Partially Observable Reference Policy Programming*

Edward Kim and **Hanna Kurniawati**
Australian National University
{edward.kim, hanna.kurniawati}@anu.edu.au

Abstract

This paper proposes Partially Observable Reference Policy Programming, a novel anytime online approximate POMDP solver which samples meaningful future histories very deeply while simultaneously forcing a gradual policy update. We provide theoretical guarantees for the algorithm's underlying scheme which say that the performance loss is bounded by the *average* of the sampling approximation errors rather than the usual maximum; a crucial requirement given the sampling sparsity of online planning. Empirical evaluations on two large-scale problems with dynamically evolving environments—including a helicopter emergency scenario in the Corsica region requiring approximately 150 planning steps—corroborate the theoretical results and indicate that our solver considerably outperforms current online benchmarks.

1 Introduction

Planning under uncertainty in non-deterministic and partially observable scenarios is critical for many (semi-)autonomous systems. Such problems can be systematically framed as a Partially Observable Markov Decision Process (POMDP) [Kaebling *et al.*, 1998]. Although solving infinite-horizon POMDPs in the worst case is undecidable [Madani *et al.*, 2003], the past decade has seen tremendous advances in the practicality of approximate POMDP solvers [Kurniawati, 2022]. Most of these solvers are online sampling-based methods that numerically compute estimates of the expected total reward of performing different actions before optimising over these estimates. Due to difficulties in computing gradients, such solvers *exhaustively enumerate* over the entire action space, which massively hinders fast computation of a close-to-optimal solution for problems with large action spaces and long horizons. This problem is even worse when the environment is also dynamically changing at each execution step.

The core difficulty is the curse of history where the set of possible futures branches by the size of the action space and grows exponentially with respect to the horizon. Most existing methods try to abstract the problem into a simpler one by either reducing the size of the action space [Wang *et al.*, 2018] or relying on *macro actions*—i.e. a set of open-loop action sequences—to reduce the planning horizon [Theocharous and Kaelbling, 2003; He *et al.*, 2010; Kurniawati *et al.*, 2011; Flaspohler *et al.*, 2020; Lee *et al.*, 2021]. Still, the fundamental problem—i.e. exhaustive action enumeration—remains.

Recently, [Kim *et al.*, 2023; Liang *et al.*, 2024] have softened this requirement by introducing the notion of a Reference-Based POMDP (RBPOMDP) which is a reformulation of a POMDP whose objective is penalised by the Kullback-Leibler (KL) divergence between a chosen and nominal *reference policy*. As such, a solution can be viewed as a KL-regularised improvement of the reference policy. The form of objective allows analytical action optimisation so that the value can be approximated by estimating expectations under the reference policy. This property accommodates solvers that have been shown to perform effectively on certain long-horizon tasks. However, the RBPOMDP formulation comes at the cost that the solution has a baked-in commitment to the reference policy. In general, it is unclear a priori which reference policies yield near optimal policies for the original POMDP of interest, so the performance of the computed solution is vulnerable to mis-specification.

The aim of this paper is to build on the advantages of the RBPOMDP framework while, in tandem, bolstering any vulnerabilities to mis-specification. To this end, our contribution is an exact iterative scheme (Sect. 3.2) whose successive policies can be viewed as solutions of a sequence of RBPOMDPs—i.e. KL-constrained policy improvements. Theoretical analysis shows that the performance loss of the exact scheme is bounded by the *average* of the sampling errors, which means the algorithm is less sensitive to large approximation errors (Theorem 1). We also contribute an explicit approximate scheme (Sect. 3.3) and provide a POMDP-specialised high-probability bound for the performance loss (Theorem 2). Finally, the scheme is practically implemented in our proposed algorithm Partially Observable Reference Policy Programming (PORPP)—an anytime online POMDP-solver—and tested on two non-trivial long-horizon POMDPs, one of which has a dynamically evolving environment. Experimental results indicate that PORPP substantially outperforms current state-of-the-art online POMDP benchmarks.

*Technical details and proofs are contained in the Supplementary Material (https://github.com/RDLLab/pomdp-py-porpp).

2 Background and Related Work

2.1 POMDP Preliminaries

An infinite-horizon POMDP is defined as the tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{Z}, \mathcal{T}, R, \gamma, b_0 \rangle$ where the sets of all possible agent states, actions and observations are denoted by $\mathcal{S}$, $\mathcal{A}$ and $\mathcal{O}$ respectively. For clarity, our presentation is for countable sets. The *transition model* $\mathcal{T}$ is such that $\mathcal{T}(s' \,|\, a, s)$ is the conditional probability that $s' \in \mathcal{S}$ occurs after performing $a \in \mathcal{A}$ from $s \in \mathcal{S}$. The *observation model* $\mathcal{Z}$ is such that $\mathcal{Z}(o \,|\, s', a)$ is the conditional probability that the agent perceives $o \in \mathcal{O}$ when it is in state $s' \in \mathcal{S}$ after performing $a \in \mathcal{A}$. The *reward model* is a real-valued function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ such that $|R(s,a)| \leq R_{\max} < \infty$ for all s, a and $\gamma \in (0,1)$ is the discount factor.

The agent does not know the true state, but it maintains a *belief* about its state—i.e. a probability distribution b on the space $\mathcal{S}$. Let $\mathcal{B}$ be the set of all such distributions. Starting with a given initial belief b_0, the agent's next belief b' after taking the action a and perceiving some observation o is given by $b'(s') \propto \mathcal{Z}(o \,|\, s', a) \sum_{s \in \mathcal{S}} \mathcal{T}(s' \,|\, a, s) b(s)$ and we write $b' = \tau(b, a, o)$ with the *belief update operator* τ. We denote the set of *reachable beliefs* by $\mathcal{R}_{b_0} \subset \mathcal{B}$; i.e. the set of beliefs reachable from b_0 under some policy. For any given belief b and action a the expected immediate reward is given by $R(b,a) := \sum_{\mathcal{S}} R(s,a) b(s)$. The probability that the agent perceives an observation $o \in \mathcal{O}$ having performed the action $a \in \mathcal{A}$ under the belief b is given by

$$P(o \,|\, a, b) := \sum_{s' \in \mathcal{S}} \mathcal{Z}(o \,|\, s', a) \sum_{s \in \mathcal{S}} \mathcal{T}(s' \,|\, a, s) b(s). \tag{1}$$

A (stochastic) *policy* is a mapping $\pi : \mathcal{R}_{b_0} \to \Delta(\mathcal{A})$. We denote its distribution for any given input $b \in \mathcal{R}_{b_0}$ by $\pi(\cdot \,|\, b)$. Let Π be the class of all stochastic policies. For any $(b,a) \in \mathcal{R}_{b_0} \times \mathcal{A}$, define the *action-value function* $Q^\pi : \mathcal{R}_{b_0} \times \mathcal{A} \to \mathbb{R}$ recursively according to

$$\begin{aligned} Q^\pi(b,a) &= R(b,a) \\ &\quad + \gamma \sum_{a', o} Q^\pi\big(\tau(b,a,o), a'\big) P(o \,|\, a', b)\, \pi(a' \,|\, b). \end{aligned} \tag{2}$$

Given the reward is uniformly bounded, for any policy $\pi \in \Pi$, we have $|Q^\pi(b,a)| \leq Q_{\max} := R_{\max}/(1-\gamma)$ for all pairs $(b,a) \in \mathcal{R}_{b_0} \times \mathcal{A}$. A *solution* to the POMDP is a policy $\pi^* \in \Pi$ satisfying $Q^*(b,a) := \sup_{\pi \in \Pi} Q^\pi(b,a) = Q^{\pi^*}(b,a)$ for all $(b,a) \in \mathcal{R}_{b_0} \times \mathcal{A}$.

