 2\sum_{s'\in S}\left|b'(s')\epsilon(o|h,b,a)-\epsilon(s',o|h,b,a)\right|\\
&\leqslant 2\left|\epsilon(o|h,b,a)\right|+2\sum_{s'\in S}\left|\epsilon(s',o|h,b,a)\right|\leqslant 4\sum_{s'\in S}\left|\epsilon(s',o|h,b,a)\right|.
\end{aligned}
$$

Thus, together with Eq. (C.5), we have

$$
Q(o|h,b,a)\|b'-\hat{b}'\|\leqslant 4\sum_{s'\in S}\left|\epsilon(s',o|h,b,a)\right| \tag{C.6}
$$

for any $h\in H$, first-order belief b, $a\in A$, and $o\in\Theta$.

Note that we have

$$
\begin{aligned}
&\sum_{o\in\Theta}\sum_{s'\in S}\left|\epsilon(s',o|h,b,a)\right|=\sum_{o\in\Theta}\sum_{s'\in S}\left|Q(s',o|h,b,a)-\hat{Q}(s',o|h,b,a)\right|\\
&\quad=\sum_{o\in\Theta}\sum_{s'\in S}\left|\int_{m_h=(T_h,O_h)\in M} O_h(s',a,o)\sum_{s\in S}T_h(s,a,s')b(s)b_h^M(m_h)\,\mathrm{d}m_h\right.\\
&\qquad\left.-\int_{m_h=(T_h,O_h)\in M} O_h(s',a,o)\sum_{s\in S}T_h(s,a,s')b(s)\hat{b}_{h,a,o}^M(m_h)\,\mathrm{d}m_h\right|\\
&\quad\leqslant\sum_{o\in\Theta}\sum_{s'\in S}\sum_{s\in S}b(s)\max_{(T^1,O^1),(T^2,O^2)\in M}\left|O^1(s',a,o)T^1(s,a,s')-O^2(s',a,o)T^2(s,a,s')\right|\\
&\quad=\sum_{o\in\Theta}\sum_{s'\in S}\sum_{s\in S}b(s)\max_{(T^0+\epsilon_T^1,O^0+\epsilon_O^1),(T^0+\epsilon_T^2,O^0+\epsilon_O^2)\in M}\left|\big(O^0(s',a,o)+\epsilon_O^1(s',a,o)\big)\big(T^0(s,a,s')+\epsilon_T^1(s,a,s')\big)\right.\\
&\qquad\left.-\big(O^0(s',a,o)+\epsilon_O^2(s',a,o)\big)\big(T^0(s,a,s')+\epsilon_T^2(s,a,s')\big)\right|\\
&\quad\text{(where we let } (T^0,\ O^0) \text{ be an arbitrary model in } M)\\
&\quad=\sum_{o\in\Theta}\sum_{s'\in S}\sum_{s\in S}b(s)\max_{(T^0+\epsilon_T^1,O^0+\epsilon_O^1),(T^0+\epsilon_T^2,O^0+\epsilon_O^2)\in M}\left|O^0(s',a,o)\big(\epsilon_T^1(s,a,s')-\epsilon_T^2(s,a,s')\big)\right.\\
&\qquad\left.+T^0(s,a,s')\big(\epsilon_O^1(s',a,o)-\epsilon_O^2(s',a,o)\big)+\epsilon_O^1(s',a,o)\epsilon_T^1(s,a,s')-\epsilon_O^2(s',a,o)\epsilon_T^2(s,a,s')\right|\\
&\quad\leqslant\sum_{o\in\Theta}\sum_{s'\in S}\sum_{s\in S}b(s)\big(O^0(s',a,o)\epsilon_{\max}^T+T^0(s,a,s')\epsilon_{\max}^O+\epsilon_{\max}^T\epsilon_{\max}^O+\epsilon_{\max}^T\epsilon_{\max}^O\big)\\
&\quad=|S|\epsilon_{\max}^T+|\Theta|\epsilon_{\max}^O+2|S||\Theta|\epsilon_{\max}^T\epsilon_{\max}^O:=d.
\end{aligned}
\tag{C.7}
$$

The desired result is obtained by combining this result with Eqs. (C.4) and (C.6). □

Next, we provide the following error bound on the first-order beliefs after any time $t\geqslant 0$. Let $E_{h\in H_t}^{\mu}\{\cdot\}$ be the expectation over all the t-length histories $h\in H_t$, each of which occurs with the probability $P(h|\emptyset,p_0,t,\mu)$.

