

Approximate information state for approximate planning and reinforcement learning in partially observed systems

Jayakumar Subramanian*

JASUBRAM@ADOBE.COM

*Media and Data Science Research Lab, Digital Experience Cloud,
Adobe Systems India Private Limited, Noida, Uttar Pradesh, India*

Amit Sinha

AMIT.SINHA@MAIL.MCGILL.CA

Raihan Seraj

RAIHAN.SERAJ@MAIL.MCGILL.CA

Aditya Mahajan

ADITYA.MAHAJAN@MCGILL.CA

*Department of Electrical and Computer Engineering
McGill University, Montreal, QC, Canada*

Editor:

Abstract

We propose a theoretical framework for approximate planning and learning in partially observed systems. Our framework is based on the fundamental notion of information state. We provide two equivalent definitions of information state—i) a function of history which is sufficient to compute the expected reward and predict its next value; ii) equivalently, a function of the history which can be recursively updated and is sufficient to compute the expected reward and predict the next observation. An information state always leads to a dynamic programming decomposition. Our key result is to show that if a function of the history (called approximate information state (AIS)) approximately satisfies the properties of the information state, then there is a corresponding approximate dynamic program. We show that the policy computed using this is approximately optimal with bounded loss of optimality. We show that several approximations in state, observation and action spaces in literature can be viewed as instances of AIS. In some of these cases, we obtain tighter bounds. A salient feature of AIS is that it can be learnt from data. We present AIS based multi-time scale policy gradient algorithms. and detailed numerical experiments with low, moderate and high dimensional environments.

Keywords: Partially observed reinforcement learning, partially observable Markov decision processes, approximate dynamic programming, information state, approximate information state.

1. Introduction

Reinforcement learning (RL) provides a conceptual framework for designing agents which learn to act optimally in an unknown environment. RL has been successfully used in various applications ranging from robotics, industrial automation, finance, healthcare, and natural language processing. The success of RL is based on a solid foundation of combining the theory of exact and approximate Markov decision processes (MDPs) with iterative algorithms that are guaranteed to learn an exact or approximate action-value function

*. This work was done when Jayakumar Subramanian was at McGill University.

and/or an approximately optimal policy (Sutton and Barto, 2018; Bertsekas and Tsitsiklis, 1996). However, for the most part, the research on RL theory is focused primarily on systems with full state observations.

In various applications including robotics, finance, and healthcare, the agent only gets a partial observation of the state of the environment. Such partially observed systems are mathematically modeled as partially observable Markov decision processes (POMDPs) and there is a fairly good understanding of how to identify optimal or approximately optimal policies for POMDPs when the system model is known to the agent.

Since the initial work on POMDPs (Aström, 1965), it is known that POMDPs can be modeled as fully observed MDPs by considering the belief state (i.e., the posterior belief of the unobserved state given all the observations made by the agent) as an information state. Therefore, the theory and algorithms for exact and approximate planning for MDPs are also applicable to POMDPs. One computational challenge is that the belief state is continuous valued. However, the value function based on the belief state has a nice property—it is piecewise linear and a convex function of the belief state—which can be exploited to develop efficient algorithms to identify the optimal policy. Building on the one-pass algorithm of (Smallwood and Sondik, 1973), various such algorithms have been proposed in the literature including the linear support algorithm (Cheng, 1988), the witness algorithm (Cassandra et al., 1994), incremental pruning (Zhang and Liu, 1996; Cassandra et al., 1997), the duality based approach (Zhang, 2009), and others. Since POMDPs are PSPACE-complete (Papadimitriou and Tsitsiklis, 1999), the worst case complexity of such algorithms is exponential in the size of the unobserved state space. To overcome the worst case complexity of finding an optimal policy, various point-based methods have been proposed in the literature which obtain an approximate solution by sampling from the belief space (Pineau et al., 2003; Smith and Simmons, 2004; Spaan and Vlassis, 2005; Shani et al., 2007; Kurniawati et al., 2008; Poupart et al., 2011); see Shani et al. (2013) for an overview and comparison.

However, the exact and approximate planning results are of limited value for partially observed reinforcement learning (PORL) because they are based on the belief state, constructing which requires the knowledge of the system model. So, when an agent is operating in an unknown environment, it cannot construct a belief state based on its observations. An attempt to circumvent this difficulty was to use memoryless policies (i.e., choose the action based only on the current observation) (Littman, 1994; Loch and Singh, 1998; Jaakkola et al., 1995; Williams and Singh, 1999; Li et al., 2011; Azizzadenesheli et al., 2016). A related idea is to choose the action based on k recent observations (Littman, 1994; Loch and Singh, 1998) or choose the action based on a memory which is updated using a finite state machine (Whitehead and Lin, 1995; McCallum, 1993; Hansen, 1997; Meuleau et al., 1999; Amato et al., 2010). Such finite memory policies are also amenable to policy search methods (Hansen, 1998; Baxter and Bartlett, 2001; Poupart and Boutilier, 2004). However, there are no approximation guarantees available for such methods.

Another approach taken in the literature is to use a Bayesian RL framework (Ross et al., 2008; Poupart and Vlassis, 2008; Ross et al., 2011; Katt et al., 2019) where a posterior distribution over the models of the environment is maintained; at each step, a model is sampled from the posterior and the corresponding optimal policy is executed. Approximation error bounds in using such methods are derived in Ross et al. (2011).

A completely different class of model-based RL algorithms are methods using predictive state representations (PSRs) (Littman et al., 2002; Singh et al., 2003). PSRs are constructed only based on observational data so they can easily be adapted to the RL setup. There have been a number of papers which use PSRs to propose model based RL algorithms (James et al., 2004; Rosencrantz et al., 2004; Boots et al., 2011; Hamilton et al., 2014; Kulesza et al., 2015b,a; Jiang et al., 2016).

Inspired by the recent successes of deep reinforcement learning, there are many recent results which suggest using RNNs (Recurrent Neural Networks (Rumelhart et al., 1986)) or LSTMs (Long Short-Term Memories (Hochreiter and Schmidhuber, 1997)) for modeling the action-value function and/or the policy function (Bakker, 2002; Wierstra et al., 2007, 2010; Hausknecht and Stone, 2015; Heess et al., 2015; Zhu et al., 2017; Ha and Schmidhuber, 2018; Baisero and Amato, 2018; Igl et al., 2018; Zhang et al., 2019). It is shown that these approaches perform well on empirical benchmarks, but there are no approximation guarantees available for such methods.

Our main contribution is to present a rigorous approach for PORL which is based on a principled theory of approximate planning for POMDPs that we develop. In particular:

1. In Sec. 2, we formalize the notion of information state for partially observed systems and provide equivalent methods of identifying information states.
2. In Secs. 3 and 4, we present the notion of an approximate information state (AIS) as a compression of history which approximately satisfies the properties of an information state. The two equivalent formulations of information state lead to two equivalent formulations of AIS. We present bounds on the loss in performance (compared to the optimal history dependent policy) when planning using an AIS. We generalize these results to cover approximation in action spaces as well. We show that various existing approximation results for MDPs and POMDPs in the literature may be viewed as special cases of AIS (and in some cases, our bounds are tighter than those in the literature).
3. In Sec. 5, we present a theory for approximate planning for decentralized (i.e., multi-agent) partially observed systems using a common-information based AIS.
4. In Secs. 6 and 7, we then present policy gradient based online RL algorithms for PORL which learn an AIS representation using multi-timescale stochastic gradient descent. We provide detailed numerical experiments on several classes of partially observed environments ranging from classical low-dimensional toy environments, to moderate-dimensional environments, and high-dimensional grid-world environments.

2. Preliminaries: Information state and dynamic programming decomposition for partially observed systems

2.1 General model for a partially observed system

Traditionally, partially observed systems are modeled as partially observable Markov decision processes (POMDPs) (Aström, 1965; Smallwood and Sondik, 1973), where there is a controlled state and an agent which makes noise corrupted observations of the state.

However, for the purpose of understanding approximation for partially observed systems, it is conceptually cleaner to start with an input-output model of the system as described below.

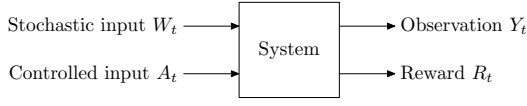


Figure 1: A stochastic input-output system

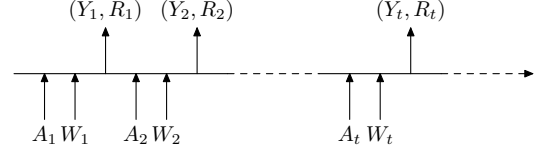


Figure 2: The timing diagram of the input-output system.

We view a partially observed system as a black-box input-output system shown in Fig. 1. At each time t , the system has two inputs and generates two outputs. The inputs to the system are a control input (also called an action) $A_t \in \mathbf{A}$ and a disturbance $W_t \in \mathbf{W}$. The outputs of the system are an observation $Y_t \in \mathbf{Y}$ and a reward $R_t \in \mathbb{R}$. For the ease of exposition, we assume that \mathbf{A} , \mathbf{W} , and \mathbf{Y} are finite sets. The analysis extends to general spaces under appropriate technical conditions. The order in which the input and output variables are generated is shown in Fig. 2.

As stated before, we do not impose a state space model on the system. Therefore, all we can say is that the outputs (Y_t, R_t) at time t are some function of all the inputs $(A_{1:t}, W_{1:t})$ up to time t , i.e.,

$$Y_t = f_t(A_{1:t}, W_{1:t}) \quad \text{and} \quad R_t = r_t(A_{1:t}, W_{1:t}),$$

where $\{f_t: \mathbf{A}^t \times \mathbf{W}^t \rightarrow \mathbf{Y}\}_{t=1}^T$ are called the system output functions and $\{r_t: \mathbf{A}^t \times \mathbf{W}^t \rightarrow \mathbb{R}\}_{t=1}^T$ are called the system reward functions.

There is an agent which observes the output Y_t and generates a control input or the action A_t as a (possibly stochastic) function of the history $H_t = (Y_{1:t-1}, A_{1:t-1})$ of the past observations and actions, i.e.,

$$A_t \sim \pi_t(H_t),$$

where $\pi := (\pi_t)_{t \geq 1}$ is a (history-dependent and possibly stochastic) policy. We use \mathbf{H}_t to denote the space of all histories up to time t . Then the policy π_t is a mapping from \mathbf{H}_t to $\Delta(\mathbf{A})$ (which denotes the space of probability measures on \mathbf{A}). We will use $\pi_t(a_t|h_t)$ to denote the probability of choosing action a_t at time t given history h_t and use $\text{Supp}(\pi_t(h_t))$ to denote the support of π_t (i.e., the set of actions chosen with positive probability).

We assume that the disturbance $\{W_t\}_{t \geq 1}$ is a sequence of independent random variables defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Thus, if the control input process $\{A_t\}_{t \geq 1}$ is specified, then the output processes $\{Y_t, R_t\}_{t \geq 1}$ are random variables on $(\Omega, \mathcal{F}, \mathbb{P})$. Specifying a policy π for the agent induces a probability measure on the output processes $\{Y_t, R_t\}_{t \geq 1}$, which we denote by \mathbb{P}^π .

We start our discussion by looking at the planning problem faced by the agent when the system runs for a finite horizon T . We will generalize our results to the infinite horizon discounted reward setup later. In the finite horizon setup, the performance of any policy π

is given by

$$J(\pi) := \mathbb{E}^\pi \left[\sum_{t=1}^T R_t \right], \quad (1)$$

where \mathbb{E}^π denotes the expectation with respect to the probability measure \mathbb{P}^π .

We assume that the agent knows the system dynamics $\{f_t\}_{t \geq 1}$, the reward functions $\{r_t\}_{t \geq 1}$, and the probability measure \mathbb{P} on the primitive random variables $\{W_t\}_{t \geq 1}$. The objective of the agent is to choose a policy π which maximizes the expected total reward $J(\pi)$.

Since all system variables are assumed to be finite valued and the system runs for a finite horizon, there are only a finite number of policies π . So, an optimal policy always exists and the important question is to determine an efficient algorithm to compute the optimal policy.

In Sec. 2.2, we start by presenting a trivial dynamic programming decomposition which uses the entire history of observations as a state. Such a history-dependent dynamic program is not an efficient method to compute the optimal policy; rather it serve as a reference with which we compare the more efficient exact and approximate dynamic programs that we derive later.

In Sec. 2.3, we present sufficient conditions to identify an information state for dynamic programming. Our main result, presented in Secs. 3 and 4, is to identify a notion of approximate information state and derive approximation bounds when an approximate policy is computed using an approximate information state.

2.2 A dynamic programming decomposition

To obtain a dynamic program to identify an optimal policy for (1), we can view the history H_t as a “state” of a Markov decision process (MDP) with transition probability

$$\mathbb{P}(H_{t+1} = (h'_t, a'_t, y_t) \mid H_t = h_t, A_t = a_t) = \begin{cases} \mathbb{P}(Y_t = y_t \mid H_t = h_t, A_t = a_t), & \text{if } h'_t = h_t \text{ \& } a'_t = a_t \\ 0, & \text{otherwise} \end{cases}$$

and per-step reward $\mathbb{E}[R_t \mid H_t, A_t]$. Therefore, from standard results from Markov decision processes Bellman (1957), we can recursively compute the performance of a given policy as well as the best possible performance using “standard” dynamic program.

Proposition 1 (Policy evaluation) *For any given (history dependent) policy π , define the reward-to-go function for any time t and realization h_t of history H_t as*

$$V_t^\pi(h_t) := \mathbb{E}^\pi \left[\sum_{s=t}^T R_s \mid H_t = h_t \right]. \quad (2)$$

The reward-to-go functions defined above satisfy the following recursion. Define $V_{T+1}^\pi(h_{T+1}) = 0$ and for any $t \in \{T, \dots, 1\}$,

$$V_t^\pi(h_t) = \mathbb{E}^\pi [R_t + V_{t+1}^\pi(H_{t+1}) \mid H_t = h_t]. \quad (3)$$

The reward-to-go function $V_t^\pi(h_t)$ denotes the expected cumulative rewards obtained in the future when starting from history h_t at time t and following policy π . Note that $V_t^\pi(h_t)$

only depends on the policy π only through the choice of the future policy (π_t, \dots, π_T) and therefore can be computed without the knowledge of the past policy $(\pi_1, \dots, \pi_{t-1})$.

Note that $h_1 = \emptyset$ and the performance $J(\pi)$ defined in (1) equals $V^\pi(h_1)$. Thus, Proposition 1 gives a recursive method to evaluate the performance of any history dependent policy π . Following the standard argument for Markov decision processes, we can modify the recursion (3) to obtain a dynamic program to identify an optimal policy as follows.

Proposition 2 (Dynamic programming) *Recursively define value functions $\{V_t: \mathbf{H}_t \rightarrow \mathbb{R}\}_{t=1}^{T+1}$ as follows. $V_{T+1}(H_{T+1}) := 0$ and for $t \in \{T, \dots, 1\}$,*

$$V_t(h_t) := \max_{a_t \in \mathbf{A}} \mathbb{E}[R_t + V_{t+1}(H_{t+1}) \mid H_t = h_t, A_t = a_t]. \quad (4)$$

Then, a stochastic policy $\pi = (\pi_1, \dots, \pi_T)$ is optimal if and only if for all $t \in \{1, \dots, T\}$ it satisfies

$$\text{Supp}(\pi_t(h_t)) \subseteq \arg \max_{a_t \in \mathbf{A}} \mathbb{E}[R_t + V_{t+1}(H_{t+1}) \mid H_t = h_t, A_t = a_t]. \quad (5)$$

Note that the expectation in (4) is with respect to the probability measure \mathbb{P} on (Ω, \mathcal{F}) and can be computed without the knowledge of the policy π .

2.3 Information state and simplified dynamic programs

The dynamic program of Proposition 2 uses the entire history as state and may not be efficient for identifying an optimal policy. In this section, we present a general class of dynamic programming decompositions which may be more efficient. This class of dynamic programs is based on the notion of information state, which we describe next.

Definition 3 *Let $\{Z_t\}_{t=1}^T$ be a pre-specified collection of Banach spaces. A collection $\{\sigma_t: \mathbf{H}_t \rightarrow Z_t\}_{t=1}^T$ of history compression functions is called an information state generator if the process $\{Z_t\}_{t=1}^T$, where $Z_t = \sigma_t(H_t)$, satisfies the following properties:*

(P1) Sufficient for performance evaluation, *i.e., for any time t , any realization h_t of H_t and any choice a_t of A_t , we have*

$$\mathbb{E}[R_t \mid H_t = h_t, A_t = a_t] = \mathbb{E}[R_t \mid Z_t = \sigma_t(h_t), A_t = a_t].$$

(P2) Sufficient to predict itself, *i.e., for any time t , any realization h_t of H_t and any choice a_t of A_t , we have that for any Borel subset \mathbf{B} of Z_{t+1} ,*

$$\mathbb{P}(Z_{t+1} \in \mathbf{B} \mid H_t = h_t, A_t = a_t) = \mathbb{P}(Z_{t+1} \in \mathbf{B} \mid Z_t = \sigma_t(h_t), A_t = a_t).$$

In the sequel, we will sometimes use the phrase “let $\{Z_t\}_{t=1}^T$ be an information state” to specify an information state and will implicitly assume that the corresponding information state spaces are $\{Z_t\}_{t=1}^T$ and the corresponding compression functions are $\{\sigma_t\}_{t=1}^T$.

Note that both the probabilities in Property (P2) can be computed without the knowledge of the policy π . Furthermore, there are no restrictions on the spaces $\{Z_t\}_{t=1}^T$ although in practice an information state is useful only when these spaces are “small” in an appropriate sense.

Condition (P1) is easy to verify but condition (P2) can be a bit abstract. For some models, instead of (P2), it is easier to verify the following stronger conditions:

(P2a) Evolves in a state-like manner, i.e., there exist measurable functions $\{\varphi_t\}_{t=1}^T$ such that for any time t and any realization h_{t+1} of H_{t+1} , we have

$$\sigma_{t+1}(h_{t+1}) = \varphi_t(\sigma_t(h_t), y_t, a_t).$$

Informally, the above condition may be written as $Z_{t+1} = \varphi_t(Z_t, Y_t, A_t)$.

(P2b) Is sufficient for predicting future observations, i.e., for any time t , any realization h_t of H_t and any choice a_t of A_t , we have that for any subset D of Y ,

$$\mathbb{P}(Y_t \in D \mid H_t = h_t, A_t = a_t) = \mathbb{P}(Y_t \in D \mid Z_t = \sigma_t(h_t), A_t = a_t).$$

Proposition 4 (P2a) and (P2b) imply (P2).

Proof For any Borel subset D of Z_{t+1} , we have

$$\begin{aligned} \mathbb{P}(Z_{t+1} \in D \mid H_t = h_t, A_t = a_t) & \\ & \stackrel{(a)}{=} \sum_{y_t \in Y} \mathbb{P}(Y_t = y_t, Z_{t+1} \in D \mid H_t = h_t, A_t = a_t) \\ & \stackrel{(b)}{=} \sum_{y_t \in Y} \mathbb{1}\{\varphi_t(\sigma_t(h_t), y_t, a_t) \in D\} \mathbb{P}(Y_t = y_t \mid H_t = h_t, A_t = a_t) \\ & \stackrel{(c)}{=} \sum_{y_t \in Y} \mathbb{1}\{\varphi_t(\sigma_t(h_t), y_t, a_t) \in D\} \mathbb{P}(Y_t = y_t \mid Z_t = \sigma_t(h_t), A_t = a_t) \\ & \stackrel{(d)}{=} \mathbb{P}(Z_{t+1} \in D \mid Z_t = \sigma_t(h_t), A_t = a_t) \end{aligned}$$

where (a) follows from the law of total probability, (b) follows from (P2a), (c) follows from (P2b) and (d) from the law of total probability. \blacksquare

The following example illustrates how (P2a) and (P2b) are stronger conditions than (P2). Consider a Markov decision process (MDP) with state $(S_t^1, S_t^2) \in \mathcal{S}^1 \times \mathcal{S}^2$ and action $A_t \in \mathcal{A}$, where the dynamics of the two components of the state are conditionally independent given the action, i.e.,

$$\begin{aligned} \mathbb{P}(S_{t+1}^1 = s_+^1, S_{t+1}^2 = s_+^2 \mid S_t^1 = s^1, S_t^2 = s^2, A_t = a) \\ = \mathbb{P}(S_{t+1}^1 = s_+^1 \mid S_t^1 = s^1, A_t = a) \mathbb{P}(S_{t+1}^2 = s_+^2 \mid S_t^2 = s^2, A_t = a). \end{aligned}$$

Furthermore, suppose the reward R_t at any time t is given by $R_t = r_t(S_t^1, A_t)$. Since the model is an MDP, the observation at time t is the same as the state. For this model, the component $\{S_t^1\}_{t \geq 1}$ of the state satisfies properties (P1) and (P2). Therefore, $\{S_t^1\}_{t \geq 1}$ is an information state process. However, $\{S_t^1\}_{t \geq 1}$ is not sufficient to predict the next observation (S_{t+1}^1, S_{t+1}^2) . Therefore, $\{S_t^1\}_{t \geq 1}$ does not satisfy property (P2b). This shows that properties (P2a) and (P2b) are stronger than property (P2). The above example may be considered as an instance of what is called the Noisy-TV problem (Burda et al., 2018).

Next, we show that an information state is useful because it is always possible to write a dynamic program based on the information state. To explain this dynamic programming

decomposition, we first write the history-based dynamic programs of Proposition 1 and 2 in a more compact manner as follows: Let $V_{T+1}(h_{T+1}) := 0$ and for $t \in \{T, \dots, 1\}$, define

$$Q_t(h_t, a_t) := \mathbb{E}[R_t + V_{t+1}(H_{t+1}) \mid H_t = h_t, A_t = a_t], \quad (6a)$$

$$V_t(h_t) := \max_{a_t \in \mathbf{A}} Q_t(h_t, a_t). \quad (6b)$$

The function $Q_t(h_t, a_t)$ is called the action-value function. Moreover, for a given stochastic policy $\pi = (\pi_1, \dots, \pi_T)$, where $\pi_t: \mathbf{H}_t \rightarrow \Delta(\mathbf{A}_t)$, let $V_{T+1}^\pi(h_{T+1}) = 0$ and for $t \in \{T, \dots, 1\}$, define

$$Q_t^\pi(h_t, a_t) := \mathbb{E}[R_t + V_{t+1}^\pi(H_{t+1}) \mid H_t = h_t, A_t = a_t], \quad (7a)$$

$$V_t^\pi(h_t) := \sum_{a_t \in \mathbf{A}} \pi_t(a_t \mid h_t) \cdot Q_t^\pi(h_t, a_t). \quad (7b)$$

Theorem 5 *Let $\{Z_t\}_{t=1}^T$ be an information state. Recursively define value functions $\{\bar{V}_t: \mathbf{Z}_t \rightarrow \mathbb{R}\}_{t=1}^{T+1}$, as follows: $\bar{V}_{T+1}(z_{T+1}) := 0$ and for $t \in \{T, \dots, 1\}$:*

$$\bar{Q}_t(z_t, a_t) := \mathbb{E}[R_t + \bar{V}_{t+1}(Z_{t+1}) \mid Z_t = z_t, A_t = a_t], \quad (8a)$$

$$\bar{V}_t(z_t) := \max_{a_t \in \mathbf{A}} \bar{Q}_t(z_t, a_t). \quad (8b)$$

Then, we have the following:

1. *For any time t , history h_t , and action a_t , we have that*

$$Q_t(h_t, a_t) = \bar{Q}_t(\sigma_t(h_t), a_t) \text{ and } V_t(h_t) = \bar{V}_t(\sigma_t(h_t)). \quad (9)$$

2. *Let $\bar{\pi} = (\bar{\pi}_1, \dots, \bar{\pi}_T)$, where $\bar{\pi}_t: \mathbf{Z}_t \rightarrow \Delta(\mathbf{A})$, be a stochastic policy. Then, the policy $\pi = (\pi_1, \dots, \pi_T)$ given by $\pi_t = \bar{\pi}_t \circ \sigma_t$ is optimal if and only if for all t and all realizations z_t of information states Z_t , $\text{Supp}(\bar{\pi}_t(z_t)) \subseteq \arg \max_{a_t \in \mathbf{A}} \bar{Q}_t(z_t, a_t)$.*

Proof We prove the result by backward induction. By construction, (9) is true at time $T+1$. This forms the basis of induction. Assume that (9) is true at time $t+1$ and consider the system at time t . Then,

$$\begin{aligned} Q_t(h_t, a_t) &= \mathbb{E}[R_t + V_{t+1}(H_{t+1}) \mid H_t = h_t, A_t = a_t] \\ &\stackrel{(a)}{=} \mathbb{E}[R_t + \bar{V}_{t+1}(\sigma_{t+1}(H_{t+1})) \mid H_t = h_t, A_t = a_t] \\ &\stackrel{(b)}{=} \mathbb{E}[R_t + \bar{V}_{t+1}(Z_{t+1}) \mid Z_t = \sigma_t(h_t), A_t = a_t] \\ &\stackrel{(c)}{=} \bar{Q}_t(\sigma_t(h_t), a_t), \end{aligned}$$

where (a) follows from the induction hypothesis, (b) follows from the properties (P1) and (P2) of information state, and (c) follows from the definition of \bar{Q} . This shows that the action-value functions are equal. By maximizing over the actions, we get that the value functions are also equal. The optimality of the policy follows immediately from (9). \blacksquare

2.4 Examples of information state

For a general model, it is not immediately evident that a non-trivial information state exists. The question of existence will depend on the specifics of the observation and reward functions $\{f_t, r_t\}_{t \geq 1}$ as well as the properties of the probability measure on the primitive random variables $\{W_t\}_{t \geq 1}$. We do not pursue the question of existence in this paper, but present various specific models where information state exists and show that the corresponding results for these models in the literature may be viewed as a special case of Theorem 5.

1. For any partially observed model, the history H_t is always a trivial information state. Therefore, the dynamic program of Proposition 2 may be viewed as a special case of Theorem 5.
2. MARKOV DECISION PROCESS (MDP): Consider a Markov decision process (MDP) with state $S_t \in \mathcal{S}$ and action $A_t \in \mathcal{A}$ (Bellman, 1957). At each time, the state evolves in a controlled Markovian manner with

$$\mathbb{P}(S_{t+1} = s_{t+1} \mid S_{1:t} = S_{1:t}, A_{1:t} = A_{1:t}) = \mathbb{P}(S_{t+1} = s_{t+1} \mid S_t = S_t, A_t = A_t).$$

The observation of the agent is $Y_t = S_{t+1}$ and the reward output is $R_t = r(S_t, A_t)$. An information state for an MDP is given by the current state S_t (the corresponding compression function is $\sigma_t(S_{1:t}, A_{1:t-1}) = S_t$). The standard dynamic program for MDPs may be viewed as a special case of Theorem 5.

3. EVEN MDPs: Consider an MDP where the state space \mathcal{S} is either \mathbb{R} or a symmetric subset of \mathbb{R} of the form $[-B, B]$, the controlled transition matrix is even, i.e., for every $a \in \mathcal{A}$ and $s, s' \in \mathcal{S}$,

$$\mathbb{P}(S_{t+1} = s' \mid S_t = s, A_t = a) = \mathbb{P}(S_{t+1} = -s' \mid S_t = -s, A_t = a),$$

and for every $a \in \mathcal{A}$, the per-step reward function $r(s, a)$ is even in s . Such MDPs are called *even* MDPs (Chakravorty and Mahajan, 2018) and an information state for such MDPs is given by the absolute value state $|S_t|$ (the corresponding compression function is $\sigma_t(S_{1:t}, A_{1:t-1}) = |S_t|$). The dynamic program for even MDPs derived in Chakravorty and Mahajan (2018) may be viewed as a special case of Theorem 5.

4. MDP WITH IRRELEVANT COMPONENTS: Consider an MDP with state space $\mathcal{S} = \mathcal{S}^1 \times \mathcal{S}^2$, action space \mathcal{A} , transition matrix $P(s_+^1, s_+^2 \mid s^1, s^2, a) = P^1(s_+^1 \mid s^1, a)P^2(s_+^2 \mid s^1, s^2, a)$, and per-step reward $r(s^1, a)$, which does not depend on the second component of the state. As explained in Feinberg (2005), such models arise in control of queues and transformation of continuous time Markov decision processes to discrete time MDPs using uniformization. An information state for such MDPs is given by the first component S_t^1 (the corresponding compression function is $\sigma_t(S_{1:t}^1, S_{1:t}^2, A_{1:t}) = S_t^1$). The qualitative properties of optimal policies for such models derived in Feinberg (2005) may be viewed as a special case of Theorem 5.
5. MDP WITH DELAYED STATE OBSERVATION: Consider an MDP where the observation Y_t of the agent is the δ -step delayed state $S_{t-\delta+1}$ of the system (Altman and Nain,

1992). An information state for such MDPs is given by the vector $(S_{t-\delta+1}, U_{t-\delta+1:t-1})$. The dynamic program for such models derived in Altman and Nain (1992) may be viewed as a special case of Theorem 5.

6. **PARTIALLY OBSERVABLE MARKOV DECISION PROCESSES (POMDPs):** Consider a partially observable Markov decision process (POMDP) where there is a state space model as for an MDP but the observation Y_t is some function of the state and the disturbance, i.e., $Y_t = f_t^y(S_t, W_t)$ (Aström, 1965; Smallwood and Sondik, 1973). An information state for the POMDP is given by the belief state $B_t \in \Delta(\mathcal{S})$ which is given by $B_t(s) = \mathbb{P}(S_t = s \mid H_t = h_t)$. The corresponding compression function may be identified via the update functions $\{\varphi_t\}_{t=1}^T$ of Property (P2a), which are the standard belief update functions for non-linear filtering. The standard belief state dynamic program for POMDPs (Aström, 1965; Smallwood and Sondik, 1973) may be viewed as a special case of Theorem 5.
7. **LINEAR QUADRATIC AND GAUSSIAN (LQG) MODELS:** Consider a POMDP where the state and action spaces are Euclidean spaces, the system dynamics $\mathbb{P}(S_{t+1} \mid S_t, A_t)$ and the observation $f_t^y(S_t, W_t)$ are linear, the disturbance W_t is Gaussian, and the per-step *cost* is a quadratic function of the state and action (Aström, 1970). For such a *linear-quadratic-and-Gaussian* POMDP, an information state is given by the state estimate $\hat{S}_t = \mathbb{E}[S_t \mid H_t = h_t]$. The corresponding compression function may be identified via the update functions $\{\varphi_t\}_{t=1}^T$ of Property (P2a), which in this case are Kalman filtering update equations. The standard conditional estimate based dynamic program for LQG models (Aström, 1970) may be viewed as a special case of Theorem 5.
8. **POMDPs WITH DELAYED OBSERVATIONS:** Consider a POMDP where the observation is delayed by δ time steps (Bander and White, 1999). For such a system the belief on δ step delayed state based on the δ -step delayed observations and control, as well as the vector of last δ control actions is an information state. The structure of the optimal policy and the dynamic program derived in Bander and White (1999) may be viewed as a special case of Theorem 5.
9. **MACHINE MAINTENANCE:** Consider the following model for machine maintenance (Eckles, 1968). A machine can be in one of n ordered states where the first state is the best and the last state is the worst. The production cost increases with the state of the machine. The state evolves in a Markovian manner. At each time, an agent has the option to either run the machine or stop and inspect it for a cost. After inspection, the agent may either repair it (at a cost that depends on the state) or replace it (at a fixed cost). The objective is to identify a maintenance policy to minimize the cost of production, inspection, repair, and replacement.

Let τ denote the time of last inspection and S_τ denote the state of the machine after inspection, repair, or replacement. Then, it can be shown that $(S_\tau, t - \tau)$ is an information state for the system. This is an instance of an incrementally expanding representation for a POMDP described in Arabneydi and Mahajan (2015).

The above examples show that there are generic information states for certain class of models (e.g., MDPs, MDPs with delays, POMDPs, POMDPs with delays) as well as specific

information states tuned to the model (e.g., even MDPs, MDPs with irrelevant components, LQG models, machine repair).

2.5 Discussion and related work

Although we are not aware of a previous result which formally defines an information state and shows that an information state always implies a dynamic programming decomposition (Theorem 5), yet the notion of information state is not new and has always existed in the stochastic control literature. Information state may be viewed as a generalization of the traditional notion of state (Nerode, 1958), which is defined as a statistic (i.e., a function of the observations) sufficient for input-output mapping. In contrast, we define an information state as a statistic sufficient for performance evaluation (and, therefore, for dynamic programming). Such a definition is hinted in Witsenhausen (1976). The notion of information state is also related to sufficient statistics for optimal control defined in Striebel (1965) for systems with state space models.

As far as we are aware, the informal definition of information state was first proposed by Kwakernaak (1965) for adaptive control systems. Formal definitions for linear control systems were given by Bohlin (1970) for discrete time systems and by Davis and Varaiya (1972) for continuous time systems. Kumar and Varaiya (1986) define an information state as a compression of past history which satisfies property (P2a) but do not formally show that such an information state always leads to a dynamic programming decomposition. A formal definition of information state appears in our previous work (Mahajan and Mannan, 2016) where the result of Theorem 5 is asserted without proof. Properties of information states for multi-agent teams were asserted in Mahajan (2008). Adlakha et al. (2012) provide a definition which is stronger than our definition. They require that in a POMDP with unobserved state $S_t \in \mathcal{S}$, $\sigma_t(h_t)$ should satisfy (P1) and (P2) as well be sufficient to predict S_t , i.e., for any Borel subset \mathcal{B} of \mathcal{S} and any realization h_t of H_t ,

$$\mathbb{P}(S_t \in \mathcal{B} \mid H_t = h_t) = \mathbb{P}(S_t \in \mathcal{B} \mid \hat{Z}_t = \sigma_t(h_t)).$$

A similar definition is also used in Francois-Lavet et al. (2019). We had presented a definition similar to Definition 3 in the preliminary version of this paper (Subramanian and Mahajan, 2019).

The notion of information state is also related to Γ -trace equivalence for MDPs and POMDPs defined by Castro et al. (2009). For MDPs, Γ -trace equivalence takes a partition of the state space and returns a finer partition such that for any choice of future actions any two states in the same cell of the finer partition have the same distribution on future states and rewards. Castro et al. (2009) show that recursive applications of Γ -trace equivalence has a fixed point, which is equivalent to bisimulation based partition (Givan et al., 2003) of the state space of the MDP. Similar results were shown for MDPs in Ferns et al. (2004, 2011).

Castro et al. (2009) extend the notion of trace equivalence for MDPs to belief trajectory equivalence for POMDPs. In particular, two belief states are said to be belief trajectory equivalent if for any choice of future actions, they generate the same distribution on future observations and rewards. Such belief trajectory equivalence is related to predictive state representation (PSR) (Littman et al., 2002; Singh et al., 2003; Izadi and Precup, 2003; James et al., 2004; Rosencrantz et al., 2004; Wolfe et al., 2005) and observable operator models

(OOM) (Jaeger, 2000; Jaeger et al., 2006), which are a compression of the past history which is sufficient to predict the future observations (but not necessarily rewards). Information state may be viewed as a “Markovianized” version of belief trajectory equivalence and PSRs, which has the advantage that both (P1) and (P2) are defined in terms of “one-step” equivalence while belief trajectory equivalence and PSR are defined in terms of “entire future trajectory” equivalence. It should be noted that PSR and bisimulation based equivalences are defined for infinite horizon models, while the information state is defined for both finite and infinite horizon models (see Sec. 4).

Another related notion is the notion of causal states (or ε -machines) used in computational mechanics (Crutchfield and Young, 1989; Shalizi and Crutchfield, 2001). and forecasting in dynamical systems (Grassberger, 1986, 1988). These definitions are for uncontrolled Markov chains and the emphasis is on the minimal state representation for time-invariant infinite-horizon systems.