2.2 POMDP Packing and Covering Numbers

For a Markov Decision Process (MDP) with finite state and action spaces, the usual input for complexity is the set cardinality $|\mathcal{S}||\mathcal{A}|$ where it is generally assumed that the spaces are finite. However, for the POMDP, the reachable belief space $\mathcal{R}_{b_0}$ is an uncountable subset even if $\mathcal{S}$ is finite so the notion of set cardinality is no longer a sensible complexity input. A more reasonable approach is to choose a metric in $\mathbb{R}^{|\mathcal{S}|}$, and estimate a "finite volume" of $\mathcal{R}_{b_0}$ via the dual concepts of a *δ-packing* or *δ-covering number*. While these are theoretical quantities, they can be explicitly computed in certain cases and highlight key properties relating to the POMDP's complexity [Lee *et al.*, 2007].

The interested reader can refer to Sect. 1 of the Supplementary Material for a more thorough review of their formal definitions and properties. In words, the *δ-covering number* $\mathcal{C}_\delta(\mathcal{R}_{b_0})$ is the minimum number of balls of radius δ needed to cover the set $\mathcal{R}_{b_0}$. If in addition, all the centres of the balls are required to belong to $\mathcal{R}_{b_0}$ then we call such a number the *internal δ-covering number* and denote it by $\mathcal{C}^\circ_\delta(\mathcal{R}_{b_0})$. The *$\delta$-packing number* $\mathcal{P}_\delta(\mathcal{R}_{b_0})$ is the maximum number of points that can be packed inside $\mathcal{R}_{b_0}$ such that all points are at least δ distance apart. The concepts are closely related and, importantly, are finite if and only if $\mathcal{R}_{b_0}$ is *totally bounded* (see Remark 1 in Supplementary Material). For instance, it suffices to assume that $\mathcal{S}$ is finite. To ensure the δ-covering number is always finite, we will make the following standing assumption for the remainder of this paper.

Assumption 1. *The reachable belief space $\mathcal{R}_{b_0}$ is totally bounded.*

2.3 KL-Penalisation and POMDPs

The idea of using KL-penalisation in fully observable MDPs started with a series of works on *Linearly Solvable MDPs* [Todorov, 2006; Todorov, 2009a; Todorov, 2009b; Dvijotham and Todorov, 2012]. The main idea is to find a control conditional distribution $p(s' \,|\, s)$ to a stochastic control problem where the control cost increases with the relative entropy between $p(\cdot \,|\, s)$ and some benchmark $\bar{p}(\cdot \,|\, s)$. The formulation results in a Bellman backup which can be optimised analytically and yields efficient methods to solve a special class of fully-controllable MDPs.

These works were reformulated over stochastic actions by [Rawlik *et al.*, 2012] and related to general MDPs by [Azar *et al.*, 2011; Azar *et al.*, 2012] who introduced *Dynamic Policy Programming*. This can be interpreted as a policy iteration scheme where each iterate π_k is a solution to a specialised MDP whose reward decreases with the relative entropy $\mathrm{KL}(\pi_{k+1} \,\|\, \pi_k)$. The scheme can be shown to converge to the solution of the MDP; indeed, the gradual update forced by the KL-penalty yields performance bounds which depend on the *average* accumulated error as opposed to the usual maximum, suggesting robustness to approximation errors.

The extension of the idea of KL-penalised MDPs to POMDPs was provided by [Kim *et al.*, 2023] who introduced the concept of a *Reference-Based POMDP* (RBPOMDP). In essence, the formulation can be viewed as a *Belief-MDP* [Kaebling *et al.*, 1998] with policies $U(b' \,|\, b)$ that control transitions between POMDP beliefs where the reward is penalised by the relative entropy $\mathrm{KL}\left(U(\cdot \,|\, b) \,\|\, \bar{U}(\cdot \,|\, b)\right)$ for some *reference policy* $\bar{U}$. Their empirical results suggest that approximate solvers for RBPOMDPs can outperform state-of-the-art benchmarks on large POMDPs for certain choices of $\bar{U}(\cdot \,|\, b)$. However, the authors did not provide a systematic procedure to determine this choice. This current work addresses this gap by providing a systematic procedure, in a similar vein to [Azar *et al.*, 2012],

3 PORPP

PORPP is an anytime online POMDP solver which approximates the solution of a policy iteration scheme whose successive policies are forced to be close to each other. Specifically, each policy iterate is a solution to a RBPOMDP over stochastic actions whose reference policy is the previous policy in the sequence and can therefore be viewed as a KL-constrained policy improvement. While this procedure converges more slowly, its advantage is that it yields a performance bound which is given by the *average* of approximation errors, suggesting that it is less prone to over-commitment—a useful feature given the scarcity of samples generated by an online planner. In what follows, $\|\cdot\|_\infty$ denotes the usual *supremum-norm* for bounded functions.