Lemma 6. *Let b_h be the correct belief of the hypothetical POMDP. Let $\hat{b}_h$ be the possibly-correct belief*; *let $\hat{b}_h$ be an arbitrary probability function on S if it cannot be calculated* (*i.e., if the denominator is zero in Eq.* (14)). *For any $t\geqslant 0$, we have*

$$
E_{h\in H_t}^{\mu}\|b_h-\hat{b}_h\|\leqslant 16dt.
$$

Proof. First, we define a *generalized* possibly-correct first-order belief, $\hat{P}(s'|h',h,b)$, which is calculated by Bayes-updating b repeatedly with the possibly-correct second-order beliefs, assuming that b is the belief after history h. That is, let $\hat{P}(s'|\emptyset,h,b)=b(s')$ hold for any $s'\in S$, $h\in H$, and first-order belief b. Also, if we have $\hat{b}':=\tau(b,a,o,\hat{b}_{h,a,o}^M)$, then let $\hat{P}(s'|\langle a,o\rangle;h'',h,b)=\hat{P}(s'|h'',h;\langle a,o\rangle,\hat{b}')$ hold, for any $s'\in S$, $h,h''\in H$, $a\in A$ and $o\in\Theta$, where

$\langle a, o\rangle; h''$ is the action-observation history in which h'' follows a and o. If we cannot calculate $\tau(b, a, o, \hat{b}^M_{h,a,o})$ due to a zero denominator in Eq. (14), then let $\hat{P}(s'|\langle a, o\rangle; h'', h, b)$ be an arbitrary probability function on S.

Next, note that

$$E^{\mu}_{h\in H_t}\|b_h - \hat{b}_h\| = \sum_{h\in H_t} P(h|\emptyset, p_0, t, \mu)\big\|P(s'|h, \emptyset, p_0) - \hat{P}(s'|h, \emptyset, p_0)\big\|$$

holds. Thus, we need to prove that the right-hand side of this equation is not larger than $16dt$. We prove here a more general equation:

$$\sum_{h'\in H_t} P(h'|h, b, t, \mu)\big\|P(s'|h', h, b) - \hat{P}(s'|h', h, b)\big\| \leqslant 16dt \tag{C.8}$$

for any $h \in H$, first-order belief b, $t \geqslant 0$, and policy μ. The proof is by induction on t. Eq. (C.8) holds for $t = 0$, since we have $P(s'|\emptyset, h, b) = \hat{P}(s'|\emptyset, h, b) = b(s')$. Assume that Eq. (C.8) holds for t. For $t + 1$, we have

$$\begin{aligned}
&\sum_{h'\in H_{t+1}} P(h'|h, b, t+1, \mu)\big\|P(s'|h', h, b) - \hat{P}(s'|h', h, b)\big\| \\
&= \sum_{a\in A,\ o\in\Theta} P\big(\langle a, o\rangle|h, b, 1, \mu\big) \sum_{h''\in H_t} P\big(h''|h; \langle a, o\rangle, b', t, \mu\big) \\
&\quad\times\big\|P\big(s'|h'', h; \langle a, o\rangle, b'\big) - \hat{P}\big(s'|h'', h; \langle a, o\rangle, \hat{b}'\big)\big\| \\
&\quad(\because \text{Divide } h' \text{ as } h' = \langle a, o\rangle; h''.\ \text{Define } b' \text{ and } \hat{b}' \text{ as in Lemma 5.}) \\
&= \sum_{a\in A,\ o\in\Theta} P\big(\langle a, o\rangle|h, b, 1, \mu\big) \sum_{h''\in H_t} P\big(h''|h; \langle a, o\rangle, b', t, \mu\big) \\
&\quad\times\big\|P\big(s'|h'', h; \langle a, o\rangle, b'\big) - P\big(s'|h'', h; \langle a, o\rangle, \hat{b}'\big)\big\| \\
&\quad+ \sum_{a\in A,\ o\in\Theta} P\big(\langle a, o\rangle|h, b, 1, \mu\big) \sum_{h''\in H_t} P\big(h''|h; \langle a, o\rangle, b', t, \mu\big) \\
&\quad\times\big\|P\big(s'|h'', h; \langle a, o\rangle, \hat{b}'\big) - \hat{P}\big(s'|h'', h; \langle a, o\rangle, \hat{b}'\big)\big\| \\
&\leqslant 2 \sum_{a\in A,\ o\in\Theta} P\big(\langle a, o\rangle|h, b, 1, \mu\big)\|b' - \hat{b}'\| \quad(\because \text{Lemma 4}) \\
&\quad+ \sum_{a\in A,\ o\in\Theta} P\big(\langle a, o\rangle|h, b, 1, \mu\big) \\
&\quad\times \sum_{h''\in H_t} \big|P\big(h''|h; \langle a, o\rangle, b', t, \mu\big) - P\big(h''|h; \langle a, o\rangle, \hat{b}', t, \mu\big)\big| \\
&\quad\times\big\|P\big(s'|h'', h; \langle a, o\rangle, \hat{b}'\big) - \hat{P}\big(s'|h'', h; \langle a, o\rangle, \hat{b}'\big)\big\| \\
&\quad+ \sum_{a\in A,\ o\in\Theta} P\big(\langle a, o\rangle|h, b, 1, \mu\big) \sum_{h''\in H_t} P\big(h''|h; \langle a, o\rangle, \hat{b}', t, \mu\big) \\
&\quad\times\big\|P\big(s'|h'', h; \langle a, o\rangle, \hat{b}'\big) - \hat{P}\big(s'|h'', h; \langle a, o\rangle, \hat{b}'\big)\big\| \\
&\leqslant 2 \sum_{a\in A,\ o\in\Theta} P\big(\langle a, o\rangle|h, b, 1, \mu\big)\|b' - \hat{b}'\| \\
&\quad+ 2 \sum_{a\in A,\ o\in\Theta} P\big(\langle a, o\rangle|h, b, 1, \mu\big)\|b' - \hat{b}'\| \quad(\because \text{Lemmas 1 and 3}) \\
&\quad+ \sum_{a\in A,\ o\in\Theta} P\big(\langle a, o\rangle|h, b, 1, \mu\big) \sum_{h''\in H_t} P\big(h''|h; \langle a, o\rangle, \hat{b}', t, \mu\big) \\
&\quad\times\big\|P\big(s'|h'', h; \langle a, o\rangle, \hat{b}'\big) - \hat{P}\big(s'|h'', h; \langle a, o\rangle, \hat{b}'\big)\big\| \\
&\leqslant 8d + 8d \quad(\because \text{Lemma 5})
\end{aligned}$$