3. Approximate planning in partially observed systems

Our key insight is that information states provide a principled approach to approximate planning and learning in partially observed systems. To illustrate this, reconsider the machine maintenance example presented earlier in Sec. 2.4. Theorem 5 implies that we can write a dynamic program for that model using the information state $(S_\tau, t - \tau)$, which takes values in a countable set. This countable state dynamic program is considerably simpler than the standard belief state dynamic program typically used for that model. Moreover, it is possible to approximate the countable state model by a finite-state model by truncating the state space, which provides an approximate planning solution to the problem. Furthermore, the information state $(S_\tau, t - \tau)$ does not depend on the transition probability of the state of the machine or the cost of inspection or repair. Thus, if these model parameters were unknown, we can use a standard reinforcement learning algorithm to find an optimal policy which maps $(S_\tau, t - \tau)$ to current action.

Given these benefits of a good information state, it is natural to consider a data-driven approach to identify an information state. An information state identified from data will not be exact and it is important to understand what is the loss in performance when using an approximate information state. Theorem 5 shows that a compression of the history which satisfies properties (P1) and (P2) is sufficient to identify a dynamic programming decomposition. Would a compression of history that approximately satisfied properties (P1) and (P2) lead to an approximate dynamic program? In this section, we show that the answer to this question is yes. First, we need to precisely define what we mean by “approximately satisfy properties (P1) and (P2)”. For that matter, we need to fix a distance metric on probability spaces. There are various metrics on probability space and it turns out that the appropriate distance metric for our purposes is the integral probability metric (IPM) (Müller, 1997).

3.1 Integral probability metrics (IPM)

Definition 6 *Let (X, \mathcal{G}) be a measurable space and \mathfrak{F} denote a class of uniformly bounded measurable functions on (X, \mathcal{G}) . The integral probability metric (IPM) between two probability*

distributions $\mu, \nu \in \Delta(\mathbf{X})$ with respect to the function class \mathfrak{F} is defined as

$$d_{\mathfrak{F}}(\mu, \nu) := \sup_{f \in \mathfrak{F}} \left| \int_{\mathbf{X}} f d\mu - \int_{\mathbf{X}} f d\nu \right|.$$

In the literature, IPMs are also known as probability metrics with a ζ -structure; see e.g., Zolotarev (1983); Rachev (1991). They are useful to establish weak convergence of probability measures. Methods for estimating IPM from samples are discussed in Sriperumbudur et al. (2012).

EXAMPLES OF INTEGRAL PROBABILITY METRICS (IPMs)

When $(\mathbf{X}, \mathcal{G})$ is a metric space, then various commonly used distance metrics on $(\mathbf{X}, \mathcal{G})$ lead to specific instances of IPM for a particular choice of function space \mathfrak{F} . We provide some examples below:

1. **TOTAL VARIATION DISTANCE:** If \mathfrak{F} is chosen as $\{f : \|f\|_{\infty} \leq 1\}$, then $d_{\mathfrak{F}}$ is the total variation distance.¹
2. **KOLMOGOROV DISTANCE:** If $\mathbf{X} = \mathbb{R}^m$ and \mathfrak{F} is chosen as $\{\mathbf{1}_{(-\infty, t]} : t \in \mathbb{R}^m\}$, then $d_{\mathfrak{F}}$ is the Kolmogorov distance.
3. **KANTOROVICH METRIC OR WASSERSTEIN DISTANCE:** Let $\|f\|_{\text{Lip}}$ denote the Lipschitz semi-norm of a function. If \mathfrak{F} is chosen as $\{f : \|f\|_{\text{Lip}} \leq 1\}$, then $d_{\mathfrak{F}}$ is the Kantorovich metric. When \mathbf{X} is separable, the Kantorovich metric is the dual representation of the Wasserstein distance via the Kantorovich-Rubinstein duality (Villani, 2008).
4. **BOUNDED-LIPSCHITZ METRIC:** If \mathfrak{F} is chosen as $\{f : \|f\|_{\infty} + \|f\|_{\text{Lip}} \leq 1\}$, then $d_{\mathfrak{F}}$ is the bounded-Lipschitz (or Dudley) metric.
5. **MAXIMUM MEAN DISCREPANCY (MMD):** Let \mathcal{H} be a reproducing kernel Hilbert space (RKHS) of real valued functions on \mathbf{X} and let $\mathfrak{F} = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq 1\}$, then $d_{\mathfrak{F}}$ is the maximum mean discrepancy² (Sriperumbudur et al., 2008). The energy distance

1. In particular, if μ and ν are absolutely continuous with respect to some measure λ and let $p = d\mu/d\lambda$ and $q = d\nu/d\lambda$, then

$$\left| \int_{\mathbf{X}} f d\mu - \int_{\mathbf{X}} f d\nu \right| = \left| \int_{\mathbf{X}} f(x) p(x) \lambda(dx) - \int_{\mathbf{X}} f(x) q(x) \lambda(dx) \right| \leq \|f\|_{\infty} \int_{\mathbf{X}} |p(x) - q(x)| \lambda(dx).$$

In this paper, we are defining total variation distance as $\int_{\mathbf{X}} |p(x) - q(x)| \lambda(dx)$. Typically, it is defined as half of that quantity. Note that it is possible to get a tighter bound than above where $\|f\|_{\infty}$ is replaced by $\frac{1}{2} \text{span}(f) = \frac{1}{2}(\max(f) - \min(f))$.

2. One of features of MMD is that the optimizing f can be identified in closed form. In particular, if k is the kernel of the RKHS, then (see Gretton et al. (2006); Sriperumbudur et al. (2012) for details)

$$\begin{aligned} d_{\mathfrak{F}}(\mu, \nu) &= \left\| \int_{\mathbf{X}} k(\cdot, x) d\mu(x) - \int_{\mathbf{X}} k(\cdot, x) d\nu(x) \right\|_{\mathcal{H}} \\ &= \left[\int_{\mathbf{X}} \int_{\mathbf{X}} k(x, y) \mu(dx) \mu(dy) + \int_{\mathbf{X}} \int_{\mathbf{X}} k(x, y) \nu(dx) \nu(dy) - 2 \int_{\mathbf{X}} \int_{\mathbf{X}} k(x, y) \mu(dx) \nu(dy) \right]^{1/2}. \end{aligned}$$

We use an MMD as a IPM in the PORL algorithms proposed in Sec. 6, where we exploit this property.

studied in statistics (Székely and Rizzo, 2004) is a special case of maximum mean discrepancy; see Sejdinovic et al. (2013) for a discussion.

We say that \mathfrak{F} is a closed set if it is closed under the topology of pointwise convergence. We say that \mathfrak{F} is a convex set if $f_1, f_2 \in \mathfrak{F}$ implies that for any $\lambda \in (0, 1)$, $\lambda f_1 + (1 - \lambda)f_2 \in \mathfrak{F}$. Note that all the above function classes are convex and all except Kolmogorov distance are closed.

We now list some useful properties of IPMs, which immediately follow from definition.

1. Given a function class \mathfrak{F} and a function f (not necessarily in \mathfrak{F}),

$$\left| \int_{\mathbf{X}} f d\mu - \int_{\mathbf{X}} f d\nu \right| \leq \rho_{\mathfrak{F}}(f) \cdot d_{\mathfrak{F}}(\mu, \nu), \quad (10)$$

where $\rho_{\mathfrak{F}}(f)$ is the Minkowski functional with respect to \mathfrak{F} given by

$$\rho_{\mathfrak{F}}(f) := \inf\{\rho \in \mathbb{R}_{>0} : \rho^{-1}f \in \mathfrak{F}\}. \quad (11)$$

For the total variation distance, $|\int_{\mathbf{X}} f d\mu - \int_{\mathbf{X}} f d\nu| \leq \frac{1}{2} \text{span}(f) d_{\mathfrak{F}}(\mu, \nu)$. Thus, for total variation, $\rho_{\mathfrak{F}}(f) = \frac{1}{2} \text{span}(f)$. For the Kantorovich metric, $|\int_{\mathbf{X}} f d\mu - \int_{\mathbf{X}} f d\nu| \leq \|f\|_{\text{Lip}} d_{\mathfrak{F}}(\mu, \nu)$. Thus, for Kantorovich metric, $\rho_{\mathfrak{F}}(f) = \|f\|_{\text{Lip}}$. For the maximum mean discrepancy, $|\int_{\mathbf{X}} f d\mu - \int_{\mathbf{X}} f d\nu| \leq \|f\|_{\mathcal{H}} d_{\mathfrak{F}}(\mu, \nu)$. Thus, for maximum mean discrepancy, $\rho_{\mathfrak{F}}(f) = \|f\|_{\mathcal{H}}$.

2. Let \mathbf{X} and \mathbf{Y} be Banach spaces and let $\mathfrak{F}_{\mathbf{X}}$ and $\mathfrak{F}_{\mathbf{Y}}$ denote the function class for $d_{\mathfrak{F}}$ with domain \mathbf{X} and \mathbf{Y} , respectively. Then, for any $\ell: \mathbf{X} \rightarrow \mathbf{Y}$, any real-valued function $f \in \mathfrak{F}_{\mathbf{Y}}$ and any measures μ and ν on $\Delta(\mathbf{X})$, we have

$$\left| \int_{\mathbf{X}} f(\ell(x)) \mu(dx) - \int_{\mathbf{X}} f(\ell(x)) \nu(dx) \right| \leq \rho_{\mathfrak{F}_{\mathbf{X}}}(f \circ \ell) d_{\mathfrak{F}_{\mathbf{X}}}(\mu, \nu).$$

We define the contraction factor of the function ℓ as

$$\kappa_{\mathfrak{F}_{\mathbf{X}}, \mathfrak{F}_{\mathbf{Y}}}(\ell) = \sup_{f \in \mathfrak{F}_{\mathbf{Y}}} \rho_{\mathfrak{F}_{\mathbf{X}}}(f \circ \ell). \quad (12)$$

Therefore, we can say that for any $f \in \mathfrak{F}_{\mathbf{Y}}$,

$$\left| \int_{\mathbf{X}} f(\ell(x)) \mu(dx) - \int_{\mathbf{X}} f(\ell(x)) \nu(dx) \right| \leq \kappa_{\mathfrak{F}_{\mathbf{X}}, \mathfrak{F}_{\mathbf{Y}}}(\ell) d_{\mathfrak{F}_{\mathbf{X}}}(\mu, \nu). \quad (13)$$

For the total variation distance, $\frac{1}{2} \text{span}(f \circ \ell) \leq \|f \circ \ell\|_{\infty} \leq \|f\|_{\infty} \leq 1$. Thus, $\kappa_{\mathfrak{F}}(\ell) \leq 1$. For the Kantorovich metric, $\|f \circ \ell\|_{\text{Lip}} \leq \|f\|_{\text{Lip}} \|\ell\|_{\text{Lip}}$. Thus, $\kappa_{\mathfrak{F}}(\ell) \leq \|\ell\|_{\text{Lip}}$.

3.2 Approximate information state (AIS) and approximate dynamic programming

Now we define a notion of AIS as a compression of the history of observations and actions which approximately satisfies properties (P1) and (P2).

Definition 7 Let $\{\hat{Z}_t\}_{t=1}^T$ be a pre-specified collection of Banach spaces, \mathfrak{F} be a function class for IPMs, and $\{(\varepsilon_t, \delta_t)\}_{t=1}^T$ be pre-specified positive real numbers. A collection $\{\hat{\sigma}_t: \mathbf{H}_t \rightarrow \hat{Z}_t\}_{t=1}^T$ of history compression functions, along with approximate update kernels $\{\hat{P}_t: \hat{Z}_t \times \mathbf{A} \rightarrow \Delta(\hat{Z}_{t+1})\}_{t=1}^T$ and reward approximation functions $\{\hat{r}_t: \hat{Z}_t \times \mathbf{A} \rightarrow \mathbb{R}\}_{t=1}^T$, is called an $\{(\varepsilon_t, \delta_t)\}_{t=1}^T$ -AIS generator if the process $\{\hat{Z}_t\}_{t=1}^T$, where $\hat{Z}_t = \hat{\sigma}_t(H_t)$, satisfies the following properties:

(AP1) Sufficient for approximate performance evaluation, i.e., for any time t , any realization h_t of H_t and any choice a_t of A_t , we have

$$|\mathbb{E}[R_t \mid H_t = h_t, A_t = a_t] - \hat{r}_t(\hat{\sigma}_t(h_t), a_t)| \leq \varepsilon_t.$$

(AP2) Sufficient to predict itself approximately. i.e., for any time t , any realization h_t of H_t , any choice a_t of A_t , and for any Borel subset B of \hat{Z}_{t+1} , define $\mu_t(B) := \mathbb{P}(\hat{Z}_{t+1} \in B \mid H_t = h_t, A_t = a_t)$ and $\nu_t(B) := \hat{P}_t(B \mid \hat{\sigma}_t(h_t), a_t)$; then,

$$d_{\mathfrak{F}}(\mu_t, \nu_t) \leq \delta_t.$$

We use the phrase “ (ε, δ) -AIS” when ε_t and δ_t do not depend on time.

Similar to Proposition 4, we can provide an alternative characterization of an AIS where we replace (AP2) with the following approximations of (P2a) and (P2b).

(AP2a) Evolves in a state-like manner, i.e., there exist measurable update functions $\{\hat{\varphi}_t: \hat{Z}_t \times \mathbf{Y} \times \mathbf{A}\}_{t=1}^T$ such that for any realization h_{t+1} of H_{t+1} , we have

$$\hat{\sigma}_{t+1}(h_{t+1}) = \hat{\varphi}_t(\hat{\sigma}_t(h_t), y_t, a_t).$$

(AP2b) Is sufficient for predicting future observations approximately, i.e., there exist measurable observation prediction kernels $\{\hat{P}_t^y: \hat{Z}_t \times \mathbf{A} \rightarrow \Delta(\mathbf{Y})\}_{t=1}^T$ such that for any time t , any realization h_t of H_t , any choice a_t of A_t , and for any Borel subset B of \mathbf{Y} define, $\mu_t^y(B) := \mathbb{P}(Y_t \in B \mid H_t = h_t, A_t = a_t)$ and $\nu_t^y(B) = \hat{P}_t^y(B \mid \hat{\sigma}_t(h_t), a_t)$; then,

$$d_{\mathfrak{F}}(\mu_t^y, \nu_t^y) \leq \delta / \kappa_{\mathfrak{F}}(\hat{\varphi}_t),$$

where $\kappa_{\mathfrak{F}}(\hat{\varphi}_t)$ is defined as $\sup_{h_t \in \mathbf{H}_t, a_t \in \mathbf{A}_t} \kappa_{\mathfrak{F}}(\hat{\varphi}_t(\hat{\sigma}_t(h_t), \cdot, a_t))$. Note that for the total variation distance $\kappa_{\mathfrak{F}}(\hat{\varphi}_t) = 1$; for the Kantorovich distance $\kappa_{\mathfrak{F}}(\hat{\varphi}_t)$ is equal to the Lipschitz uniform bound on the Lipschitz constant of $\hat{\varphi}_t$ with respect to y_t .

Proposition 8 (AP2a) and (AP2b) imply (AP2) holds with transition kernels $\{\hat{P}_t^y\}_{t=1}^T$ defined as follows: for any Borel subset B of \hat{Z} ,

$$\hat{P}_t(B \mid \hat{\sigma}_t(h_t), a_t) = \int_{\mathbf{Y}} \mathbb{1}_B(\hat{\varphi}_t(\hat{\sigma}_t(h_t), y_t, a_t)) \hat{P}_t^y(dy_t \mid \hat{\sigma}_t(h_t), a_t).$$

Therefore, we can alternatively define an $\{(\varepsilon_t, \delta_t)\}_{t=1}^T$ -AIS generator as a tuple $\{(\hat{\sigma}_t, \hat{r}_t, \hat{\varphi}_t, \hat{P}_t^y)\}_{t=1}^T$ which satisfies (AP1), (AP2a), and (AP2b).

Proof Note that by the law of total probability, μ_t and ν_t defined in (AP2) are

$$\begin{aligned}\mu_t(B) &= \int_Y \mathbf{1}_B(\hat{\varphi}_t(\hat{\sigma}_t(h_t), y_t, a_t)) \mu_t^y(dy_t), \\ \nu_t(B) &= \int_Y \mathbf{1}_B(\hat{\varphi}_t(\hat{\sigma}_t(h_t), y_t, a_t)) \nu_t^y(dy_t).\end{aligned}$$

Thus, for any function $f: \hat{Z}_{t+1} \rightarrow \mathbb{R}$,

$$\begin{aligned}\int_{\hat{Z}_{t+1}} f d\mu_t &= \int_{Y_t} f(\hat{\varphi}_t(\hat{\sigma}_t(h_t), y_t, a_t)) \mu_t^y(dy_t), \\ \int_{\hat{Z}_{t+1}} f d\nu_t &= \int_{Y_t} f(\hat{\varphi}_t(\hat{\sigma}_t(h_t), y_t, a_t)) \nu_t^y(dy_t).\end{aligned}$$

The result then follows from (13). ■

Our main result is to establish that any AIS gives rise to an approximate dynamic program.

Theorem 9 *Suppose $\{\hat{\sigma}_t, \hat{P}_t, \hat{r}_t\}_{t=1}^T$ is an $\{(\varepsilon_t, \delta_t)\}_{t=1}^T$ -AIS generator. Recursively define approximate action-value functions $\{\hat{Q}_t: \hat{Z}_t \times A \rightarrow \mathbb{R}\}_{t=1}^T$ and value functions $\{\hat{V}_t: \hat{Z}_t \rightarrow \mathbb{R}\}_{t=1}^T$ as follows: $\hat{V}_{T+1}(\hat{z}_{T+1}) := 0$ and for $t \in \{T, \dots, 1\}$:*

$$\hat{Q}_t(\hat{z}_t, a_t) := \hat{r}_t(\hat{z}_t, a_t) + \int_{\hat{Z}_{t+1}} \hat{V}_{t+1}(\hat{z}_{t+1}) \hat{P}_t(d\hat{z}_{t+1} \mid \hat{z}_t, a_t), \quad (14a)$$

$$\hat{V}_t(\hat{z}_t) := \max_{a_t \in A} \hat{Q}_t(\hat{z}_t, a_t). \quad (14b)$$

Then, we have the following:

1. **Value function approximation:** For any time t , realization h_t of H_t , and choice a_t of A_t , we have

$$|Q_t(h_t, a_t) - \hat{Q}_t(\hat{\sigma}_t(h_t), a_t)| \leq \alpha_t \quad \text{and} \quad |V_t(h_t) - \hat{V}_t(\hat{\sigma}_t(h_t))| \leq \alpha_t, \quad (15)$$

where α_t satisfies the following recursion: $\alpha_{T+1} = 0$ and for $t \in \{T, \dots, 1\}$,

$$\alpha_t = \varepsilon_t + \rho_{\mathfrak{F}}(\hat{V}_{t+1})\delta_t + \alpha_{t+1}.$$

Therefore,

$$\alpha_t = \varepsilon_t + \sum_{\tau=t+1}^T [\rho_{\mathfrak{F}}(\hat{V}_{\tau})\delta_{\tau-1} + \varepsilon_{\tau}].$$

2. **Approximately optimal policy:** Let $\hat{\pi} = (\hat{\pi}_1, \dots, \hat{\pi}_T)$, where $\hat{\pi}_t: \hat{Z}_t \rightarrow \Delta(A)$, be a stochastic policy that satisfies

$$\text{Supp}(\hat{\pi}(\hat{z}_t)) \subseteq \arg \max_{a_t \in A} \hat{Q}_t(\hat{z}_t, a_t). \quad (16)$$

Define policy $\pi = (\pi_1, \dots, \pi_T)$, where $\pi_t: H_t \rightarrow \Delta(A)$ by $\pi_t := \hat{\pi}_t \circ \hat{\sigma}_t$. Then, for any time t , realization h_t of H_t , and choice a_t of A_t , we have

$$|Q_t(h_t, a_t) - Q_t^{\pi}(h_t, a_t)| \leq 2\alpha_t \quad \text{and} \quad |V_t(h_t) - V_t^{\pi}(h_t)| \leq 2\alpha_t. \quad (17)$$

Proof We prove both parts by backward induction. We start with value function approximation. Eq. (15) holds at $T + 1$ by definition. This forms the basis of induction. Assume that (15) holds at time $t + 1$ and consider the system at time t . We have that

$$\begin{aligned}
 & |Q_t(h_t, a_t) - \hat{Q}_t(\hat{\sigma}_t(h_t), a_t)| \\
 & \stackrel{(a)}{\leq} |\mathbb{E}[R_t \mid H_t = h_t, A_t = a_t] - \hat{r}_t(\hat{\sigma}_t(h_t), a_t)| \\
 & \quad + \mathbb{E}[|V_{t+1}(H_{t+1}) - \hat{V}_{t+1}(\hat{\sigma}_{t+1}(H_{t+1}))| \mid H_t = h_t, A_t = a_t] \\
 & \quad + \left| \mathbb{E}[\hat{V}_{t+1}(\hat{\sigma}_{t+1}(H_{t+1})) \mid H_t = h_t, A_t = a_t] - \int_{\hat{Z}_{t+1}} \hat{V}_{t+1}(\hat{z}_{t+1}) \hat{P}_t(d\hat{z}_{t+1} \mid \hat{\sigma}_t(h_t), a_t) \right| \\
 & \stackrel{(b)}{\leq} \varepsilon_t + \alpha_{t+1} + \rho_{\mathfrak{F}}(\hat{V}_{t+1})\delta_t = \alpha_t
 \end{aligned}$$

where (a) follows from triangle inequality and (b) follows from (AP1), the induction hypothesis, (AP2) and (10). This proves the first part of (15). The second part follows from

$$|V_t(h_t) - \hat{V}_t(\hat{\sigma}_t(h_t))| \stackrel{(a)}{\leq} \max_{a_t \in \mathbf{A}} |Q_t(h_t, a_t) - \hat{Q}_t(\hat{\sigma}_t(h_t), a_t)| \leq \alpha_t,$$

where (a) follows from the inequality $\max f(x) \leq \max |f(x) - g(x)| + \max g(x)$.

To prove the policy approximation, we first prove an intermediate result. For policy $\hat{\pi}$ recursively define $\{\hat{Q}_t^{\hat{\pi}}: \hat{Z} \times \mathbf{A} \rightarrow \mathbb{R}\}_{t=1}^T$ and $\{\hat{V}_t^{\hat{\pi}}: \hat{Z} \rightarrow \mathbb{R}\}_{t=1}^{T+1}$ as follows: $\hat{V}_{T+1}^{\hat{\pi}}(\hat{z}_{T+1}) := 0$ and for $t \in \{T, \dots, 1\}$:

$$\hat{Q}_t^{\hat{\pi}}(\hat{z}_t, a_t) := \hat{r}_t(\hat{z}_t, a_t) + \int_{\hat{Z}_{t+1}} \hat{V}_{t+1}^{\hat{\pi}}(\hat{z}_{t+1}) \hat{P}_t(d\hat{z}_{t+1} \mid \hat{z}_t, a_t) \quad (18a)$$

$$\hat{V}_t^{\hat{\pi}}(\hat{z}_t) := \sum_{a_t \in \mathbf{A}} \hat{\pi}_t(a_t \mid \hat{z}_t) \cdot \hat{Q}_t^{\hat{\pi}}(\hat{z}_t, a_t). \quad (18b)$$

Note that (16) implies that

$$\hat{Q}_t^{\hat{\pi}}(\hat{z}_t, a_t) = \hat{Q}_t(\hat{z}_t, a_t) \quad \text{and} \quad \hat{V}_t^{\hat{\pi}}(\hat{z}_t) = \hat{V}_t(\hat{z}_t). \quad (19)$$

Now, we prove that

$$|Q_t^{\pi}(h_t, a_t) - \hat{Q}_t^{\hat{\pi}}(\hat{\sigma}_t(h_t), a_t)| \leq \alpha_t \quad \text{and} \quad |V_t^{\pi}(h_t) - \hat{V}_t^{\hat{\pi}}(\hat{\sigma}_t(h_t))| \leq \alpha_t. \quad (20)$$

We prove the result by backward induction. By construction, Eq. (20) holds at time $T + 1$. This forms the basis of induction. Assume that (20) holds at time $t + 1$ and consider the system at time t . We have

$$\begin{aligned}
 & |Q_t^{\pi}(h_t, a_t) - \hat{Q}_t^{\hat{\pi}}(\hat{\sigma}_t(h_t), a_t)| \\
 & \stackrel{(a)}{\leq} |\mathbb{E}[R_t \mid H_t = h_t, A_t = a_t] - \hat{r}_t(\hat{\sigma}_t(h_t), a_t)| \\
 & \quad + \mathbb{E}[|V_{t+1}^{\pi}(H_{t+1}) - \hat{V}_{t+1}^{\hat{\pi}}(\hat{\sigma}_{t+1}(H_{t+1}))| \mid H_t = h_t, A_t = a_t] \\
 & \quad + \left| \mathbb{E}[\hat{V}_{t+1}^{\hat{\pi}}(\hat{\sigma}_{t+1}(H_{t+1})) \mid H_t = h_t, A_t = a_t] - \int_{\hat{Z}_{t+1}} \hat{V}_{t+1}^{\hat{\pi}}(\hat{z}_{t+1}) \hat{P}_t(d\hat{z}_{t+1} \mid \hat{\sigma}_t(h_t), a_t) \right| \\
 & \stackrel{(b)}{\leq} \varepsilon_t + \alpha_{t+1} + \rho_{\mathfrak{F}}(\hat{V}_{t+1})\delta_t = \alpha_t
 \end{aligned}$$

where (a) follows from triangle inequality and (b) follows from (AP1), the induction hypothesis, (AP2) and (10). This proves the first part of (20). The second part follows from the triangle inequality:

$$|V_t^\pi(h_t) - \hat{V}_t^{\hat{\pi}}(\hat{\sigma}_t(h_t))| \leq \sum_{a_t \in \mathbf{A}} \hat{\pi}_t(a_t | \hat{\sigma}_t(h_t)) |Q_t^\pi(h_t, a_t) - \hat{Q}_t^{\hat{\pi}}(\hat{\sigma}_t(h_t), a_t)| \leq \alpha_t.$$

Now, to prove the policy approximation, we note that

$$|Q_t(h_t, a_t) - Q_t^\pi(h_t, a_t)| \leq |Q_t(h_t, a_t) - \hat{Q}_t^{\hat{\pi}}(\hat{\sigma}_t(h_t), a_t)| + |Q_t^\pi(h_t, a_t) - \hat{Q}_t^{\hat{\pi}}(\hat{\sigma}_t(h_t), a_t)| \leq \alpha_t + \alpha_t,$$

where the first inequality follows from the triangle inequality, the first part of the second inequality follows from (15) and (19) and the second part follows from (20). This proves the first part of (17). The second part of (17) follows from the same argument. \blacksquare

An immediate implication of Theorems 5 and 9 is the following.

Corollary 10 *Let $\{\sigma_t\}_{t=1}^T$ be an information state generator and $\{(\hat{\sigma}_t, \hat{P}_t, \hat{r}_t)\}_{t=1}^T$ be an AIS generator. Then, for any time t , realization h_t of history H_t , and choice a_t of action A_t , we have*

$$|\bar{Q}_t(\sigma_t(h_t), a_t) - \hat{Q}_t(\hat{\sigma}_t(h_t), a_t)| \leq \alpha_t \quad \text{and} \quad |\bar{V}_t(\sigma_t(h_t)) - \hat{V}_t(\hat{\sigma}_t(h_t))| \leq \alpha_t,$$

where \bar{Q}_t and \bar{V}_t are defined as in Theorem 5.

Remark 11 *It is possible to derive a tighter bound in Theorem 9 and show that*

$$\alpha_t = \varepsilon_t + \Delta_t^*(\hat{V}_{t+1}) + \alpha_{t+1}$$

where

$$\Delta_t^*(\hat{V}_{t+1}) = \sup_{h_t, a_t} \left| \mathbb{E}[\hat{V}_{t+1}(\hat{\sigma}_{t+1}(H_{t+1})) \mid H_t = h_t, A_t = a_t] - \int_{\hat{\mathbf{Z}}_{t+1}} \hat{V}_{t+1}(\hat{z}_{t+1}) \hat{P}_t(d\hat{z}_{t+1} \mid \hat{\sigma}_t(h_t), a_t) \right|$$

The bound presented in Theorem 9 can be then thought of as an upper bound on $\Delta_t^*(\hat{V}_{t+1}) \leq \rho_{\mathfrak{F}}(\hat{V}_{t+1})\delta$ using (10).

Remark 12 *In part 1 of Theorem 9, it is possible to derive an alternative bound*

$$|Q_t(h_t, a_t) - \hat{Q}_t(\hat{\sigma}_t(h_t), a_t)| \leq \alpha'_t \quad \text{and} \quad |V_t(h_t) - \hat{V}_t(\hat{\sigma}_t(h_t))| \leq \alpha'_t$$

where α'_t satisfies the recursion: $\alpha'_{T+1} = 0$ and for $t \in \{T, \dots, 1\}$,

$$\alpha'_t = \varepsilon_t + \rho_{\mathfrak{F}}(V_{t+1})\delta_t + \alpha'_{t+1}.$$

This is because while using the triangle inequality in step (a) in the proof of Theorem 9, we could have alternatively added and subtracted the term $\mathbb{E}[V_{t+1}^\pi(H_{t+1}) \mid H_t = h_t, A_t = a_t]$ instead of $\mathbb{E}[\hat{V}_{t+1}^{\hat{\pi}}(\hat{\sigma}_{t+1}(H_{t+1})) \mid H_t = h_t, A_t = a_t]$. Using this bound, we can also derive an alternative bound for part 2 of the Theorem and show that

$$|Q_t(h_t, a_t) - Q_t^\pi(h_t, a_t)| \leq \alpha_t + \alpha'_t \quad \text{and} \quad |V_t(h_t) - V_t^\pi(h_t)| \leq \alpha_t + \alpha'_t.$$

3.3 Examples of approximate information states

We now present various examples of information state and show that many existing results in the literature may be viewed as a special case of Theorem 9. Some of these examples are for infinite horizon discounted reward version of Theorem 9 (with discount factor $\gamma \in (0, 1)$), which we prove later in Theorem 27.

1. **MODEL APPROXIMATION IN MDPs:** Consider an MDP with state space \mathbf{S} , action space \mathbf{A} , transition kernel $P_t: \mathbf{S} \times \mathbf{A} \rightarrow \Delta(\mathbf{S})$, and per-step reward $r_t: \mathbf{S} \times \mathbf{A} \rightarrow \mathbb{R}$. Consider an approximate model defined on the same state and action spaces with transition kernel $\hat{P}_t: \mathbf{S} \times \mathbf{A} \rightarrow \Delta(\mathbf{S})$ and per-step reward $\hat{r}_t: \mathbf{S} \times \mathbf{A} \rightarrow \mathbb{R}$. Define $\hat{\sigma}_t(S_{1:t}, A_{1:t-1}) = S_t$. Then $\{(\hat{\sigma}_t, \hat{P}_t, \hat{r}_t)\}_{t=1}^T$ is an AIS with

$$\varepsilon_t := \sup_{s \in \mathbf{S}, a \in \mathbf{A}} |r_t(s, a) - \hat{r}_t(s, a)| \quad \text{and} \quad \delta_t = \sup_{s \in \mathbf{S}, a \in \mathbf{A}} d_{\mathfrak{F}}(P_t(\cdot|s, a), \hat{P}_t(\cdot|s, a)).$$

A result similar in spirit to Theorem 9 for this setup for general $d_{\mathfrak{F}}$ is given in Theorem 4.2 of Müller (1997). When $d_{\mathfrak{F}}$ is the Kantorovich metric, a bound for model approximation for infinite horizon setup is provided in Theorem 2 of Asadi et al. (2018). This is similar to our result generalization of Theorem 9 to infinite horizon, which is given in Theorem 27; a bound on $\rho_{\mathfrak{F}}(\hat{V})$ in this case can be obtained using results of Hinderer (2005); Rachelson and Lagoudakis (2010).

2. **STATE ABSTRACTION IN MDPs:** Consider an MDP with state space \mathbf{S} , action space \mathbf{A} , transition kernel $P_t: \mathbf{S} \times \mathbf{A} \rightarrow \Delta(\mathbf{S})$, and per-step reward $r_t: \mathbf{S} \times \mathbf{A} \rightarrow \mathbb{R}$. Consider an abstract model defined over a state space $\hat{\mathbf{S}}$ (which is “smaller” than \mathbf{S}) and the same action space with transition kernel $\hat{P}_t: \hat{\mathbf{S}} \times \mathbf{A} \rightarrow \Delta(\hat{\mathbf{S}})$ and per-step reward $\hat{r}_t: \hat{\mathbf{S}} \times \mathbf{A} \rightarrow \mathbb{R}$. Suppose there is an abstraction function $q: \mathbf{S} \rightarrow \hat{\mathbf{S}}$ and, in state $S \in \mathbf{S}$, we choose an action based on $q(S)$. For such a model, define $\hat{\sigma}_t(S_{1:t}, A_{1:t-1}) = q(S_t)$. Then $\{(\hat{\sigma}_t, \hat{P}_t, \hat{r}_t)\}_{t=1}^T$ is an AIS with

$$\varepsilon_t := \sup_{s \in \mathbf{S}, a \in \mathbf{A}} |r_t(s, a) - \hat{r}_t(q(s), a)| \quad \text{and} \quad \delta_t := \sup_{s \in \mathbf{S}, a \in \mathbf{A}} d_{\mathfrak{F}}(\mu_t(\cdot|s, a), \hat{P}_t(\cdot|q(s), a)),$$

where for any Borel subset B of $\hat{\mathbf{S}}$, $\mu_t(B|s, a) := P_t(q^{-1}(B)|s, a)$.

There is a rich literature on state abstraction starting with Bertsekas (1975) and Whitt (1978), but the error bounds in those papers are of a different nature. There are some recent papers which derive error bounds similar to Theorem 9 for the infinite horizon setup with state abstraction. We generalize Theorem 9 to infinite horizon later in Theorem 27.

When $d_{\mathfrak{F}}$ is the Kantorovich metric, a bound on $\rho_{\mathfrak{F}}(\hat{V}) = \|\hat{V}\|_{\text{Lip}}$ can be obtained using results of Hinderer (2005); Rachelson and Lagoudakis (2010). Substituting this bound in Theorem 27 gives us the following bound on the policy approximation error by using AIS.

$$|V(s) - V^{\pi}(s)| \leq \frac{2\varepsilon}{(1-\gamma)} + \frac{2\gamma\delta\|\hat{V}\|_{\text{Lip}}}{(1-\gamma)}.$$

Similar bound has been obtained in Theorem 5 of Gelada et al. (2019). A detailed comparison with this model is presented in Appendix B.

When $d_{\mathfrak{F}}$ is the total variation distance, a bound on $d_{\mathfrak{F}}(\hat{V})$ is given by $\text{span}(r)/(1-\gamma)$. Substituting this in Theorem 27, we get that

$$|V(s) - V^\pi(s)| \leq \frac{2\varepsilon}{(1-\gamma)} + \frac{\gamma\delta \text{span}(r)}{(1-\gamma)^2}.$$

A $\mathcal{O}(1/(1-\gamma)^3)$ bound on the policy approximation error in this setup was obtained in Lemma 2 and Theorem 2 of Abel et al. (2016). **Directly using the AIS bound of Theorems 9 and 27 gives a factor of $1/(1-\gamma)$ improvement in the error bound of Abel et al. (2016).** See Appendix A for a detailed comparison.