3.1 RBPOMDPs over Stochastic Actions

In [Kim *et al.*, 2023], the reliance on *belief-to-belief transitions* $U(b' \,|\, b)$ implicitly allows the agent to control the choice of observation, which may not be valid in general. We will consider a more natural formulation over *stochastic actions* which will form the building blocks for the required systematic procedure. Namely, a ***RBPOMDP over stochastic actions*** is a tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{T}, \mathcal{Z}, R, \gamma, \eta, \pi_0, b_0 \rangle$. Its value $\mathcal{V}$, for a given $b \in \mathcal{R}_{b_0}$, is specified by the recursive equation

$$\mathcal{V}(b) = \sup_{\pi \in \Pi} \Big[\sum_{a \in \mathcal{A}} R(b,a)\pi(a \,|\, b) - \frac{1}{\eta} \mathrm{KL}(\pi \,\|\, \pi_0) + \gamma \sum_{a,o} P(o \,|\, a, b)\pi(a \,|\, b)\mathcal{V}\big(\tau(b,a,o)\big) \Big]. \quad (3)$$

Intuitively its solution is a stochastic policy that tries to respect the reference policy π_0 unless deviating substantially leads to greater rewards where the trade-off is balanced by the temperature parameter $\eta > 0$. The right-hand-side can be optimised analytically so that (3) is equivalent to

$$\mathcal{V}(b) = \frac{1}{\eta} \log \Big[\sum_{a \in \mathcal{A}} \pi_0(a \,|\, b) \exp \Big\{ \eta \big[R(b,a) + \gamma \sum_o P(o \,|\, a, b)\mathcal{V}\big(\tau(b,a,o)\big) \big] \Big\} \Big]. \quad (4)$$

In fact, we can represent the Bellman equation (4) in a slightly different way by introducing *preferences* Ψ over belief-action pairs. More specifically, let

$$\Psi(b,a) := \frac{1}{\eta} \log \big(\pi_0(a \,|\, b)\big) + R(b,a) + \gamma \sum_o P(o \,|\, a, b)\mathcal{V}\big(\tau(b,a,o)\big). \quad (5)$$

This yields $\mathcal{V}(b) = \frac{1}{\eta} \log \Big[\sum_a \exp[\eta \Psi(b,a)] \Big] =: [\mathcal{L}_\eta \Psi](b)$ where $\mathcal{L}_\eta$ is the *log-sum-exp* operator [Blanchard *et al.*, 2021; Asadi and Littman, 2017] and eq. (4) stated with respect to preferences becomes

$$\Psi(b,a) = \frac{1}{\eta} \log[\pi_0(a \,|\, b)] + R(b,a) + \gamma \sum_o P(o \,|\, a, b)[\mathcal{L}_\eta \Psi]\big(\tau(b,a,o)\big). \quad (6)$$

If Ψ^* satisfies (6), the solution of the RBPOMDP is

$$\pi^*(a \,|\, b) = \frac{\exp[\eta \Psi^*(b,a)]}{\sum_{a'} \exp[\eta \Psi^*(b,a')]}. \quad (7)$$

being the exact maximiser of (3).

3.2 Exact Scheme

We are now in a position to describe the exact iterative scheme that relates the RBPOMDP to that of the standard POMDP. Taking inspiration from (7), the scheme implicitly represents a reference policy π_k by maintaining *action preferences* $\Psi_k : \mathcal{R}_{b_0} \times \mathcal{A} \to \mathbb{R}$ according to the equation

$$\pi_k(a \,|\, b) := \frac{\exp[\eta \Psi_k(b,a)]}{\sum_{a'} \exp[\eta \Psi_k(b,a')]}. \quad (8)$$

The policy is then updated *gradually* by asserting that π_{k+1} is the solution to a RBPOMDP whose reference policy is π_k. That is,

$$\begin{aligned} \Psi_{k+1}(b,a) &= \frac{1}{\eta} \log[\pi_k(a \,|\, b)] + R(b,a) \\ &\quad + \gamma \sum_o P(o \,|\, a, b)[\mathcal{L}_\eta \Psi_k]\big(\tau(b,a,o)\big) \\ &= \Psi_k(b,a) - [\mathcal{L}_\eta \Psi_k](b) + R(b,a) \\ &\quad + \gamma \sum_o P(o \,|\, a, b)[\mathcal{L}_\eta \Psi_k]\big(\tau(b,a,o)\big) \\ &=: [\mathscr{L}_\eta \Psi_k](b,a). \end{aligned} \quad (9)$$

The exact scheme indeed converges to the action-value Q^* of the POMDP. To show this, let $\mathscr{L}_\eta$ be the exact function operator defined by (9) and consider a sequence of *approximate preferences* $(\hat{\Psi}_k)_{k \geq 0}$ such that $\hat{\Psi}_{k+1} \approx \mathscr{L}_\eta \hat{\Psi}_k$. For arbitrary $(b,a) \in B \times \mathcal{A}$, let

$$\epsilon_k(b,a) := \begin{cases} \hat{\Psi}_k(b,a) - [\mathscr{L}_\eta \hat{\Psi}_{k-1}](b,a) & \text{if } k \geq 1 \\ 0 & \text{if } k = 0 \end{cases} \quad (10)$$

and $E_k(b,a) := \sum_{j=0}^k \epsilon_j(b,a)$ and define the approximating policy to be

$$\hat{\pi}_k(a \,|\, b) := \frac{\exp[\eta \hat{\Psi}_k(b,a)]}{\sum_{a'} \exp[\eta \hat{\Psi}_k(b,a')]}. \quad (11)$$

We have the following general error bound which says that the total error is bounded by the *average* of approximation errors at each iteration. Since the exact scheme has $E_k = 0$ for all k, the result also validates the asymptotic convergence of the exact scheme.

Theorem 1. *Suppose* $\|\hat{\Psi}_0\|_\infty \leq Q_{\max}$. *Then*

$$\|Q^* - Q^{\hat{\pi}_k}\|_\infty \leq \frac{2}{(1-\gamma)(k+1)} \Big[\frac{\gamma(4Q_{\max} + \frac{\log(|\mathcal{A}|)}{\eta})}{(1-\gamma)} + \sum_{j=0}^k \gamma^{k-j} \|E_j\|_\infty \Big]. \quad (12)$$

Proof. See Supplementary Material. □

3.3 Explicit Sampling-Based Approximate Scheme

We will now introduce explicit synchronous and asynchronous sampling-based approximate schemes and prove their asymptotic optimality. In both cases, we prove specialised bounds with respect to the POMDP's δ-covering numbers. The asynchronous scheme is especially important, as it forms the basis for the design of our online planning algorithm.

We will need some setting up to introduce the sampling-based scheme that approximates (9). Let ρ be a metric on $\mathcal{B}$ and B be some well-ordered[1] subset of $\mathcal{B}$. Let $\tilde{\tau}_{B,\rho} : \mathcal{B} \times \mathcal{A} \times \mathcal{O} \to B$ be the mapping which takes an arbitrary belief $b \in \mathcal{B}$ to the least element of

$$\arg\min_{b' \in B} \rho\big(b', \tau(b, a, o)\big). \tag{13}$$

Intuitively, $\tilde{\tau}_{B,\rho}$ finds the set of points in B nearest to $\tau(b, a, o)$ (it is not necessarily a singleton set) and has a rule to break ties so that the mapping is well-defined.