$$+ 16dt \quad (\because \text{assumption})$$
$$= 16d(t+1).$$

Thus, Eq. (C.8) holds for $t+1$. This completes the proof. □

Next, we provide an error bound on the values of a given policy. For a policy μ, let $V^\mu(h)$ be the *value* of history h for the hypothetical POMDP, which is defined as the solution to

$$V^\mu(h) = \sum_{a\in A} \Pr(a|h,\mu)\left\{\rho(h,a) + \gamma \sum_{o\in\Theta} P(o|h,a) V^\mu\big(h;\langle a,o\rangle\big)\right\},$$

where $\rho(h,a)$ and $P(o|h,a)$ are defined as Eqs. (17) and (18), respectively. Similarly, define the *quasi-value* of history h, which we denote by $\hat{V}^\mu(h)$, as the solution to

$$\hat{V}^\mu(h) = \sum_{a\in A} \Pr(a|h,\mu)\left\{\hat{\rho}(h,a) + \gamma \sum_{o\in\Theta} \hat{P}(o|h,a) \hat{V}^\mu\big(h;\langle a,o\rangle\big)\right\},$$

where $\hat{\rho}(h,a)$ and $\hat{P}(o|h,a)$ are defined as Eqs. (23) and (24), respectively.

Let us call $V^\mu := V^\mu(\emptyset)$ the *value of policy* μ, and $\hat{V}^\mu := \hat{V}^\mu(\emptyset)$ the *quasi-value of policy* μ. Let $R_{\max}$ and $\hat{V}^\mu_{\max}$ be defined as Eq. (44) and Eq. (45), respectively. For any policy $\mu : H \to A$, define W^μ as

$$W^\mu := \frac{((1-\gamma)|S|\epsilon^T_{\max} + 16\gamma d)R_{\max} + (1+15\gamma)\gamma d\hat{V}^\mu_{\max}}{(1-\gamma)^2}.$$

Then we have the following.

Lemma 7. *For any policy μ, we have*

$$|V^\mu - \hat{V}^\mu| \leqslant W^\mu.$$

Proof. First, we prove that

$$E^\mu_{h\in H_t}\big|V^\mu(h) - \hat{V}^\mu(h)\big| \leqslant \gamma E^\mu_{h'\in H_{t+1}}\big|V^\mu(h') - \hat{V}^\mu(h')\big| + 16d(R_{\max} + \gamma\hat{V}^\mu_{\max})t + |S|\epsilon^T_{\max}R_{\max} + \gamma d\hat{V}^\mu_{\max}. \tag{C.9}$$