3. **BELIEF APPROXIMATION IN POMDPs:** Consider a POMDP with state space \mathbf{S} , action space \mathbf{A} , observation space \mathbf{Y} , and a per-step reward function $r_t: \mathbf{S} \times \mathbf{A} \rightarrow \mathbb{R}$. Let $b_t(\cdot|H_t) \in \Delta(\mathbf{S})$ denote the belief of the current state given the history, i.e., $b_t(s|H_t) = \mathbb{P}(S_t = s | H_t)$. Suppose there are history compression functions $\{\phi_t: H_t \rightarrow \Phi_t\}_{t=1}^T$ (where Φ_t is some arbitrary space) along with belief approximation functions $\{\hat{b}_t: \Phi_t \rightarrow \Delta(\mathbf{S})\}_{t=1}^T$, such that for any time t and any realization h_t of H_t , we have

$$\|\hat{b}_t(\cdot | \phi_t(h_t)) - b_t(\cdot | h_t)\|_1 \leq \varepsilon.$$

Such a $\{(\phi_t, \hat{b}_t)\}_{t=1}^T$ was called an ε -sufficient statistic in Francois-Lavet et al. (2019). An example of ε -sufficient statistic is belief quantization, where the belief is quantized to the nearest point in the *type lattice* (here $m = |\mathbf{S}|$)

$$Q_n := \{(p_1, \dots, p_m) \in \Delta(\mathbf{S}) : np_i \in \mathbb{Z}_{\geq 0}\}.$$

An efficient algorithm to find the nearest point in Q_n for any given belief $b_t \in \Delta(\mathbf{S})$ is presented in Reznik (2011). Under such a quantization, the maximum ℓ_1 distance between a belief vector and its quantized value is given by $2\lfloor m/2 \rfloor \lceil m/2 \rceil / mn \approx m/2n$ (see Proposition 2 of Reznik (2011)). Thus, by taking $n > m/2\varepsilon$, we get an ε -sufficient statistic.

Francois-Lavet et al. (2019) showed that the bias of using the optimal policy based on $\hat{b}_t(h_t)$ in the original model is $2\varepsilon\|r\|_\infty/(1-\gamma)^3$. This result uses the same proof argument as Abel et al. (2016) discussed in the previous bullet point, which is not tight. By metricizing the belief space using total variation distance and using the bounded-Lipschitz metric on the space of probability measures on beliefs, we can show that an ε -sufficient statistic induces a $(\varepsilon \text{span}(r), 3\varepsilon)$ -AIS. When $d_{\mathfrak{F}}$ is the bounded-Lipschitz metric, a bound on $\rho_{\mathfrak{F}}(\hat{V})$ is given by $2\|r\|_\infty/(1-\gamma)$. Substituting this in Theorem 27, we get that

$$|V(s) - V^\pi(s)| \leq \frac{2\varepsilon\|r\|_\infty}{(1-\gamma)} + \frac{6\gamma\varepsilon\|r\|_\infty}{(1-\gamma)^2}.$$

Thus, **directly using the AIS bound of Theorems 9 and 27 gives a factor of $1/(1-\gamma)$ improvement in the error bound of Francois-Lavet et al. (2019).** See Appendix C for details.

In a slightly different vein, belief quantization in POMDPs with finite or Borel valued unobserved state was investigated in Saldi et al. (2018), who showed that

under appropriate technical conditions the value function and optimal policies for the quantized model converge to the value function and optimal policy of the true model. However Saldi et al. (2018) did not provide approximation error for a fixed quantization level.

3.4 Approximate policy evaluation

In some settings, we are interested in comparing the performance of an arbitrary policy in an approximate model with its performance in the real model. The bounds of Theorem 9 can be adapted to such a setting as well.

Theorem 13 *Suppose $\{\hat{\sigma}_t, \hat{P}_t, \hat{r}_t\}_{t=1}^T$ is an $\{(\varepsilon_t, \delta_t)\}_{t=1}^T$ -AIS generator. Let $\hat{\pi}^\# = (\hat{\pi}_1^\#, \dots, \hat{\pi}_T^\#)$, where $\hat{\pi}_t^\# : \hat{Z}_t \rightarrow \Delta(\mathbf{A})$, be an arbitrary stochastic policy. Recursively define approximate policy action-value functions $\{\hat{Q}_t^{\hat{\pi}^\#} : \hat{Z}_t \times \mathbf{A} \rightarrow \mathbb{R}\}_{t=1}^T$ and value functions $\{\hat{V}_t^{\hat{\pi}^\#} : \hat{Z}_t \rightarrow \mathbb{R}\}_{t=1}^T$ as follows: $\hat{V}_{T+1}^{\hat{\pi}^\#}(\hat{z}_{T+1}) := 0$ and for $t \in \{T, \dots, 1\}$:*

$$\hat{Q}_t^{\hat{\pi}^\#}(\hat{z}_t, a_t) := \hat{r}_t(\hat{z}_t, a_t) + \int_{\hat{Z}_{t+1}} \hat{V}_t^{\hat{\pi}^\#}(\hat{z}_{t+1}) \hat{P}_t(d\hat{z}_{t+1} \mid \hat{z}_t, a_t) \quad (21a)$$

$$\hat{V}_t^{\hat{\pi}^\#}(\hat{z}_t) := \sum_{a_t \in \mathbf{A}} \hat{\pi}_t^\#(a_t \mid \hat{z}_t) \cdot \hat{Q}_t^{\hat{\pi}^\#}(\hat{z}_t, a_t). \quad (21b)$$

Define policy $\pi^\# = (\pi_1^\#, \dots, \pi_T^\#)$, where $\pi_t^\# : \mathbf{H}_t \rightarrow \Delta(\mathbf{A})$ by $\pi_t^\# := \hat{\pi}_t^\# \circ \hat{\sigma}_t$. Then, for any time t , realization h_t of H_t , and choice a_t of A_t , we have:

$$|Q_t^{\pi^\#}(h_t, a_t) - \hat{Q}_t^{\hat{\pi}^\#}(\hat{\sigma}_t(h_t), a_t)| \leq \alpha_t^\# \quad \text{and} \quad |V_t^{\pi^\#}(h_t) - \hat{V}_t^{\hat{\pi}^\#}(\hat{\sigma}_t(h_t))| \leq \alpha_t^\#, \quad (22)$$

where $\alpha_t^\#$ satisfies the following recursion: $\alpha_{T+1}^\# = 0$ and for $t \in \{T, \dots, 1\}$,

$$\alpha_t^\# = \varepsilon_t + \rho_{\mathfrak{F}}(\hat{V}_{t+1}^{\hat{\pi}^\#})\delta_t + \alpha_{t+1}^\#.$$

Therefore,

$$\alpha_t^\# = \varepsilon_t + \sum_{\tau=t+1}^T [\rho_{\mathfrak{F}}(\hat{V}_\tau^{\hat{\pi}^\#})\delta_{\tau-1} + \varepsilon_\tau].$$

Proof The proof proceeds by backward induction along the same lines as the proof of Theorem 9. By construction, Eq. (22) holds at time $T+1$. This forms the basis of induction. Assume that (22) holds at time $t+1$ and consider the system at time t . We have

$$\begin{aligned} & |Q_t^{\pi^\#}(h_t, a_t) - \hat{Q}_t^{\hat{\pi}^\#}(\hat{\sigma}_t(h_t), a_t)| \\ & \stackrel{(a)}{\leq} |\mathbb{E}[R_t \mid H_t = h_t, A_t = a_t] - \hat{r}_t(\hat{\sigma}_t(h_t), a_t)| \\ & \quad + \mathbb{E}[|V_{t+1}^{\pi^\#}(H_{t+1}) - \hat{V}_{t+1}^{\hat{\pi}^\#}(\hat{\sigma}_{t+1}(H_{t+1}))| \mid H_t = h_t, A_t = a_t] \\ & \quad + \left| \mathbb{E}[\hat{V}_{t+1}^{\hat{\pi}^\#}(\hat{\sigma}_{t+1}(H_{t+1})) \mid H_t = h_t, A_t = a_t] - \int_{\hat{Z}_{t+1}} \hat{V}_{t+1}^{\hat{\pi}^\#}(\hat{z}_{t+1}) \hat{P}_t(d\hat{z}_{t+1} \mid \hat{\sigma}_t(h_t), a_t) \right| \\ & \stackrel{(b)}{\leq} \varepsilon_t + \alpha_{t+1}^\# + \rho_{\mathfrak{F}}(\hat{V}_{t+1}^{\hat{\pi}^\#})\delta_t = \alpha_t^\# \end{aligned}$$

where (a) follows from triangle inequality and (b) follows from (AP1), the induction hypothesis, (AP2) and (10). This proves the first part of (22). The second part follows from the fact that $\pi^\#(a_t|h_t) = \hat{\pi}^\#(a_t|\hat{\sigma}_t(h_t))$ and the triangle inequality:

$$|V_t^{\pi^\#}(h_t) - \hat{V}_t^{\hat{\pi}^\#}(\hat{\sigma}_t(h_t))| \leq \sum_{a_t \in \mathbf{A}} \hat{\pi}_t^\#(a_t|\hat{\sigma}_t(h_t)) |Q^{\pi^\#}(h_t, a_t) - \hat{Q}_t^{\hat{\pi}^\#}(\hat{\sigma}_t(h_t), a_t)| \leq \alpha_t^\#.$$

■

3.5 Stochastic AIS

We have so far assumed that the history compression functions $\hat{\sigma}_t: \mathbf{H}_t \rightarrow \hat{\mathbf{Z}}_t$ are deterministic functions. When learning a discrete-valued AIS from data, it is helpful to consider stochastic mappings of history, so that quality of the mapping may be improved via stochastic gradient descent. In general, the definition of deterministic AIS also covers the case of stochastic AIS because a stochastic function from \mathbf{H}_t to $\hat{\mathbf{Z}}_t$ may be viewed as a deterministic function from \mathbf{H}_t to $\Delta(\hat{\mathbf{Z}}_t)$. However, a more explicit characterization is also possible, which we present next.

Definition 14 Let $\{\hat{\mathbf{Z}}_t\}_{t=1}^T$ be a pre-specified collection of Banach spaces, \mathfrak{F} be a function class for IPMs, and $\{(\varepsilon_t, \delta_t)\}_{t=1}^T$ be pre-specified positive real numbers. A collection $\{\hat{\sigma}_t^s: \mathbf{H}_t \rightarrow \Delta(\hat{\mathbf{Z}}_t)\}_{t=1}^T$ of stochastic history compression functions, along with approximate update kernels $\{\hat{P}_t: \hat{\mathbf{Z}}_t \times \mathbf{A} \rightarrow \Delta(\hat{\mathbf{Z}}_{t+1})\}_{t=1}^T$ and reward approximation functions $\{\hat{r}_t: \hat{\mathbf{Z}}_t \times \mathbf{A} \rightarrow \mathbb{R}\}_{t=1}^T$, is called an $\{(\varepsilon_t, \delta_t)\}_{t=1}^T$ -stochastic AIS generator if the process $\{\hat{\mathbf{Z}}_t\}_{t=1}^T$, where $\hat{\mathbf{Z}}_t = \hat{\sigma}_t(H_t)$, satisfies the following properties:

(AP1) Sufficient for approximate performance evaluation, i.e., for any time t , any realization h_t of H_t and any choice a_t of A_t , we have

$$|\mathbb{E}[R_t | H_t = h_t, A_t = a_t] - \mathbb{E}_{\hat{\mathbf{Z}}_t \sim \hat{\sigma}_t^s(h_t)}[\hat{r}_t(\hat{\mathbf{Z}}_t, a_t)]| \leq \varepsilon_t.$$

(AP2) Sufficient to predict itself approximately. i.e., for any time t , any realization h_t of H_t , any choice a_t of A_t , and for any Borel subset \mathbf{B} of $\hat{\mathbf{Z}}_{t+1}$, define $\mu_t(\mathbf{B}) := \mathbb{P}(\hat{\mathbf{Z}}_{t+1} \in \mathbf{B} | H_t = h_t, A_t = a_t)$ and $\nu_t(\mathbf{B}) := \mathbb{E}_{\hat{\mathbf{Z}}_t \sim \hat{\sigma}_t^s(h_t)}[\hat{P}_t(\mathbf{B} | \hat{\mathbf{Z}}_t, a_t)]$; then,

$$d_{\mathfrak{F}}(\mu_t, \nu_t) \leq \delta_t.$$

Similar to Theorem 9, we then have the following result.

Theorem 15 Given a stochastic AIS generator $\{\hat{\sigma}_t^s, \hat{P}_t, \hat{r}_t\}_{t=1}^T$, define value functions $\{\hat{V}_t: \hat{\mathbf{Z}}_t \rightarrow \mathbb{R}\}_{t=1}^T$ and action-value functions $\{\hat{Q}_t: \hat{\mathbf{Z}}_t \times \mathbf{A} \rightarrow \mathbb{R}\}_{t=1}^T$ as in Theorem 9. Then, we have the following:

1. **Value function approximation:** For any time t , realization h_t of H_t , and choice a_t of A_t we have

$$|Q_t(h_t, a_t) - \mathbb{E}_{\hat{\mathbf{Z}}_t \sim \hat{\sigma}_t^s(h_t)}[\hat{Q}_t(\hat{\mathbf{Z}}_t, a_t)]| \leq \alpha_t \quad \text{and} \quad |V_t(h_t) - \mathbb{E}_{\hat{\mathbf{Z}}_t \sim \hat{\sigma}_t^s(h_t)}[\hat{V}_t(\hat{\mathbf{Z}}_t)]| \leq \alpha_t, \quad (23)$$

where α_t is defined as in Theorem 9.

2. **Approximately optimal policy:** Let $\hat{\pi} = (\hat{\pi}_1, \dots, \hat{\pi}_T)$, where $\hat{\pi}_t: \hat{Z}_t \rightarrow \Delta(\mathbf{A})$, be a stochastic policy that satisfies

$$\text{Supp}(\hat{\pi}(\hat{z}_t)) \subseteq \arg \max_{a_t \in \mathbf{A}} \hat{Q}_t(\hat{z}_t, a_t). \quad (24)$$

Define policy $\pi = (\pi_1, \dots, \pi_T)$, where $\pi_t: H_t \rightarrow \Delta(\mathbf{A})$ by $\pi_t(h_t) = \mathbb{E}_{\hat{Z}_t \sim \hat{\sigma}_t^s(h_t)}[\hat{\pi}_t(\hat{Z}_t)]$. Then, for any time t , realization h_t of H_t , and choice a_t of A_t , we have

$$|Q_t(h_t, a_t) - Q_t^\pi(h_t, a_t)| \leq 2\alpha_t \quad \text{and} \quad |V_t(h_t) - V_t^\pi(h_t)| \leq 2\alpha_t. \quad (25)$$

Proof The proof is almost the same as the proof of Theorem 9. The main difference is that for the value and action-value functions of the stochastic approximation state, we take an additional expectation over the realization of the stochastic AIS. We only show the details of the proof of the first part of the result (value approximation). The second part (policy approximation) follows along similar lines.

Eq. (23) holds at $T+1$ by definition. This forms the basis of induction. Assume that (23) holds at time $t+1$ and consider the system at time t . We have that

$$\begin{aligned} & |Q_t(h_t, a_t) - \mathbb{E}_{\hat{Z}_t \sim \hat{\sigma}_t^s(h_t)}[\hat{Q}_t(\hat{Z}_t, a_t)]| \\ & \stackrel{(a)}{\leq} |\mathbb{E}[R_t \mid H_t = h_t, A_t = a_t] - \mathbb{E}_{\hat{Z}_t \sim \hat{\sigma}_t^s(h_t)}[\hat{r}_t(\hat{Z}_t, a_t)]| \\ & \quad + \mathbb{E}[|V_{t+1}(H_{t+1}) - \mathbb{E}_{\hat{Z}_{t+1} \sim \hat{\sigma}_{t+1}^s(h_{t+1})}[\hat{V}_{t+1}(\hat{Z}_{t+1})]| \mid H_t = h_t, A_t = a_t] \\ & \quad + \left| \mathbb{E}[\hat{V}_{t+1}(\hat{\sigma}_{t+1}^s(H_{t+1})) \mid H_t = h_t, A_t = a_t] - \mathbb{E}_{\hat{Z}_t \sim \hat{\sigma}_t^s(h_t)} \left[\int_{\hat{Z}_t} \hat{V}_{t+1}(\hat{z}_{t+1}) \hat{P}_t(d\hat{z}_{t+1} \mid \hat{Z}_t, a_t) \right] \right| \\ & \stackrel{(b)}{\leq} \varepsilon_t + \alpha_{t+1} + \rho_{\mathfrak{F}}(\hat{V}_{t+1})\delta_t = \alpha_t \end{aligned}$$

where (a) follows from triangle inequality and (b) follows from (AP1), the induction hypothesis, (AP2) and (10). This proves the first part of (23). The second part follows from

$$|V_t(h_t) - \hat{V}_t(\hat{\sigma}_t^s(h_t))| \stackrel{(a)}{\leq} \max_{a_t \in \mathbf{A}} |Q_t(h_t, a_t) - \hat{Q}_t(\hat{\sigma}_t^s(h_t), a_t)| \leq \alpha_t,$$

where (a) follows from the inequality $\max f(x) \leq \max |f(x) - g(x)| + \max g(x)$. This completes the proof of value approximation. The proof of policy approximation is similar to that of Theorem 9 adapted in the same manner as above. \blacksquare

3.6 AIS with action compression

So far we have assumed that the action space for the AIS is the same as the action space for the original model. In some instances, for example, for continuous or large action spaces, it may be desirable to quantize or compress the actions as well. In this section, we generalize the notion of AIS to account for action compression.

Definition 16 As in the definition of AIS, suppose $\{\hat{Z}_t\}_{t=1}^T$ are pre-specified collection of Banach spaces, \mathfrak{F} be a function class for IPMs, and $\{(\varepsilon_t, \delta_t)\}_{t=1}^T$ be pre-specified positive real numbers. In addition, suppose we have a subset $\hat{A} \subset A$ of quantized actions. Then, a collection $\{\hat{\sigma}_t: H_t \rightarrow \hat{Z}_t\}_{t=1}^T$ of history compression functions, along with action quantization function $\psi: A \rightarrow \hat{A}$, approximate update kernels $\{\hat{P}_t: \hat{Z}_t \times \hat{A} \rightarrow \Delta(\hat{Z}_{t+1})\}_{t=1}^T$ and reward approximation functions $\{\hat{r}_t: \hat{Z}_t \times \hat{A} \rightarrow \mathbb{R}\}_{t=1}^T$, is called an $\{(\varepsilon_t, \delta_t)\}_{t=1}^T$ -action-quantized AIS generator if the process $\{\hat{Z}_t\}_{t=1}^T$, where $\hat{Z}_t = \hat{\sigma}_t(H_t)$, satisfies the following properties:

(AQ1) Sufficient for approximate performance evaluation, i.e., for any time t , any realization h_t of H_t and any choice a_t of A_t , we have

$$|\mathbb{E}[R_t \mid H_t = h_t, A_t = a_t] - \hat{r}_t(\hat{\sigma}_t(h_t), \psi(a_t))| \leq \varepsilon_t.$$

(AQ2) Sufficient to predict itself approximately. i.e., for any time t , any realization h_t of H_t , any choice a_t of A_t , and for any Borel subset B of \hat{Z}_{t+1} , define $\mu_t(B) := \mathbb{P}(\hat{Z}_{t+1} \in B \mid H_t = h_t, A_t = a_t)$ and $\nu_t(B) := \hat{P}_t(B \mid \hat{\sigma}_t(h_t), \psi(a_t))$; then,

$$d_{\mathfrak{F}}(\mu_t, \nu_t) \leq \delta_t.$$

Similar to Theorem 9, we show that an action-quantized AIS can be used to determine an approximately optimal policy.

Theorem 17 Suppose $\{\hat{\sigma}_t, \psi, \hat{P}_t, \hat{r}_t\}_{t=1}^T$ is an action-quantized AIS generator. Recursively define approximate action-value functions $\{\hat{Q}_t: \hat{Z}_t \times \hat{A} \rightarrow \mathbb{R}\}_{t=1}^T$ and value functions $\{\hat{V}_t: \hat{Z}_t \rightarrow \mathbb{R}\}_{t=1}^T$ as follows: $\hat{V}_{T+1}(\hat{z}_{T+1}) := 0$ and for $t \in \{T, \dots, 1\}$:

$$\hat{Q}_t(\hat{z}_t, \hat{a}_t) := \hat{r}_t(\hat{z}_t, \hat{a}_t) + \int_{\hat{Z}_t} \hat{V}_{t+1}(\hat{z}_{t+1}) \hat{P}_t(d\hat{z}_{t+1} \mid \hat{z}_t, \hat{a}_t), \quad (26a)$$

$$\hat{V}_t(\hat{z}_t) := \max_{\hat{a}_t \in \hat{A}} \hat{Q}_t(\hat{z}_t, \hat{a}_t). \quad (26b)$$

Then, we have the following:

1. **Value function approximation:** For any time t , realization h_t of H_t , and choice a_t of A_t , we have

$$|Q_t(h_t, a_t) - \hat{Q}_t(\hat{\sigma}_t(h_t), \psi(a_t))| \leq \alpha_t \quad \text{and} \quad |V_t(h_t) - \hat{V}_t(\hat{\sigma}_t(h_t))| \leq \alpha_t, \quad (27)$$

where α_t is defined as in Theorem 9.

2. **Approximately optimal policy:** Let $\hat{\pi} = (\hat{\pi}_1, \dots, \hat{\pi}_T)$, where $\hat{\pi}_t: \hat{Z}_t \rightarrow \Delta(\hat{A})$, be a stochastic policy that satisfies

$$\text{Supp}(\hat{\pi}_t(\hat{z}_t)) \subseteq \arg \max_{\hat{a}_t \in \hat{A}} \hat{Q}_t(\hat{z}_t, \hat{a}_t). \quad (28)$$

Define policy $\pi = (\pi_1, \dots, \pi_T)$, where $\pi_t: H_t \rightarrow \Delta(A)$ by $\pi_t := \hat{\pi}_t \circ \hat{\sigma}_t$. Then, for any time t , realization h_t of H_t , and choice a_t of A_t , we have

$$|Q_t(h_t, a_t) - Q_t^\pi(h_t, \psi(a_t))| \leq 2\alpha_t \quad \text{and} \quad |V_t(h_t) - V_t^\pi(h_t)| \leq 2\alpha_t. \quad (29)$$

Proof The proof is similar to the proof of Theorem 9. We only show the details of the first part (value approximation). The second part (policy approximation) follows along similar lines.

As before, we prove the result by backward induction. Eq. (27) holds at $T + 1$ by definition. This forms the basis of induction. Assume that (27) holds at time $t + 1$ and consider the system at time t . We have that

$$\begin{aligned}
 & |Q_t(h_t, a_t) - \hat{Q}_t(\hat{\sigma}_t(h_t), \psi(a_t))| \\
 & \stackrel{(a)}{\leq} |\mathbb{E}[R_t \mid H_t = h_t, A_t = a_t] - \hat{r}_t(\hat{\sigma}_t(h_t), \psi(a_t))| \\
 & \quad + \mathbb{E}[|V_{t+1}(H_{t+1}) - \hat{V}_{t+1}(\hat{\sigma}_{t+1}(H_{t+1}))| \mid H_t = h_t, A_t = a_t] \\
 & \quad + \left| \mathbb{E}[\hat{V}_{t+1}(\hat{\sigma}_{t+1}(H_{t+1})) \mid H_t = h_t, A_t = a_t] - \int_{\hat{\mathbf{Z}}_t} \hat{V}_{t+1}(\hat{z}_{t+1}) \hat{P}_t(d\hat{z}_{t+1} \mid \hat{\sigma}_t(h_t), \hat{a}_t) \right| \\
 & \stackrel{(b)}{\leq} \varepsilon_t + \alpha_{t+1} + \rho_{\mathfrak{F}}(\hat{V}_{t+1})\delta_t = \alpha_t
 \end{aligned}$$

where (a) follows from triangle inequality and (b) follows from (AQ1), the induction hypothesis, (AQ2) and (10). This proves the first part of (27). The second part follows from

$$|V_t(h_t) - \hat{V}_t(\hat{\sigma}_t(h_t))| \stackrel{(a)}{\leq} \max_{a_t \in \mathbf{A}} |Q_t(h_t, a_t) - \hat{Q}_t(\hat{\sigma}_t(h_t), \psi(a_t))| \leq \alpha_t,$$

where (a) follows from the inequality $\max f(x) \leq \max |f(x) - g(x)| + \max g(x)$. We have also used the fact that if ψ is an onto function, then $\max_{\hat{a} \in \hat{\mathbf{A}}} \hat{Q}_t(\hat{z}_t, \hat{a}_t) = \max_{a \in \mathbf{A}} \hat{Q}_t(\hat{z}_t, \psi(a_t))$. This completes the proof of value approximation. The proof of policy approximation is similar to that of Theorem 9 adapted in the same manner as above. \blacksquare

Action quantization in POMDPs with finite or Borel valued unobserved state was investigated in Saldi et al. (2018), who showed that under appropriate technical conditions the value function and optimal policies for the quantized model converge to the value function and optimal policy of the true model. However Saldi et al. (2018) did not provide approximation error for a fixed quantization level.

Simplification for perfectly observed case: The approximation bounds for action compression derived in Theorem 17 can be simplified when the system is perfectly observed. In particular, consider an MDP with state space \mathbf{S} , action space \mathbf{A} , transition probability $P: \mathbf{S} \times \mathbf{A} \rightarrow \Delta(\mathbf{S})$, per-step reward function $r: \mathbf{S} \times \mathbf{A} \rightarrow \mathbb{R}$, and discount factor γ .

For MDPs, we can simplify the definition of action quantized AIS-generator as follows.

Definition 18 *Given an MDP as defined above, let \mathfrak{F} be a function class for IPMs, and (ε, δ) be pre-specified positive real numbers. In addition, suppose we have a subset $\hat{\mathbf{A}} \subset \mathbf{A}$ of quantized actions. Then, an action quantization function $\psi: \mathbf{A} \rightarrow \hat{\mathbf{A}}$, where $\hat{\mathbf{A}} \subset \mathbf{A}$, is called an (ε, δ) -action-quantizer if the following properties are satisfied:*

(AQM1) Sufficient for approximate performance evaluation, i.e., for any $s \in \mathbf{S}$ and $a \in \mathbf{A}$, we have

$$|r(s, a) - r(s, \psi(a))| \leq \varepsilon.$$

(AQM2) Sufficient to predict the next state approximately. *i.e., for any $s \in \mathcal{S}$ and $a \in \mathcal{A}$,*

$$d_{\mathcal{F}}(P(\cdot|s, a), P(\cdot|s, \psi(a))) \leq \delta.$$

Then, the approximation in Theorem 17 simplifies for an MDP as follows.

Corollary 19 *Suppose ψ is an (ε, δ) -action-quantizer. Recursively define approximate action-value functions $\{\hat{Q}_t: \mathcal{S} \times \hat{\mathcal{A}} \rightarrow \mathbb{R}\}$ and value functions $\{\hat{V}_t: \mathcal{S} \rightarrow \mathbb{R}\}$ as follows: $\hat{V}_{T+1}(s_{T+1}) := 0$ and for $t \in \{T, \dots, 1\}$:*

$$\hat{Q}_t(s_t, \hat{a}_t) := r(s_t, \hat{a}_t) + \int_{\mathcal{S}} \hat{V}_{t+1}(s_{t+1}) P(ds_{t+1} | s_t, \hat{a}_t), \quad (30a)$$

$$\hat{V}_t(s_t) := \max_{\hat{a}_t \in \hat{\mathcal{A}}} \hat{Q}_t(s_t, \hat{a}_t). \quad (30b)$$

Then, we have the following:

1. **Value function approximation:** *For any time t , $s \in \mathcal{S}$ and $a \in \mathcal{A}$, we have*

$$|Q_t(s, a) - \hat{Q}_t(s, \psi(a))| \leq \alpha_t \quad \text{and} \quad |V_t(s) - \hat{V}_t(s)| \leq \alpha_t, \quad (31)$$

where α_t is defined as in Theorem 9.

2. **Approximately optimal policy:** *Let $\hat{\pi} = (\hat{\pi}_1, \dots, \hat{\pi}_T)$, where $\hat{\pi}_t: \mathcal{S} \rightarrow \Delta(\hat{\mathcal{A}})$, be a stochastic policy that satisfies*

$$\text{Supp}(\hat{\pi}_t(s_t)) \subseteq \arg \max_{\hat{a}_t \in \hat{\mathcal{A}}} \hat{Q}_t(s_t, \hat{a}_t). \quad (32)$$

Since $V_t^{\hat{\pi}}(s_t) = \hat{V}_t(s_t)$ and $Q_t^{\hat{\pi}}(s, \hat{a}_t) = \hat{Q}_t(s, \hat{a}_t)$, we have

$$|Q_t(s_t, a_t) - Q_t^{\hat{\pi}}(s_t, \psi(a_t))| \leq \alpha_t \quad \text{and} \quad |V_t(s_t) - V_t^{\hat{\pi}}(s_t)| \leq \alpha_t. \quad (33)$$

Proof The proof follows in a straightforward manner from the proof of Theorem 17. ■

Note that in contrast to Theorem 17, the final approximation bounds (33) in Corollary 19 do not have an additional factor of 2. This is because the approximate policy $\hat{\pi}$ can be directly executed in the original MDP because $\hat{\mathcal{A}} \subset \mathcal{A}$.

Approximation bounds similar to Corollary 19 are used to derive bounds for lifelong learning in Chandak et al. (2020). We show that similar bounds may be obtained using Corollary 19 in Appendix D.

3.7 AIS with observation compression

In applications with high-dimensional observations such as video input, it is desirable to pre-process the video frames into a low-dimensional representation before passing them on to a planning or learning algorithm. In this section, we generalize the notion of AIS to account for such observation compression.

Definition 20 *As in the definition of AIS, suppose $\{\hat{Z}_t\}_{t=1}^T$ are a pre-specified collection of Banach spaces, \mathfrak{F} be a function class for IPMs, and $\{(\varepsilon_t, \delta_t)\}_{t=1}^T$ be pre-specified positive real numbers. In addition, suppose we have a set \hat{Y} of compressed observations and a compression function $q: Y \rightarrow \hat{Y}$. Let \hat{H}_t denote the history $(\hat{Y}_{1:t-1}, A_{1:t-1})$ of compressed observations and actions and \hat{H}_t denote the space of realizations of such compressed histories. Then, a collection $\{\hat{\sigma}_t: \hat{H}_t \rightarrow \hat{Z}_t\}_{t=1}^T$ of history compression functions, along with observation compression function $q: Y \rightarrow \hat{Y}$, approximate update kernels $\{\hat{P}_t: \hat{Z}_t \times A \rightarrow \Delta(\hat{Z}_{t+1})\}_{t=1}^T$ and reward approximation functions $\{\hat{r}_t: \hat{Z}_t \times A \rightarrow \mathbb{R}\}_{t=1}^T$, is called an $\{(\varepsilon_t, \delta_t)\}_{t=1}^T$ -observation-compressed AIS generator if the process $\{\hat{Z}_t\}_{t=1}^T$, where $\hat{Z}_t = \hat{\sigma}_t(\hat{H}_t)$, satisfies properties (AP1) and (AP2).*

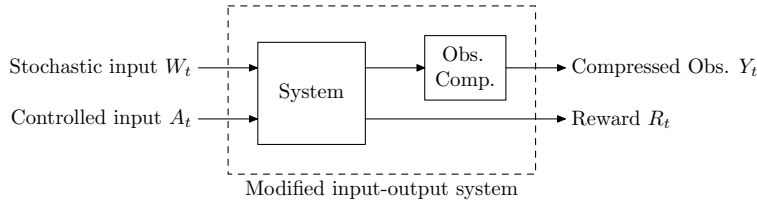


Figure 3: A stochastic input-output system with observation compression

In essence, we can view observation compression as a new input-output system whose outputs are (\hat{Y}_t, R_t) instead of (Y_t, R_t) as shown in Fig. 3. A construction similar to observation-compressed AIS is proposed in Ha and Schmidhuber (2018), where it is shown that such a construction performs well empirically, but there was no analysis of the approximation guarantees of such a construction.

An immediate implication of the above definition is the following:

Corollary 21 *Let $\{\hat{\sigma}_t, q, \hat{P}_t, \hat{r}_t\}_{t \geq 1}$ be an $\{(\varepsilon_t, \delta_t)\}_{t=1}^T$ -observation-compression AIS. Then, the bounds of Theorem 9 hold.*

3.8 Discussion and related work

AIS may be viewed as a generalization of state discretization (Bertsekas, 1975) or state aggregation (Whitt, 1978) in MDPs. As illustrated by the examples in Sec. 3.3, many of the recent results on approximation bounds for state aggregation and latent state embedding in MDPs are specific instances of AIS and, in some instances, using the approximation bounds of Theorem 9 or its generalization to infinite horizon (Theorem 27) provide tighter bounds than those in the literature. A detailed comparison with these results is presented in the Appendices. We had presented a simpler definition of AIS and the approximation bounds in the preliminary version of this paper (Subramanian and Mahajan, 2019).

As mentioned in Sec. 2.5 while discussing the related literature on information states, there are two other methods for identifying “states” for POMDPs: bisimulation-based methods and predictive state representations (PSRs). Approximation techniques for both these methods have been proposed in the literature.

State aggregation techniques based on bisimulation metrics have been proposed in Ferns et al. (2004, 2011) for MDPs and Castro et al. (2009) for POMDPs. The key insight of these

papers is to define a semi-metric called bisimulation metric on the state space of an MDP or the belief space of a POMDP as the unique fixed point of an operator on the space of semi-metrics on the state space of the MDP or the belief space of the POMDP. It is then shown that the value function is Lipschitz with respect to this metric. Then, they propose state aggregation based on the bisimulation metric. Although the basic building blocks of bisimulation metrics are the same as those of an AIS, the approximation philosophies are different. The bisimulation-metric based approximations are a form of state aggregation, while AIS need not be a state aggregation.

Various methods for learning low dimensional approximations of PSRs have been proposed in the literature, including approaches which use spectral learning algorithms (Rosencrantz et al., 2004; Boots et al., 2011; Hamilton et al., 2014; Kulesza et al., 2015b,a; Jiang et al., 2016), and stochastic gradient descent (Jiang et al., 2016). Error bounds for using an approximate PSR were derived in Wolfe et al. (2008); Hamilton et al. (2014). These approximation methods for PSRs rely on the specific structure of PSRs and are conceptually different from the approximation methods used in AIS.

4. Infinite-horizon discounted reward setup

So far, we have restricted attention to the finite horizon setup. In this section, we show how to generalize the notions of information state and approximate information state to the infinite horizon discounted reward setup.

4.1 System model and problem formulation

We consider the same model as described in Sec. 2.1 but assume that the system runs for an infinite horizon. The performance of any (history dependent and possibly stochastic) policy $\pi := (\pi_1, \pi_2, \dots)$, where $\pi_t: \mathbf{H}_t \rightarrow \Delta(\mathbf{A})$, is given by

$$J(\pi) := \liminf_{T \rightarrow \infty} \mathbb{E}^\pi \left[\sum_{t=1}^T \gamma^{t-1} R_t \right],$$

where $\gamma \in (0, 1)$ is the discount factor. As before, we assume that the agent knows the system dynamics $\{f_t\}_{t \geq 1}$, the reward functions $\{r_t\}_{t \geq 1}$, and the probability measure \mathbb{P} on the primitive random variables $\{W_t\}_{t \geq 1}$. The objective of the agent is to choose a policy π that maximizes the expected discounted total reward $J(\pi)$.

Note that we use \liminf rather than \lim in the above definition because in general the limit might not exist. We later assume that the rewards are uniformly bounded (see Assumption 1) which, together with the finiteness of the action space, implies that the limit is well defined. When the action space is uncountable, we need to impose appropriate technical conditions on the model to ensure that an appropriate measurable selection condition holds (Hernández-Lerma and Lasserre, 2012).