Let $Q^{\pi}_{B,\rho} : \mathcal{R}_{b_0} \times \mathcal{A} \to \mathbb{R}$ be the *action-value approximation* on any subset $B \subset \mathcal{R}_{b_0}$ which is the unique solution to the recursion

$$\begin{aligned} Q^{\pi}(b, a) &= R(b, a) \\ &+ \gamma \sum_{a', o} Q^{\pi}\big(\tilde{\tau}_{B,\rho}(b, a, o), a'\big) P(o \,|\, a', b)\, \pi(a' \,|\, b). \end{aligned} \tag{14}$$

The difference between (2) and (14) is that the next belief is forced to a nearest belief in $B \subset \mathcal{R}_{b_0}$ in the latter, whereas the belief update for the former is the natural one. As such, we expect the two quantities to differ according to the precision of B in approximating $\mathcal{R}_{b_0}$. In fact, it can be shown that if B is a δ-covering of $\mathcal{R}_{b_0}$ the approximation becomes negligible for the optimal policy π^* as $\delta \downarrow 0$ (see Proposition 3 in the Supplementary Material).

It is clear from (14) that it suffices to evaluate $Q^{\pi}_{B,\rho}$ on the subset $B \times \mathcal{A}$. The *synchronous* scheme therefore updates action preference approximations according to the rule

$$\begin{aligned} \hat{\Psi}_{k+1}(b, a) := \hat{\Psi}_k(b, a) - [\mathcal{L}_\eta \hat{\Psi}_k](b) + \sum_{i=1}^{N_k(b,a)} \frac{R(s_i, a)}{N_k(b, a)} \\ + \gamma \sum_{j=1}^{M_k(b,a)} \frac{[\mathcal{L}_\eta \hat{\Psi}_k]\big(\tilde{\tau}_{B,\rho}(b, a, o_j)\big)}{M_k(b, a)} \end{aligned} \tag{15}$$

for all $(b, a) \in B \times \mathcal{A}$ where $s_i \sim b$ and $o_j \sim P(\cdot \,|\, a, b)$ and generic increasing sequences N_k and M_k having the property that $N_k(b, a) \uparrow \infty$, $M_k(b, a) \uparrow \infty$ as $k \uparrow \infty$. The scheme is *synchronous* in the sense that, at each step k, it samples $\{s_{N_{k-1}(b,a)+1}, \ldots, s_{N_k(b,a)}, o_{M_{k-1}(b,a)+1}, \ldots, o_{M_k(b,a)}\}$ *for each* (b, a) and updates the action preferences according to (15). The approximate stochastic policy $\hat{\pi}_k$ is then fully specified by the approximate preferences according to (11).

The synchronous scheme yields the following high-probability bound when B is an internal δ-covering $\mathcal{E}_\delta$ of $\mathcal{R}_{b_0}$ for the metric ρ_1 induced by the 1-norm—i.e. $\rho_1(x, y) := \|x - y\|_1$ for $x, y \in \mathbb{R}^{|\mathcal{S}|-1}$.[2]

Theorem 2. *Let $\mathcal{C}^{\circ}_{\delta} = |\mathcal{E}_\delta|$ be the internal δ-covering number of $\mathcal{R}_{b_0}$ for a given $\delta > 0$. If $\|\hat{\Psi}_0\| \leq Q_{\max}$ then, for any $\alpha \in (0, 1)$, we have with probability at least $1 - \alpha$*

$$\|Q^* - Q^{\hat{\pi}_k}_{\mathcal{E}_\delta, \rho_1}\|_\infty \leq \frac{K_1}{k+1} + \frac{K_2}{\sqrt{k+1}} + \frac{\gamma \delta Q_{\max}}{1-\gamma} \tag{16}$$

where

$$K_1 := \frac{2\gamma}{(1-\gamma)^2}\big[\log(|\mathcal{A}|)/\eta + 4Q_{\max}\big] \tag{17}$$

and

$$K_2 := \Big[\frac{4\gamma \log(|\mathcal{A}|)}{\eta(1-\gamma)^3} + \frac{2Q_{\max}}{1-\gamma}\Big]\sqrt{2 \log\Big\{\frac{2|\mathcal{A}|\mathcal{C}^{\circ}_{\delta}}{\alpha}\Big\}}. \tag{18}$$

Proof. See Supplementary Material. □

Although the precision of the bound gets more precise after every synchronous update, the error can still be large if the covering $\mathcal{E}_\delta$ is not a good representation of $\mathcal{R}_{b_0}$—i.e. δ is large. In general, $\mathcal{E}_\delta$ may be required to be extremely large and performing even one synchronous update can be an exorbitantly expensive task.

To mitigate this fundamental problem, PORPP employs a heuristic action sampler $\tilde{\pi}$ to bias towards a selection of promising beliefs and asynchronously updates preference approximations on the selection. The underpinning assumption for optimality of this procedure is that the selection grows to include the set of beliefs reachable under the optimal policy π^*—which is not known a priori—while simultaneously being small enough to be tractable for online planning.

More precisely, let $\mathcal{E}_\delta$ be an internal δ-covering of $\mathcal{R}_{b_0}$ and let $\Omega_k := \big((b_1, a_1), (b_2, a_2), \ldots, (b_k, a_k)\big)$ be the sequence of pairs in $\mathcal{E}_\delta \times \mathcal{A}$ traversed by $\tilde{\pi}$ after k steps. Then, by definition, our *asynchronous* scheme updates action preference approximations according to

$$\begin{aligned} \hat{\Psi}_{k+1}(b_k, a_k) := \hat{\Psi}_k(b_k, a_k) \\ - [\mathcal{L}_\eta \hat{\Psi}_k](b_k) + \sum_{i=1}^{N(b_k,a_k)} \frac{R(s_i, a_k)}{N(b_k, a_k)} \\ + \gamma \sum_{j=1}^{N(b_k,a_k)} \frac{[\mathcal{L}_\eta \hat{\Psi}_k]\big(\tilde{\tau}_{\mathcal{E}_\delta,\rho_1}(b_k, a_k, o_j)\big)}{N(b_k, a_k)} \end{aligned} \tag{19}$$

where $s_i \sim b_k$ and $o_j \sim P(\cdot \,|\, a_k, b_k)$ and $N(b_k, a_k)$ is the number of times $\tilde{\pi}$ has visited (b_k, a_k). Let $\mathcal{R}^*_{b_0}$ be the set of beliefs reachable under the optimal policy π^* of the POMDP and denote by κ_k the number of times that $\tilde{\pi}$ has visited Ω_∞ after k steps. Then, provided $\tilde{\pi}$ traverses Ω_∞ infinitely often and $\{b : (b, a) \in \Omega_\infty\} \supset \mathcal{R}^*_{b_0} \cap \mathcal{E}_\delta$, the bounds of Theorem 2 hold with (16) replaced by

$$\|Q^* - Q^{\hat{\pi}_k}_{\mathcal{E}_\delta, \rho_1}\|_\infty \leq \frac{K_1}{\kappa_k + 1} + \frac{K_2}{\sqrt{\kappa_k + 1}} + \frac{\gamma \delta Q_{\max}}{1-\gamma} \tag{20}$$

[1] It suffices for B to be finite.