To prove this, note that we have

$$\begin{aligned}
&E^\mu_{h\in H_t}\big|V^\mu(h) - \hat{V}^\mu(h)\big| \\
&= E^\mu_{h\in H_t}\sum_{a\in A}\Pr(a|h,\mu)\left|\rho(h,a) + \gamma\sum_{o\in\Theta}P(o|h,a)V^\mu\big(h;\langle a,o\rangle\big) - \hat{\rho}(h,a) + \gamma\sum_{o\in\Theta}\hat{P}(o|h,a)\hat{V}^\mu\big(h;\langle a,o\rangle\big)\right| \\
&\leqslant E^\mu_{h\in H_t}\sum_{a\in A}\Pr(a|h,\mu)\big|\rho(h,a) - \hat{\rho}(h,a)\big| \\
&\quad + \gamma E^\mu_{h\in H_t}\sum_{a\in A}\Pr(a|h,\mu)\left|\sum_{o\in\Theta}P(o|h,a)V^\mu\big(h;\langle a,o\rangle\big) - \sum_{o\in\Theta}P(o|h,a)\hat{V}^\mu\big(h;\langle a,o\rangle\big)\right| \\
&\quad + \gamma E^\mu_{h\in H_t}\sum_{a\in A}\Pr(a|h,\mu)\left|\sum_{o\in\Theta}P(o|h,a)\hat{V}^\mu\big(h;\langle a,o\rangle\big) - \sum_{o\in\Theta}\hat{P}(o|h,a)\hat{V}^\mu\big(h;\langle a,o\rangle\big)\right| \\
&\leqslant E^\mu_{h\in H_t}\sum_{a\in A}\Pr(a|h,\mu)\big|\rho(h,a) - \hat{\rho}(h,a)\big| \\
&\quad + \gamma E^\mu_{h\in H_t}\sum_{a\in A}\Pr(a|h,\mu)\sum_{o\in\Theta}P(o|h,a)\big|V^\mu\big(h;\langle a,o\rangle\big) - \hat{V}^\mu\big(h;\langle a,o\rangle\big)\big| \\
&\quad + \gamma E^\mu_{h\in H_t}\sum_{a\in A}\Pr(a|h,\mu)\left|\sum_{o\in\Theta}P(o|h,a) - \sum_{o\in\Theta}\hat{P}(o|h,a)\right|\hat{V}^\mu_{\max},
\end{aligned}$$

for which we have

$$
\begin{aligned}
&E^{\mu}_{h\in H_t}\sum_{a\in A}\Pr(a|h,\mu)\big|\rho(h,a)-\hat{\rho}(h,a)\big| \\
&\quad= E^{\mu}_{h\in H_t}\sum_{a\in A}\Pr(a|h,\mu)\Bigg|\int\limits_{m_h=(T_h,O_h)\in M}\sum_{s\in S}\sum_{s'\in S}R(s,a,s')T_h(s,a,s')b_h(s)\hat{b}^M_h(m_h)\,\mathrm{d}m_h \\
&\qquad-\int\limits_{m_h=(T_h,O_h)\in M}\sum_{s\in S}\sum_{s'\in S}R(s,a,s')T_h(s,a,s')\hat{b}_h(s)\hat{b}^M_h(m_h)\,\mathrm{d}m_h\Bigg| \\
&\quad\leqslant R_{\max}E^{\mu}_{h\in H_t}\sum_{a\in A}\Pr(a|h,\mu)\sum_{s\in S}\sum_{s'\in S}\Bigg|b_h(s)\int\limits_{m_h=(T_h,O_h)\in M}T_h(s,a,s')b^M_h(m_h)\,\mathrm{d}m_h \\
&\qquad-\hat{b}_h(s)\int\limits_{m_h=(T_h,O_h)\in M}T_h(s,a,s')\hat{b}^M_h(m_h)\,\mathrm{d}m_h\Bigg| \\
&\quad\leqslant R_{\max}E^{\mu}_{h\in H_t}\sum_{a\in A}\Pr(a|h,\mu)\sum_{s\in S}\sum_{s'\in S}\Bigg|b_h(s)\int\limits_{m_h=(T_h,O_h)\in M}T_h(s,a,s')b^M_h(m_h)\,\mathrm{d}m_h \\
&\qquad-\hat{b}_h(s)\int\limits_{m_h=(T_h,O_h)\in M}T_h(s,a,s')b^M_h(m_h)\,\mathrm{d}m_h\Bigg| \\
&\qquad+R_{\max}E^{\mu}_{h\in H_t}\sum_{a\in A}\Pr(a|h,\mu)\sum_{s\in S}\sum_{s'\in S}\Bigg|\hat{b}_h(s)\int\limits_{m_h=(T_h,O_h)\in M}T_h(s,a,s')b^M_h(m_h)\,\mathrm{d}m_h \\
&\qquad-\hat{b}_h(s)\int\limits_{m_h=(T_h,O_h)\in M}T_h(s,a,s')\hat{b}^M_h(m_h)\,\mathrm{d}m_h\Bigg| \\
&\quad\leqslant R_{\max}E^{\mu}_{h\in H_t}\|b_h-\hat{b}_h\|+R_{\max}E^{\mu}_{h\in H_t}\sum_{a\in A}\Pr(a|h,\mu)\sum_{s\in S}\hat{b}_h(s)\sum_{s'\in S}\Bigg|\int\limits_{m_h=(T_h,O_h)\in M}T_h(s,a,s')b^M_h(m_h)\,\mathrm{d}m_h \\
&\qquad-\int\limits_{m_h=(T_h,O_h)\in M}T_h(s,a,s')\hat{b}^M_h(m_h)\,\mathrm{d}m_h\Bigg| \\
&\quad\leqslant \big(16dt+|S|\epsilon^T_{\max}\big)R_{\max}\quad(\because\text{ by Lemma 6 and by a proof similar to Eq. (C.7)),}
\end{aligned}
$$