4.2 A dynamic programming decomposition

In the finite-horizon setup, we started with a dynamic program to evaluate the performance $\{V_t^\pi\}_{t=1}^T$ for any history dependent policy π . We then identified an upper-bound $\{V_t\}_{t=1}^T$

on $\{V_t^\pi\}_{t=1}^T$ and showed that this upper bound is tight and achieved by any optimal policy. The subsequent analysis of the information state and the approximate information state based dynamic programs was based on comparison with $\{V_t\}_{t=1}^T$.

One conceptual difficulty with the infinite horizon setup is that we cannot write a general dynamic program to evaluate the performance $\{V_t^\pi\}_{t \geq 1}$ of an arbitrary history dependent policy π and therefore identify a tight upper-bound $\{V_t\}_{t \geq 1}$. In traditional MDP models, this conceptual difficulty is resolved by restricting attention to Markov strategies and then establishing that the performance of a Markov strategy can be evaluated by solving a fixed point equation. For partially observed MDPs, a similar resolution works because one can view the belief state as an information state. However, for general partially observed models as considered in this paper, there is no general methodology to identify a time-homogeneous information state. So, we follow a different approach and identify a dynamic program which bounds the performance of a general history dependent policy. We impose the following mild assumption on the model.

Assumption 1 *The reward process $\{R_t\}_{t \geq 1}$ is uniformly bounded and takes values inside a finite interval $[R_{\min}, R_{\max}]$.*

Given any (history dependent) policy π , we define the *reward-to-go* function for any time t and any realization h_t of H_t as

$$V_t^\pi(h_t) := \mathbb{E}^\pi \left[\sum_{s=t}^{\infty} \gamma^{s-t} R_s \mid H_t = h_t \right]. \quad (34)$$

Define the corresponding action value function as:

$$Q_t^\pi(h_t, a_t) := \mathbb{E}^\pi [R_t + \gamma V_{t+1}^\pi(H_{t+1}) \mid H_t = h_t, A_t = a_t]. \quad (35)$$

As stated above, we cannot identify a dynamic program to recursively compute $\{V_t^\pi\}_{t \geq 1}$. Nonetheless, we show that under Assumption 1 we can identify arbitrarily precise upper and lower bounds for $\{V_t^\pi\}_{t \geq 1}$ which can be recursively computed.

Proposition 22 *Arbitrarily pick a horizon T and define $\{J_{t,T}^\pi: \mathbf{H}_t \rightarrow \mathbb{R}\}_{t=1}^T$ as follows: $J_{T,T}^\pi(h_T) = 0$ and for $t \in \{T-2, \dots, 1\}$,*

$$J_{t,T}^\pi(h_t) := \mathbb{E}^\pi [R_t + \gamma J_{t+1,T}^\pi(H_{t+1}) \mid H_t = h_t]. \quad (36)$$

Then, for any time $t \in \{1, \dots, T\}$ and realization h_t of H_t , we have

$$J_{t,T}^\pi(h_t) + \frac{\gamma^{T-t}}{1-\gamma} R_{\min} \leq V_t^\pi(h_t) \leq J_{t,T}^\pi(h_t) + \frac{\gamma^{T-t}}{1-\gamma} R_{\max}. \quad (37)$$

Proof The proof follows from backward induction. Note that for $t = T$, $R_t \in [R_{\min}, R_{\max}]$ implies that

$$\frac{R_{\min}}{1-\gamma} \leq V_T^\pi(h_T) \leq \frac{R_{\max}}{1-\gamma}.$$

This forms the basis of induction. Now assume that (37) holds for time $t + 1$ and consider the model for time t :

$$\begin{aligned}
 V_t^\pi(h_t) &= \mathbb{E}^\pi \left[\sum_{s=t}^{\infty} \gamma^{s-t} R_s \mid H_t = h_t \right] \\
 &\stackrel{(a)}{=} \mathbb{E}^\pi \left[R_t + \gamma \mathbb{E}^\pi \left[\sum_{s=t+1}^{\infty} \gamma^{s-(t+1)} R_s \mid H_{t+1} \right] \mid H_t = h_t \right] \\
 &\stackrel{(b)}{\leq} \mathbb{E}^\pi \left[R_t + \gamma \mathbb{E}^\pi \left[J_{t+1,T}^\pi(H_{t+1}) + \frac{\gamma^{T-(t+1)}}{1-\gamma} R_{\max} \mid H_{t+1} \right] \mid H_t = h_t \right] \\
 &\stackrel{(c)}{=} J_{t,T}^\pi(h_t) + \frac{\gamma^{T-t}}{1-\gamma} R_{\max},
 \end{aligned}$$

where (a) follows from the smoothing property of conditional expectation, (b) follows from the induction hypothesis, and (c) follows from the definition of $J_{t,T}^\pi(\cdot)$. This establishes one side of (37). The other side can be established in a similar manner. Therefore, the result holds by the principle of induction. \blacksquare

Note that Proposition 22 gives a recursive method to approximately evaluate the performance of any history dependent policy π . We can modify the recursion in (36) to obtain policy independent upper bound on performance of an arbitrary policy. For that matter, define value functions $\{V_t: \mathbf{H}_t \rightarrow \mathbb{R}\}_{t \geq 1}$ as follows:

$$V_t(h_t) = \sup_{\pi} V_t^\pi(h_t), \quad (38)$$

where the supremum is over all history dependent policies. Furthermore, define action-value functions $\{Q_t: \mathbf{H}_t \times \mathbf{A} \rightarrow \mathbb{R}\}_{t \geq 1}$ as follows:

$$Q_t(h_t, a_t) = \mathbb{E}[R_t + \gamma V_{t+1}(H_{t+1}) \mid H_t = h_t, A_t = a_t]. \quad (39)$$

Then, we have the following.

Proposition 23 *Arbitrarily pick a horizon T and define $\{J_{t,T}: \mathbf{H}_t \rightarrow \mathbb{R}\}$ as follows: $J_{T,T}(h_T) = 0$ and for $t \in \{T-2, \dots, 1\}$,*

$$J_{t,T}(h_t) := \max_{a_t \in \mathbf{A}} \mathbb{E}[R_t + \gamma J_{t+1,T}(H_{t+1}) \mid H_t = h_t, A_t = a_t]. \quad (40)$$

Then, for any time $t \in \{1, \dots, T\}$ and realization h_t of H_t ,

$$V_t^\pi(h_t) \leq J_{t,T}(h_t) + \frac{\gamma^{T-t}}{1-\gamma} R_{\max}. \quad (41)$$

Therefore,

$$J_{t,T}(h_t) + \frac{\gamma^{T-t}}{1-\gamma} R_{\min} \leq V_t(h_t) \leq J_{t,T}(h_t) + \frac{\gamma^{T-t}}{1-\gamma} R_{\max}. \quad (42)$$

Note that $J_{t,T}(h_t)$ is the optimal value function for a finite horizon system with the discounted reward criterion that runs for horizon $T - 1$.

Proof By following almost the same argument as Proposition 2, we can establish that for any history dependent policy π , $J_{t,T}^\pi(h_t) \leq J_{t,T}(h_t)$, which immediately implies (41).

Maximizing the left hand side of (41) gives us the upper bound in (42). For the lower bound in (42), observe that

$$\begin{aligned}
 V_t(h_t) &= \sup_{\pi} \mathbb{E}^\pi \left[\sum_{s=t}^{\infty} \gamma^{s-t} R_s \mid H_t = h_t \right] \\
 &\stackrel{(a)}{\geq} \sup_{\pi} \mathbb{E}^\pi \left[\sum_{s=t}^{T-1} \gamma^{s-t} R_s + \sum_{s=T}^{\infty} \gamma^{s-t} R_{\min} \mid H_t = h_t \right] \\
 &= \sup_{\pi} \mathbb{E}^\pi \left[\sum_{s=t}^{T-1} \gamma^{s-t} R_s \mid H_t = h_t \right] + \frac{\gamma^{T-t}}{1-\gamma} R_{\min} \\
 &\stackrel{(b)}{=} J_{t,T}(h_t) + \frac{\gamma^{T-t}}{1-\gamma} R_{\min}.
 \end{aligned}$$

where (a) follows from the fact that $R_s \geq R_{\min}$ and (b) follows from the definition of $J_{t,T}(h_t)$. This complete the proof of (42). \blacksquare

4.3 Time-homogeneous information state and simplified dynamic program

Definition 24 Given a Banach space \mathbf{Z} , an information state generator $\{\sigma_t: \mathbf{H}_t \rightarrow \mathbf{Z}\}$ is said to be time-homogeneous if, in addition to (P1) and (P2), it satisfies the following:

(S) The expectation $\mathbb{E}[R_t | Z_t = \sigma_t(H_t), A_t = a_t]$ and the transition kernel $\mathbb{P}(Z_{t+1} \in B | Z_t = \sigma_t(H_t), A_t = a_t)$ are time-homogeneous.

Note that all except the first example of information state presented in Sec. 2.4 are time-homogeneous. However, in general, a time-homogeneous information state may not exist for all partially observed models and it is important to understand conditions under which such an information state exists. However, we do not pursue that direction in this paper.

For any time-homogeneous information state, define the Bellman operator $\mathcal{B}: [\mathbf{Z} \rightarrow \mathbb{R}] \rightarrow [\mathbf{Z} \rightarrow \mathbb{R}]$ as follows: for any uniformly bounded function $\bar{V}: \mathbf{Z} \rightarrow \mathbb{R}$

$$[\mathcal{B}\bar{V}](z) = \max_{a \in \mathbf{A}} \mathbb{E}[R_t + \gamma \bar{V}(Z_{t+1}) \mid Z_t = z, A_t = a], \quad (43)$$

where $\gamma \in (0, 1)$ is the discount factor. Because of (S), the expectation on the right hand side does not depend on time. Due to discounting, the operator \mathcal{B} is a contraction and therefore, under Assumption 1, the fixed point equation

$$\bar{V} = \mathcal{B}\bar{V} \quad (44)$$

has a unique bounded solution (due to the Banach fixed point theorem). Let \bar{V}^* be the fixed point and π^* be any policy such that $\pi^*(z)$ achieves the arg max in the right hand side

of (43) for $[\mathcal{B}\bar{V}^*](z)$. It is easy to see that \bar{V}^* is the performance of the time homogeneous policy (π^*, π^*, \dots) . However, it is not obvious that \bar{V}^* equals to the optimal performance V_1 defined in (38), because the proof of Theorem 5 relies on backward induction and is not applicable to infinite horizon models. So, we present an alternative proof below which uses the performance bounds of Proposition 23.

Theorem 25 *Let $\{Z_t\}_{t \geq 1}$ be a time-homogeneous information state process with generator $\{\sigma_t: \mathbf{H}_t \rightarrow \mathbf{Z}\}_{t \geq 1}$. Suppose Assumption 1 holds and let \bar{V}^* be the unique bounded fixed point of (43). Then, for any time t and realization h_t of H_t , we have*

$$V_t(h_t) = \bar{V}^*(\sigma_t(h_t)).$$

Furthermore, let $\pi^*: \mathbf{Z} \rightarrow \Delta(\mathbf{A})$ be a time-homogeneous (stochastic) policy such that $\text{Supp}(\pi^*(z))$ is a subset of the arg max of the right hand side of (43). Then, the time-homogeneous policy $\pi^* := (\pi^*, \pi^*, \dots)$ is optimal.

Proof Consider the following sequence of value functions: $\bar{V}^{(0)}(z) = 0$ and for $n \geq 0$, define $\bar{V}^{(n+1)} = \mathcal{B}\bar{V}^{(n)}$. Now fix a horizon T and consider the finite-horizon discounted reward problem of horizon $T - 1$. As argued earlier, $J_{t,T}(h_t)$ is the optimal value-function for this finite horizon discounted problem. Moreover, note that $\{Z_t\}_{t=1}^T$ is an information state for this finite horizon discounted problem. Therefore, from using the result of Theorem 5, we get that for any time $t \in \{1, \dots, T\}$, and realization h_t of H_t ,

$$J_{t,T}(h_t) = \bar{V}^{(T-t)}(\sigma_t(h_t)).$$

Substituting (42) from Proposition 23 in the above, we get

$$\bar{V}^{(T-t)}(\sigma_t(h_t)) + \frac{\gamma^{T-t}}{1-\gamma} R_{\min} \leq V_t(h_t) \leq \bar{V}^{(T-t)}(\sigma_t(h_t)) + \frac{\gamma^{T-t}}{1-\gamma} R_{\max}.$$

The result follows from taking limit $T \rightarrow \infty$ and observing that $\bar{V}^{(T-t)}(z)$ converges to $\bar{V}^*(z)$. \blacksquare

4.4 Time-homogeneous AIS and approximate dynamic programming

Definition 26 *Given a Banach space $\hat{\mathbf{Z}}$, a function class \mathfrak{F} for IPMs, and positive real numbers (ε, δ) , we say that a collection $\{\hat{\sigma}_t: \mathbf{H}_t \rightarrow \hat{\mathbf{Z}}\}_{t \geq 1}$ along with a time-homogeneous update kernel $\hat{P}: \hat{\mathbf{Z}} \times \mathbf{A} \rightarrow \Delta(\hat{\mathbf{Z}})$ and a time-homogeneous reward approximation function $\hat{r}: \hat{\mathbf{Z}} \times \mathbf{A} \rightarrow \mathbb{R}$ is a (ε, δ) time homogeneous AIS generator if the process $\{\hat{\mathbf{Z}}_t\}_{t \geq 1}$, where $\hat{\mathbf{Z}}_t = \hat{\sigma}_t(H_t)$, satisfies (AP1) and (AP2) where \hat{r}_t , \hat{P}_t , ε_t and δ_t in the definition of (AP1) and (AP2) are replaced by their time-homogeneous counterparts.*

For any time-homogeneous AIS, define the approximate Bellman operator $\hat{\mathcal{B}}: [\hat{\mathbf{Z}} \rightarrow \mathbb{R}] \rightarrow [\hat{\mathbf{Z}} \rightarrow \mathbb{R}]$ as follows: for any uniformly bounded function $\hat{V}: \hat{\mathbf{Z}} \rightarrow \mathbb{R}$,

$$[\hat{\mathcal{B}}\hat{V}](\hat{z}) = \max_{a \in \mathbf{A}} \left\{ \hat{r}(\hat{z}, a) + \gamma \int_{\hat{\mathbf{Z}}} \hat{V}(\hat{z}') \hat{P}(d\hat{z}' | \hat{z}, a) \right\}. \quad (45)$$

Note that the expectation on the right hand side does not depend on time. Due to discounting, the operator $\hat{\mathcal{B}}$ is a contraction, and therefore, under Assumption 1, the fixed point equation

$$\hat{V} = \hat{\mathcal{B}}\hat{V} \quad (46)$$

has a unique bounded solution (due to the Banach fixed point theorem). Let \hat{V}^* be the fixed point and $\hat{\pi}^*$ be any policy such that $\hat{\pi}^*(\hat{z})$ achieves the arg max in the right hand side of (45) for $[\hat{\mathcal{B}}\hat{V}^*](\hat{z})$. It is not immediately clear if \hat{V}^* is close to the performance of policy $\pi = (\pi_1, \pi_2, \dots)$, where $\pi_t = \pi^* \circ \hat{\sigma}_t$, or if \hat{V}^* is close to the optimal performance. The proof of Theorem 9 relies on backward induction and is not immediately applicable to the infinite horizon setup. Nonetheless, we establish results similar to Theorem 9 by following the proof idea of Theorem 25.

Theorem 27 *Suppose $(\{\hat{\sigma}_t\}_{t \geq 1}, \hat{P}, \hat{r})$ is a time-homogeneous (ε, δ) -AIS generator. Consider the fixed point equation (46), which we rewrite as follows:*

$$\hat{Q}(\hat{z}, a) := \hat{r}(\hat{z}, a) + \gamma \int_{\hat{Z}} \hat{V}(\hat{z}') \hat{P}(d\hat{z}' | \hat{z}, a), \quad (47a)$$

$$\hat{V}(\hat{z}) := \max_{a \in \mathbf{A}} \hat{Q}(\hat{z}, a). \quad (47b)$$

Let \hat{V}^* denote the fixed point of (47) and \hat{Q}^* denote the corresponding action-value function. Then, we have the following:

1. **Value function approximation:** For any time t , realization h_t of H_t , and choice a_t of A_t , we have

$$|Q_t(h_t, a_t) - \hat{Q}^*(\hat{\sigma}_t(h_t), a_t)| \leq \alpha \quad \text{and} \quad |V_t(h_t) - \hat{V}^*(\hat{\sigma}_t(h_t))| \leq \alpha, \quad (48)$$

where

$$\alpha = \frac{\varepsilon + \gamma \rho_{\mathcal{F}}(\hat{V}^*) \delta}{1 - \gamma}$$

2. **Approximately optimal policy:** Let $\hat{\pi}^*: \hat{Z} \rightarrow \Delta(\mathbf{A})$ be a stochastic policy that satisfies

$$\text{Supp}(\hat{\pi}^*(\hat{z})) \subseteq \arg \max_{a \in \mathbf{A}} \hat{Q}^*(\hat{z}, a). \quad (49)$$

Define policy $\pi = (\pi_1, \pi_2, \dots)$, where $\pi_t: H_t \rightarrow \Delta(\mathbf{A})$ is defined by $\pi_t := \hat{\pi}^* \circ \hat{\sigma}_t$. Then, for any time t , realization h_t of H_t , and choice a_t of A_t , we have

$$|Q_t(h_t, a_t) - Q_t^\pi(h_t, a_t)| \leq 2\alpha \quad \text{and} \quad |V_t(h_t) - V_t^\pi(h_t)| \leq 2\alpha. \quad (50)$$

Proof The proof follows by combining ideas from Theorem 9 and 25. We provide a detailed proof of the value approximation. The proof argument for policy approximation is similar.

Consider the following sequence of value functions: $\hat{V}^{(0)}(\hat{z}) = 0$ and for $n \geq 0$, define $\hat{V}^{(n+1)} = \hat{\mathcal{B}}\hat{V}^{(n)}$. Now fix a horizon T and consider the finite-horizon discounted reward problem of horizon $T - 1$. As argued earlier, $J_{t,T}(h_t)$ is the optimal value-function for this finite horizon discounted problem. Moreover, note that $\{\hat{Z}_t\}_{t=1}^T$ is an (ε, δ) -AIS for this finite

horizon discounted problem. Therefore, from using the result of Theorem 9, we get that for any time $t \in \{1, \dots, T\}$, and realization h_t of H_t ,

$$|J_{t,T}(h_t) - \hat{V}^{(T-t)}(\hat{\sigma}_t(h_t))| \leq \alpha_t,$$

where

$$\alpha_t = \varepsilon + \sum_{\tau=t+1}^{T-1} \gamma^{\tau-t} [\rho_{\mathfrak{F}}(\hat{V}^{(T-\tau)})\delta + \varepsilon].$$

Substituting (42) from Proposition 23 in the above, we get that

$$\hat{V}^{(T-t)}(\hat{\sigma}_t(h_t)) - \alpha_t + \frac{\gamma^{T-t}}{1-\gamma} R_{\min} \leq V_t(h_t) \leq \hat{V}^{(T-t)}(\hat{\sigma}_t(h_t)) + \alpha_t + \frac{\gamma^{T-t}}{1-\gamma} R_{\max}.$$

Since $\hat{\mathcal{B}}$ is a contraction, from the Banach fixed point theorem we know that $\lim_{T \rightarrow \infty} \hat{V}^{(T-t)} = \hat{V}^*$. Therefore, by continuity of $\rho_{\mathfrak{F}}(\cdot)$, we have $\lim_{T \rightarrow \infty} \rho_{\mathfrak{F}}(\hat{V}^{(T-t)}) = \rho_{\mathfrak{F}}(\hat{V}^*)$. Consequently, $\lim_{T \rightarrow \infty} \alpha_t = \alpha$. Therefore, taking the limit $T \rightarrow \infty$ in the above equation, we get

$$\hat{V}^*(\hat{\sigma}_t(h_t)) - \alpha \leq V_t(h_t) \leq \hat{V}^*(\hat{\sigma}_t(h_t)) + \alpha,$$

which establishes the bound on the value function in (48). The bound on the action-value function in (48) follows from a similar argument. \blacksquare

Theorem 27 shows how the result of Theorem 9 generalizes to infinite horizon. We can similarly extend the results for approximate policy evaluation (as in Sec. 3.4), the stochastic AIS case (as in Sec. 3.5), the action compression case (as in Sec. 3.6), and the observation compression case (as in Sec. 3.7).

5. An AIS-based approximate dynamic programming for Dec-POMDPs

The theory of approximation for partially observed systems presented in the previous section is fairly general and is applicable to other models of decision making as well. As an example, in this section we show how to use the same ideas to obtain approximation results for decentralized (i.e., multi-agent) partially observed models.

There is a rich history of research on these models in multiple research disciplines. Decentralized multi-agent systems have been studied in Economics and Organizational Behavior since the mid 1950s (Marschak, 1954; Radner, 1962; Marschak and Radner, 1972) under the heading of team theory. Such models have been studied in systems and control since the mid 1960s under the heading of decentralized stochastic control (Witsenhausen, 1968, 1971; Sandell et al., 1978). Such models have also been studied in Artificial Intelligence since the 2000s (Bernstein et al., 2005; Szer et al., 2005; Seuken and Zilberstein, 2007; Carlin and Zilberstein, 2008) under the heading of Dec-POMDPs. In the interest of space, we do not provide a detailed overview of this rich area; instead we refer the reader to the comprehensive survey articles of Mahajan et al. (2012); Liu et al. (2016) for a detailed overview from the perspective of Systems and Control and Artificial Intelligence, respectively.

We briefly state the facts about this literature which are pertinent to the discussion below. The general Dec-POMDP problem is NEXP complete (Bernstein et al., 2002),

so it is not possible to derive an efficient algorithm to compute the optimal solution. Nonetheless, considerable progress has been made in identifying special cases where a dynamic programming decomposition is possible (Walrand and Varaiya, 1983; Aicardi et al., 1987; Ooi et al., 1997; Mahajan and Teneketzis, 2009a,b; Mahajan et al., 2008; Nayyar, 2011; Nayyar et al., 2013; Mahajan, 2013; Arabneydi and Mahajan, 2014; Oliehoek and Amato, 2015; Dibangoye et al., 2016; Boularias and Chaib-Draa, 2008; Kumar and Zilberstein, 2009). A high level approach which encapsulates many of these special cases is the common information approach of Nayyar et al. (2013) which shows that the Dec-POMDP problem with a specific but relatively general information structure can be converted into a single agent, partially observed problem from the point of view of a virtual agent which knows the information commonly known to all agents and chooses prescriptions (or partially evaluated policies) which map the local information at each agent to their respective actions. We summarize these results in the next subsection and then show how we can identify an AIS for such models.

5.1 Model of a Dec-POMDP

A Dec-POMDP is a tuple $\langle K, S, (Y^k)_{k \in K}, (A_t^k)_{k \in K}, P_1, P, P^y, r \rangle$ where

- $K = \{1, \dots, K\}$ is the set of agents.
- S is the state space. $Y^k, A^k, k \in K$, are the observation and action spaces of agent k . Let $Y = \prod_{k \in K} Y^k$ and $A = \prod_{k \in K} A^k$. We use $S_t \in S$, $Y_t := (Y_t^k)_{k \in K} \in Y$, and $A_t := (A_t^k)_{k \in K} \in A$, to denote the system state, observations, and actions at time t .
- $P_1 \in \Delta(S)$ is the initial distribution of the initial state S_1 .
- $P: S \times A \rightarrow \Delta(S)$ denotes the transition probability of the system, i.e.,

$$\begin{aligned} \mathbb{P}(S_{t+1} = s_{t+1} \mid S_{1:t} = s_{1:t}, A_{1:t} = a_{1:t}) &= \mathbb{P}(S_{t+1} = s_{t+1} \mid S_t = s_t, A_t = a_t) \\ &= P(s_{t+1} \mid s_t, a_t). \end{aligned}$$

- $P^y: S \times A \rightarrow \Delta(Y)$ denotes the observation probability of the system, i.e.,

$$\begin{aligned} \mathbb{P}(Y_t = y_t \mid S_{1:t} = s_{1:t}, A_{1:t-1} = a_{1:t-1}) &= \mathbb{P}(Y_t = y_t \mid S_t = s_t, A_{t-1} = a_{t-1}) \\ &= P^y(y_t \mid s_t, a_{t-1}). \end{aligned}$$

- $r: S \times A \times S \rightarrow \mathbb{R}$ denotes the per-step reward function. The team receives a reward $R_t = r(S_t, A_t, S_{t+1})$ at time t .

Information structure: A critical feature of a Dec-POMDP is the *information structure* which captures the knowledge of who knows what about the system and when. We use I_t^k to denote the information known to agent k at time t . In general, I_t^k is a subset of the total information $(Y_{1:t}, A_{1:t-1}, R_{1:t-1})$ known to all agents in the system. We use I_t^k to denote the space of the information available to agent k at time t . Note that, in general, the information available to agent k increases with time. So, I_t^k are sets that are increasing with time. Some examples of information structures are:

- **Delayed sharing:** $I_t^k = \{Y_{1:t-d}, A_{1:t-d}, Y_{t-d+1:t}^k, A_{t-d+1:t-1}^k\}$. This models systems where agents broadcast their information and communication has delay of d . Planning for models where $d = 1$ has been considered in Sandell and Athans (1974); Yoshikawa (1975) and for general d has been considered in Nayyar et al. (2011).
- **Periodic sharing:** $I_t^k = \{Y_{1:t-\tau}, A_{1:t-\tau}, Y_{t-\tau+1:t}^k, A_{t-\tau+1:t-1}^k\}$, where $\tau = p \lfloor \frac{t}{p} \rfloor$. This models systems where agents periodically broadcast their information every p steps. Planning for this model has been considered in Ooi et al. (1997).
- **Control sharing:** $I_t^k = \{Y_{1:t}^k, A_{1:t-1}\}$. This models systems where control actions are observed by everyone (which is the case for certain communication and economic applications). Planning for variations of this model has been considered in Bismut (1972); Sandell and Athans (1974); Mahajan (2013).
- **Mean-field sharing:** $I_t^k = \{S_{1:t}^k, A_{1:t-1}^k, M_{1:t}\}$, where the state S_t is (S_t^1, \dots, S_t^K) , the observation of agent k is S_t^k , and $M_t = (\sum_{k \in K} \delta_{S_t^k})/K$ denotes the empirical distribution of the states. This models systems where mean-field is observed by all agents (which is the case for smart grid and other large-scale systems). Planning for variations of this model has been considered in Arabneydi and Mahajan (2014).

Policy: The policy of agent k is a collection $\pi^k = (\pi_1^k, \pi_2^k, \dots)$, where $\pi_t^k: \mathcal{I}_t^k \rightarrow \Delta(\mathcal{A}^i)$. We use $\pi = (\pi^k)_{k \in K}$ to denote the policy for all agents. The performance of a policy π is given by

$$J(\pi) = \mathbb{E}^\pi \left[\sum_{t=1}^T R_t \right]. \quad (51)$$

The objective is to find a (possibly time-varying) policy π that maximizes the performance $J(\pi)$ defined in (51).

5.2 Common information based planning for Dec-POMDPs

As mentioned earlier, in general, finding the optimal plan for multi-agent teams is NEXP-complete (Bernstein et al., 2002). However, it is shown in Nayyar et al. (2013) that when the information structure is of a particular form (known as partial history sharing), it is possible to reduce the multi-agent planning problem to a single agent planning problem from the point of view of a virtual agent called the coordinator. We summarize this approach below.

Common and local information: Define

$$C_t = \bigcap_{s \geq t} \bigcap_{k \in K} I_s^k \quad \text{and} \quad L_t^k = I_t^k \setminus C_t, \quad k \in K.$$

C_t denotes the *common information*, i.e., the information that is common to all agents all the time in the future and L_t^k denotes the *local information* at agent k . By construction, $I_t^k = \{C_t, L_t^k\}$. Let \mathcal{C}_t and \mathcal{L}_t^k denote the space of realizations of C_t and L_t^k and let $L_t = (L_t^k)_{k \in K}$ and $\mathcal{L}_t = \prod_{k \in K} \mathcal{L}_t^k$. By construction, $C_t \subseteq C_{t+1}$. Let $C_{t+1}^{\text{NEW}} = C_{t+1} \setminus C_t$ denote the new common information at time t . Then, C_t may be written as $C_{1:t}^{\text{NEW}}$.

Definition 28 *The information structure is called partial history sharing if for any Borel subset B of \mathcal{L}_{t+1}^k and any realization c_t of C_t , ℓ_t^k of L_t^k , a_t^k of A_t^k and y_{t+1}^k of Y_{t+1}^k , we have*

$$\begin{aligned} \mathbb{P}(L_{t+1}^k \in B \mid C_t = c_t, L_t^k = \ell_t^k, A_t^k = a_t^k, Y_{t+1}^k = y_{t+1}^k) \\ = \mathbb{P}(L_{t+1}^k \in B \mid L_t^k = \ell_t^k, A_t^k = a_t^k, Y_{t+1}^k = y_{t+1}^k). \end{aligned}$$

The main intuition behind this definition is as follows. For any system, the information available to the agents can always be split into common and local information such that $I_t^k = \{C_t, L_t^k\}$. A partial history sharing information structure satisfies the property that at any time t and for any agent k , the updated value L_{t+1}^k of the local information is a function of only the current local information L_t^k , the current local action A_t^k and the next local observation Y_{t+1}^k . Consequently, the common information C_t is not needed to keep track of the update of the local information. This ensures that compressing the common information into an information state or an approximate information state does not impact the update of the local information.

Prescriptions: Given a policy $\pi = (\pi^k)_{k \in K}$ and a realized trajectory (c_1, c_2, \dots) of the common information, the prescription $\hat{\xi}_t^k$ is the partial application of c_t to π_t^k , i.e., $\hat{\xi}_t^k = \pi_t^k(c_t, \cdot)$, $k \in K$. Note that $\hat{\xi}_t^k$ is a function from \mathcal{L}_t^k to $\Delta(A_t^k)$. Let $\hat{\xi}_t$ denote $(\hat{\xi}_t^k)_{k \in K}$ and let \mathcal{X} denote the space of all such prescriptions for time t .

The reason for constructing prescriptions is as follows. Prescriptions encode the information about the policies of all agents needed to evaluate the conditional expected per-step reward given the common information, i.e., $\mathbb{E}[R_t \mid C_t, (\pi^k)_{k \in K}]$ can be written as a function of C_t and $(\hat{\xi}_t^k)_{k \in K}$, say $\hat{r}_t(C_t, (\hat{\xi}_t^k)_{k \in K})$. This allows us to construct a virtual single-agent optimization problem where a decision maker (which we call the virtual coordinator) observes the common information C_t and chooses the prescriptions $(\hat{\xi}_t^k)_{k \in K}$ to maximize the sum of rewards $\hat{r}_t(C_t, (\hat{\xi}_t^k)_{k \in K})$. The details of this virtual coordinated system are presented next.

A virtual coordinated system: The key idea of Nayyar et al. (2013) is to construct a virtual single agent planning problem which they call a coordinated system. The environment of the virtual coordinated system consists of two components: the first component is the same as the environment of the original multi-agent system which evolves according to dynamics P ; the second component consists of K *passive agents*, whose operation we will describe later. There is a virtual coordinator who observes the common information C_t and chooses *prescriptions* $\hat{\Xi}_t = (\hat{\Xi}_t^k)_{k \in K}$, where $\hat{\Xi}_t^k : \mathcal{L}^k \rightarrow \Delta(A^k)$ using a *coordination rule* ψ_t , i.e., $\hat{\Xi}_t \sim \psi_t(C_t)$. In general, the coordination rule can be stochastic. Let $\hat{\xi}_t$ denote the realization of $\hat{\Xi}_t$. Each agent in the virtual coordinated system is a passive agent and agent k uses the prescription $\hat{\Xi}_t^k$ to sample an action $A_t^k \sim \hat{\Xi}_t^k(L_t^k)$.

A key insight of Nayyar et al. (2013) is that the virtual coordinated system is equivalent to the original multi-agent system in the following sense.

Theorem 29 (Nayyar et al. (2013)) *Consider a Dec-POMDP with a partial history sharing information structure. Then, for any policy $\pi = (\pi^k)_{k \in K}$, where $\pi^k = (\pi_1^k, \dots, \pi_T^k)$ for the Dec-POMDP, define a coordination policy $\psi = (\psi_1, \dots, \psi_T)$ for the virtual coordinated system given by $\psi_t(c_t) = (\pi_t^k(c_t, \cdot))_{k \in K}$. Then, the performance of the virtual coordinated system with policy ψ is the same as the performance of the Dec-POMDP with policy π .*

Conversely, for any coordination policy $\psi = (\psi_1, \dots, \psi_T)$ for the virtual coordinated system, define a policy $\pi = (\pi^k)_{k \in \mathcal{K}}$ with $\pi^k = (\pi_1^k, \dots, \pi_T^k)$ for the Dec-POMDP given by $\pi_t^k(c_t, \ell_t^k) = \psi_t^k(c_t)(\ell_t^k)$. Then, the performance of the Dec-POMDP with policy π is the same as that of the virtual coordinated system with policy ψ .

Dynamic program: Theorem 29 implies that the problem of finding optimal decentralized policies in a Dec-POMDP is equivalent to a centralized (single-agent) problem of finding the optimal coordination policy for the virtual coordinated system. The virtual coordinated system is a POMDP with unobserved state $(S_t, L_t^1, \dots, L_t^K)$, observation C_t^{NEW} , and actions $\hat{\Xi}_t$. The corresponding history of observations is $(C_{1:t}^{\text{NEW}}, \hat{\Xi}_{1:t-1})$ and therefore we can write a history dependent dynamic program similar to the one presented in Proposition 2. Nayyar et al. (2013) presented a simplified dynamic program which used the belief state as an information state; however, it is clear from the above discussion that any other choice of information state will also lead to a dynamic programming decomposition.

5.3 Common-information based AIS and approximate dynamic programming

Since the coordinated system is a POMDP, we can simply adapt the definition of AIS Dec-POMDPs and obtain an approximate dynamic program with approximation guarantees. Let \mathcal{X}_t denote the space of realization of $\hat{\Xi}_t$. Then, we have the following.

Definition 30 Let $\{\hat{Z}_t\}_{t=1}^T$ be a pre-specified collection of Banach spaces, \mathfrak{F} be a function class for IPMs, and $\{(\varepsilon_t, \delta_t)\}_{t=1}^T$ be pre-specified positive real numbers. A collection $\{\hat{\sigma}_t: (C_t, \hat{\Xi}_{1:t-1}) \mapsto \hat{Z}_t\}_{t=1}^T$ of history compression functions, along with approximate update kernels $\{\hat{P}_t: \hat{Z}_t \times \mathcal{X}_t \rightarrow \Delta(\hat{Z}_{t+1})\}_{t=1}^T$ and reward approximation functions $\{\hat{r}_t: \hat{Z}_t \times \mathcal{X}_t \rightarrow \mathbb{R}\}_{t=1}^T$, is called an $\{(\varepsilon_t, \delta_t)\}_{t=1}^T$ -AIS generator if the process $\{\hat{Z}_t\}_{t=1}^T$, where $\hat{Z}_t = \hat{\sigma}_t(C_t, \hat{\Xi}_{1:t-1})$, satisfies the following properties:

(DP1) Sufficient for approximate performance evaluation, i.e., for any time t , any realization c_t of C_t and any choice $\hat{\xi}_{1:t}$ of $\hat{\Xi}_{1:t}$, we have

$$|\mathbb{E}[R_t \mid C_t = c_t, \hat{\Xi}_{1:t} = \hat{\xi}_{1:t}] - \hat{r}_t(\hat{\sigma}_t(c_t, \hat{\xi}_{1:t-1}), \hat{\xi}_t)| \leq \varepsilon_t.$$

(DP2) Sufficient to predict itself approximately. i.e., for any time t , any realization c_t of C_t , any choice $\hat{\xi}_{1:t}$ of $\hat{\Xi}_{1:t}$, and for any Borel subset \mathcal{B} of \hat{Z}_{t+1} , define $\mu_t(\mathcal{B}) := \mathbb{P}(\hat{Z}_{t+1} \in \mathcal{B} \mid C_t = c_t, \hat{\Xi}_{1:t} = \hat{\xi}_{1:t})$ and $\nu_t(\mathcal{B}) := \hat{P}_t(\mathcal{B} \mid \hat{\sigma}_t(c_t, \hat{\xi}_{1:t-1}), \hat{\xi}_t)$; then,

$$d_{\mathfrak{F}}(\mu_t, \nu_t) \leq \delta_t.$$

Similar to Proposition 4, we can provide an alternative characterization of an AIS where we replace (DP2) with approximations of (P2a) and (P2b) and we can prove a proposition similar to Proposition 8 for the virtual coordinated system.