[2] Note that the Euclidean space under consideration can, in theory, be infinite-dimensional under Assumption 1.

Algorithm 1 PORPP

Input: Root node h_0 of T equipped with belief particles b
Output: T

1: **while** steps remaining **do**
2: **while** time permitting **do**
3: Sample belief particle s from h_0
4: SIMULATE$(h_0, s, 0)$
5: **end while**
6: $\mathbf{a} \leftarrow \arg\max_{\mathbf{a} \in \text{children}(h_0)} \hat{\Psi}(h_0\mathbf{a})$
7: Execute macro action $\mathbf{a}$ in environment
8: Receive macro observation $\mathbf{o}$ from environment
9: Update history $h_0 \leftarrow h_0\mathbf{ao}$
10: Resample new state particles and add to h_0
11: **end while**

for Q^* and $Q^{\hat{\pi}_k}_{\mathcal{E}_\delta, \rho_1}$ being functions defined on Ω_∞ and $\mathcal{C}^\circ_\delta$ now being the δ-covering number of Ω_∞. As such, we would like to ensure that Ω_∞ is as small as possible without compromising optimality.

3.4 Algorithm: PORPP

We propose Partially Observable Reference Policy Programming (PORPP), a specific online implementation of the asynchronous scheme discussed above. PORPP represents beliefs as nodes h in a tree where each node is associated with a history of action-observation pairs and maintains a belief estimate by progressively sampling a richer set of state particles at each node. With enough time, the planner grows a rich tree (i.e. δ small) and improves preference estimates by sampling sequences of action-observation histories up to a required depth $D_{\max}$ and backing up estimates according to the sampling-based scheme.

Specifically, at each history, PORPP's heuristic action sampler SAMPLECANDIDATEACTION(h, s) uses domain specific knowledge about the problem to propose a (macro) action $\mathbf{a}$—i.e. a sequence of primitive actions—to add to the tree. The aim of the sampler is to sample actions that cover the optimal policy while avoiding counterproductive ones—see Sec. 3.6 for examples. The action is added to the tree if it has not been already and the progressive widening threshold $\kappa_{\mathcal{A}} N(h)^{\alpha_{\mathcal{A}}}$ (e.g. [Sunberg and Kochenderfer, 2018]) has not been exceeded. PORPP then selects an action by sampling the softmax distribution given by the current preferences—cf. (11)—before sampling a new state s', (macro) observation $\mathbf{o}$ and (macro) reward $r(s, \mathbf{a}; \gamma)$ using a generative model. The observation is then added to the tree, and the procedure continues recursively until the depth exceeds $D_{\max}$. At this point the value is estimated from the sampled state using a value heuristic and the information is propagated back up to the root node via (19) (lines 18 to 23 in Algorithm 2).

This planning procedure continues until timeout (lines 2 to 5 in Algorithm 1) after which the algorithm executes the action with the best sampled preference in the environment. Upon receiving an observation, particles that are consistent with the realised action-observation pair are resampled and added to the associated node (line 10 in Algorithm 1). This planning-execution loop continues until a step budget is reached, at which point the algorithm terminates.

Algorithm 2 SIMULATE(h, s, depth)

Parameters: $\kappa_{\mathcal{A}} \geq 0, \alpha_{\mathcal{A}} \in (0, 1), D_{\max} \geq 1, \eta > 0$.

1: **if** depth $> D_{\max}$ **then**
2: **return** VALUEHEURISTIC(h, s)
3: **end if**
4: **if** depth > 0 **then**
5: $b(h) \leftarrow b(h) \cup \{s\}$
6: **end if**
7: $N(h) \leftarrow N(h) + 1$
8: **if** $|\text{children}(h)| < \kappa_{\mathcal{A}} N(h)^{\alpha_{\mathcal{A}}}$ **then**
9: $\mathbf{a} \leftarrow$ SAMPLECANDIDATEACTION(h, s)
10: **if** $h\mathbf{a} \notin T$ **then**
11: Add $h\mathbf{a}$ to T
12: **end if**
13: **end if**
14: $\mathbf{a} \leftarrow$ SAMPLEPREFSOFTMAX$(h; \eta)$
15: Resample s from $b(h)$
16: Sample $(s', \mathbf{o}, r(s, \mathbf{a}; \gamma))$ from gen. model $\mathcal{G}(s, \mathbf{a})$
17: Create nodes for $h\mathbf{ao}$ if not created already
18: $N(h\mathbf{a}) \leftarrow N(h\mathbf{a}) + 1$
19: $R(h\mathbf{a}) \leftarrow R(h\mathbf{a}) + \frac{r(s, \mathbf{a}; \gamma) - R(h\mathbf{a})}{N(h\mathbf{a})}$
20: $D(h\mathbf{a}) \leftarrow D(h\mathbf{a}) + \frac{\text{SIMULATE}(h\mathbf{ao}, s', \text{depth} + |\mathbf{a}|) - D(h\mathbf{a})}{N(h\mathbf{a})}$
21: $\hat{\Psi}(ha) \leftarrow \hat{\Psi}(ha) - \mathcal{V}(h) + R(h\mathbf{a}) + \gamma^{|\mathbf{a}|} D(h\mathbf{a})$
22: $\mathcal{V}(h) \leftarrow \log\left\{ \sum_{a \in \text{children}(h)} \exp[\eta \hat{\Psi}(ha)] \right\} / \eta$
23: **return** $\mathcal{V}(h)$

3.5 Problem Scenarios

We evaluated the performance of PORPP on two challenging long-horizon POMDPs.

3D Maze with Poor Localisation. A 3-dimensional holonomic cuboid drone needs to navigate to one of two goal regions in a closed maze with very poor localisation (Figure 1). The state of the robot is represented by a continuous 3-dimensional co-ordinate for its centre of mass, and the robot can move continuously in any direction of fixed magnitude (i.e. $v = 1$) plus some mean zero (Gaussian) noise with covariance matrix $\mathbf{I} \times 0.02 \times v$ and any movement conforms to the "walls" of the environment. However, the robot does not know its true state and only knows that it can spawn at two starting positions with equal probability (Figure 1). The robot can only localise its co-ordinate if it comes in contact with a landmark where it receives an observation of its true position; otherwise, it receives no feedback about its position. The scenario terminates if the robot comes in contact with a danger zone—which incurs a penalty of -500—or reaches the goal—which yields a reward of 2000. A step penalty of -5 is incurred in all other cases. This is a long-horizon problem requiring 100 steps to reach the goal while simultaneously navigating around danger zones.