and

$$
\gamma E^{\mu}_{h\in H_t}\sum_{a\in A}\Pr(a|h,\mu)\sum_{o\in\Theta}P(o|h,a)\big|V^{\mu}\big(h;\langle a,o\rangle\big)-\hat{V}^{\mu}\big(h;\langle a,o\rangle\big)\big|=\gamma E^{\mu}_{h'\in H_{t+1}}\big|V^{\mu}(h')-\hat{V}^{\mu}(h')\big|,
$$

and also

$$
\begin{aligned}
&E^{\mu}_{h\in H_t}\sum_{a\in A}\Pr(a|h,\mu)\Bigg|\sum_{o\in\Theta}P(o|h,a)-\sum_{o\in\Theta}\hat{P}(o|h,a)\Bigg| \\
&\quad= E^{\mu}_{h\in H_t}\sum_{a\in A}\Pr(a|h,\mu)\Bigg|\sum_{o\in\Theta}\int\limits_{m_h=(T_h,O_h)\in M}\sum_{s'\in S}O_h(s',a,o)\sum_{s\in S}T_h(s,a,s')b_h(s)b^M_h(m_h)\,\mathrm{d}m_h \\
&\qquad-\sum_{o\in\Theta}\int\limits_{m_h=(T_h,O_h)\in M}\sum_{s'\in S}O_h(s',a,o)\sum_{s\in S}T_h(s,a,s')\hat{b}_h(s)\hat{b}^M_h(m_h)\,\mathrm{d}m_h\Bigg| \\
&\quad\leqslant E^{\mu}_{h\in H_t}\sum_{a\in A}\Pr(a|h,\mu)\Bigg|\sum_{o\in\Theta}\int\limits_{m_h=(T_h,O_h)\in M}\sum_{s'\in S}O_h(s',a,o)\sum_{s\in S}T_h(s,a,s')b_h(s)b^M_h(m_h)\,\mathrm{d}m_h
\end{aligned}
$$

$$
\begin{aligned}
&- \sum_{o\in\Theta} \int_{m_h=(T_h,O_h)\in M} \sum_{s'\in S} O_h(s',a,o) \sum_{s\in S} T_h(s,a,s')\hat{b}_h(s) b_h^M(m_h)\,\mathrm{d}m_h \Big| \\
&+ E^{\mu}_{h\in H_t} \sum_{a\in A} \Pr(a|h,\mu) \Big| \sum_{o\in\Theta} \int_{m_h=(T_h,O_h)\in M} \sum_{s'\in S} O_h(s',a,o) \sum_{s\in S} T_h(s,a,s')\hat{b}_h(s) b_h^M(m_h)\,\mathrm{d}m_h \\
&- \sum_{o\in\Theta} \int_{m_h=(T_h,O_h)\in M} \sum_{s'\in S} O_h(s',a,o) \sum_{s\in S} T_h(s,a,s')\hat{b}_h(s) \hat{b}_h^M(m_h)\,\mathrm{d}m_h \Big| \\
\leqslant\ & E^{\mu}_{h\in H_t} \|b_h - \hat{b}_h\| \\
&+ E^{\mu}_{h\in H_t} \sum_{a\in A} \Pr(a|h,\mu) \sum_{o\in\Theta} \sum_{s'\in S} \Big| \int_{m_h=(T_h,O_h)\in M} O_h(s',a,o) \sum_{s\in S} T_h(s,a,s')\hat{b}_h(s) b_h^M(m_h)\,\mathrm{d}m_h \\
&- \int_{m_h=(T_h,O_h)\in M} O_h(s',a,o) \sum_{s\in S} T_h(s,a,s')\hat{b}_h(s) \hat{b}_h^M(m_h)\,\mathrm{d}m_h \Big| \\
\leqslant\ & 16dt + d \quad (\because \text{by Lemma 6 and by a proof similar to Eq. (C.7)}).
\end{aligned}
$$