We can now establish a result similar to Theorem 9 that any AIS gives rise to an approximate dynamic program. In this discussion, h_t denotes $(c_t, \hat{\xi}_{1:t-1})$ and \mathcal{H}_t denotes the space of realization of h_t .

Theorem 31 Suppose $\{\hat{\sigma}_t, \hat{P}_t, \hat{r}_t\}_{t=1}^T$ is an $\{(\varepsilon_t, \delta_t)\}_{t=1}^T$ -AIS generator. Recursively define approximate action-value functions $\{\hat{Q}_t: \hat{\mathbf{Z}}_t \times \mathcal{X}_t \rightarrow \mathbb{R}\}_{t=1}^T$ and value functions $\{\hat{V}_t: \hat{\mathbf{Z}}_t \rightarrow \mathbb{R}\}_{t=1}^T$ as follows: $\hat{V}_{T+1}(\hat{z}_{T+1}) := 0$ and for $t \in \{T, \dots, 1\}$:

$$\hat{Q}_t(\hat{z}_t, \hat{\xi}_t) := \hat{r}_t(\hat{z}_t, \hat{\xi}_t) + \int_{\hat{\mathbf{Z}}_{t+1}} \hat{V}_{t+1}(\hat{z}_{t+1}) \hat{P}_t(d\hat{z}_{t+1} \mid \hat{z}_t, \hat{\xi}_t), \quad (52a)$$

$$\hat{V}_t(\hat{z}_t) := \max_{\hat{\xi}_t \in \mathcal{X}_t} \hat{Q}_t(\hat{z}_t, \hat{\xi}_t). \quad (52b)$$

Then, we have the following:

1. **Value function approximation:** For any time t , realization h_t of H_t , and choice $\hat{\xi}_t$ of $\hat{\Xi}_t$, we have

$$|Q_t(h_t, \hat{\xi}_t) - \hat{Q}_t(\hat{\sigma}_t(h_t), \hat{\xi}_t)| \leq \alpha_t \quad \text{and} \quad |V_t(h_t) - \hat{V}_t(\hat{\sigma}_t(h_t))| \leq \alpha_t, \quad (53)$$

where

$$\alpha_t = \varepsilon_t + \sum_{\tau=t+1}^T [\rho_{\mathfrak{F}}(\hat{V}_\tau) \delta_{\tau-1} + \varepsilon_\tau].$$

2. **Approximately optimal policy:** Let $\hat{\psi} = (\hat{\psi}_1, \dots, \hat{\psi}_T)$, where $\hat{\psi}_t: \hat{\mathbf{Z}}_t \rightarrow \Delta(\mathcal{X}_t)$, be a coordination rule that satisfies

$$\text{Supp}(\hat{\psi}(\hat{z}_t)) \subseteq \arg \max_{\hat{\xi}_t \in \mathcal{X}_t} \hat{Q}_t(\hat{z}_t, \hat{\xi}_t). \quad (54)$$

Define coordination rule $\psi = (\psi_1, \dots, \psi_T)$, where $\psi_t := \hat{\psi}_t \circ \hat{\sigma}_t$. Then, for any time t , realization h_t of H_t , and choice $\hat{\xi}_t$ of $\hat{\Xi}_t$, we have

$$|Q_t(h_t, \hat{\xi}_t) - Q_t^\psi(h_t, \hat{\xi}_t)| \leq 2\alpha_t \quad \text{and} \quad |V_t(h_t) - V_t^\psi(h_t)| \leq 2\alpha_t. \quad (55)$$

Proof The proof is similar to the proof of Theorem 9. ■

We can extend the approximation results for the virtual coordinated system to the approximate policy evaluation case (as in Sec. 3.4), infinite horizon case (as in Sec. 4), the stochastic AIS case (as in Sec. 3.5), the action compression case (as in Sec. 3.6), and the observation compression case (as in Sec. 3.7) in a straightforward manner.

6. Reinforcement learning for partially observed systems using AIS

In this section, we present a policy gradient based reinforcement learning (RL) algorithm for infinite horizon partially observed systems. The algorithm learns a time-homogeneous AIS generator $(\hat{\sigma}_t, \hat{r}, \hat{P})$ which satisfies (AP1) and (AP2) or a time-homogeneous AIS generator $(\hat{\sigma}_t, \hat{r}, \hat{\varphi}, \hat{P}^y)$ which satisfies (AP1), (AP2a), and (AP2b). The key idea is to represent each component of the AIS generator using a parametric family of functions/distributions and use a multi time-scale stochastic gradient descent algorithm (Borkar, 1997) which learns AIS generator at a faster time-scale than the policy and/or the action-value function.

Then, for the ease of exposition, we first assume that the policy is fixed and describe how to learn the AIS generator using stochastic gradient descent. To specify an AIS, we must pick an IPM \mathfrak{F} as well. Although, in principle, we can choose any IPM, in practice, we want to choose an IPM such that the distance $d_{\mathfrak{F}}(\mu_t, \nu_t)$ in (AP2) or (AP2b) can be computed efficiently. We discuss the choice of IPMs in Sec. 6.1 and then discuss the stochastic gradient descent algorithm to learn the AIS-generator for a fixed policy in Sec. 6.2. Then we describe how to simultaneously learn the AIS generator and the policy using a multi-time scale algorithm, first for an actor only framework and then for an actor-critic framework in Sec. 6.3.

6.1 The choice of an IPM

As we will explain in the next section in detail, our general *modus operandi* is to assume that the stochastic kernel \hat{P} or \hat{P}^y that we are trying to learn belongs to a parametric family and then update the parameters of the distribution to either minimize $d_{\mathfrak{F}}(\mu, \nu)$ defined in (AP2) or minimize $d_{\mathfrak{F}}(\mu^y, \nu^y)$ defined in (AP2b). Just to keep the discussion concrete, we focus on (AP2). Similar arguments apply to (AP2b) as well. First note that for a particular choice of parameters, we know the distribution ν in closed form, but we do not know the distribution μ in closed form and only have samples from that distribution. One way to estimate the IPM between a distribution and samples from another distribution is to use duality and minimize $|\int_{\mathcal{Z}} f d\mu - \int_{\mathcal{Z}} f d\nu|$ over the choice of function f such that $f \in \mathfrak{F}$. When $d_{\mathfrak{F}}$ is equal to the total variation distance or the Wasserstein distance, this optimization problem may be solved using a linear program (Sriperumbudur et al., 2012). However, solving a linear program at each step of the stochastic gradient descent algorithm can become a computational bottleneck. We propose two alternatives here. The first is to use the total variation distance or the Wasserstein distance but instead of directly working with them, we use a KL divergence based upper bound as a surrogate loss. The other alternative is to work with RKHS-based MMD (maximum mean discrepancy) distance, which can be computed from samples without solving an optimization problem (Sriperumbudur et al., 2012). It turns out that for the AIS-setup, a specific form of MMD known as distance-based MMD is particularly convenient as we explain below.

KL-divergence based upper bound for total variation or Wasserstein distance.

Recall that the KL-divergence between two densities μ and ν on $\Delta(\mathbf{X})$ is defined as

$$D_{\text{KL}}(\mu \parallel \nu) = \int_{\mathbf{X}} \log \mu(x) \mu(dx) - \int_{\mathbf{X}} \log \nu(x) \mu(dx).$$

The total variation distance can be upper bounded by the KL-divergence using Pinsker's inequality (Csiszar and Körner, 2011) (see footnote 1 for the difference in constant factor from the standard Pinsker's inequality):

$$d_{\text{TV}}(\mu, \nu) \leq \sqrt{2D_{\text{KL}}(\mu \parallel \nu)}. \quad (56)$$

As we will explain in the next section, we consider the setup where we know the distribution ν but only obtain samples from the distribution μ . Since there are two losses—the reward prediction loss ε and the AIS/observation prediction loss δ , we work with minimizing

the weighted square average $\lambda\varepsilon^2 + (1 - \lambda)\delta^2$, where $\lambda \in [0, 1]$ is a hyper-parameter. Pinsker's inequality (56) suggests that instead of $d_{\text{TV}}(\mu, \nu)^2$, we can use the surrogate loss function

$$\int_{\mathbf{X}} \log \nu(x) \mu(dx)$$

where we have dropped the term that does not depend on ν . Note that the above expression is the same as the cross-entropy between μ and ν which can be efficiently computed from samples. In particular, if we get T i.i.d. samples X_1, \dots, X_T from μ , then

$$\frac{1}{T} \sum_{t=1}^T \log \nu(X_t) \quad (57)$$

is an unbiased estimator of $\int_{\mathbf{X}} \log \nu(x) \mu(dx)$.

Finally, if \mathbf{X} is a bounded space with diameter D , then

$$d_{\text{Wass}}(\mu, \nu) \leq D d_{\text{TV}}(\mu, \nu).$$

So, using cross-entropy as a surrogate loss also works for Wasserstein distance.

Distance-based MMD. The key idea behind using a distance-based MMD is the following results.

Proposition 32 (Theorem 22 of Sejdinovic et al. (2013)) *Let $\mathbf{X} \subseteq \mathbb{R}^m$ and $d_{\mathbf{X},p}: \mathbf{X} \times \mathbf{X} \rightarrow \mathbb{R}_{\geq 0}$ be a metric given by $d_{\mathbf{X},p}(x, x') = \|x - x'\|_2^p$, for $p \in (0, 2]$. Let $k_p: \mathbf{X} \times \mathbf{X} \rightarrow \mathbb{R}$ be any kernel given*

$$k_p(x, x') = \frac{1}{2} [d_{\mathbf{X},p}(x, x_0) + d_{\mathbf{X},p}(x', x_0) - d_{\mathbf{X},p}(x, x')],$$

where $x_0 \in \mathbf{X}$ is arbitrary, and let \mathcal{H}_p be a RKHS with kernel k_p and $\mathfrak{F}_p = \{f \in \mathcal{H}_p : \|f\|_{\mathcal{H}_p} \leq 1\}$. Then, for any distributions $\mu, \nu \in \Delta(\mathbf{X})$, the IPM $d_{\mathfrak{F}_p}(\mu, \nu)$ can be expressed as follows:

$$d_{\mathfrak{F}_p}(\mu, \nu) = \sqrt{\mathbb{E}[d_{\mathbf{X},p}(X, W)] - \frac{1}{2}\mathbb{E}[d_{\mathbf{X},p}(X, X')] - \frac{1}{2}\mathbb{E}[d_{\mathbf{X},p}(W, W')]}, \quad (58)$$

where $X, X' \sim \mu$, $W, W' \sim \nu$ and (X, X', W, W') are all independent.

We call d_p defined above as a *distance-based MMD*. For $p = 1$ (for which $d_{\mathbf{X}}$ corresponds to the L_2 distance), the expression inside the square root in (58) is called the Energy distance in the statistics literature (Székely and Rizzo, 2004). In Sejdinovic et al. (2013), the above result is stated for a general semimetric of a negative type. Our statement of the above result is specialized to the semimetric $d_{\mathbf{X},p}$. See Proposition 3 and Example 15 of Sejdinovic et al. (2013) for details.

As explained in the previous section, we work with minimizing the weighted square average $\lambda\varepsilon^2 + (1 - \lambda)\delta^2$, where λ is a hyper-parameter. Proposition 32 suggests that instead of $d_{\mathfrak{F}_p}(\mu, \nu)^2$, we can use a surrogate loss function

$$\int_{\mathbf{X}} \int_{\mathbf{X}} \|x - w\|_2^p \mu(dx) \nu(dw) - \frac{1}{2} \int_{\mathbf{X}} \int_{\mathbf{X}} \|w - w'\|_2^p \nu(dw) \nu(dw') \quad (59)$$

for $p \in (0, 2]$, where we have dropped the term that does not depend on ν . It is possible to compute the surrogate loss efficiently from samples as described in Sriperumbudur et al. (2012). In particular, if we get T i.i.d. samples X_1, \dots, X_T from μ , then

$$\frac{1}{T} \sum_{t=1}^T \int_{\mathbf{X}} \|X_t - w\|_2^p \nu(dw) - \frac{1}{2} \int_{\mathbf{X}} \int_{\mathbf{X}} \|w - w'\|_2^p \nu(dw) \nu(dw') \quad (60)$$

is an unbiased estimator of (59).

In our numerical experiments, we use the surrogate loss (60) for $p = 2$, which simplifies as follows.

Proposition 33 *Consider the setup of Proposition 32 for $p = 2$. Suppose ν_ξ is a known parameterized distribution with mean M_ξ and X is a sample from μ . Then, the gradient of*

$$(M_\xi - 2X)^\top M_\xi \quad (61)$$

with respect to ξ in an unbiased estimator of $\nabla_\xi d_{\mathfrak{H}_2}(\mu, \nu_\xi)^2$.

Proof For $p = 2$, we have that

$$d_{\mathfrak{H}_2}(\mu, \nu_\xi)^2 = \mathbb{E}[\|X - W\|_2^2] - \frac{1}{2} \mathbb{E}[\|X - X'\|_2^2] - \frac{1}{2} \mathbb{E}[\|W - W'\|_2^2],$$

where $X, X' \sim \mu$ and $W, W' \sim \nu_\xi$. Simplifying the right hand side, we get that

$$d_{\mathfrak{H}_2}(\mu, \nu_\xi)^2 = \|\mathbb{E}[X]\|_2^2 - 2\mathbb{E}[X]^\top \mathbb{E}[W] + \|\mathbb{E}[W]\|_2^2.$$

Note that the term $\|\mathbb{E}[X]\|_2^2$ does not depend on the distribution ν_ξ . Thus, the expression (61) captures all the terms which depend on ξ . \blacksquare

The implication of Proposition 33 is if we use MMD with the RKHS \mathcal{H}_2 defined in Proposition 32, then we can use the expression in (61) as a surrogate loss function for $d_{\mathfrak{H}_2}(\mu, \nu_\xi)^2$.

Now we show how to compute the surrogate loss (61) for two types of parameterized distributions ν_ξ .

1. **SURROGATE LOSS FOR PREDICTING DISCRETE VARIABLES:** When predicting a discrete-valued random variable, say a discrete-valued AIS \hat{Z}_{t+1} in (AP2) or a discrete-valued observation Y_t in (AP2b), we view the discrete random variable as a continuous-valued random vector by representing it as a one-hot encoded vector. In particular, if the discrete random variable, which we denote by V , takes m values, then its one-hot encoded representation, which we denote by X , takes values in the corner points of the simplex on \mathbb{R}^m . Now, suppose ν_ξ is any parameterized distribution on the discrete set $\{1, \dots, m\}$ (e.g., the softmax distribution). Then, in the one-hot encoded representation, the mean M_ξ is given by

$$M_\xi = \sum_{i=1}^m \nu_\xi(i) e_i = \begin{bmatrix} \nu_\xi(1) \\ \vdots \\ \nu_\xi(m) \end{bmatrix},$$

where e_i denotes the m -dimensional unit vector with 1 in the i -th location. Thus, when we one-hot encode discrete AIS or discrete observations, the “mean” M_ξ is same as the probability mass function (PMF) ν_ξ . Thus, effectively, $d_{\mathfrak{F}_2}(\mu, \nu)^2$ is equivalent to $\|\mu - \nu\|_2^2$ and (61) is an unbiased estimator where we have removed the terms that do not depend on ν .

2. **SURROGATE LOSS FOR PREDICTING CONTINUOUS VARIABLES:** When predicting a continuous-valued random variable, say a continuous-valued AIS \hat{Z}_{t+1} in (AP2) or a continuous-valued observation Y_t in (AP2b), we can immediately use the surrogate loss (61) as long as the parameterized distribution ν_ξ is such that its mean M_ξ is given in closed form. Note that the surrogate loss (61) only depends on the mean of the distribution and not one any other moment. So, any two distributions ν and ν' that have the same mean, the surrogate loss between any distribution μ and ν is same as the surrogate loss between μ and ν' . Thus, using the surrogate loss (61) for predicting continuous variables only makes sense when we expect the true distribution to be close to a deterministic function.

6.2 Learning an AIS for a fixed policy

The definition of AIS suggests that there are two ways to construct an information state from data: we either learn a time-homogeneous AIS-generator $(\hat{\sigma}, \hat{r}, \hat{P})$ that satisfies (AP1) and (AP2) or we learn a time-homogeneous AIS-generator $(\hat{\sigma}, \hat{r}, \hat{\varphi}, \hat{P}^y)$ that satisfies (AP1), (AP2a), and (AP2b). In either case, there are three types of components of AIS-generators: (i) regular functions such as \hat{r} and $\hat{\varphi}$; (ii) history compression functions $\{\hat{\sigma}_t\}_{t \geq 1}$; and (iii) stochastic kernels \hat{P} and \hat{P}^y . To learn these components from data, we must choose parametric class of functions for all of these. In this section, we do not make any assumption about how these components are chosen. In particular, \hat{r} and $\hat{\varphi}$ could be represented by any class of function approximators (such as a multi-layer perceptron); $\hat{\sigma}$ could be represented by any class of time-series approximators (such as a RNN or its refinements such as LSTM or GRU); and \hat{P} and \hat{P}^y could be represented by any class of stochastic kernel approximators (such as softmax distribution or mixture of Gaussians). We use ξ_t to denote the corresponding parameters.

There are two losses in the definition of an AIS: the reward loss $|R_t - \hat{r}(\hat{z}_t, a_t)|$ and the prediction loss $d_{\mathfrak{F}}(\mu_t, \nu_t)$ or $d_{\mathfrak{F}}(\mu_t^y, \nu_t^y)$. We combine these into a single criterion and minimize the combined loss function

$$\frac{1}{T} \sum_{t=1}^T \left[\lambda |R_t - \hat{r}(\hat{Z}_t, A_t)|^2 + (1 - \lambda) d_{\mathfrak{F}}(\mu_t, \nu_t)^2 \right]$$

where T is the length of the episode or the rollout horizon and $\lambda \in [0, 1]$ may be viewed as a hyper-parameter.

As described in Section 6.1, there are two possibilities to efficiently compute $d_{\mathfrak{F}}(\mu_t, \nu_t)^2$: use total-variation distance or Wasserstein distance as the IPM and use surrogate loss (57); or use distance-based MMD as the IPM and use the surrogate loss (61).

In particular, to choose an AIS that satisfies (AP1) and (AP2), we either minimize the surrogate loss

$$\frac{1}{T} \sum_{t=1}^T [\lambda |R_t - \hat{r}(\hat{Z}_t, A_t)|^2 + (1 - \lambda) \log(\nu_t(\hat{Z}_{t+1}))] \quad (62)$$

or we minimize the surrogate loss (specialized for $p = 2$)

$$\frac{1}{T} \sum_{t=1}^T [\lambda |R_t - \hat{r}(\hat{Z}_t, A_t)|^2 + (1 - \lambda)(M_t - 2\hat{Z}_{t+1})^\top M_t] \quad (63)$$

where M_t is the mean of the distribution ν_t .

Similarly, in order to choose an AIS that satisfies (AP1), (AP2a) and (AP2b), we minimize the surrogate loss

$$\frac{1}{T} \sum_{t=1}^T [\lambda |R_t - \hat{r}(\hat{Z}_t, A_t)|^2 + (1 - \lambda) \log(\nu_t^y(Y_t))] \quad (64)$$

or we minimize the surrogate loss (specialized for $p = 2$)

$$\frac{1}{T} \sum_{t=1}^T [\lambda |R_t - \hat{r}(\hat{Z}_t, A_t)|^2 + (1 - \lambda)(M_t^y - 2Y_t)^\top M_t^y] \quad (65)$$

where M_t^y is the mean of the distribution ν_t^y .

We use $\bar{\xi}$ to denote the parameters of the AIS-generator, i.e., the parameters of $(\hat{\sigma}, \hat{P}, \hat{r})$ when using (AP1) and (AP2) or the parameters of $(\hat{\sigma}, \hat{\varphi}, \hat{P}^y, \hat{r})$ when using (AP1), (AP2a), (AP2b). We then use $\mathcal{L}(\bar{\xi})$ to denote the corresponding loss (62), (63), (64), or (65). Then, we can learn the parameters $\bar{\xi}$ using stochastic gradient descent:

$$\bar{\xi}_{k+1} = \bar{\xi}_k - a_k \nabla_{\bar{\xi}} \mathcal{L}(\bar{\xi}_k), \quad (66)$$

where the learning rates $\{a_k\}_{k \geq 0}$ satisfy the standard conditions $\sum a_k = \infty$ and $\sum a_k^2 < \infty$.

6.3 AIS-based PORL

Given the stochastic gradient descent algorithm to learn an AIS-generator for a fixed policy, we can simultaneously learn a policy and AIS-generator by following a multi time-scale stochastic gradient descent (Borkar, 1997), where we learn the AIS-generator at a faster learning rate than the policy.

In particular, let $\pi_\theta: \hat{Z} \rightarrow \Delta(\mathbf{A})$ be a parameterized stochastic policy with parameters θ . Let $J(\bar{\xi}, \theta)$ denote the performance of policy π_θ . From the policy gradient theorem (Sutton et al., 2000; Baxter and Bartlett, 2001), we know that

$$\nabla_\theta J(\bar{\xi}, \theta) = \mathbb{E} \left[\sum_{t=1}^{\infty} \left(\sum_{\tau=1}^t \nabla_\theta \log \pi_\theta(A_\tau | \hat{Z}_\tau) \right) \gamma^{t-1} R_t \right] \quad (67)$$

which can be estimated from a sampled trajectory with a rollout horizon of T using the G(PO)MDP gradient (Baxter and Bartlett, 2001)

$$\widehat{\nabla}_{\theta} J(\bar{\xi}, \theta) = \sum_{t=1}^T \left(\sum_{\tau=1}^t \nabla_{\theta} \log \pi_{\theta}(A_t | \hat{Z}_t) \right) \gamma^{t-1} R_t. \quad (68)$$

We can iteratively update the parameters $\{(\bar{\xi}_k, \theta_k)\}_{k \geq 1}$ of both the AIS-generator and policy as follows. We start with an initial choice $(\bar{\xi}_1, \theta_1)$, update both parameters after a rollout of T as follows

$$\bar{\xi}_{k+1} = \bar{\xi}_k - a_k \nabla_{\bar{\xi}} \mathcal{L}(\bar{\xi}_k) \quad \text{and} \quad \theta_{k+1} = \theta_k + b_k \widehat{\nabla}_{\theta} J(\bar{\xi}_k, \theta_k) \quad (69)$$

where the learning rates $\{a_k\}_{k \geq 1}$ and $\{b_k\}_{k \geq 1}$ satisfy the standard conditions on multi time-scale learning: $\sum_k a_k = \infty$, $\sum_k b_k = \infty$, $\sum_k a_k^2 < \infty$, $\sum_k b_k^2 < \infty$, and $\lim_{k \rightarrow \infty} b_k/a_k = 0$, which ensures that AIS-generator learns at a faster rate than the policy.

A similar idea can be used for an actor-critic algorithm. Suppose we have a parameterized policy $\pi_{\theta}: \hat{Z} \rightarrow \Delta(\mathbf{A})$ and a parameterized critic $\hat{Q}_{\zeta}: \hat{Z} \times \mathbf{A} \rightarrow \mathbb{R}$, where θ denotes the parameters of the policy and ζ denotes the parameters of the critic. Let $J(\bar{\xi}, \theta, \zeta)$ denote the performance of the policy. From the policy gradient theorem (Sutton et al., 2000; Konda and Tsitsiklis, 2003), we know that

$$\nabla_{\theta} J(\bar{\xi}, \theta, \zeta) = \frac{1}{1-\gamma} \mathbb{E}[\nabla_{\theta} \log \pi_{\theta}(A_t | \hat{Z}_t) Q_{\zeta}(\hat{Z}_t, A_t)] \quad (70)$$

which can be estimated from a sampled trajectory with a rollout horizon of T by

$$\widehat{\nabla}_{\theta} J(\bar{\xi}, \theta, \zeta) = \frac{1}{(1-\gamma)T} \sum_{t=1}^T \nabla_{\theta} \log \pi_{\theta}(A_t | \hat{Z}_t) \hat{Q}_{\zeta}(\hat{Z}_t, A_t). \quad (71)$$

For the critic, we use the temporal difference loss

$$\mathcal{L}_{\text{TD}}(\bar{\xi}, \theta, \zeta) = \frac{1}{T} \sum_{t=1}^T \text{smoothL1}(\hat{Q}_{\zeta}(\hat{Z}_t, A_t) - R_t - \gamma \hat{Q}_{\zeta}(\hat{Z}_{t+1}, A_{t+1})) \quad (72)$$

where **smoothL1** is the smooth L_1 distance given by

$$\text{smoothL1}(x) = \begin{cases} \frac{1}{2}x^2 & \text{if } |x| < 1 \\ |x| - \frac{1}{2} & \text{otherwise.} \end{cases}$$

We can iteratively update the parameters $\{(\bar{\xi}_k, \theta_k, \zeta_k)\}_{k \geq 1}$ of the AIS-generator, policy, and critic as follows. We start with an initial choice $(\bar{\xi}_1, \theta_1, \zeta_1)$, and update all the parameters after a rollout of T steps as follows

$$\bar{\xi}_{k+1} = \bar{\xi}_k - a_k \nabla_{\bar{\xi}} \mathcal{L}(\bar{\xi}_k), \quad \theta_{k+1} = \theta_k + b_k \widehat{\nabla}_{\theta} J(\bar{\xi}_k, \theta_k, \zeta_k) \quad \text{and} \quad \zeta_{k+1} = \zeta_k - c_k \nabla_{\zeta} \mathcal{L}_{\text{TD}}(\bar{\xi}_k, \theta_k, \zeta_k) \quad (73)$$

where the learning rates $\{a_k\}_{k \geq 1}$, $\{b_k\}_{k \geq 1}$, $\{c_k\}_{k \geq 1}$ satisfy the standard conditions on multi time-scale learning: $\sum_k a_k = \infty$, $\sum_k b_k = \infty$, $\sum_k c_k = \infty$, $\sum_k a_k^2 < \infty$, $\sum_k b_k^2 < \infty$,

Algorithm 1 AIS-based PORL algorithm

Input: Initial AIS-Generator: $(\hat{\sigma}, \hat{P}, \hat{r})_{\bar{\xi}_0}$, Initial Policy: π_{θ_0} , Discount factor: γ ,
 Reward weight: λ , Number of episodes: K , AIS-LR: $a_{k=1}^K$, Policy-LR: $b_{k=1}^K$.

Output: Learned policy: π_{θ_K} , Learned AIS-generator: $(\hat{\sigma}, \hat{P}, \hat{r})_{\bar{\xi}_K}$

```

1: procedure AIS-BASED PORL
2:   for all  $k \in \{1, \dots, K\}$  do
3:     Reset environment and perform an episode using  $\pi_{\theta_{k-1}}, (\hat{\sigma}, \hat{P}, \hat{r})_{\bar{\xi}_{k-1}}$ .
4:      $A_{1:T}, Y_{1:T}, R_{1:T} \leftarrow$  Actions, observations, and rewards for episode  $k$ .
5:     Compute AIS loss using  $A_{1:T}, Y_{1:T}, R_{1:T}, \lambda, (\hat{\sigma}, \hat{P}, \hat{r})_{\bar{\xi}_{k-1}}$  using Eq. (64) or (65)
6:     Compute policy loss using  $A_{1:T}, Y_{1:T}, R_{1:T}, \gamma, \pi_{\theta_{k-1}}, (\hat{\sigma})_{\bar{\xi}_{k-1}}$  using Eq. (68)
7:     Update AIS parameters  $\bar{\xi}_{k-1}$  and policy parameters  $\pi_{\theta_{k-1}}$  using Eq. (69)

```

$\sum_k c_k^2 < \infty$, $\lim_{k \rightarrow \infty} c_k/a_k = 0$, and $\lim_{k \rightarrow \infty} b_k/c_k = 0$, which ensures that AIS-generator learns at a faster rate than the critic, and the critic learns at a faster rate than the policy. The complete algorithm is shown in Algorithm 1.

Under standard technical conditions (see Theorem 23 of Borkar (1997) or Page 35 of Leslie (2004)), we can show that iterations (69) and (73) will converge to a stationary point of the corresponding ODE limits. At convergence, depending on ε and δ for the quality of AIS approximation, we can obtain approximation guarantees corresponding to Theorem 27. For a more detailed discussion on convergence, please refer to Appendix E.

We conclude this discussion by mentioning that algorithms similar to the AIS-based PORL have been proposed in the literature including Bakker (2002); Wierstra et al. (2007, 2010); Hausknecht and Stone (2015); Heess et al. (2015); Zhu et al. (2017); Ha and Schmidhuber (2018); Baisero and Amato (2018); Igl et al. (2018); Zhang et al. (2019). However, these papers only discuss the empirical performance of the proposed algorithms but do not derive performance bounds.

7. Experiments

We perform numerical experiments to check the effectiveness of AIS-based PORL algorithms proposed in the previous section. The code for all AIS experiments is available in Subramanian et al. (2020). We consider three classes of POMDP environments, which have increasing difficulty in terms of the dimension of their state and observation spaces:

1. Low-dimensional environments (Tiger, Voicemail, and Cheese Maze)
2. Moderate-dimensional environments (Rock Sampling and Drone Surveillance)
3. High-dimensional environments (different variations of MiniGrid)

For each environment, we use the actor only framework and learn an AIS based on (AP1), (AP2a), and (AP2b). There are four components of the corresponding AIS-generator: the history compression function $\hat{\sigma}$, the AIS update function $\hat{\varphi}$, the reward prediction function \hat{r} , and the observation prediction kernel \hat{P}^y . We model the $\hat{\sigma}$ as an LSTM, where the memory update unit of LSTM acts as $\hat{\varphi}$. We model \hat{r} , \hat{P}^y , and the policy $\hat{\pi}$ as feed-forward neural

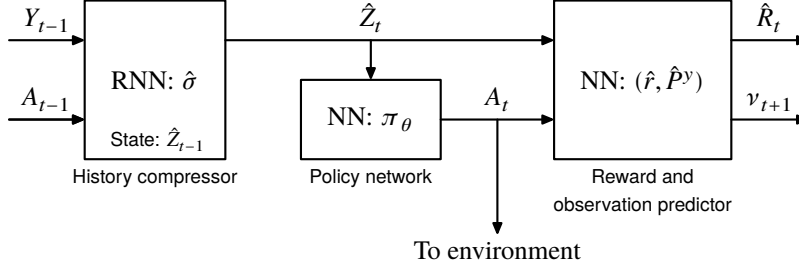


Figure 4: Network architecture for PORL using AIS.

networks. A block diagram of the network architecture is shown in Fig. 4 and the details of the networks and the hyperparameters are presented in Appendix F. To avoid over-fitting, we use the same network architecture and hyperparameters for all environments in the same difficulty class.

We repeat each experiment for multiple random seeds and plot the median value along with the uncertainty band from the first to the third quartile. For all environments, we compare our performance with a baseline which uses an actor-critic algorithm where both the actor and critic are modeled using LSTM and the policy parameters are updated using PPO. This architecture was proposed as a baseline for the Minigrid environments in Chevalier-Boisvert et al. (2018a). The details of the baseline architecture are presented in Appendix F.

To evaluate the performance of the policy while training for AIS-based PORL, a separate set of rollouts is carried out at fixed intervals of time steps and the mean of these rollouts is considered. For the PPO baseline a number of parallel actors are used during training, and once the episodes are completed, their returns are stored in a list. A fixed number (based on the number of parallel actors) of past episodes are considered to evaluate the mean performance of the current policy during training. See Appendix F for details.

For the low and moderate dimensional environments, we compare the performance with the best performing planning solution obtained from the JuliaPOMDP repository (Egorov et al., 2017). For the high-dimensional environments, finding a planning solution is intractable, so we only compare with the PPO baseline mentioned previously.

7.1 Low-dimensional environments

In these POMDP environments, the size of the unobserved state space is less than about 10 and the planning solution can be easily obtained using standard POMDP solvers.

1. **TIGER:** The Tiger environment is a sequential hypothesis testing task proposed in Kaelbling et al. (1998). The environment consists of two doors, with a tiger behind one door and a treasure behind the other. The agent can either perform a LISTEN action, which has a small negative reward of -1 and gives a noisy observation about the location of the tiger, or the agent can open one of the doors. Opening the door with the treasure gives a reward of $+10$ while opening the door with a tiger gives a large negative reward of -100 . After opening a door, the environment is reset.

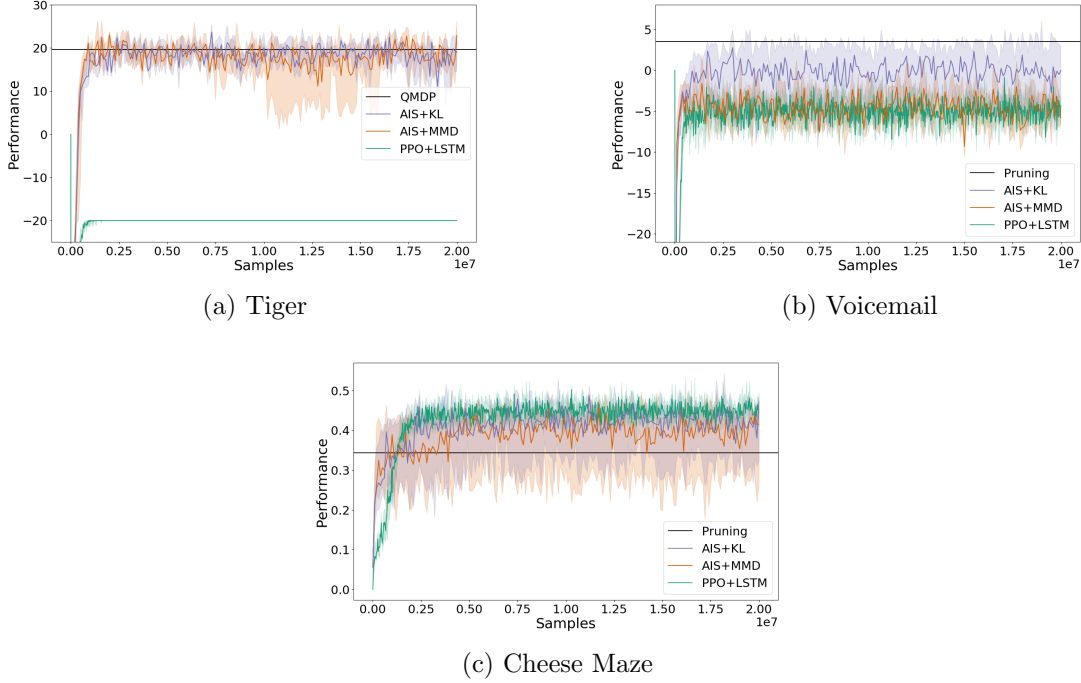


Figure 5: Comparison of AIS-based actor only PORL algorithm with LSTM+PPO baseline for low-dimensional environments (for 10 random seeds).