HEMS Mission with Evolving No-Fly-Zones. We considered a Helicopter Emergency Medical Service (HEMS) mission set on the Cap Corse peninsula in Corsica (Figure 2). The mesh used to generate the terrain was extracted from X-Plane 12. The mission objective is to navigate a holonomic helicopter starting from the west end of the island (ar-

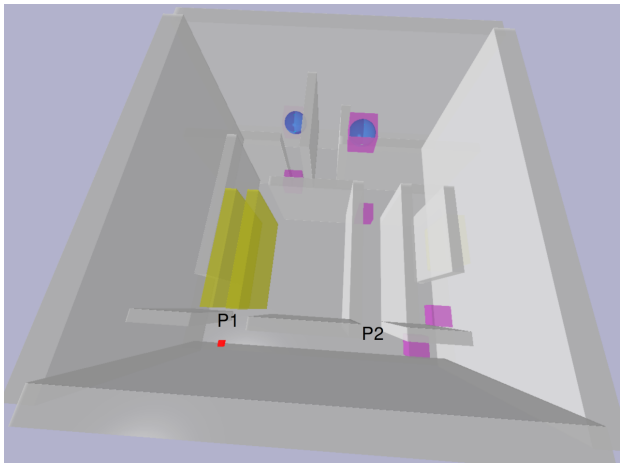
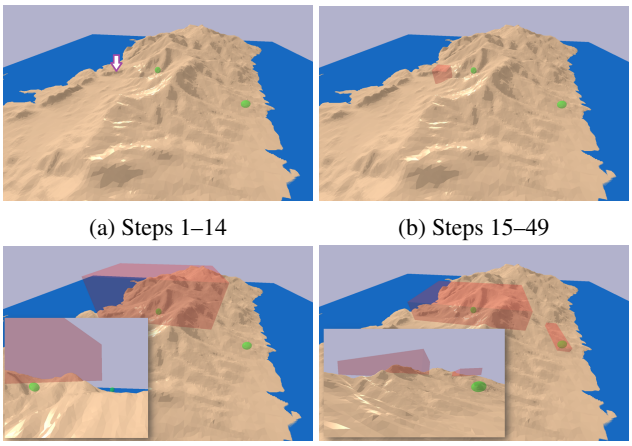


Figure 1: *3D Maze with Poor Localisation*. The environment is a closed box and walls are **grey**; danger zones are **yellow**; landmarks are **purple**; goal region is labelled **blue**. The robot spawns in two positions P1 and P2 with equal likelihood and the robot does not receive any initial feedback about its position. If the robot spawns at P1, the direct route to the goal has a high likelihood of collision with a danger zone so the robot must localise first and take a safer route.



(a) Steps 1–14 (b) Steps 15–49

(c) Steps 50–99 (d) Steps 99+

Figure 2: Corsica Rescue Mission with Evolving NFZs. The starting position is indicated by the arrow in (a); objectives are **green**; NFZs are **red**. The environment evolves at preset time-steps that are unknown to the agent. The agent should react to avoid NFZs but may elect not to do so in order to evade a greater catastrophe.

row in Figure 2 (a)) to two unordered objectives—i.e. the victim's locations (green balls in Figure 2) where the agent receives a reward of 2000 for each new objective achieved. The mission ends if there is a collision (which incurs a reward penalty of -2000) or both objectives are achieved—i.e. the mission is accomplished—which yields an additional reward of 20000. The scenario is complicated by the fact that no-fly-zones (NFZs) evolve at fixed time steps that are unknown to the agent (see Figure 2). The agent need not avoid NFZs entirely, but incurs an additional penalty of -20 for each step inside a NFZ. We assume that the agent has no predictive model of when NFZs will appear; hence, the agent only re-plans with respect to reward changes due to NFZ evolutions. To encourage the agent to achieve the objective, a step penalty of -5 is incurred at each time step. The state of the helicopter is fully specified by a continuous 3-dimensional co-ordinate representing the helicopter's centre of mass (its orientation is always fixed)—notice that fuel and weight of the craft are not considerations—and actions are the continuous directional vectors of a fixed magnitude $v = 2$ (i.e. the helicopter's speed) representing the agent's intended direction of movement. Transitions in the intended direction and readings of the true state of the helicopter are subject to Gaussian noise with covariance matrices $\mathbf{I} \times 0.25 \times v$ and $\mathbf{I} \times 0.2$ respectively. This problem is a long-horizon problem often requiring a minimum of 150 steps to accomplish the mission without consideration of NFZs.

3.6 Heuristic Action Sampler

One crucial factor in the overall performance of PORPP is the heuristic action sampler $\textsc{SampleCandidateAction}(h, s)$. We stress that the heuristic action sampler is *not* a solution to the POMDP; indeed, the heuristic sampler need not account for uncertainty being a function of a determined state. Rather, its fundamental purpose is to exploit domain-specific knowledge to propose promising actions to explore given a belief.

In both environments, our specific implementation of this subroutine relies on an offline-generated Probabilistic Roadmap (PRM) [Choset *et al.*, 2005] to represent the environment's collision-free configuration space. Based on the input particle an *objective* in the environment's configuration space is sampled and collision-free paths to the sampled objective are queried from the PRM. That is,

- For the 3D maze, a random landmark or goal region is sampled and targeted and the shortest path on the PRM starting from the position given by the state particle to the target is returned.
- For the Corsica map, the state s records which victims have been visited. Accordingly, a simple homotopic collision-free path starting from the helicopter's position (as recorded in s) and ending at a random unvisited victim location is sampled.

The returned paths are then truncated at a fixed length, and a macro action which traces the path is returned.

3.7 Benchmark Methods

The benchmarks used for comparison are:

- *RefPol*. This simply samples a state particle and executes the action returned by the heuristic action sampler without further POMDP planning.
- *RefSolver*. The solver from [Kim *et al.*, 2023] a RBPOMDP which uses the heuristic action sampler as its *reference policy*.
- *POMCP*—[Silver and Veness, 2010]. The canonical benchmark to beat for online POMDP planning. For a fair comparison, it expands 16 macro actions composed of equally spaced directional vectors.