Taken together, we obtain

$$
E^{\mu}_{h\in H_t} \big| V^{\mu}(h) - \hat{V}^{\mu}(h) \big| \leqslant \gamma E^{\mu}_{h'\in H_{t+1}} \big| V^{\mu}(h') - \hat{V}^{\mu}(h') \big| + \big(16dt + |S|\epsilon^T_{\max}\big) R_{\max} + \gamma(16dt + d)\hat{V}^{\mu}_{\max}
$$

and hence Eq. (C.9).

Next, let us re-write this equation by defining $Y_t := E^{\mu}_{h\in H_t} |V^{\mu}(h) - \hat{V}^{\mu}(h)|$, $A := 16d(R_{\max} + \gamma \hat{V}^{\mu}_{\max})$, and $B := |S|\epsilon^T_{\max} R_{\max} + \gamma d \hat{V}^{\mu}_{\max}$. Then we have

$$
\begin{aligned}
Y_0 &\leqslant \gamma Y_1 + B \leqslant \gamma(\gamma Y_2 + A + B) + B \leqslant \gamma\big(\gamma(\gamma Y_3 + 2A + B) + A + B\big) + B \leqslant \cdots \\
&\leqslant A(\gamma + 2\gamma^2 + \cdots) + B(1 + \gamma + \gamma^2 + \cdots) = A\left(\frac{1}{(1-\gamma)^2} - \frac{1}{(1-\gamma)}\right) + B\left(\frac{1}{1-\gamma}\right).
\end{aligned}
$$

Substituting back $A = 16d(R_{\max} + \gamma \hat{V}^{\mu}_{\max})$ and $B = |S|\epsilon^T_{\max} R_{\max} + \gamma d \hat{V}^{\mu}_{\max}$ into this equation proves the Lemma (note that $|V^{\mu} - \hat{V}^{\mu}|$ equals to Y_0). □

Now, we can prove Theorem 1 as

$$
\begin{aligned}
V^{\mu^*} - V^{\hat{\mu}^*} &\leqslant V^{\mu^*} - \hat{V}^{\hat{\mu}^*} + W^{\hat{\mu}^*} \quad (\because \text{Lemma 7}) \\
&\leqslant V^{\mu^*} - \hat{V}^{\mu^*} + W^{\hat{\mu}^*} \quad (\because \text{Definition of } \hat{\mu}^*) \\
&\leqslant V^{\mu^*} - V^{\mu^*} + W^{\hat{\mu}^*} + W^{\mu^*} \quad (\because \text{Lemma 7}) \\
&= W^{\hat{\mu}^*} + W^{\mu^*}. \qquad \square
\end{aligned}
$$

References

[1] D. Aberdeen, J. Baxter, Scaling internal-state policy-gradient methods for POMDPs, in: International Conference on Machine Learning (ICML-02), Sydney, Australia, July 2002, pp. 1–12.

[2] K.J. Aström, Optimal control of Markov decision processes with incomplete state estimation, Journal of Mathematical Analysis and Applications 10 (1965) 174–205.

[3] T. Augustin, On the suboptimality of the generalized Bayes rule and robust Bayesian procedures from the decision theoretic point of view—a cautionary note on updating imprecise priors, in: Proceedings of 3rd International Symposium on Imprecise Probabilities and their Applications (ISIPTA-03), 2003.

[4] R. Bellman, Dynamic Programming, Princeton Univ. Press, Princeton, NJ, 1957.

[5] J.M. Bernard, T. Seidenfeld, M. Zaffalon (Eds.), Proceedings of the Third International Symposium in Imprecise Probabilities and its Applications, Carleton Scientific, 2003.

[6] D.P. Bertsekas, Dynamic Programming and Optimal Control, vol. 2, second ed., Athena Scientific, Belmont, MA, 2001.

[7] B. Bonet, An epsilon-optimal grid-based algorithm for partially observable Markov decision processes, in: Proc. 19th International Conf. on Machine Learning (ICML-02), Morgan Kaufmann, 2002, pp. 51–58.