2. VOICEMAIL: The Voicemail environment is also a sequential hypothesis testing task proposed in Williams and Young (2007). This environment models a dialog system for managing voicemails. The agent can either perform an ASK action, which has a small negative reward of -1 and gives a noisy observation about the intent of the user, or the agent can execute SAVE or DELETE. Choosing a SAVE/DELETE action which matches the intent of the user gives a reward of $+5$. The agent receives a negative reward of -20 for action DELETE when the user intent is SAVE, while choosing action SAVE when the user intent is DELETE gives a smaller but still significant negative reward of -10 . Since the user prefers SAVE more than DELETE, the initial belief is given by $[0.65, 0.35]$ for SAVE and DELETE respectively. After taking a SAVE/DELETE action, the agent moves on to the next voicemail message.
3. CHEESEMAZE: The CheeseMaze environment is a POMDP with masked states proposed in McCallum (1993). The environment consists of 11 states and 7 observations as shown on the right. The objective is to reach the goal state, which is indicated by observation 7. The agent only receives a reward of $+1$, when the goal state is reached.

1	2	3	2	4
5		5		5
6		7		6

For all three environments, we compare the performance of AIS-based PORL with the LSTM+PPO baseline, described earlier. We also compare with the best performing planning solution from the JuliaPOMDP repository (Egorov et al., 2017). The results are presented in

Fig. 5, which shows both AIS-based PORL and LSTM+PPO converge close to the planning solutions relatively quickly.³

7.2 Moderate-dimensional environments

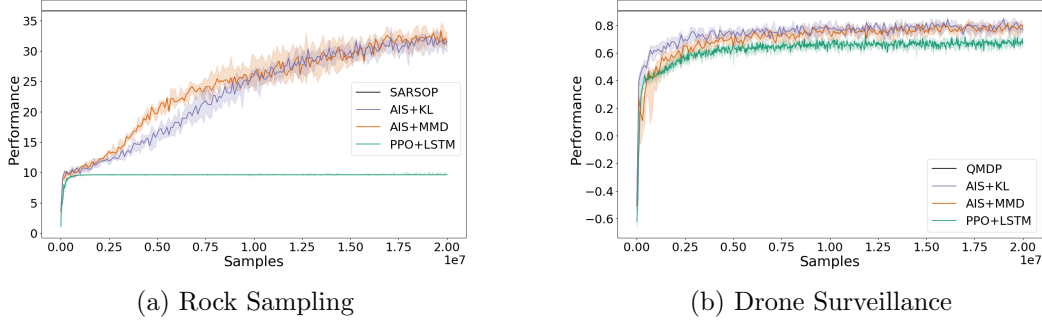
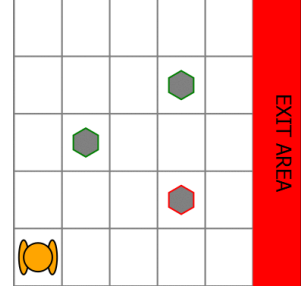


Figure 6: Comparison of AIS-based actor only PORL algorithm with LSTM+PPO baseline for moderate-dimensional environments (for 10 random seeds).

In these environments, the size of the unobserved state is moderately large (of the order of 10^2 to 10^3 unobserved states) and the optimal planning solution cannot be easily obtained using standard POMDP solvers. However, an approximate planning solution can be easily obtained using standard approximation algorithms for POMDPs.

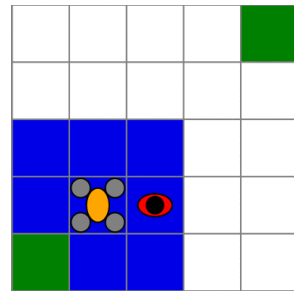
1. **ROCKSAMPLE**: RockSample is a scalable POMDP environment introduced in Smith and Simmons (2004) which models the rover science exploration. The $\text{RockSample}(n, k)$ environment consists of a $n \times n$ grid with k rocks. The rocks are at known positions. Some of the rocks which are labeled as GOOD rocks have scientific values; other rocks which are labeled as BAD rocks do not. Sampling a rock is expensive and the agent has a noisy long-range sensor to help determine if a rock is GOOD before choosing to approach and sample it.



At each stage, the agent can choose from $k + 5$ actions: NORTH, SOUTH, EAST, WEST, SAMPLE, $\text{CHECK}_1, \dots, \text{CHECK}_k$. The first four are deterministic single-step motion actions. The SAMPLE action samples the rock at the current location; if the rock is GOOD, there is a reward of +20 and the rock becomes BAD (so that no further reward can be gained from sampling it); if the rock is BAD, there is a negative reward of -10. The right edge of the map is a terminal state and reaching it gives a reward of +10. In our experiments, we use a $\text{RockSample}(5, 3)$ environment.

3. The performance of all learning algorithms for CHEESEMAZE are better than the best planning solution. We solved the CHEESEMAZE model with other solvers available in the JuliaPOMDP Egorov et al. (2017), and all these solution performed worse than the solution obtained by incremental pruning presented here.

2. **DRONESURVEILLANCE**: DroneSurveillance is a POMDP model of deploying an autonomous aerial vehicle in a partially observed, dynamic, indoor environment introduced in Svoreňová et al. (2015). The environment is a 5×5 grid with two agents: a ground agent which moves randomly and an aerial agent, whose motion has to be controlled. The aerial agent starts at the bottom-left cell and has to reach the upper-right cell (the goal state) without being in the same location as the ground agent. The ground agent cannot enter the start or goal states. The aerial agent has a downward facing camera which can view a 3×3 grid centered at its current location and it can perfectly see the location of the ground agent if it is in this view. At each stage, the aerial agent may choose from 5 actions: NORTH, SOUTH, EAST, WEST, HOVER. The first four are deterministic single-step motion actions and the HOVER action keeps the aerial vehicle at its current position. Reaching the goal gives a reward of +1 and ends the episode. If both agents are in the same cell, there is a negative reward of -1 and the episode ends.

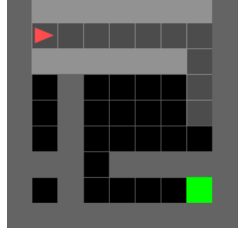
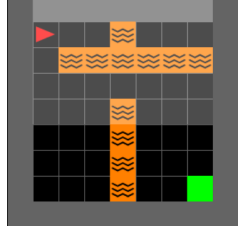
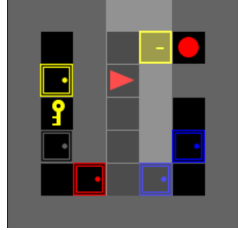
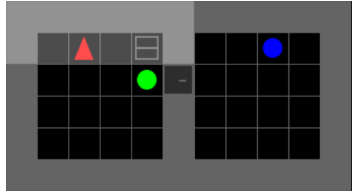


The visualizations above are taken from the JuliaPOMDP environments (Egorov et al., 2017). For both environments, we compare the performance of AIS-based PORL with the LSTM+PPO baseline described earlier. We also compare with the best performing planning solution from the JuliaPOMDP repository (Egorov et al., 2017). The results are shown in Fig. 6 which shows that both AIS-based PORL algorithms converge close to the best planning solution in both environments. The performance of LSTM+PPO is similar in DRONESURVEILLANCE but LSTM+PPO gets stuck in a local minima in ROCKSAMPLE.

7.3 High-dimensional environments

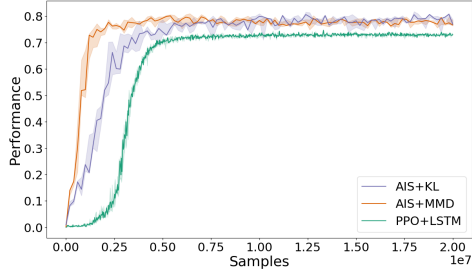
We use the MiniGrid environments from the BabyAI platform (Chevalier-Boisvert et al., 2018b), which are partially observable 2D grid environments which has tasks of increasing complexity level. The environment has multiple entities (agent, walls, lava, boxes, doors, and keys); objects can be picked up, dropped, and moved around by the agent; doors can be unlocked via keys of the same color (which might be hidden inside boxes). The agents can see a 7×7 view in front of it but it cannot see past walls and closed doors. At each time, it can choose from the following actions: {MOVE FORWARD, TURN LEFT, TURN RIGHT, OPEN DOOR/BOX, PICK UP ITEM, DROP ITEM, DONE}. The agent can only hold one item at a time. The objective is to reach a goal state in the quickest amount of time (which is captured by assigning to the goal state a reward which decays over time).

Most of the environments have a certain theme, and we cluster the environments accordingly. The visualizations below are taken from the Gym Minigrid environments (Chevalier-Boisvert et al., 2018b).

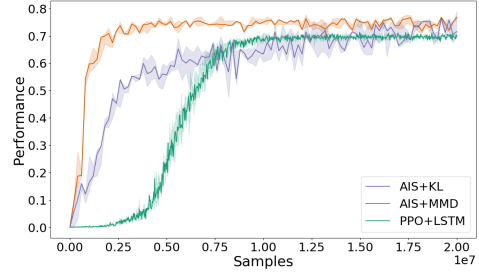
1. **SIMPLE CROSSING:** A simple crossing environment is a 2D grid with columns of walls with an opening (or a crossing). The agent can traverse the wall only through the openings and needs to find a path from the start to the goal state. There are four such environments (MGSCS9N1, MGSCS9N2, MGSCS9N3, and MGSCS11N5) where the label S_nN_m means that the size of the environment is $n \times n$ and there are m columns of walls. 
2. **LAVA CROSSING:** The lava crossing environments are similar to the simple crossing environments, but the walls are replaced by lava. If the agent steps on to the lava block then it dies and the episode ends. Therefore, exploration is more difficult in lava crossing as compared to simple crossing. There are two such environments (MGLCS9N1 and MGLCS9N2) where the label S_nN_m has the same interpretation as simple crossing. 
3. **KEY CORRIDOR:** The key corridor environments consist of a central corridor which has rooms on the left and right sides which can be accessed through doors. When the door is locked it can be opened using a key of the same color. The agent has to move to the location of the key, pick it up, move to the location of the correct door, open the door, drop the key, and pick up the colored ball. There are three such environments (MGKCS3R1, MGKCS3R2, and MGKCS3R3), where the label S_nR_m means that the size of the grid is proportional to n and the number of rooms present is $2m$. 
4. **OBSTRUCTED MAZE:** The obstructed maze environments are similar to key corridor environments but the key is inside a box and the box has to be opened to find the key. We consider two such environments (MGOM1Dl and MGOM1Dlh). In MGOM1Dl box is already open while in MGOM1Dlh the box is closed. There is an additional such environment in the BabyAI platform (MGOM1Dlhb), which is more suitable for continual learning algorithms so we exclude it here. 

The number of observations in a given Minigrid environment is discrete but is too large to model it as a one-hot encoded discrete observation as done in the previous environments. Instead we compress the observations as described in Section 3.7 by using an autoencoder to convert a large discrete space to a continuous space with a tractable size. A separate autoencoder is trained for each environment using a dataset that is created by performing random rollouts. Once the autoencoder is trained over the fixed dataset for several epochs, it is fixed and used to generate the observations for learning the AIS. This is very similar to Ha and Schmidhuber (2018), where they learn the autoencoder in a similar fashion and then fix it, following which their training procedure for the next observation distribution prediction and policy takes place.

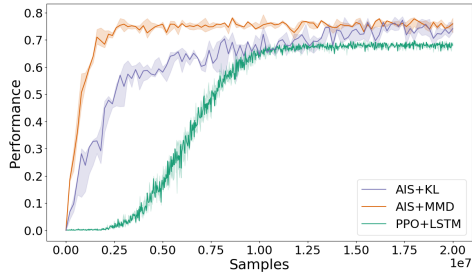
Note that the output of the autoencoder is a continuous variable and we are using MMD with $p = 2$ as an IPM. As explained in Section 6.1, $d_{\mathfrak{F}_2}(\mu, \nu)^2$ only depends on the



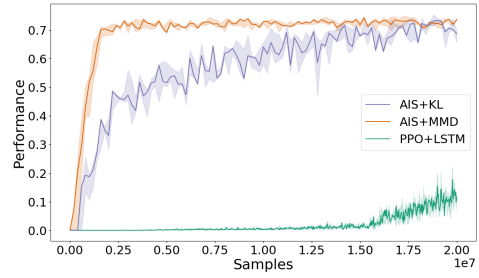
(a) MGSCS9N1



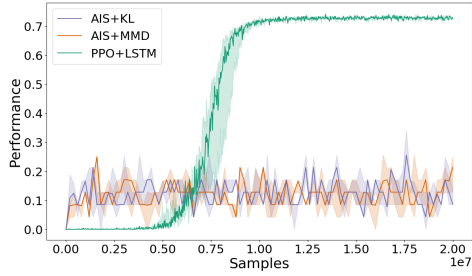
(b) MGSCS9N2



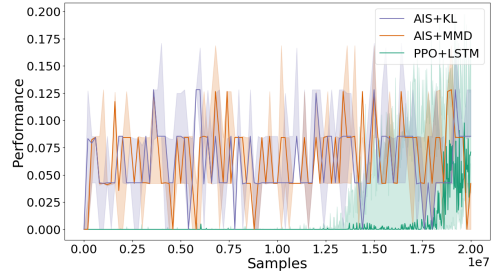
(c) MGSCS9N3



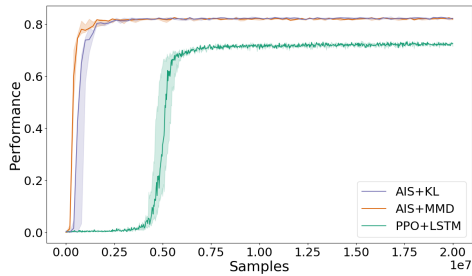
(d) MGSCS11N5



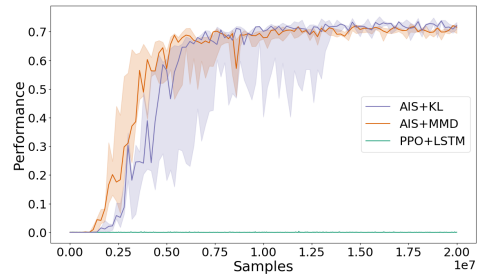
(e) MGLCS9N1



(f) MGLCS9N2



(g) MGKCS3R1



(h) MGKCS3R2

Figure 7: Comparison of AIS-based actor only PORL algorithm with LSTM+PPO baseline for high-dimensional environments (for 5 random seeds).

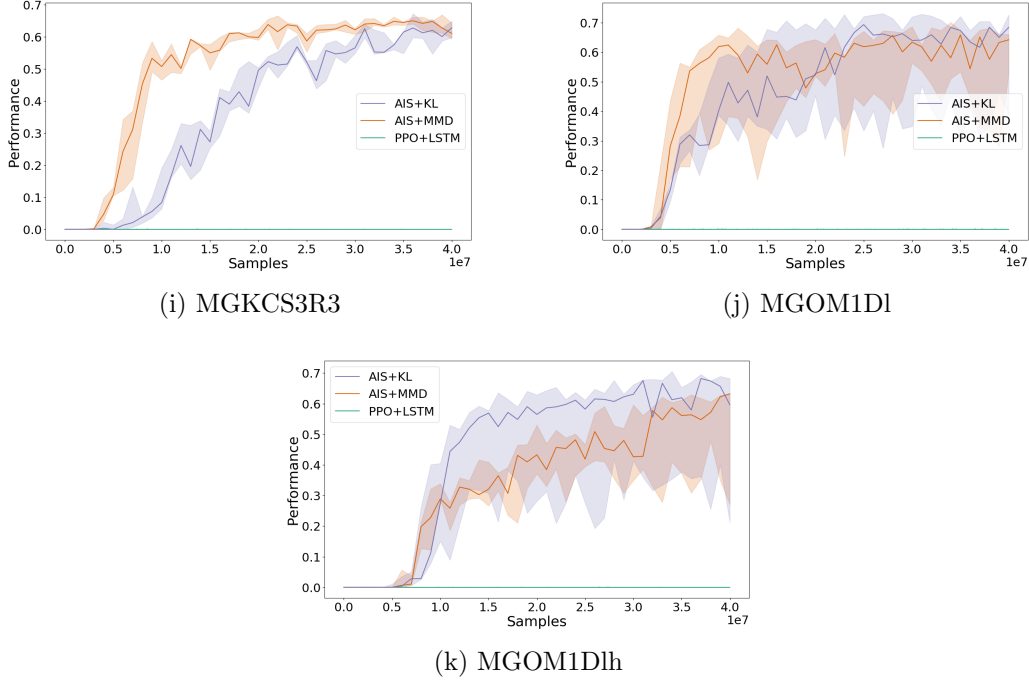


Figure 7 (continued): Comparison of AIS-based actor only PORL algorithm with LSTM+PPO baseline for high-dimensional environments (for 5 random seeds).

mean of μ and ν . So, to simplify the computations, we assume that ν is a Dirac delta distribution centered at its mean. Thus, effectively, we are predicting the mean of the next observation. In general, simply predicting the mean of the observations may not lead to a good representation, but in the Minigrid environments, the transitions are deterministic and the only source of stochasticity in the observations is due to the initial configuration of the environment. So, in practice, simply predicting the mean of the next observation works reasonably well. We emphasize that for other more general environments with truly stochastic observations, such a choice of IPM may not work well and it may be better to choose the MMD $d_{\mathfrak{F}_p}$ defined in Proposition 32 for a different value of p , say $p = 1$ (which corresponds to the energy distance (Székely and Rizzo, 2004)).

For all minigrid environments, we compare the performance of AIS-based PORL with the LSTM+PPO baseline proposed in Chevalier-Boisvert et al. (2018a). The results are shown in Fig. 7 which shows that for most environments AIS-based PORL converges to better performance values. Note that AIS-based PORL fails to learn in the LAVA CROSSING environments (MGLCS9N1 and MGLCS9N2) while LSTM+PPO fails to learn in the larger KEY CROSSING environments (MGKCS3R2 and MGKCS3R3) and in the OBSTRUCTED MAZE environments (MGOM1Dl and MGOM1Dlh).

The results indicate that one IPM does not necessarily lead to better performance than others in all cases. The performance of a particular IPM depends on whether the observation and AIS spaces are discrete or continuous, on the size of these spaces, and the stochasticity of the environment. The fact that we are approximating the policy using non-linear neural

networks makes it difficult to quantify the impact of the choice of IPM on the accuracy of learning. It will be important to understand this impact in more detail and develop guidelines on how to choose an IPM based on the features of the environment.

8. Conclusion

In this paper, we present a theoretical framework for approximate planning and learning in partially observed system. Our framework is based on the fundamental notion of information state. We provide two equivalent definitions of information state. An information state is a function of history which is sufficient to compute the expected reward and predict its next value. Equivalently, an information state is a function of the history which can be recursively updated and is sufficient to compute the expected reward and predict the next observation. We show that an information state always leads to a dynamic programming decomposition and provide several examples of simplified dynamic programming decompositions proposed in the literature which may be viewed as specific instances of information states.

We then relax the definition of an information state to describe an approximate information state (AIS), which is a function of the history that approximately satisfies the properties of the information state. We show that an AIS can be used to identify an approximately optimal policy with the approximation error specified in terms of the “one-step” approximation errors in the definition of the AIS. We present generalizations of AIS to setups with observation and action compression as well as to multi-agent systems. We show that various approximation approaches for both fully and partially observed setups proposed in the literature may be viewed as special cases of AIS.

One of the salient features of the AIS is that it is defined in terms of properties that can be estimated from data, and hence the corresponding AIS generators can be learnt from data. These can then be used as history representations in partially observed reinforcement learning (PORL) algorithms. We build up on this idea to present policy gradient algorithms which learn an AIS representation and an optimal policy and/or action-value function using multi time-scale stochastic gradient descent.

We present detailed numerical experiments which compare the performance of AIS-based PORL algorithms with a state-of-the-art PORL algorithm for three classes of partially observed problems—small, medium and large scale problems—and find out that AIS-based PORL outperforms the state-of-the-art baseline in most cases.

We conclude by observing that in this paper we restricted attention to the simplest classes of algorithms but the same idea can be extended to develop AIS-based PORL algorithms which uses value-based approaches such as Q-learning and its improved variants such as DQN, DDQN, distributional RL, etc. Finally, we note that the AIS representation includes a model of the system, so it can be used as a component of model-based reinforcement learning algorithms such as Dyna (Sutton and Barto, 2018, Sec 8.2, page 161). Such an approach will provide anytime guarantees on the approximation error which will depend on the “one-step” approximation error of the current AIS-representation. Therefore, we believe that AIS presents a systematic framework to reason about learning in partially observed environments.

Acknowledgment

The authors are grateful to Demosthenis Teneketzis, Peter Caines, and Dileep Kalathil for useful discussions and feedback. The work of JS and AM was supported in part by the Natural Science and Engineering Research Council of Canada through Discovery Grant RGPIN-2016-05165. The work of AS, RS, and AM was supported in part by the Innovation for Defence Excellence and Security (IDEaS) Program of the Canadian Department of National Defence through grant CFPMN2-037. AS was also supported by an FRQNT scholarship. The numerical experiments were enabled in part by support provided by Calcul Québec and Compute Canada.

References

- D. Abel, D. Hershkowitz, and M. Littman. Near optimal behavior via approximate state abstraction. In M. F. Balcan and K. Q. Weinberger, editors, *International Conference on Machine Learning (ICML)*, volume 48, pages 2915–2923, New York, New York, USA, June 2016.
- S. Adlakha, S. Lall, and A. Goldsmith. Networked Markov decision processes with delays. *IEEE Transactions on Automatic Control (TAC)*, 57(4):1013–1018, Apr. 2012.
- M. Aicardi, F. Davoli, and R. Minciardi. Decentralized optimal control of Markov chains with a common past information set. *IEEE Trans. Autom. Control*, 32(11):1028–1031, 1987.
- E. Altman and P. Nain. Closed-loop control with delayed information. *ACM SIGMETRICS Performance Evaluation Review*, 20(1):193–204, June 1992.
- C. Amato, D. S. Bernstein, and S. Zilberstein. Optimizing fixed-size stochastic controllers for POMDPs and decentralized POMDPs. *Autonomous Agents and Multi-Agent Systems (AAMAS)*, 21(3):293–320, 2010.
- J. Arabneydi and A. Mahajan. Team optimal control of coupled subsystems with mean-field sharing. In *IEEE Conference on Decision and Control (CDC)*, pages 1669–1674. IEEE, 2014.
- J. Arabneydi and A. Mahajan. Reinforcement learning in decentralized stochastic control systems with partial history sharing. In *American Control Conference (ACC)*, pages 5449–5456. IEEE, July 2015.
- K. Asadi, D. Misra, and M. Littman. Lipschitz continuity in model-based reinforcement learning. In *International Conference on Machine Learning (ICML)*, pages 264–273, Stockholm, Sweden, 10–15 Jul 2018.
- K. J. Aström. Optimal control of Markov processes with incomplete state information. *Journal of mathematical analysis and applications*, 10(1):174–205, 1965.
- K. J. Aström. *Introduction to Stochastic Control Theory*. Dover, 1970.

- K. Azizzadenesheli, A. Lazaric, and A. Anandkumar. Reinforcement learning of POMDPs using spectral methods. In *Annual Conference on Learning Theory (COLT)*, volume 49, pages 193–256, 2016.
- A. Baisero and C. Amato. Learning internal state models in partially observable environments;. *Reinforcement Learning under Partial Observability, NeurIPS Workshop*, 2018.
- B. Bakker. Reinforcement learning with long short-term memory. In *Neural Information Processing Systems (NIPS)*, 2002.
- J. L. Bander and C. C. White. Markov decision processes with noise-corrupted and delayed state observations. *Journal of the Operational Research Society*, 50(6):660–668, June 1999.
- J. Baxter and P. L. Bartlett. Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 15:319–350, 2001.
- R. Bellman. *Dynamic Programming*. Princeton University Press, 1957.
- D. S. Bernstein, R. Givan, N. Immerman, and S. Zilberstein. The complexity of decentralized control of markov decision processes. *Mathematics of operations research*, 27(4):819–840, 2002.
- D. S. Bernstein, E. A. Hansen, and S. Zilberstein. Bounded policy iteration for decentralized pomdps. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 52–57, 2005.
- D. Bertsekas. Convergence of discretization procedures in dynamic programming. *IEEE Transactions on Automatic Control (TAC)*, 20:415–419, 1975.
- D. Bertsekas and J. Tsitsiklis. *Neuro-dynamic Programming*. Anthropological Field Studies. Athena Scientific, 1996. ISBN 9781886529106.
- J.-M. Bismut. An example of interaction between information and control: The transparency of a game. *IEEE Trans. Autom. Control*, 18(5):518–522, Oct. 1972.
- T. Bohlin. Information pattern for linear discrete-time models with stochastic coefficients. *IEEE Transactions on Automatic Control (TAC)*, 15(1):104–106, Feb. 1970.
- B. Boots, S. M. Siddiqi, and G. J. Gordon. Closing the learning-planning loop with predictive state representations. *The International Journal of Robotics Research*, 30(7):954–966, 2011.
- V. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press, 2008.
- V. S. Borkar. Stochastic approximation with two time scales. *Systems & Control Letters*, 29(5):291–294, 1997.

- A. Boularias and B. Chaib-Draa. Exact dynamic programming for decentralized POMDPs with lossless policy compression. In *International Conference on Automated Planning and Scheduling (ICAPS)*, pages 20–27. AAAI Press, 2008.
- Y. Burda, H. Edwards, A. Storkey, and O. Klimov. Exploration by random network distillation. *arXiv preprint arXiv:1810.12894*, 2018.
- A. Carlin and S. Zilberstein. Observation compression in Dec-POMDP policy trees. In *International Conference on Autonomous Agents and Multi-agent Systems (AAMAS)*, pages 31–45, 2008.
- A. Cassandra, M. L. Littman, and N. L. Zhang. Incremental pruning: A simple, fast, exact method for partially observable Markov decision processes. In *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*, 1997.
- A. R. Cassandra, L. P. Kaelbling, and M. L. Littman. Acting optimally in partially observable stochastic domains. In *AAAI Conference on Artificial Intelligence*, 1994.
- P. S. Castro, P. Panangaden, and D. Precup. Equivalence relations in fully and partially observable Markov decision processes. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1653–1658, 2009.
- J. Chakravorty and A. Mahajan. Sufficient conditions for the value function and optimal strategy to be even and quasi-convex. *IEEE Transactions on Automatic Control (TAC)*, 63(11):3858–3864, Nov. 2018.
- Y. Chandak, G. Theodorou, C. Nota, and P. S. Thomas. Lifelong learning with a changing action set. In *AAAI Conference on Artificial Intelligence*, pages 3373–3380, 2020.
- H.-T. Cheng. *Algorithms for Partially Observable Markov Decision Processes*. PhD thesis, University of British Columbia, Vancouver, BC, 1988.
- M. Chevalier-Boisvert, D. Bahdanau, S. Lahlou, L. Willems, C. Saharia, T. H. Nguyen, and Y. Bengio. Babyai: A platform to study the sample efficiency of grounded language learning. In *International Conference on Learning Representations (ICLR)*, 2018a.
- M. Chevalier-Boisvert, L. Willems, and S. Pal. Minimalistic gridworld environment for openai gym. <https://github.com/maximecb/gym-minigrid>, 2018b.
- J. P. Crutchfield and K. Young. Inferring statistical complexity. *Physical Review Letters*, 63: 105–108, July 1989.
- I. Csiszar and J. Körner. *Information theory: coding theorems for discrete memoryless systems*. Cambridge University Press, 2011.
- M. Davis and P. Varaiya. Information states for linear stochastic systems. *Journal of Mathematical Analysis and Applications*, 37(2):384–402, Feb. 1972.
- J. S. Dibangoye, C. Amato, O. Buffet, and F. Charpillat. Optimally solving dec-pomdp as continuous-state mdps. *Journal of Artificial Intelligence Research*, 55:443–497, 2016.

- J. E. Eckles. Optimum maintenance with incomplete information. *Operations Research*, 16(5):1058–1067, 1968.
- M. Egorov, Z. N. Sunberg, E. Balaban, T. A. Wheeler, J. K. Gupta, and M. J. Kochenderfer. POMDPs.jl: A framework for sequential decision making under uncertainty. *Journal of Machine Learning Research (JMLR)*, 18(26):1–5, 2017. URL <http://jmlr.org/papers/v18/16-300.html>.
- E. A. Feinberg. On essential information in sequential decision processes. *Mathematical Methods of Operations Research*, 62(3):399–410, Nov. 2005.
- N. Ferns, P. Panangaden, and D. Precup. Metrics for finite Markov decision processes. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 162–169. AUAI Press, 2004.
- N. Ferns, P. Panangaden, and D. Precup. Bisimulation metrics for continuous Markov decision processes. *SIAM Journal on Computing*, 40(6):1662–1714, 2011.
- V. Francois-Lavet, G. Rabusseau, J. Pineau, D. Ernst, and R. Fonteneau. On overfitting and asymptotic bias in batch reinforcement learning with partial observability. *Journal of Artificial Intelligence Research (JAIR)*, 65:1–30, 2019.
- C. Gelada, S. Kumar, J. Buckman, O. Nachum, and M. G. Bellemare. Deepmdp: Learning continuous latent space models for representation learning. In *International Conference on Machine Learning (ICML)*, 2019.
- R. Givan, T. Dean, and M. Greig. Equivalence notions and model minimization in Markov decision processes. *Artificial Intelligence*, 147(1-2):163–223, July 2003.
- P. Grassberger. Toward a quantitative theory of self-generated complexity. *International Journal of Theoretical Physics*, 25(9):907–938, Sept. 1986.
- P. Grassberger. *Complexity and forecasting in dynamical systems*. Springer Berlin Heidelberg, 1988. ISBN 978-3-540-45968-2.
- A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola. A kernel method for the two-sample-problem. In *Neural Information Processing Systems (NIPS)*, page 513–520, 2006.
- D. Ha and J. Schmidhuber. World models. *arXiv:1803.10122*, 2018.
- W. Hamilton, M. M. Fard, and J. Pineau. Efficient learning and planning with compressed predictive states. *The Journal of Machine Learning Research (JMLR)*, 15(1):3395–3439, 2014.
- E. A. Hansen. An improved policy iteration algorithm for partially observable MDPs. In *Neural Information Processing Systems (NIPS)*, page 1015–1021. MIT Press, 1997.
- E. A. Hansen. Solving POMDPs by searching in policy space. In *Uncertainty in Artificial Intelligence (UAI)*, pages 211–219, 1998.

- M. Hausknecht and P. Stone. Deep recurrent Q-learning for partially observable MDPs. In *AAAI Fall Symposium Series*, 2015.
- N. Heess, J. J. Hunt, T. P. Lillicrap, and D. Silver. Memory-based control with recurrent neural networks, 2015. arXiv:1512.04455.
- O. Hernández-Lerma and J. B. Lasserre. *Discrete-time Markov control processes: basic optimality criteria*. Springer, 2012.
- K. Hinderer. Lipschitz continuity of value functions in Markovian decision processes. *Mathematical Methods of Operations Research*, 62(1):3–22, Sep 2005. ISSN 1432-5217.
- S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.
- M. Igl, L. Zintgraf, T. A. Le, F. Wood, and S. Whiteson. Deep variational reinforcement learning for POMDPs. In *International Conference on Machine Learning (ICML)*, pages 2117–2126, 2018.
- M. T. Izadi and D. Precup. A planning algorithm for predictive state representations. In G. Gottlob and T. Walsh, editors, *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1520–1521. Morgan Kaufmann, 2003.
- T. Jaakkola, S. P. Singh, and M. I. Jordan. Reinforcement learning algorithm for partially observable markov decision problems. In *Neural Information Processing Systems (NIPS)*, pages 345–352. 1995.
- H. Jaeger. Observable operator models for discrete stochastic time series. *Neural computation*, 12(6):1371–1398, 2000.
- H. Jaeger, M. Zhao, and A. Kolling. Efficient estimation of OOMs. In *Advances in Neural Information Processing Systems*, pages 555–562, 2006.
- M. James, S. Singh, and M. Littman. Planning with predictive state representations. In *International Conference on Machine Learning and Applications (ICMLA)*, 2004.
- N. Jiang, A. Kulesza, and S. P. Singh. Improving predictive state representations via gradient descent. In *AAAI Conference on Artificial Intelligence*, pages 1709–1715, 2016.
- L. P. Kaelbling, M. L. Littman, and A. R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998.
- S. Katt, F. A. Oliehoek, and C. Amato. Bayesian reinforcement learning in factored POMDPs. In *Autonomous Agents and MultiAgent Systems (AAMAS)*, page 7–15, 2019.
- V. R. Konda and J. N. Tsitsiklis. On actor-critic algorithms. *SIAM Journal on Control and Optimization*, 42(4):1143–1166, 2003.
- A. Kulesza, N. Jiang, and S. Singh. Low-rank spectral learning with weighted loss functions. In *Artificial Intelligence and Statistics*, pages 517–525, 2015a.

- A. Kulesza, N. Jiang, and S. P. Singh. Spectral learning of predictive state representations with insufficient statistics. In *AAAI Conference on Artificial Intelligence*, pages 2715–2721, 2015b.
- A. Kumar and S. Zilberstein. Constraint-based dynamic programming for decentralized POMDPs with structured interactions. In C. Sierra, C. Castelfranchi, K. S. Decker, and J. S. Sichman, editors, *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 561–568. IFAAMAS, 2009.
- P. R. Kumar and P. Varaiya. *Stochastic Systems: Estimation, Identification and Adaptive Control*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1986. ISBN 0-13-846684-X.
- H. Kurniawati, D. Hsu, and W. S. Lee. Sarsop: Efficient point-based pomdp planning by approximating optimally reachable belief spaces. In *Robotics: Science and systems*, 2008.
- H. Kwakernaak. *Theory of Self-Adaptive Control Systems*, chapter Admissible Adaptive Control, pages 14–18. Springer, 1965.
- W. S. Lee, N. Rong, and D. Hsu. What makes some POMDP problems easy to approximate? In *Neural Information Processing Systems (NIPS)*, pages 689–696, 2008.
- D. S. Leslie. *Reinforcement learning in games*. PhD thesis, The University of Bristol, 2004.
- Y. Li, B. Yin, and H. Xi. Finding optimal memoryless policies of POMDPs under the expected average reward criterion. *European Journal of Operational Research*, 211(3): 556–567, 2011.
- M. L. Littman. Memoryless policies: Theoretical limitations and practical results. In *International conference on simulation of adaptive behavior*, volume 3, page 238, 1994.
- M. L. Littman, R. S. Sutton, and S. P. Singh. Predictive representations of state. In *Neural Information Processing Systems (NIPS)*, 2002.
- M. Liu, C. Amato, E. P. Anesta, J. D. Griffith, and J. P. How. Learning for decentralized control of multiagent systems in large, partially-observable stochastic environments. In *AAAI Conference on Artificial Intelligence*, pages 2523–2529, 2016.
- J. Loch and S. P. Singh. Using eligibility traces to find the best memoryless policy in partially observable markov decision processes. In *ICML*, pages 323–331, 1998.
- A. Mahajan. *Sequential decomposition of sequential dynamic teams: applications to real-time communication and networked control systems*. PhD thesis, Univ. Michigan, Ann Arbor, MI, Sept. 2008.
- A. Mahajan. Optimal decentralized control of coupled subsystems with control sharing. *IEEE Transactions on Automatic Control (TAC)*, 58(9):2377–2382, Sept. 2013.
- A. Mahajan and M. Mannan. Decentralized stochastic control. *Annals of Operations Research*, 241:109–126, June 2016.