Planners	Time (s)	Succ. %	E[Tot. Reward]
PORPP	1	71	570.3 ± 183.5
	2	75	628.9 ± 191.3
	3	80	625.0 ± 215.2
	5	81	688.3 ± 200.1
	10	88	873.4 ± 172.1
	15	94	983.1 ± 168.0
RefSolver	2	39	-244.6 ± 224.5
	3	38	-278.9 ± 216.5
	5	26	-544.9 ± 204.6
	10	30	-384.0 ± 213.3
POMCP	2	10	-786.3 ± 378.0
	3	9	-1637.3 ± 256.8
	5	7	-2150.8 ± 161.0
	10	13	-1897.5 ± 240.1
RefPol	N/A	29	-572.2 ± 231.1

Table 1: Results for 3D Maze with Poor Localisation (100 runs; maximum macro action length = 10)

3.8 Experimental Setup

All experiments were performed on a desktop computer with 128GB DDR4 RAM and an 8 Core Intel Xeon Silver 4110 Processor. All solvers were implemented in the pomdp_py library [H2RLab, 2024] and Cythonised for a fair comparison. The discount factor for all environments was $\gamma = 0.99$.[3]

3.9 Results and Discussion

Results are summarised in Table 1 and Table 2. In both scenarios we ran RefPol to corroborate our claim that the heuristic action sampler is significantly sub-optimal. Still, PORPP was able leverage the heuristic action sampler to significantly outperform both benchmarks yielding very high success rates with $>$10 seconds of planning time. As expected from our theoretical analysis, the results improve in trend with the planning time. Notably, RefSolver does not improve quite as much PORPP which seems consistent with the idea that RefSolver is converging to a policy which is somewhere in between the reference policy and the optimal policy of the POMDP. POMCP, meanwhile, was myopic in both scenarios and could not take advantage of deep rewards even when helped by macro actions because of the need to exhaustively enumerate. Interestingly, in the HEMS mission, we typically observe the PORPP policy trace non-trivially adapting to the environment (Figure 3).

4 Summary

This paper presents PORPP an online particle-based anytime POMDP solver which provably approximates a gradual KL-constrained iterative scheme making it robust to large approximation errors. Empirical results indicate the feasibility of our planner for large-scale POMDPs showing that it outperforms existing benchmarks for the long-horizon POMDPs with evolving environments presented in this paper.

[3]See https://github.com/RDLLab/pomdp-py-porpp for the code and parameters used to run the experiments.

Planners	Time (s)	Succ. %	E[Tot. Reward]
PORPP	1	58	11393.5 ± 1588.4
	2	75	15408.8 ± 1399.3
	3	78	16207.7 ± 1316.7
	5	78	16231.6 ± 1320.2
	10	90	19393.5 ± 967.9
RefSolver	2	2	-1453.9 ± 947.8
	3	4	-860.6 ± 1297.7
	5	28	3514.9 ± 3043.1
	10	22	2258.7 ± 2809.2
POMCP	2	2	-410.5 ± 900.0
	3	0	-942.5 ± 181.1
	5	2	-421.8 ± 928.1
	10	0	-839.6 ± 227.4
RefPol	N/A	0	-6584.3 ± 379.5

Table 2: Results for HEMS Mission with Evolving NFZs (100 runs; maximum macro action length = 15)

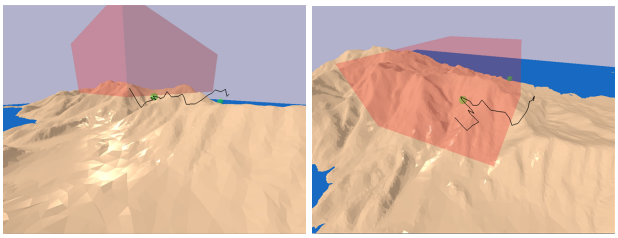


Figure 3: Two perspectives of the PORPP trajectory trace of the helicopter in the HEMS mission during steps 50–149. At the beginning of the trace the helicopter initially descends to avoid the new NFZ and targets the nearest objective. Once this objective is achieved, the helicopter successful navigates a path around the NFZ and surrounding terrain rather than taking the shortest path through the NFZ to the next objective.

For future work, we would like to examine the solver on non-holonomic problems (realistic ODE approximations of helicopter dynamics, robotic manipulators, etc.) with more complex domains (e.g. HEMS fire and flood rescue scenarios). We would also like to systematically stress test PORPP with respect to different parameter settings and choices of heuristic samplers.

Acknowledgments

This work was supported by Safran Electronics & Defense Australia Pty Ltd and Safran Group under the ARC Linkage project LP200301612.

References

[Asadi and Littman, 2017] Kavosh Asadi and Michael Littman. An alternative softmax operator for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 243–252, 2017.

[Azar *et al.*, 2011] Mohammed Gheshlaghi Azar, Vincenç Gómez, and Hilbert Kappen. Dynamic policy program-

ming with function approximation. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 119–127, 2011.

[Azar *et al.*, 2012] Mohammed Gheshlaghi Azar, Vincenç Gómez, and Hilbert Kappen. Dynamic policy programming. *Journal of Machine Learning Research*, 13:3207–3245, 2012.

[Blanchard *et al.*, 2021] Pierre Blanchard, Desmond J. Higham, and Nicholas J. Higham. Accurately computing the log-sum-exp and softmax functions. *IMA Journal of Numerical Analysis*, 41:2311–2330, 2021.

[Choset *et al.*, 2005] Howie Choset, Kevin Lynch, Seth Hutchinson, George Kantor, Wolfram Burgard, Lydia Kavraki, and Sebastian Thrun. *Principles of Robot Motion: Theory, Algorithms and Implementation*. MIT Press, 2005.

[Dvijotham and Todorov, 2012] Krishnamurthy Dvijotham and Emmanuel Todorov. Linearly solvable optimal control. *Reinforcement learning and approximate dynamic programming for feedback control*, pages 119–141, 2012.

[Flaspohler *et al.*, 2020] Genevieve Flaspohler, Nicholas Roy, and John W. Fisher III. Belief-dependent macro-action discovery in POMDPs using the value of information. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 11108–11118. Curran Associates, Inc., 2020.

[H2RLab, 2024] H2RLab. pomdp_py. https://h2r.github.io/pomdp-py, 2024. Accessed: 2025-06-06.

[He *et al.*, 2010] Ruijie He, Emma Brunskill, and Nicholas Roy. PUMA: planning under uncertainty with macro-actions. In Maria Fox and David Poole, editors, *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010, Atlanta, Georgia, USA, July 11-15, 2010*, pages 1089–1095. AAAI Press, 2010.

[Kaebling *et al.*, 1998] Leslie Kaebling, Michael Littman, and Anthony Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101:99–134, 1998.

[Kim *et al.*, 2023] Edward Kim, Yohan Karunanayake, and Hanna Kurniawati. Reference-based POMDPs. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 40659–40675. Curran Associates, Inc., 2023.

[Kurniawati *et al.*, 2011] Hanna Kurniawati, Yanzhu Du, David Hsu, and Wee Sun Lee. Motion planning under uncertainty for robotic tasks with long time horizons. *International Journal of Robotics Research*, 30(3):308–323, 2011.

[Kurniawati, 2022] Hanna Kurniawati. Partially observed Markov decision processes and robotics. *Annual Review of Control, Robotics, and Autonomous Systems*, 5:254–277, 2022.

[Lee *et al.*, 2007] Wee Lee, Nan Rong, and David Hsu. What makes some POMDP problems easy to approximate? In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007.