[8] C. Boutilier, T. Dean, S. Hanks, Decision-theoretic planning: Structural assumptions and computational leverage, Journal of Artificial Intelligence Research 11 (1999) 1–94.
[9] C. Boutilier, D. Poole, Computing optimal policies for partially observable decision processes using compact representations, in: Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI-96), Portland, OR, AAAI Press/The MIT Press, 1996, pp. 1168–1175.
[10] J. Breese, K. Fertig, Decision making with interval influence diagrams, in: Proceedings of the 6th Annual Conference on Uncertainty in Artificial Intelligence (UAI-91), New York, Elsevier Science, 1991, pp. 467–478.
[11] A. Cassandra, M.L. Littman, N.L. Zhang, Incremental Pruning: A simple, fast, exact method for partially observable Markov decision processes, in: D. Geiger, P.P. Shenoy (Eds.), Proceedings of the Thirteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI-97), San Francisco, CA, Morgan Kaufmann, 1997, pp. 54–61.
[12] L. Chrisman, Independence with lower and upper probabilities, in: Proceedings of the 12th Annual Conference on Uncertainty in Artificial Intelligence (UAI-96), San Francisco, CA, Morgan Kaufmann, 1996, pp. 169–177.
[13] F.G. Cozman, Credal networks, Artificial Intelligence 120 (2000) 199–233.
[14] F.G. Cozman, E. Krotkov, Quasi-Bayesian strategies for efficient plan generation: application to the 'planning to observe' problem, in: E. Horvitz, F.V. Jensen (Eds.), Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence (UAI-96), San Francisco, CA, Morgan Kaufmann, 1996, pp. 186–193.
[15] A. Drake, Observation of a Markov process through a noisy channel, PhD thesis, Massachusetts Institute of Technology, 1962.
[16] Z. Feng, E.A. Hansen, Approximate planning for factored POMDPs, in: Proceedings of the 6th European Conference on Planning (ECP-01), Toledo, Spain, September 2001.
[17] K. Fertig, J. Breese, Interval influence diagrams, in: Proceedings of the 5th Annual Conference on Uncertainty in Artificial Intelligence (UAI-90), New York, Elsevier Science, 1990, pp. 149–161.
[18] K.W. Fertig, J.S. Breese, Probability intervals over influence diagrams, IEEE Transactions on Pattern Analysis and Machine Intelligence 15 (3) (1993) 280–286.
[19] H. Gaifman, A theory of higher order probabilities, in: Proceedings of the 1986 Conference on Theoretical Aspects of Reasoning about Knowledge, Morgan Kaufmann, 1986, pp. 275–292.
[20] R. Givan, S.M. Leach, T. Dean, Bounded-parameter Markov decision processes, Artificial Intelligence 122 (1–2) (2000) 71–109.
[21] I.J. Good, Good Thinking: The Foundations of Probability and its Applications, University of Minnesota Press, Minneapolis, 1983.
[22] A.J. Grove, J.Y. Halpern, Updating sets of probabilities, in: Proceedings of the 14th Annual Conference on Uncertainty in Artificial Intelligence (UAI-98), San Francisco, CA, Morgan Kaufmann, 1998, pp. 173–182.
[23] V. Ha, P. Haddawy, Theoretical foundations for abstraction-based probabilistic planning, in: Proceedings of the 12th Annual Conference on Uncertainty in Artificial Intelligence (UAI-96), San Francisco, CA, Morgan Kaufmann, 1996, pp. 291–298.
[24] E.A. Hansen, Solving POMDPs by searching in policy space, in: Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI-98), 1998, pp. 211–219.
[25] E.A. Hansen, Z. Feng, Dynamic programming for POMDPs using a factored state representation, in: Artificial Intelligence Planning Systems (AIPS-00), 2000, pp. 130–139.
[26] E.A. Hansen, R. Zhou, Synthesis of hierarchical finite-state controllers for POMDPs, in: Thirteenth International Conference on Automated Planning and Scheduling (ICAPS-03), June 2003.
[27] D. Harmanec, Generalizing Markov decision processes to imprecise probabilities, Journal of Statistical Planning and Inference 105 (2002) 199–213.
[28] M. Hauskrecht, Value-function approximations for partially observable Markov decision processes, Journal of Artificial Intelligence Research 13 (2000) 33–94.
[29] M. Hauskrecht, H. Fraser, Planning treatment of ischemic heart disease with partially observable Markov decision processes, Artificial Intelligence in Medicine 18 (2000) 221–244.
[30] L.P. Kaelbling, M.L. Littman, A.R. Cassandra, Planning and acting in partially observable stochastic domains, Artificial Intelligence 101 (1999) 99–134.
[31] N. Karmarkar, A new polynomial-time algorithm for linear programming, Combinatorica 4 (1984) 373–395.
[32] P.E. Lehner, K.B. Laskey, D. Dubois, An introduction to issues in higher order uncertainty, IEEE Transactions on Systems, Man and Cybernetics, Part A 26 (3) (1996) 289–293.
[33] I. Levi, On indeterminate probabilities, Journal of Philosophy 71 (1974) 391–418.
[34] I. Levi, The Enterprise of Knowledge, MIT Press, Cambridge, MA, 1980.
[35] W.S. Lovejoy, A survey of algorithmic methods for partially observed Markov decision processes, Annals of Operations Research 28 (1991) 47–66.
[36] C. Lusena, J. Goldsmith, M. Mundhenk, Nonapproximability results for partially observable Markov decision processes, Journal of Artificial Intelligence Research 14 (2001) 83–103.
[37] D.A. McAllester, S. Singh, Approximate planning for factored POMDPs using belief state simplification, in: Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI-99), 1999, pp. 409–416.
[38] M. Montemerlo, J. Pineau, N. Roy, S. Thrun, V. Verma, Experiences with a mobile robotic guide for the elderly, in: Proceedings of the National Conference of Artificial Intelligence (AAAI-02), Edmonton, AB, July 2002, pp. 587–592.
[39] A. Nilim, L. El-Ghaoui, Robustness in Markov decision problems with uncertain transition matrices, in: Advances in Neutral Information Processing Systems 16 (NIPS-03), MIT Press, Cambridge, MA, 2004.
[40] A. Nilim, L. El-Ghaoui, Robust control of Markov decision processes with uncertain transition matrices, Operations Research 53 (2005) 780–798.
[41] G. Paaß, Second order probabilities for uncertain and conflicting evidence, in: Proceedings of the 6th Annual Conference on Uncertainty in Artificial Intelligence (UAI-91), New York, Elsevier Science, 1991, pp. 447–456.