- A. Mahajan and D. Teneketzis. Optimal performance of networked control systems with non-classical information structures. *SIAM Journal of Control and Optimization*, 48(3): 1377–1404, May 2009a.
- A. Mahajan and D. Teneketzis. Optimal design of sequential real-time communication systems. *IEEE Trans. Inf. Theory*, 55(11):5317–5338, Nov. 2009b.
- A. Mahajan, A. Nayyar, and D. Teneketzis. Identifying tractable decentralized control problems on the basis of information structure. In *Proc. 46th Annual Allerton Conf. Communication, Control, and Computing*, pages 1440–1449, Monticello, IL, Sept. 2008.
- A. Mahajan, N. C. Martins, M. C. Rotkowitz, and S. Yüksel. Information structures in optimal decentralized control. In *IEEE Conference on Decision and Control (CDC)*, pages 1291–1306. IEEE, 2012.
- J. Marschak. *Decision Processes*, chapter Towards an Economic Theory of Organization and Information. Wiley, New York, 1954.
- J. Marschak and R. Radner. *Economic theory of teams*. Yale University Press, 1972.
- R. A. McCallum. Overcoming incomplete perception with utile distinction memory. In *International Conference on Machine Learning (ICML)*, 1993.
- N. Meuleau, L. Peshkin, K.-E. Kim, and L. P. Kaelbling. Learning finite-state controllers for partially observable environments. In *Uncertainty in Artificial Intelligence (UAI)*, page 427–436, 1999.
- A. Müller. How does the value function of a Markov decision process depend on the transition probabilities? *Mathematics of Operations Research*, 22(4):872–885, 1997.
- A. Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.
- A. Nayyar. *Sequential Decision Making in Decentralized Systems*. PhD thesis, University of Michigan, 2011.
- A. Nayyar, A. Mahajan, and D. Teneketzis. Optimal control strategies in delayed sharing information structures. *IEEE Trans. Autom. Control*, 56(7):1606–1620, July 2011.
- A. Nayyar, A. Mahajan, and D. Teneketzis. Decentralized stochastic control with partial history sharing: A common information approach. *IEEE Transactions on Automatic Control (TAC)*, 58(7):1644–1658, 2013.
- A. Nerode. Linear automaton transformations. *Proceedings of American Mathematical Society*, 9:541–544, 1958.
- F. A. Oliehoek and C. Amato. *A concise introduction to decentralized POMDPs*, volume 1. Springer, 2015.

- J. M. Ooi, S. M. Verbout, J. T. Ludwig, and G. W. Wornell. A separation theorem for periodic sharing information patterns in decentralized control. *IEEE Trans. Autom. Control*, 42(11):1546–1550, Nov. 1997. ISSN 0018-9286. doi: 10.1109/9.649699.
- C. H. Papadimitriou and J. N. Tsitsiklis. The complexity of optimal queuing network control. *Mathematics of Operations Research*, 24(2):293–305, 1999.
- J. Pineau, G. Gordon, S. Thrun, et al. Point-based value iteration: An anytime algorithm for POMDPs. In *IJCAI*, volume 3, pages 1025–1032, 2003.
- P. Poupart and C. Boutilier. Bounded finite state controllers. In *Neural Information Processing Systems (NIPS)*, pages 823–830, 2004.
- P. Poupart and N. Vlassis. Model-based bayesian reinforcement learning in partially observable domains. In *Symp. on Artificial Intelligence and Mathematics*, 2008.
- P. Poupart, K.-E. Kim, and D. Kim. Closing the gap: Improved bounds on optimal POMDP solutions. In *Twenty-First International Conference on Automated Planning and Scheduling*, 2011.
- E. Rachelson and M. G. Lagoudakis. On the locality of action domination in sequential decision making. In *International Symposium on Artificial Intelligence and Mathematics*, Jan. 2010.
- S. T. Rachev. *Probability Metrics and the Stability of Stochastic Models*. Wiley, 1991.
- R. Radner. Team decision problems. *The Annals of Mathematical Statistics*, 33(3):857–881, 1962.
- Y. A. Reznik. An algorithm for quantization of discrete probability distributions. In *2011 Data Compression Conference*. IEEE, mar 2011.
- M. Rosencrantz, G. Gordon, and S. Thrun. Learning low dimensional predictive representations. In *International Conference on Machine Learning (ICML)*, 2004.
- S. Ross, B. Chaib-draa, and J. Pineau. Bayes-adaptive POMDPs. In *Neural Information Processing Systems (NIPS)*, pages 1225–1232, 2008.
- S. Ross, J. Pineau, B. Chaib-draa, and P. Kreitmann. A bayesian approach for learning and planning in partially observable markov decision processes. *Journal of Machine Learning Research (JMLR)*, 12(5), 2011.
- D. E. Rumelhart, G. E. Hinton, R. J. Williams, et al. Learning representations by back-propagating errors. *Nature*, 323(9):533–536, 1986.
- N. Saldi, T. Linder, and S. Yüksel. *Finite approximations in discrete-time stochastic control*. Springer, 2018.
- N. Sandell and M. Athans. Solution of some nonclassical lqg stochastic decision problems. *IEEE Trans. Autom. Control*, 19:108–116, 1974.

- N. Sandell, P. Varaiya, M. Athans, and M. Safonov. Survey of decentralized control methods for large scale systems. *IEEE Transactions on automatic Control (TAC)*, 23(2):108–128, 1978.
- I. Sason. On data-processing and majorization inequalities for f-divergences with applications. *Entropy*, 21(10):1022, oct 2019.
- D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5): 2263–2291, Oct. 2013.
- S. Seuken and S. Zilberstein. Memory-bounded dynamic programming for Dec-POMDPs. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2009–2015, 2007.
- C. R. Shalizi and J. P. Crutchfield. Computational Mechanics: Pattern and prediction, structure and simplicity. *Journal of Statistical Physics*, 104(3):817–879, Aug. 2001.
- G. Shani, R. I. Brafman, and S. E. Shimony. Forward search value iteration for POMDPs. In *IJCAI*, pages 2619–2624, 2007.
- G. Shani, J. Pineau, and R. Kaplow. A survey of point-based POMDP solvers. *International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, 27:1–51, 2013.
- S. P. Singh, M. L. Littman, N. K. Jong, D. Pardoe, and P. Stone. Learning predictive state representations. In *International Conference on Machine Learning (ICML)*, 2003.
- R. D. Smallwood and E. J. Sondik. The optimal control of partially observable Markov processes over a finite horizon. *Operations research*, 21(5):1071–1088, 1973.
- T. Smith and R. Simmons. Heuristic search value iteration for POMDPs. In *UA*, pages 520–527, 2004.
- M. T. Spaan and N. Vlassis. Perseus: Randomized point-based value iteration for POMDPs. *Journal of artificial intelligence research*, 24:195–220, 2005.
- B. K. Sriperumbudur, A. Gretton, K. Fukumizu, G. R. G. Lanckriet, and B. Schölkopf. Injective Hilbert space embeddings of probability measures. In *Conference on Learning Theory*, 2008.
- B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G. R. G. Lanckriet. On integral probability metrics, ϕ -divergences and binary classification, 2009. arXiv:0901.2698.
- B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G. R. G. Lanckriet. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6(0):1550–1599, 2012.
- C. Striebel. Sufficient statistics in the optimal control of stochastic systems. *Journal of Mathematical Analysis and Applications*, 12:576–592, 1965.
- J. Subramanian and A. Mahajan. Approximate information state for partially observed system. In *IEEE Conference on Decision and Control (CDC)*, Dec. 2019.

- J. Subramanian, A. Sinha, R. Seraj, and A. Mahajan. Approximate information state for reinforcement learning in partially observed systems. <https://github.com/info-structures/ais>, 2020.
- R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT Press, 2018.
- R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Neural Information Processing Systems (NIPS)*, pages 1057–1063, Nov. 2000.
- M. Svorenřová, M. Chmelík, K. Leahy, H. F. Eniser, K. Chatterjee, I. Āerná, and C. Belta. Temporal logic motion planning using POMDPs with parity objectives: Case study paper. In *International Conference on Hybrid Systems: Computation and Control*, page 233–238, 2015.
- G. J. Székely and M. L. Rizzo. Testing for equal distributions in high dimensions. *InterStat*, (5), 2004.
- D. Szer, F. Charpillet, and S. Zilberstein. MAA*: A heuristic search algorithm for solving decentralized POMDPs. In *Conference in Uncertainty in Artificial Intelligence (UAI)*, pages 576–590. AUAI Press, 2005.
- C. Villani. *Optimal transport: Old and New*. Springer, 2008.
- J. C. Walrand and P. Varaiya. Optimal causal coding-decoding problems. *IEEE Trans. Inf. Theory*, 29(6):814–820, Nov. 1983.
- S. D. Whitehead and L.-J. Lin. Reinforcement learning of non-markov decision processes. *Artificial Intelligence*, 73(1-2):271–306, 1995.
- W. Whitt. Approximations of dynamic programs, I. *Mathematics of Operations Research*, 3(3):231–243, 1978.
- D. Wierstra, A. Foerster, J. Peters, and J. Schmidhuber. Solving deep memory POMDPs with recurrent policy gradients. In *International Conference on Artificial Neural Networks (ICANN)*, 2007.
- D. Wierstra, A. Förster, J. Peters, and J. Schmidhuber. Recurrent policy gradients. *Logic Journal of the IGPL*, 18(5):620–634, 2010.
- J. D. Williams and S. Young. Partially observable Markov decision processes for spoken dialog systems. *Computer Speech & Language*, 21(2):393–422, 2007.
- J. K. Williams and S. P. Singh. Experimental results on learning stochastic memoryless policies for partially observable markov decision processes. In *Neural Information Processing Systems (NIPS)*, pages 1073–1080, 1999.
- H. S. Witsenhausen. A counterexample in stochastic optimum control. *SIAM Journal on Control*, 6(1):131–147, 1968.

- H. S. Witsenhausen. Separation of estimation and control for discrete time systems. *Proceedings of the IEEE*, 59(11):1557–1566, 1971.
- H. S. Witsenhausen. Some remarks on the concept of state. In Y. C. Ho and S. K. Mitter, editors, *Directions in Large-Scale Systems*, pages 69–75. Plenum, 1976.
- B. Wolfe, M. R. James, and S. Singh. Learning predictive state representations in dynamical systems without reset. In *International Conference on Machine Learning (ICML)*, 2005.
- B. Wolfe, M. R. James, and S. Singh. Approximate predictive state representations. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 363–370, 2008.
- T. Yoshikawa. Dynamic programming approach to decentralized stochastic control problems. *IEEE Trans. Autom. Control*, 20(6):796 – 797, Dec. 1975.
- A. Zhang, Z. C. Lipton, L. Pineda, K. Azizzadenesheli, A. Anandkumar, L. Itti, J. Pineau, and T. Furlanello. Learning causal state representations of partially observable environments, 2019. arXiv:1906.10437.
- H. Zhang. Partially observable Markov decision processes: A geometric technique and analysis. *Operations Research*, 2009.
- N. Zhang and W. Liu. Planning in stochastic domains: Problem characteristics and approximation. Technical Report HKUST-CS96-31, Hong Kong Univeristy of Science and Technology, 1996.
- P. Zhu, X. Li, P. Poupart, and G. Miao. On improving deep reinforcement learning for POMDPs, 2017. arXiv:1704.07978.
- V. M. Zolotarev. Probability metrics. *Theory of Probability & Its Applications*, 28(2): 278–302, Jan. 1983.

A. Comparison with the results of Abel et al. (2016) for state aggregation in MDPs

Abel et al. (2016) introduce four models of state aggregation and derive approximation bounds for all four. In this section, we show that one of these models, which they call *approximate model similarity* may be viewed as an AIS. We also show that the approximation bounds of Theorem 27 for this model are stronger than those derived in Abel et al. (2016) by a factor of $\mathcal{O}(1/(1-\gamma))$.

Since we follow a slightly different notation than Abel et al. (2016) and for the sake of completeness, we start by describing the notion of approximate model similarity defined in Abel et al. (2016).

Consider an infinite horizon finite-state finite-action MDP with state space \mathbf{S} , action space \mathbf{A} , transition probability matrix $P: \mathbf{S} \times \mathbf{A} \rightarrow \Delta(\mathbf{S})$, per-step reward function $r: \mathbf{S} \times \mathbf{A} \rightarrow \mathbb{R}$, and discount factor γ .

Let $\hat{\mathbf{S}}$ be an aggregated state space and it is assumed that the following two functions are available: a compression function $q: \mathbf{S} \rightarrow \hat{\mathbf{S}}$ and a weight function $w: \mathbf{S} \rightarrow [0, 1]$ such that for all $\hat{s} \in \hat{\mathbf{S}}$, $\sum_{s \in q^{-1}(\hat{s})} w(s) = 1$. Given these functions, define an aggregated MDP with state space $\hat{\mathbf{S}}$, action space \mathbf{A} , transition probability function $\hat{P}: \hat{\mathbf{S}} \times \mathbf{A} \rightarrow \hat{\mathbf{S}}$ given by

$$\hat{P}(\hat{s}'|\hat{s}, a) = \sum_{s \in q^{-1}(\hat{s})} \sum_{s' \in q^{-1}(\hat{s}')} P(s'|s, a) w(s), \quad \forall \hat{s}, \hat{s}' \in \hat{\mathbf{S}}, a \in \mathbf{A},$$

and a per-step reward $\hat{r}: \hat{\mathbf{S}} \times \mathbf{A} \rightarrow \mathbb{R}$ given by

$$\hat{r}(\hat{s}, a) = \sum_{s \in q^{-1}(\hat{s})} r(s, a) w(s), \quad \forall \hat{s} \in \hat{\mathbf{S}}, a \in \mathbf{A}.$$

Definition 34 (ε -approximate model similarity (Abel et al., 2016)) *The aggregated MDP is said to be ε -approximate model similar to the original MDP if it satisfies the following two properties:*

1. *For all $\hat{s} \in \hat{\mathbf{S}}$, $s_1, s_2 \in q^{-1}(\hat{s})$, and $a \in \mathbf{A}$, we have*

$$|r(s_1, a) - r(s_2, a)| \leq \varepsilon.$$

2. *For all $\hat{s}, \hat{s}' \in \hat{\mathbf{S}}$, $s_1, s_2 \in q^{-1}(\hat{s})$, and $a \in \mathbf{A}$, we have*

$$\left| \sum_{s' \in q^{-1}(\hat{s}')} P(s'|s_1, a) - \sum_{s' \in q^{-1}(\hat{s}')} P(s'|s_2, a) \right| \leq \varepsilon.$$

Proposition 35 (Lemma 2 of Abel et al. (2016)) *Let $\hat{\pi}: \hat{\mathbf{S}} \rightarrow \mathbf{A}$ be the (deterministic) optimal policy for the aggregated MDP. Define $\pi: \mathbf{S} \rightarrow \mathbf{A}$ by $\pi = \hat{\pi} \circ q$. Let $V: \mathbf{S} \rightarrow \mathbb{R}$ denote the optimal value function and let $V^\pi: \mathbf{S} \rightarrow \mathbb{R}$ denote the value function for policy π . Then, for all $s \in \mathbf{S}$*

$$|V(s) - V^\pi(s)| \leq \frac{2\varepsilon}{(1-\gamma)^2} + \frac{2\gamma\varepsilon|\mathbf{S}|\|r\|_\infty}{(1-\gamma)^3}.$$

Note that the result is presented slightly differently in Abel et al. (2016). They assume that $\|r\|_\infty = 1$ and simplify the above expression.

We now show an approximate model similarity is also an AIS and directly using the result of Theorem 27 for this model gives a stronger bound than Proposition 35.

Proposition 36 *Let (q, w) be such that the aggregated model is ε -approximate model similar to the true model. Then, (q, \hat{P}, \hat{r}) is an $(\varepsilon, \varepsilon|\hat{\mathbf{S}}|)$ -AIS with respect to the total variation distance.*

Proof We first establish (AP1). For any $s \in \mathbf{S}$ and $a \in \mathbf{A}$,

$$\begin{aligned} |r(s, a) - \hat{r}(q(s), a)| &\stackrel{(a)}{\leq} \left| \sum_{\tilde{s} \in q^{-1}(q(s))} w(\tilde{s})r(s, a) - \sum_{\tilde{s} \in q^{-1}(q(s))} w(\tilde{s})r(\tilde{s}, a) \right| \\ &\stackrel{(b)}{\leq} \sum_{\tilde{s} \in q^{-1}(q(s))} w(\tilde{s})|r(s, a) - r(\tilde{s}, a)| \\ &\stackrel{(c)}{\leq} \varepsilon \end{aligned}$$

where (a) follows from the basic property of the weight function w and the definition of the aggregated reward \hat{r} ; (b) follows from the triangle inequality; and (c) follows from the definition of approximate model similarity and the basic property of the weight function w . This establishes property (AP1).

Now, we establish (AP2). Let $d_{\mathfrak{F}}$ denote the total variation distance. Define probability measures μ, ν on $\Delta(\hat{\mathbf{S}})$ in the definition of (AP2), i.e., for any $s \in \mathbf{S}$, $\hat{s}' \in \hat{\mathbf{S}}$, and $a \in \mathbf{A}$,

$$\begin{aligned} \mu(\hat{s}') &:= \sum_{s' \in q^{-1}(\hat{s}')} P(s'|s, a) \\ \nu(\hat{s}') &:= \hat{P}(\hat{s}'|q(s), a) = \sum_{\tilde{s} \in q^{-1}(q(s))} \sum_{s' \in q^{-1}(\hat{s}')} P(s'|\tilde{s}, a)w(\tilde{s}) \end{aligned}$$

Now consider (see footnote 1 on page 13)

$$\begin{aligned} d_{\mathfrak{F}}(\mu, \nu) &= \sum_{\hat{s}' \in \hat{\mathbf{S}}} |\mu(\hat{s}') - \nu(\hat{s}')| \\ &= \sum_{\hat{s}' \in \hat{\mathbf{S}}} \left| \sum_{s' \in q^{-1}(\hat{s}')} P(s'|s, a) - \sum_{\tilde{s} \in q^{-1}(q(s))} \sum_{s' \in q^{-1}(\hat{s}')} P(s'|\tilde{s}, a)w(\tilde{s}) \right| \\ &\stackrel{(a)}{\leq} \sum_{\hat{s}' \in \hat{\mathbf{S}}} \sum_{\tilde{s} \in q^{-1}(q(s))} w(\tilde{s}) \left| \sum_{s' \in q^{-1}(\hat{s}')} P(s'|s, a) - \sum_{s' \in q^{-1}(\hat{s}')} P(s'|\tilde{s}, a) \right| \\ &\stackrel{(b)}{\leq} \sum_{\hat{s}' \in \hat{\mathbf{S}}} \sum_{\tilde{s} \in q^{-1}(q(s))} w(\tilde{s}) \varepsilon \stackrel{(c)}{=} \sum_{\hat{s}' \in \hat{\mathbf{S}}} \varepsilon = |\hat{\mathbf{S}}|\varepsilon, \end{aligned}$$

where (a) follows from triangle inequality, (b) follows from definition of approximate model similarity and (c) follows from the basic property of the weight function. This proves (AP2). ■

Lemma 37 *For any MDP*

$$\text{span}(V) \leq \frac{\text{span}(r)}{1 - \gamma}.$$

Therefore, when $d_{\mathfrak{F}}$ is the total variation distance, $\rho_{\mathfrak{F}}(V) \leq \frac{1}{2} \text{span}(r)/(1 - \gamma)$.

Proof This result follows immediately by observing that the per-step cost $r(S_t, A_t) \in [\min(r), \max(r)]$. Therefore, $\max(V) \leq \max(r)/(1 - \gamma)$ and $\min(V) \geq \min(r)/(1 - \gamma)$. ■

Proposition 38 *Let $\hat{\pi}$, π , V , and V^π be defined as in Proposition 36. Then, for all $s \in \mathcal{S}$,*

$$|V(s) - V^\pi(s)| \leq \frac{2\varepsilon}{(1 - \gamma)} + \frac{\gamma\varepsilon|\hat{\mathcal{S}}|\text{span}(r)}{(1 - \gamma)^2}.$$

Proof This follows immediately from Theorem 27, Proposition 36 and Lemma 37. ■

Note that the error bounds of Propositions 36 and 38 have similar structure but the key difference is that the bound of Proposition 38 is tighter than a factor of $1/(1 - \gamma)$ as compared to Proposition 36. There are other minor improvements as well ($|\hat{\mathcal{S}}|$ instead of $|\mathcal{S}|$ and $\frac{1}{2} \text{span}(r)$ instead of $\|r\|_\infty$).

B. Comparison with the results of Gelada et al. (2019) for latent space models for MDPs

Gelada et al. (2019) propose a latent space model for an MDP and show that minimizing the losses in predicting the per-step reward and repredicting the distribution over next latent space provides a bound on the quality of the representation. In this section, we show that latent space representation defined in Gelada et al. (2019) may be viewed as an instance of an AIS and show that the approximation bounds of Theorem 27 are similar to those derived in Gelada et al. (2019).

Since we follow a slightly different notation than Gelada et al. (2019) and for the sake of completeness, we start by describing the notion of latent space representation used in Gelada et al. (2019).

Consider an MDP with infinite horizon, finite-state and finite-action having state space \mathcal{S} , action space \mathcal{A} , transition probability matrix $P: \mathcal{S} \times \mathcal{A} \rightarrow \Delta(\mathcal{S})$, per-step reward function $r: \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ and discount factor γ .

Let $(\hat{\mathcal{S}}, d)$ be a Banach space and it is assumed that we are given an embedding function $\phi: \mathcal{S} \rightarrow \hat{\mathcal{S}}$, along with transition dynamics $\hat{P}: \hat{\mathcal{S}} \times \mathcal{A} \rightarrow \Delta(\hat{\mathcal{S}})$ and reward function $\hat{r}: \hat{\mathcal{S}} \times \mathcal{A} \rightarrow \mathbb{R}$. The MDP $\hat{\mathcal{M}} = (\hat{\mathcal{S}}, \mathcal{A}, \hat{P}, \hat{r}, \gamma)$ along with the embedding function ϕ is called the *latent space model* of the original MDP.

Definition 39 *The MDP $\hat{\mathcal{M}}$ is said to be (L_r, L_p) -Lipschitz if for any $\hat{s}_1, \hat{s}_2 \in \hat{\mathcal{S}}$ and $a \in \mathcal{A}$,*

$$|\hat{r}(\hat{s}_1, a) - \hat{r}(\hat{s}_2, a)| \leq L_r d(\hat{s}_1, \hat{s}_2) \quad \text{and} \quad d_{\mathfrak{F}}(\hat{P}(\cdot|\hat{s}_1, a), \hat{P}(\cdot|\hat{s}_2, a)) \leq L_p d(\hat{s}_1, \hat{s}_2),$$

where $d_{\mathfrak{F}}$ denotes the Kantorovich distance.

Given a latent space embedding, define

$$\varepsilon = \sup_{s \in \mathbf{S}, a \in \mathbf{A}} |r(s, a) - \hat{r}(\phi(s), a)| \quad \text{and} \quad \delta = \sup_{s \in \mathbf{S}, a \in \mathbf{A}} d_{\mathfrak{F}}(\mu, \hat{P}(\cdot | \phi(s), a)),$$

where $\mu \in \Delta(\hat{\mathbf{S}})$ given by $\mu(\mathbf{B}) = P(\phi^{-1}(\mathbf{B}) | s, a)$ for any Borel subset \mathbf{B} of $\hat{\mathbf{S}}$.

Proposition 40 (Theorem 5 of Gelada et al. (2019)) *Let $\hat{\pi}: \hat{\mathbf{S}} \rightarrow \mathbf{A}$ be the (deterministic) optimal policy of the latent space MDP. Define $\pi: \mathbf{S} \rightarrow \mathbf{A}$ by $\pi = \hat{\pi} \circ \phi$. Let $V: \mathbf{S} \rightarrow \mathbb{R}$ denote the optimal value function and let $V^\pi: \mathbf{S} \rightarrow \mathbb{R}$ denote the value function for policy π .*

If the latent space MDP $\widehat{\mathcal{M}}$ is (L_r, L_p) -Lipschitz, then,

$$|V(s) - V^\pi(s)| \leq \frac{2\varepsilon}{1 - \gamma} + \frac{2\gamma\delta L_r}{(1 - \gamma)(1 - \gamma L_p)}.$$

We show that a latent space model is an AIS and directly using the result of Theorem 27 gives the same approximation bound.

Proposition 41 *Let $\widehat{\mathcal{M}} = (\hat{\mathbf{S}}, \mathbf{A}, \hat{P}, \hat{r}, \gamma)$ be a latent space model with embedding function ϕ . Then, (ϕ, \hat{P}, \hat{r}) is an (ε, δ) -AIS with respect to the Kantorovich distance.*

Proof The result is an immediate consequence of the definition of ε and δ for latent space model. ■

Lemma 42 *For any (L_r, L_p) -Lipschitz MDP, if $\gamma L_p < 1$, then*

$$\|V\|_{\text{Lip}} \leq \frac{L_r}{1 - \gamma L_p}.$$

Therefore, when $d_{\mathfrak{F}}$ is the Kantorovich distance, $\rho_{\mathfrak{F}}(V) = \|V\|_{\text{Lip}} \leq L_r/(1 - \gamma L_p)$.

Proof This result follows immediately from Theorem 4.2 of Hinderer (2005). ■

Proposition 43 *Let $\hat{\pi}$, π , V , and V^π be defined in Proposition 40. The, for all $s \in \mathbf{S}$,*

$$|V(s) - V^\pi(s)| \leq \frac{2\varepsilon}{1 - \gamma} + \frac{2\gamma\delta L_r}{(1 - \gamma)(1 - \gamma L_p)}.$$

Proof This follows immediately from Theorem 27, Proposition 40, and Lemma 42. ■

Note that the error bounds in Propositions 40 and 43 are exactly the same.

C. Comparison with the results of Francois-Lavet et al. (2019) for belief approximation in POMDPs

Francois-Lavet et al. (2019) analyze the trade off between asymptotic bias and overfitting in reinforcement learning with partial observations. As part of their analysis, they express the quality of state representation in terms of the bounds on the L_1 error of the associated belief states. We show that these approximation bounds may be viewed as an instance of AIS-based bounds of Theorems 9 and 27. We also show that the bounds of Theorem 27 for this model are stronger than those derived in Francois-Lavet et al. (2019) by a factor of $\mathcal{O}(1/(1 - \gamma))$.

Since we follow a slightly different notation than Francois-Lavet et al. (2019) and for the sake of completeness, we start by describing the notion of ε -sufficient statistics defined in Francois-Lavet et al. (2019).

Consider an infinite-horizon finite-state finite-action POMDP with state space \mathbf{S} , action space \mathbf{A} , observation space \mathbf{Y} , transition probability matrix $P: \mathbf{S} \times \mathbf{A} \rightarrow \Delta(\mathbf{S})$, observation matrix $P^y: \mathbf{S} \rightarrow \Delta(\mathbf{Y})$, per-step reward $r: \mathbf{S} \times \mathbf{A} \rightarrow \mathbb{R}$, and discount factor γ .

Definition 44 (ε -sufficient statistic (Francois-Lavet et al., 2019)) *Given a family of Banach spaces $\{\Phi_t\}_{t=1}^L$, an ε -sufficient statistic is a collection of history compression function $\{\phi_t: \mathbf{H}_t \rightarrow \Phi_t\}_{t=1}^T$ and belief approximation functions $\{\hat{b}_t: \Phi_t \rightarrow \Delta(\mathbf{S})\}_{t=1}^T$ such that for any time t and any realization h_t of H_t , we have*

$$\|\hat{b}_t(\cdot|\phi_t(h_t)) - b_t(\cdot|h_t)\|_1 \leq \varepsilon.$$

Given an ε -sufficient statistic, Francois-Lavet et al. (2019) define an MDP with state space $\Delta(\mathbf{S})$, action space \mathbf{A} , transition probability kernel $\mathbb{P}(\hat{b}_{t+1}(\cdot|\phi(h_{t+1})) | \hat{b}_t(\cdot|\phi(h_t)), a_t)$ computed from the underlying POMDP, and per-step reward given by

$$\hat{r}(\hat{b}_t(h_t), a_t) = \sum_{s \in \mathbf{S}} r(s, a_t) \hat{b}_t(s|\phi(h_t)).$$

Proposition 45 (Theorem 1 of Francois-Lavet et al. (2019)) *Let $\{(\hat{b}_t, \phi_t)\}_{t=1}^T$ be an ε -sufficient statistic and $\hat{\pi} = (\hat{\pi}_1, \hat{\pi}_2, \dots)$ be an optimal policy for the MDP described above. Define a policy $\pi = (\pi_1, \pi_2, \dots)$ given by $\pi_t = \hat{\pi}_t \circ \phi_t$. Let $V_t: \mathbf{H}_t \rightarrow \mathbb{R}$ denote the optimal value functions and $\hat{V}_t^\pi: \mathbf{H}_t \rightarrow \mathbb{R}$ denote the value function for policy π . Then for any initial history $h_1 \in \mathbf{H}_1$,*

$$|V_1(h_1) - \hat{V}_1^\pi(h_1)| \leq \frac{2\varepsilon\|r\|_\infty}{(1 - \gamma)^3}.$$

We now show that an ε -sufficient statistic gives rise to an AIS and directly using the results of Theorem 27 for this model gives a stronger bound than Proposition 45.

Proposition 46 *Let $\{(\hat{b}_t, \phi_t)\}_{t=1}^T$ be an ε -sufficient statistic. Let $\hat{\mathbf{Z}}_t = \Delta(\mathbf{S})$ and define the different components of an AIS as follows:*

- history compression functions $\hat{\sigma}_t = \hat{b}_t \circ \phi_t$,

- AIS prediction kernels $\hat{P}_t(\cdot|\hat{z}_t, a_t)$ is given by

$$\hat{P}_t(\mathbf{B}|\hat{z}_t, a_t) = \sum_{y_{t+1} \in \mathbf{Y}} \psi(y_{t+1}|\hat{z}_t, a_t) \mathbb{1}_{\mathbf{B}}\{\hat{\varphi}(\hat{z}_t, y_{t+1}, a_t)\},$$

where

$$\psi(y_{t+1}|\hat{z}_t, a_t) = \sum_{s_{t+1} \in \mathbf{S}} \sum_{s_t \in \mathbf{S}} P^y(y_{t+1}|s_{t+1}) P(s_{t+1}|s_t, a_t) \hat{z}_t(s_t)$$

and

$$\hat{\varphi}(\hat{z}_t, y_{t+1}, a_t)(s_{t+1}) = \frac{\sum_{s_t \in \mathbf{S}} P^y(y_{t+1}|s_{t+1}) P(s_{t+1}|s_t, a_t) \hat{z}_t(s_t)}{\psi(y_{t+1}|\hat{z}_t, a_t)},$$

where $\hat{\varphi}$ is the same as the Bayes'-rule based update of the belief state,

- reward approximation functions $\hat{r}(\hat{z}_t, a_t) = \sum_{s \in \mathbf{S}} \hat{z}_t(s) r(s, a_t)$.

Then, $\{(\hat{\sigma}_t, \hat{P}_t, \hat{r}_t)\}_{t=1}^T$ is an $(\varepsilon\|r\|_\infty, 3\varepsilon)$ -AIS with respect to the bounded-Lipschitz metric.

Proof We need to equip $\hat{\mathbf{Z}} = \Delta(\mathbf{S})$ with a metric in order to define a bounded-Lipschitz metric over $\Delta(\hat{\mathbf{Z}})$. We use the total variation as the metric and denote it by d_{TV} . We use \mathfrak{F} to denote $\{f: \Delta(\hat{\mathbf{Z}}) \rightarrow \mathbb{R} : \|f\|_\infty + \|f\|_{\text{Lip}} \leq 1\}$ and denote the corresponding bounded-Lipschitz metric over $\Delta(\hat{\mathbf{Z}})$ by $d_{\mathfrak{F}}$.

We first establish (AP1). For any time t , realization h_t of history H_t , and action $a_t \in \mathbf{A}$, we have

$$\begin{aligned} & |\mathbb{E}[r(S_t, a_t) \mid H_t = h_t, A_t = a_t] - \hat{r}_t(\hat{\sigma}_t(h_t), a_t)| \\ &= \left| \sum_{s \in \mathbf{S}} r(s, a_t) b_t(s|h_t) - \sum_{s \in \mathbf{S}} r(s, a_t) \hat{b}_t(s|\phi(h_t)) \right| \\ &\stackrel{(a)}{\leq} \|r\|_\infty d_{\text{TV}}(b_t, \hat{b}_t) \\ &\stackrel{(b)}{\leq} \varepsilon \|r\|_\infty \end{aligned}$$

where (a) follows from (10) and the fact that for total variation distance $\rho_{\text{TV}}(r) \leq \|r\|_\infty$; and (b) follows from definition of ε -sufficient statistic.

Before establishing (AP2), we note that $\hat{\varphi}$ is the Bayes'-rule based update of the true belief; therefore,

$$b_{t+1}(\cdot|h_{t+1}) = \hat{\varphi}(b_t(\cdot|h_t), y_{t+1}, a_t).$$

For ease of notation, we use $b_t(\cdot)$ and $\hat{b}_t(\cdot)$ instead of $b_t(\cdot|h_t)$ and $\hat{b}_t(\cdot|\phi(h_t))$, when the conditioning is clear from context.