[42] C.H. Papadimitriou, J.N. Tsitsiklis, The complexity of Markov decision processes, Mathematics of Operations Research 12 (3) (1987) 441–450.
[43] J. Pineau, Tractable planning under uncertainty: Exploiting structure, PhD thesis, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, 2004.
[44] J. Pineau, G. Gordon, S. Thrun, Point-based value iteration: An anytime algorithm for POMDPs, in: Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence (IJCAI-03), AAAI Press, Menlo Park, CA, 2003.
[45] P. Poupart, Exploiting structure to efficiently solve large scale partially observable Markov decision processes, PhD thesis, Department of Computer Science, University of Toronto, Toronto, Ontario, Canada, 2005.
[46] P. Poupart, C. Boutilier, Bounded finite state controllers, in: Advances in Neural Information Processing Systems 16 (NIPS-03), MIT Press, Cambridge, MA, 2004.
[47] P. Poupart, C. Boutilier, VDCBPI: An approximate scalable algorithm for large scale POMDPs, in: Advances in Neural Information Processing Systems 17 (NIPS-04), MIT Press, Cambridge, MA, 2005.
[48] J.K. Satia, R.E. Lave, Markovian decision processes with uncertain transition probabilities, Operations Research 21 (1973) 728–740.
[49] T. Seidenfeld, M.J. Schervish, Two perspectives on consensus for (Bayesian) inference and decisions, IEEE Transactions on Systems, Man and Cybernetics 20 (2) (1990) 318–325.
[50] G. Shafer, A Mathematical Theory of Evidence, Princeton Univ. Press, Princeton, NJ, 1976.
[51] E.J. Sondik, The optimal control of partially observable Markov processes, PhD thesis, Stanford University, 1971.
[52] M.T.J. Spaan, N. Vlassis, Perseus: Randomized point-based value iteration for POMDPs, Journal of Artificial Intelligence Research 24 (2005) 195–220.
[53] N. Vlassis, M.T.J. Spaan, A fast point-based algorithm for POMDPs, in: Benelearn 2004: Proceedings of the Annual Machine Learning Conference of Belgium and the Netherlands, Brussels, Belgium, 2004, pp. 170–176.
[54] P. Walley, Statistical Reasoning with Imprecise Probabilities, Chapman and Hall, London, 1991.
[55] C.C. White, H.K. Eldeib, Parameter imprecision in finite state, finite action dynamic programs, Operations Research 34 (1986) 120–129.
[56] C.C. White, H.K. Eldeib, Markov decision processes with imprecise transition probabilities, Operations Research 43 (1994) 739–749.
[57] N.L. Zhang, W. Zhang, Speeding up the convergence of value iteration in partially observable Markov decision processes, Journal of Artificial Intelligence Research 14 (2001) 29–51.
[58] W. Zhang, N.L. Zhang, Restricted value iteration: Theory and algorithms, Journal of Artificial Intelligence Research 23 (2005) 123–165.
[59] R. Zhou, E.A. Hansen, An improved grid-based approximation algorithm for POMDPs, in: Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI-01), 2001, pp. 707–716.