Now consider μ_t and ν_t as defined in the definition of (AP2). In particular, for any Borel set \mathbf{B} ,

$$\begin{aligned} \mu_t(\mathbf{B}) &= \sum_{y_{t+1} \in \mathbf{Y}} \psi(y_{t+1}|b_t, a_t) \mathbb{1}_{\mathbf{B}}\{\hat{b}_{t+1}(\cdot|\phi(h_t, y_{t+1}, a_t))\} \\ \nu_t(\mathbf{B}) &= \hat{P}_t(\mathbf{B}|\hat{z}_t, a_t). \end{aligned}$$

We also define an additional measure ξ_t given by

$$\xi_t(\mathbf{B}) = \sum_{y_{t+1} \in \mathbf{Y}} \psi(y_{t+1}|b_t, a_t) \mathbb{1}_{\mathbf{B}}\{\hat{\varphi}(b_t, y_{t+1}, a_t)\},$$

Now, by the triangle inequality

$$d_{\mathfrak{F}}(\mu_t, \nu_t) \leq d_{\mathfrak{F}}(\mu_t, \xi_t) + d_{\mathfrak{F}}(\xi_t, \nu_t). \quad (74)$$

Now consider the first term of (74):

$$\begin{aligned} d_{\mathfrak{F}}(\mu_t, \xi_t) &= \sup_{f \in \mathfrak{F}} \left| \int_{\hat{\mathbf{Z}}} f d\mu_t - \int_{\hat{\mathbf{Z}}} f d\xi_t \right| \\ &= \sup_{f \in \mathfrak{F}} \left| \sum_{y_{t+1} \in \mathbf{Y}} f(\hat{b}_{t+1}(\cdot|\phi(h_t, y_{t+1}, a_t))) \psi(y_{t+1}|b_t, a_t) \right. \\ &\quad \left. - \sum_{y_{t+1} \in \mathbf{Y}} f(b_{t+1}(\cdot|h_t, y_{t+1}, a_t)) \psi(y_{t+1}|b_t, a_t) \right| \\ &\stackrel{(a)}{\leq} \sum_{y_{t+1} \in \mathbf{Y}} d_{\text{TV}}(\hat{b}_{t+1}(\cdot|\phi(h_{t+1})), b_{t+1}(\cdot|h_{t+1})) \psi(y_{t+1}|h_t, a_t) \\ &\stackrel{(b)}{\leq} \varepsilon \end{aligned} \quad (75)$$

where (a) follows from triangle inequality and the fact that slope of f is bounded by 1; and (b) follows from the definition of ε -sufficient statistic (see footnote 1 on page 13). Now consider the second term of (74) (for ease of notation, we use $b_t(\cdot)$ instead of $b_t(\cdot|h_t)$):

$$\begin{aligned} d_{\mathfrak{F}}(\xi_t, \nu_t) &= \sup_{f \in \mathfrak{F}} \left| \int_{\hat{\mathbf{Z}}} f d\xi_t - \int_{\hat{\mathbf{Z}}} f d\nu_t \right| \\ &= \sup_{f \in \mathfrak{F}} \left| \sum_{y_{t+1} \in \mathbf{Y}} f(\hat{\varphi}(b_t, y_{t+1}, a_t)) \psi(y_{t+1}|b_t, a_t) - \sum_{y_{t+1} \in \mathbf{Y}} f(\hat{\varphi}(\hat{z}_t, y_{t+1}, a_t)) \psi(y_{t+1}|\hat{z}_t, a_t) \right| \\ &\stackrel{(c)}{\leq} \sup_{f \in \mathfrak{F}} \left| \sum_{y_{t+1} \in \mathbf{Y}} f(\hat{\varphi}(b_t, y_{t+1}, a_t)) \psi(y_{t+1}|b_t, a_t) - \sum_{y_{t+1} \in \mathbf{Y}} f(\hat{\varphi}(\hat{z}_t, y_{t+1}, a_t)) \psi(y_{t+1}|b_t, a_t) \right| \\ &\quad + \sup_{f \in \mathfrak{F}} \left| \sum_{y_{t+1} \in \mathbf{Y}} f(\hat{\varphi}(\hat{z}_t, y_{t+1}, a_t)) \psi(y_{t+1}|b_t, a_t) - \sum_{y_{t+1} \in \mathbf{Y}} f(\hat{\varphi}(\hat{z}_t, y_{t+1}, a_t)) \psi(y_{t+1}|\hat{z}_t, a_t) \right| \\ &\stackrel{(d)}{\leq} \sum_{y_{t+1} \in \mathbf{Y}} d_{\text{TV}}(\hat{\varphi}(b_t, y_{t+1}, a_t), \hat{\varphi}(\hat{z}_t, y_{t+1}, a_t)) \psi(y_{t+1}|b_t, a_t) \\ &\quad + \kappa_{\mathfrak{F}, \text{TV}}(\hat{\varphi}(\hat{z}_t, \cdot, a_t)) d_{\text{TV}}(\psi(\cdot|b_t, a_t), \psi(\cdot|\hat{z}_t, a_t)), \end{aligned} \quad (76)$$

where (c) follows from the triangle inequality; the first step of (d) follows from an argument similar to step (a) of (75); and the second part of (d) follows from (13).

Now, we obtain bounds for both terms of (76). For $y_{t+1} \in \mathbf{Y}$, define

$$\begin{aligned}\xi_t^y(y_{t+1}) &:= \psi(y_{t+1}|b_t, a_t) = \sum_{s_{t+1} \in \mathbf{S}} \sum_{s_t \in \mathbf{S}} P^y(y_{t+1}|s_{t+1})P(s_{t+1}|s_t, a_t)b_t(s_t|h_t), \\ \nu_t^y(y_{t+1}) &:= \psi(y_{t+1}|\hat{z}_t, a_t) = \sum_{s_{t+1} \in \mathbf{S}} \sum_{s_t \in \mathbf{S}} P^y(y_{t+1}|s_{t+1})P(s_{t+1}|s_t, a_t)\hat{z}_t(s_t),\end{aligned}$$

Total variation is also an f -divergence⁴ therefore, it satisfies the strong data processing inequality.⁵ Note that the definition of both ξ_t^y and ν_t^y may be viewed as outputs of a “channel” from s_t to y_{t+1} . In case of ξ_t^y , the channel input is distributed according to $b_t(\cdot|h_t)$ and in case of ν_t^y , the channel input is distributed according to \hat{z}_t . Therefore, from the data processing inequality,

$$d_{\text{TV}}(\xi_t^y, \nu_t^y) \leq d_{\text{TV}}(b_t(\cdot|h_t), \hat{z}_t) \leq \varepsilon \quad (77)$$

where the last inequality follows from the definition of ε -sufficient statistic.

A similar argument can be used to bound $d_{\text{TV}}(\hat{\varphi}(b_t, y_{t+1}, a_t), \hat{\varphi}(\hat{z}_t, y_{t+1}, a_t))$. In particular, we can think of $\hat{\varphi}(\cdot, y_{t+1}, a_t)$ as a channel from s_t to s_{t+1} . Then, by the data processing inequality,

$$d_{\text{TV}}(\hat{\varphi}(b_t, y_{t+1}, a_t), \hat{\varphi}(\hat{z}_t, y_{t+1}, a_t)) \leq d_{\text{TV}}(b_t, \hat{z}_t) \leq \varepsilon \quad (78)$$

where the last inequality follows from the definition of ε -sufficient statistic.

The final part of (76) that needs to be characterized is $\kappa_{\mathfrak{F}, \text{TV}}(\hat{\varphi}(\hat{z}_t, \cdot, a_t))$. From (12)

$$\kappa_{\mathfrak{F}, \text{TV}}(\hat{\varphi}(\hat{z}_t, \cdot, a_t)) = \sup_{f \in \mathfrak{F}} \rho_{\text{TV}}(f \circ \hat{\varphi}(\hat{z}_t, \cdot, a_t)) \leq \sup_{f \in \mathfrak{F}} \|f \circ \hat{\varphi}(\hat{z}_t, \cdot, a_t)\|_{\infty} \leq 1.$$

Substituting this bound along with (77) and (78) in (76), we get $d_{\mathfrak{F}}(\xi_t, \nu_t) \leq 2\varepsilon$. Substituting this along with (75) in (74), we get that $d_{\mathfrak{F}}(\mu_t, \nu_t) \leq 3\varepsilon$. Hence (AP2) is satisfied. \blacksquare

Lemma 47 *For any POMDP,*

$$\rho_{\mathfrak{F}}(V) = \|V\|_{\infty} + \|V\|_{\text{Lip}} \leq \frac{2\|r\|_{\infty}}{1 - \gamma}.$$

Proof The result follows immediately from the sup norm on the value function (Lemma 37 and the bounds on the Lipschitz constant of the value function (Lemma 1 of Lee et al. (2008))). \blacksquare

4. Let $f: \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}$ be a convex function such that $f(1) = 0$. Then the f -divergence between two measures μ and ν defined on a measurable space \mathbf{X} is given by

$$D_f(\mu\|\nu) = \int_{\mathbf{X}} f\left(\frac{d\mu}{d\nu}\right) d\nu.$$

Total variation is a f -divergence with $f(x) = |x - 1|$ (also see footnote 1 on page 1). Sriperumbudur et al. (2009) showed that total variation is the only non-trivial IPM which is also an f -divergence.

5. Let \mathbf{X} and \mathbf{Y} be measurable spaces, μ and ν be measures on \mathbf{X} and $P: \mathbf{X} \rightarrow \Delta(\mathbf{Y})$ be a stochastic kernel from \mathbf{X} to \mathbf{Y} . We use μP to denote the measure $\mu_{\mathbf{Y}}$ on \mathbf{Y} given by $\mu_{\mathbf{Y}}(dy) = \int_{\mathbf{X}} P(dy|x)\mu(dx)$. Similar interpretation holds for νP . Then, the *strong data processing inequality* (Sason, 2019) states that for any f -divergence, $D_f(\mu P\|\nu P) \leq D_f(\mu\|\nu)$.

Proposition 48 *Let $\hat{\pi}$, π , V , and V^π be as defined in Proposition 45. Then, for any initial history $h_1 \in \mathbf{H}_1$,*

$$|V(h_1) - V^\pi(h_1)| \leq \frac{2\varepsilon\|r\|_\infty}{(1-\gamma)} + \frac{6\gamma\varepsilon\|r\|_\infty}{(1-\gamma)^2}.$$

Proof This follows immediately from Theorem 27, Proposition 46, and Lemma 47. \blacksquare

Note that the error bounds of Propositions 45 and 48 have similar structure but the key difference is that the bound of Proposition 48 is tighter by a factor of $1/(1-\gamma)$.

D. Comparison with the results of Chandak et al. (2020) on lifelong learning for time-varying action spaces

Lifelong learning refers to settings where a reinforcement learning agent adapts to a time-varying environment. There are various models for lifelong learning and Chandak et al. (2020) recently proposed a model where the action spaces change over time. The environment has an underlying finite state space \mathbf{S} , finite action space \mathbf{A} , and reward $r: \mathbf{S} \rightarrow \mathbb{R}$. Note that the reward depends only on the current state and not the current action.

It is assumed that there is an underlying finite dimensional representation space \mathbf{E} and for any feasible action $a \in \mathbf{A}$, there is an underlying representation $e \in \mathbf{E}$. This relationship is captured via an invertible map ϕ , i.e., $a = \phi(e)$. There is a transition kernel $P: \mathbf{S} \times \mathbf{E} \rightarrow \Delta(\mathbf{S})$ with respect to this representation space. This induces a transition kernel $P^a: \mathbf{S} \times \mathbf{A} \rightarrow \Delta(\mathbf{S})$ with respect to the action, where $P^a(s'|s, a) = P(s'|s, \phi^{-1}(a))$. It is assumed that the transition kernel P is ρ -Lipschitz, i.e., for all $s, s' \in \mathbf{S}$ and $e_i, e_j \in \mathbf{E}$,

$$\|P(s'|s, e_i) - P(s'|s, e_j)\|_1 \leq \rho\|e_i - e_j\|_1.$$

Chandak et al. (2020) consider infinite horizon discounted setup with discount factor γ .

Initially, the RL agent is not aware of the action space and learns about the actions in discrete stages indexed by $k \in \mathbb{Z}_{\geq 0}$. At stage k , the agent becomes aware of a subset \mathbf{U}_k of \mathbf{E} , where $\mathbf{U}_k \supseteq \mathbf{U}_{k-1}$. Thus, the environment at stage k may be modelled as an MDP $\mathcal{M}_k = \{\mathbf{S}, \mathbf{A}_k, P_k^a, r\}$, where $\mathbf{A}_k = \{\phi(e) : e \in \mathbf{U}_k\}$ and $P_k^a(s'|s, a) = P(s'|s, \phi^{-1}(a))$.

Two main results are established in Chandak et al. (2020). The first one is the following.

Proposition 49 (Theorem 1 of Chandak et al. (2020)) *Let π_k and V^{π_k} denote the optimal policy for MDP \mathcal{M}_k and its performance. Let V denote the value function for the hypothetical model when the agent has access to all actions. Let*

$$\eta_k = \sup_{a_i, a_j \in \mathbf{A}_k} \|\phi^{-1}(a_i) - \phi^{-1}(a_j)\|_1.$$

Then, for any $s \in \mathbf{S}$,

$$V(s) - V^{\pi_k}(s) \leq \frac{\gamma\rho\eta_k\|r\|_\infty}{(1-\gamma)^2}.$$

We now show that this result may be viewed as a corollary of Corollary 19. In particular, we have the following.

Lemma 50 *The action set \mathbf{A}_k may be viewed as a “quantization” of \mathbf{A} using a function $\psi: \mathbf{A} \rightarrow \mathbf{A}_k$, which maps any action $a = \phi(e) \in \mathbf{A}$ to action $a' = \phi(e') \in \mathbf{A}_k$ such that $e' = \arg \min_{e'' \in \mathbf{U}_k} \|e - e''\|_1$. Then, ψ is $(0, \rho\eta_k)$ -action-quantizer with respect to the total variation distance.*

Proof Since the per-step reward does not depend on the action, there is no approximation error in the reward and, therefore, $\varepsilon = 0$. Now note that for any $s \in \mathbf{S}$ and $a \in \mathbf{A}$, we have

$$d_{\text{TV}}(P^a(\cdot|s, a), P^a(\cdot|s, \psi(a))) \leq \sup_{e_i, e_j \in \mathbf{U}_k} \|P(\cdot|s, e_i) - P(\cdot|s, e_j)\|_1 \leq \rho\eta_k$$

where the last equality follows from the ρ -Lipschitz continuity of P and the definition of η_k . Thus, $\delta = \rho\eta_k$. \blacksquare

Proposition 51 *Let π_k , V^{π_k} and V be as defined in Proposition 49. Then, for any $s \in \mathbf{S}$,*

$$V(s) - V^{\pi_k}(s) \leq \frac{\gamma\rho\eta_k \text{span}(r)}{2(1 - \gamma)^2}$$

Proof The result can be established from the following observations: (i) The result of Corollary 19 continues to hold in the infinite horizon discount reward setup with α_t replaced by $(\varepsilon + \gamma\rho_{\mathfrak{F}}(\hat{V}^*)\delta)/(1 - \gamma)$. This can be established in a manner similar to Theorem 27. (ii) From Lemma 37, we know that for total variation distance $\rho_{\mathfrak{F}}(\hat{V}^*) \leq \frac{1}{2} \text{span}(r)/(1 - \gamma)$. The result follows from substituting the values of (ε, δ) from Lemma 50 and the value of $\rho_{\mathfrak{F}}(\hat{V}^*)$ from (ii) in (i). \blacksquare

Note that if the rewards $r(s)$ belongs in a symmetric interval, say $[-R_{\max}, R_{\max}]$, as is assumed in Chandak et al. (2020), the result of Proposition 51 matches that of Proposition 49.

The second result of Chandak et al. (2020) is for the setting when the mapping ϕ is not known. They assume that the agent selects some finite dimensional representation $\hat{\mathbf{E}}$ and, for every k , parameterizes the policy using two components: (i) a map $\beta: \mathbf{S} \rightarrow \Delta(\hat{\mathbf{E}})$ and (ii) an estimator $\hat{\phi}_k: \hat{\mathbf{E}} \rightarrow \Delta(\mathbf{A}_k)$. Then the action at any state $S_t \in \mathbf{S}$ is chosen by first sampling $\hat{e} \sim \beta(s)$ and then choosing the action $a \sim \hat{\phi}_k(\hat{e})$. The second main result in Chandak et al. (2020) is the following.⁶

Proposition 52 (Theorem 2 of Chandak et al. (2020)) *Let $\hat{\pi}_k$ denote the best overall policy that can be represented using the above structure, $V^{\hat{\pi}_k}$ denotes its performance, and V denote the value function when the agent has access to the complete model. Suppose there exists a $\zeta \in \mathbb{R}_{\geq 0}$, $\beta: \mathbf{S} \rightarrow \Delta(\hat{\mathbf{E}})$ and $\hat{\phi}_k: \hat{\mathbf{E}} \rightarrow \Delta(\mathbf{A}_k)$, such that for*

$$\sup_{s \in \mathbf{S}, a_k \in \mathbf{A}_k} \text{KL}(P^a(\cdot|s, a_k) \| P^a(\cdot|s, \hat{A})) \leq \zeta_k^2/2,$$

6. This result is stated slightly different in Chandak et al. (2020) using an *inverse dynamics* function $\varphi: \mathbf{S} \times \mathbf{S} \rightarrow \Delta(\hat{\mathbf{E}})$, where $e \sim \varphi(s, s')$ is a prediction of a latent action e which might have caused the transition from s to s' . However, the bounds hold for the simpler form presented here as well.

where $\hat{A} \sim \hat{\phi}_k(\hat{E})$ and $\hat{E} \sim \beta(s)$. Then, for any $s \in \mathcal{S}$,

$$V(s) - V^{\pi_k}(s) \leq \frac{\gamma(\rho\eta_k + \zeta_k)\|r\|_\infty}{(1 - \gamma)^2}.$$

We now show that this result may be viewed as a corollary of Corollary 19. In particular, we have the following.

Lemma 53 *The action set $\hat{\mathcal{E}}$ may be viewed as a “compression” of the “quantized” action set \mathcal{A}_k . In particular, let $\psi: \mathcal{A} \rightarrow \mathcal{A}_k$ be as defined in Lemma 50. Then, the function $\hat{\phi}_k^{-1} \circ \psi$ is a $(0, \rho\eta_k + \zeta_k)$ -action-quantizer with respect to the total variation distance.*

Proof As argued in the proof of Lemma 50, since the reward function does not depend on action, $\varepsilon = 0$. Now, recall that from Pinsker’s inequality, for any distribution μ and ν , $d_{\text{TV}}(\mu, \nu) \leq \sqrt{2D_{\text{KL}}(\mu\|\nu)}$. Thus,

$$\sup_{s \in \mathcal{S}, a_k \in \mathcal{A}_k} d_{\text{TV}}(P^a(\cdot|s, a_k), P^a(\cdot|s, \hat{A})) \leq \zeta_k$$

where $\hat{A} \sim \hat{\phi}_k(\hat{E})$ and $\hat{E} \sim \beta(s)$. Now, by the triangle inequality, for any $s \in \mathcal{S}$ and $a \in \mathcal{A}$

$$\begin{aligned} d_{\text{TV}}(P^a(\cdot|s, a), P^a(\cdot|s, \hat{A})) &\leq d_{\text{TV}}(P^a(\cdot|s, a), P^a(\cdot|s, \psi(a))) + d_{\text{TV}}(P^a(\cdot|s, \psi(a)), P^a(\cdot|s, \hat{A})) \\ &\leq \rho\eta_k + \zeta_k, \end{aligned}$$

Thus, $\delta = \rho\eta_k + \zeta_k$. ■

Proposition 54 *Let $\hat{\pi}_k$, $V^{\hat{\pi}_k}$ and V be as defined in Proposition 49. Then, for any $s \in \mathcal{S}$,*

$$V(s) - V^{\hat{\pi}_k}(s) \leq \frac{\gamma(\rho\eta_k + \zeta_k) \text{span}(r)}{2(1 - \gamma)^2}$$

Proof The proof is similar to the proof of Proposition 51, where we replace the values of Lemma 50 with those of Lemma 53. ■

As before, if the rewards $r(s)$ belongs in a symmetric interval, say $[-R_{\max}, R_{\max}]$, as is assumed in Chandak et al. (2020), the result of Proposition 54 matches that of Proposition 52.

E. Convergence of the PORL algorithm

In this section, we discuss the convergence of the PORL algorithm presented in Sec. 6.2 and 6.3. The proof of convergence relies on multi-timescale stochastic approximation Borkar (1997) under conditions similar to the standard conditions for convergence of policy gradient algorithms with function approximation stated below:

Assumption 2 *The following conditions are satisfied:*

1. *All network parameters $(\bar{\xi}_k, \zeta_k, \theta_k)$ lie in convex and bounded subsets of Euclidean spaces.*

2. The gradient of the loss function $\nabla_{\bar{\xi}} \mathcal{L}(\bar{\xi}_k)$ of the state approximator is Lipschitz in $\bar{\xi}_k$, the gradient of the TD loss $\nabla_{\zeta} \mathcal{L}_{TD}(\bar{\xi}_k, \theta_k, \zeta_k)$ and the policy gradient $\hat{\nabla}_{\theta_k} J(\bar{\xi}_k, \theta_k, \zeta_k)$ is Lipschitz in $(\bar{\xi}_k, \theta_k, \zeta_k)$ with respect to the sup norm.
3. All the gradients— $\nabla_{\bar{\xi}} \mathcal{L}(\bar{\xi}_k)$ at the state approximator; $\nabla_{\zeta} \mathcal{L}_{TD}(\bar{\xi}_k, \theta_k, \zeta_k)$ at the critic; and $\hat{\nabla}_{\theta_k} J(\bar{\xi}_k, \theta_k, \zeta_k)$ at the actor—are unbiased with bounded variance. Furthermore, the critic and the actor function approximators are compatible as given in Sutton et al. (2000), i.e.,

$$\frac{\partial Q_{\zeta_k}(\hat{Z}_t, A_t)}{\partial \zeta} = \frac{1}{\pi_{\theta_k}(\hat{Z}_t, A_t)} \frac{\partial \pi_{\theta_k}(\hat{Z}_t, A_t)}{\partial \theta}.$$

4. The learning rates are sequences of positive numbers $\{a_k\}_{k \geq 0}, \{b_k\}_{k \geq 0}, \{c_k\}_{k \geq 0}$ that satisfy: $\sum a_k = \infty, \sum b_k = \infty, \sum c_k = \infty, \sum a_k^2 < \infty, \sum b_k^2 < \infty, \sum c_k^2 < \infty, \lim_{k \rightarrow \infty} c_k/a_k = 0$, and $\lim_{k \rightarrow \infty} b_k/c_k = 0$.

Assumption 3 The following regularity conditions hold:

1. The ODE corresponding to θ in (73) is locally asymptotically stable.
2. The ODEs corresponding to $\bar{\xi}$ and ζ in (73) are globally asymptotically stable. In addition, the ODE corresponding to ζ has a fixed point which is Lipschitz continuous in θ .

The proposed RL framework has the following convergence guarantees.

Theorem 55 Under Assumptions 2 and 3, along any sample path, almost surely we have the following:

- (a) The iteration for $\bar{\xi}$ in (73) converges to a state estimator that minimizes the loss function $\mathcal{L}(\bar{\xi})$;
- (b) The iteration for ζ in (73) converges to a critic that minimizes the error with respect to the true Q -function;
- (c) The iteration for θ in (73) converges to a local maximum of the performance $J(\bar{\xi}^*, \zeta^*, \theta)$, where $\bar{\xi}^*$ and ζ^* are the converged values of $\bar{\xi}$ and ζ .

Proof The assumptions satisfy all the four conditions stated in (Leslie, 2004, page 35), (Borkar, 1997, Theorem 23). The proof follows from combining this two-time scale algorithm proof with the fastest third time-scale of learning the state representation. Due to the specific choice of learning rates, the state representation algorithm sees a stationary actor and critic, while the actor and critic in turn see a converged state approximator iteration due to its faster learning rate. The convergence of the state approximator follows from (Borkar, 2008, Theorem 2.2) and the fact that the model satisfies conditions (A1)–(A4) of (Borkar, 2008, pg 10–11). The Martingale difference condition (A3) of Borkar (2008) is satisfied due to the unbiasedness assumption of the state approximator. The result then follows from by combining the theorem given in (Leslie, 2004, page 35), (Borkar, 1997, Theorem 23) along with (Borkar, 2008, Theorem 2.2) and using a third fastest time scale for the state approximator. ■

F. Details about the network architecture, training, and hyperparameters

As explained in Sec. 7, the AIS-generator consists of four components: the history compression function $\hat{\sigma}$, the AIS update function $\hat{\varphi}$, the reward prediction function \hat{r} , and the observation prediction kernel \hat{P}^y . We model the first as an LSTM, where the memory update unit of LSTM acts as $\hat{\varphi}$. We model \hat{r} , \hat{P}^y , and the policy $\hat{\pi}$ as feed-forward neural networks. We describe the details for each difficulty class of environment separately. In the description below, we use $\text{Linear}(n, m)$ to denote a linear layer $\text{Tanh}(n, m)$ to denote a tanh layer, $\text{ReLU}(n, m)$ to denote a ReLU layer, and $\text{LSTM}(n, m)$ to denote an LSTM layer, where n denotes the number of inputs and m denotes the number of outputs of each layer. The size of the input of the outputs depend on the size of the observation and action spaces, which we denote by n_O and n_A , respectively as well as on the dimension of AIS and for the case of minigrid environments, the dimension of the latent space for observations, we denote by $d_{\hat{Z}}$ and d_O . We also use $\text{Conv2d}(IC, OC, (FSx, FSy))$ to denote a 2D convolutional layer with IC , OC , (FSx, FSy) represent the number of input channels, output channels and kernel size (along x and y) respectively. Note that the strides are the same as the kernel size in this case. ELU represents Exponential Linear Unit and is used to model the prediction of variance. Finally, $\text{GMM}(n_{\text{comp}})$ represents a Gaussian Mixture Model with n_{comp} Gaussian components. Most of the details are common for both the AIS+KL and the AIS+MMD cases, we make a distinction whenever they are different by indicating KL or MMD.

F.1 Details for low dimensional environments:

- ENVIRONMENT DETAILS:

Environment	Discount	No. of actions	No. of obs.
	γ	n_A	n_O
VOICEMAIL	0.95	3	2
TIGER	0.95	3	2
CHEESEMAZE	0.7	4	7

The discount factor for CHEESEMAZE is chosen to match with standard value used in that environment (McCallum, 1993).

- AIS AND NETWORK DETAILS:

- Dimensions of AIS ($d_{\hat{Z}}$) : 40
- Weight in AIS loss (λ) (KL) : 0.0001
- Weight in AIS loss (λ) (MMD) : 0.001

$\hat{\sigma}$	\hat{r}	\hat{P}^y	$\hat{\pi}$
$\text{Linear}(n_O + n_A + 1, d_{\hat{Z}})$	$\text{Linear}(n_A + d_{\hat{Z}}, \frac{1}{2}d_{\hat{Z}})$	$\text{Linear}(n_A + d_{\hat{Z}}, \frac{1}{2}d_{\hat{Z}})$	$\text{Linear}(d_{\hat{Z}}, d_{\hat{Z}})$
\Downarrow	\Downarrow	\Downarrow	\Downarrow
$\text{Tanh}(d_{\hat{Z}}, d_{\hat{Z}})$	$\text{Tanh}(\frac{1}{2}d_{\hat{Z}}, \frac{1}{2}d_{\hat{Z}})$	$\text{Tanh}(\frac{1}{2}d_{\hat{Z}}, \frac{1}{2}d_{\hat{Z}})$	$\text{Tanh}(d_{\hat{Z}}, d_{\hat{Z}})$
\Downarrow	\Downarrow	\Downarrow	\Downarrow
$\text{LSTM}(d_{\hat{Z}}, d_{\hat{Z}})$	$\text{Linear}(\frac{1}{2}d_{\hat{Z}}, 1)$	$\text{Linear}(\frac{1}{2}d_{\hat{Z}}, n_O)$	$\text{Linear}(d_{\hat{Z}}, n_A)$
		\Downarrow	\Downarrow
		Softmax	Softmax

- **TRAINING DETAILS:** As explained in Section 6.3, we update the parameters after a rollout of T , which we call a *training batch*. The choice of parameters for the training batch are as follows:

- Samples per training batch : 200
- Number of training batches : 10^5

In addition, we use the following learning rates:

- AIS learning rate : ADAM(0.003)
- Policy learning rate (KL) : ADAM(0.0006)
- Policy learning rate (MMD) : ADAM(0.0008)

In the above description, we use ADAM(α) to denote the choice of α parameter of ADAM. All other parameters have their default value.

- **EVALUATION DETAILS:**

- No. of batches after which evaluation is done : 500
- Number of rollouts per evaluation : 50

F.2 Details for moderate dimensional environments:

- **ENVIRONMENT DETAILS:**

Environment	Discount γ	No. of actions n_A	No. of obs. n_O
DRONE SURVEILLANCE	0.99	5	10
ROCK SAMPLING	0.99	8	3

- **AIS AND NETWORK DETAILS:**

- Dimensions of AIS ($d_{\hat{Z}}$) : 128
- Weight in AIS loss (λ) (KL) : 0.0001
- Weight in AIS loss (λ) (MMD) : 0.001

$\hat{\sigma}$	\hat{r}	\hat{P}^y	$\hat{\pi}$
LSTM($n_O + n_A + 1, d_{\hat{Z}}$)	Linear($n_A + d_{\hat{Z}}, \frac{1}{2}d_{\hat{Z}}$)	Linear($n_A + d_{\hat{Z}}, \frac{1}{2}d_{\hat{Z}}$)	Linear($d_{\hat{Z}}, n_A$)
	\downarrow ReLU($\frac{1}{2}d_{\hat{Z}}, \frac{1}{2}d_{\hat{Z}}$)	\downarrow ReLU($\frac{1}{2}d_{\hat{Z}}, \frac{1}{2}d_{\hat{Z}}$)	\downarrow Softmax
	\downarrow Linear($\frac{1}{2}d_{\hat{Z}}, 1$)	\downarrow Linear($\frac{1}{2}d_{\hat{Z}}, n_O$)	
		\downarrow Softmax	

- **TRAINING DETAILS:** As explained in Section 6.3, we update the parameters after a rollout of T , which we call a *training batch*. The choice of parameters for the training batch are as follows:

- Samples per training batch : 200
- Number of training batches : 10^5

In addition, we use the following learning rates:

- AIS learning rate : ADAM(0.003)
- Policy learning rate : ADAM(0.0007)

In the above description, we use $\text{ADAM}(\alpha)$ to denote the choice of α parameter of ADAM. All other parameters have their default value.

- EVALUATION DETAILS:

- No. of batches after which evaluation is done : 500
- Number of rollouts per evaluation : 100

F.3 Details for high dimensional environments:

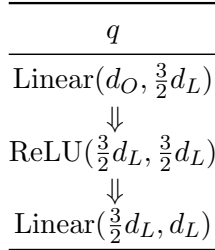
- ENVIRONMENT DETAILS:

Note that here n_O represents the number of possible observations that a general minigrid environment can have. With the actual rules of the environment plugged in, this number is smaller since some combinations of the encoded observation are not possible. The actual input that we get from the environment is a vector of size 147 (d_O) which is basically an observation grid of 7×7 with 3 channels containing characteristic information about the observation.

Environment	Discount	No. of actions	No. of obs.	Obs. dimen.
	γ	n_A	n_O	d_O
MINIGRID ENVS	0.99	7	$(6 \times 11 \times 3)^{7 \times 7}$	$7 \times 7 \times 3$

- AUTOENCODER (q) DETAILS:

- Latent space dimensions (d_L) : 64
- Type of autoencoder used : Basic autoencoder
- Reconstruction Loss Criterion Used : Mean Square Error



- AIS AND NETWORK DETAILS:

- Dimensions of AIS ($d_{\hat{Z}}$) : 128
- Weight in AIS loss (λ) : 0.1
- Number of GMM components used (n_{comp}) (only for KL) : 5

$\hat{\sigma}$	\hat{r}	\hat{P}^y	$\hat{\pi}$
LSTM($d_L + n_A + 1, d_{\hat{Z}}$)	Linear($n_A + d_{\hat{Z}}, \frac{1}{2}d_{\hat{Z}}$)	Linear($n_A + d_{\hat{Z}}, \frac{1}{2}d_{\hat{Z}}$)	Linear($d_{\hat{Z}}, d_{\hat{Z}}$)
	\Downarrow	\Downarrow	\Downarrow
	ReLU($\frac{1}{2}d_{\hat{Z}}, \frac{1}{2}d_{\hat{Z}}$)	ReLU($\frac{1}{2}d_{\hat{Z}}, \frac{1}{2}d_{\hat{Z}}$)	ReLU($d_{\hat{Z}}, d_{\hat{Z}}$)
	\Downarrow	\Downarrow	\Downarrow
	Linear($\frac{1}{2}d_{\hat{Z}}, 1$)	Linear($\frac{1}{2}d_{\hat{Z}}, d_L$)	Linear($d_{\hat{Z}}, n_A$)
			\Downarrow
			Softmax

For KL, \hat{P}^y is replaced by the following while other networks remain the same:

\hat{P}^y		
Linear($n_A + d_{\hat{Z}}, \frac{1}{2}d_{\hat{Z}}$)		
\Downarrow		
ReLU($\frac{1}{2}d_{\hat{Z}}, \frac{1}{2}d_{\hat{Z}}$)		
\Downarrow		
Linear($\frac{1}{2}d_{\hat{Z}}, d_L n_{\text{comp}}$) \swarrow	ELU(Linear($\frac{1}{2}d_{\hat{Z}}, d_L n_{\text{comp}}$)) + 1 + 10^{-6}	Softmax(Linear($\frac{1}{2}d_{\hat{Z}}, n_{\text{comp}}$)) \searrow
\searrow	\Downarrow	\swarrow
GMM(n_{comp})		

Note that the third layer generates the mean vector of each component, the diagonal vector for variance of each component and the mixture weights of each component of the GMM model in the last layer.

- **TRAINING DETAILS:** As explained in Section 6.3, we update the parameters after a rollout of T , which we call a *training batch*. The choice of parameters for the training batch are as follows:
 - Samples per training batch : 200
 - Number of training batches : 2×10^5 (MGKCS3R3, MGOM1Dl, MGOM1Dlh)
 10^5 (others)

In addition, we use the following learning rates:

- AIS learning rate : ADAM(0.001)
- Policy learning rate : ADAM(0.0007)

In the above description, we use ADAM(α) to denote the choice of α parameter of ADAM. All other parameters have their default value.

- **EVALUATION DETAILS:**
 - No. of batches after which evaluation is done : 5000 (MGKCS3R3, MGOM1Dl, MGOM1Dlh)
 1000 (others)
 - Number of rollouts per evaluation : 20

F.4 Details for PPO with LSTM and Critic:

- **ENVIRONMENT DETAILS:**

The environment details are the same as mentioned previously.

- NETWORK DETAILS:

- Low and moderate dimensionality environments:

Feature Extractor	Actor Head	Critic Head
LSTM(n_O, n_O)	Linear($n_O, 64$)	Linear($n_O, 64$)
	\Downarrow	\Downarrow
	Tanh(64, 64)	Tanh(64, 64)
	\Downarrow	\Downarrow
	Linear(64, n_A)	Linear(64, 1)
	\Downarrow	
	Softmax	

- High dimensionality environments:

- Observation tensor : $7 \times 7 \times 3$
- Embedding size (d_E) : 64

Conv. Feature Extractor	Actor Head	Critic Head
Conv2d($3, \frac{1}{4}d_E, (2, 2)$)	Linear(d_E, d_E)	Linear(d_E, d_E)
\Downarrow	\Downarrow	\Downarrow
ReLU	Tanh(d_E, d_E)	Tanh(d_E, d_E)
\Downarrow	\Downarrow	\Downarrow
MaxPool2d	Linear(d_E, n_A)	Linear($d_E, 1$)
\Downarrow	\Downarrow	
Conv2d($\frac{1}{4}d_E, \frac{1}{2}d_E, (2, 2)$)	Softmax	
\Downarrow		
ReLU		
\Downarrow		
Conv2d($\frac{1}{2}d_E, d_E, (2, 2)$)		
\Downarrow		
ReLU		
\Downarrow		
LSTM(d_E, d_E)		

- TRAINING DETAILS:

- Number of parallel actors : 64
- Number of training batches : 4×10^7 (MGKCS3R3, MGOM1Dl, MGOM1Dlh)
 2×10^7 (others)
- Epochs per training batch : 4
- Samples per training batch : 1280
- Frames per parallel actor : 40
- GAE (λ_{GAE}) : 0.99
- Trajectory recurrence length : 20

In addition, we use ADAM with the following details:

- Learning rate α : 0.0001
- ADAM parameter ϵ : 0.00001

- EVALUATION DETAILS:
 - No. of batches after which evaluation is done : 200
 - Rollouts used for evaluation : All recent episodes completed by all actors

F.5 Details about hyperparameter tuning

Hyperparameter tuning was carried out by searching a grid of values, but exhaustive grid search was not carried out due to the prohibitive computational cost. Instead, coarse values were used initially as starting points and finer tuning was done around promising values, which was essentially an iterative process of performing experiments, observing results and trying similar parameters to the ones generating good results. Hyperparameters observed in each previous environment class (low, moderate, high dimensionality) were used as a starting point for the search in the new environment class.

Performance was quite sensitive to different learning rates used for the AIS and policy in most environments. Performance generally improved or remained the same when a larger AIS State Size was used (values considered were 128, 256, 512 for moderate/high-dimensional environments and 5, 10, 20, 40 for low-dimensional environments), although in some cases, it was more unstable during training. λ values considered were between 0 and 1 and generally only made a difference (in terms of performance results) when the rewards were very large. The choice of activation function between ReLU and Tanh did not seem to make a significant difference for the considered environments.