[Lee *et al.*, 2021] Yiyuan Lee, Panpan Cai, and David Hsu. MAGIC: Learning Macro-Actions for Online POMDP Planning . In *Proceedings of Robotics: Science and Systems*, July 2021.

[Liang *et al.*, 2024] Yuanchu Liang, Edward Kim, Wil Thomason, Zachary Kingston, Hanna Kurniawati, and Lydia Kavraki. Scaling long-horizon online POMDP planning via rapid state space sampling. In *Springer Proc. in Adv. Rob. (to appear)*, 2024. arXiv:2411.07032.

[Madani *et al.*, 2003] Omid Madani, Steve Hanks, and Anne Condon. On the undecidability of probabilistic planning and related stochastic optimization problems. *Artificial Intelligence*, 147:5–34, 2003.

[Rawlik *et al.*, 2012] Konrad Rawlik, Marc Toussaint, and Sethu Vijayakumar. On stochastic optimal control and reinforcement learning by approximate inference. *Proceedings of Robotics: Science and Systems VIII*, 2012.

[Silver and Veness, 2010] David Silver and Joel Veness. Monte-Carlo planning in large POMDPs. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010.

[Sunberg and Kochenderfer, 2018] Zachary Sunberg and Mykel Kochenderfer. Online algorithms for POMDPs with continuous state, action, and observation spaces. *Proceedings of the International Conference on Automated Planning and Scheduling*, 28(1):259–263, Jun. 2018.

[Theocharous and Kaelbling, 2003] Georgios Theocharous and Leslie Kaelbling. Approximate planning in POMDPs with macro-actions. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems*, volume 16. MIT Press, 2003.

[Todorov, 2006] Emanuel Todorov. Linearly-solvable Markov decision problems. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006.

[Todorov, 2009a] Emmanuel Todorov. Efficient computation of optimal actions. *Proceedings of the National Academy of Sciences*, 106(28):11478–11483, 2009.

[Todorov, 2009b] Emmanuel Todorov. Eigenfunction approximation methods for linearly-solvable optimal control problems. In *2009 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning*, pages 161–168. IEEE, 2009.

[Wang *et al.*, 2018] Erli Wang, Hanna Kurniawati, and Dirk Kroese. An on-line planner for POMDPs with large discrete action space: a quantile-based approach. In *Proceedings of the 28th International Conference on Aut. Plan. Sched.*, pages 273–77, Palo Alto, CA, 2018. AAAI Press.

Supplementary Material

1 Packing and Covering Numbers

For completeness, we briefly review the concepts of δ-packing and -covering numbers in general metric spaces. Let $B(x, \delta)$ denote the closed ball of radius δ centred at x.

Definition 1. *Let (X, ρ) be a metric space, let Y be a subset of X. For $\delta > 0$, the set of points $\{x_1, \ldots, x_n\} \subset X$ is a* δ-covering *of Y if $Y \subset \bigcup_{i=1}^{n} B(x_i, \delta)$, or equivalently, $\forall y \in Y$, $\exists i$ such that $\rho(y, x_i) \leq \delta$. If, moreover, the set $\{x_1, \ldots x_n\}$ is a subset of Y, then we say it is an* internal δ-covering *of Y.*[1] *Respectively, the* δ-covering number *and* internal δ-covering number *are defined as*

$$\mathcal{C}_\delta(Y) := \inf\{n : \exists \text{ a } \delta\text{-covering of } Y \text{ of size } n\} \quad (1)$$

and

$$\mathcal{C}^\circ_\delta(Y) := \inf\{n : \exists \text{ an internal } \delta\text{-covering of } Y \text{ of size } n\}. \quad (2)$$

Definition 2. *Let (X, ρ) be a metric space. For $\delta > 0$ and $Y \subset X$, the set of points $\{y_1, \ldots, y_m\} \subset Y$ is a* δ-packing *of Y if, $\forall i \neq j$, $\rho(x_i, x_j) > \delta$ (notice strict inequality), or equivalently, $\bigcap_{i=1}^{m} B(y_i, \delta) = \emptyset$. The* δ-packing number *is defined as*

$$\mathcal{P}_\delta(Y) := \sup\{m : \exists a \ \delta\text{-packing of size } m\}. \quad (3)$$

Proposition 1 (Packing-Covering Duality)**.** *Let (X, ρ) be a metric space, and $Y \subset X$. Then for arbitrary $\delta > 0$*

$$\mathcal{P}_\delta(Y) \leq \mathcal{C}_{\delta/2}(Y) \leq \mathcal{C}^\circ_{\delta/2}(Y) \leq \mathcal{P}_{\delta/2}(Y). \quad (4)$$

Proof. To prove the first inequality suppose, for a contradiction, that there exists a δ-packing $\{y_1, \ldots, y_m\}$ and a $\delta/2$-covering $\{x_1, \ldots, x_n\}$ such that $m > n$. By the pigeonhole principle, there must be at least one pair y_i and y_j belonging to the same ball $B(x_k, \delta/2)$ for some k. This means that $\rho(y_i, y_j) \leq \delta$ thereby contradicting the fact that $\{y_1, \ldots, y_m\}$ is a δ-packing. Hence $m \leq n$ and the conclusion follows.

For the last inequality, let $\mathcal{E} = \{y_1, \ldots, y_m\} \subset Y$ be a maximal packing. Suppose, for a contradiction, that there is a $y \in Y \setminus \mathcal{E}$ such that $\forall i$ we have $\rho(y_i, y) > \delta/2$. But this contradicts the maximality of $\mathcal{E}$ since we can simply construct a larger packing with $y_{m+1} := y$. Hence, $\mathcal{E}$ must be an internal $\delta/2$-covering of Y. Since $\mathcal{C}^\circ_{\delta/2}(Y)$ is the minimal size of all possible $\delta/2$-coverings, we have $\mathcal{C}^\circ_{\delta/2}(Y) \leq \mathcal{P}_{\delta/2}(Y)$.

The middle inequality follows trivially by observing that the set of internal δ-coverings is a subset of the set of all δ-coverings. □

Remark 1. *The quantity $\mathcal{C}_\delta(Y)$ is finite for all $\delta > 0$ if and only if Y is* totally bounded *– i.e. for every $\delta > 0$, there is a finite covering of the space by balls of radius δ. It is a well known fact that a compact set is totally bounded. Moreover, a bounded subset of a Euclidean space of arbitrary (finite) dimension is totally bounded (see e.g. [Kolmogorov and Fomin, 1970]). Therefore, if $|\mathcal{S}| < \infty$, the set $\mathcal{R}_{b_0}$ is totally bounded, being bounded by the $(|\mathcal{S}| - 1)$-dimensional simplex in $\mathbb{R}^{|\mathcal{S}|}$. This observation, along with Proposition 1, means that the quantities $\mathcal{C}_\delta(\mathcal{R}_{b_0})$, $\mathcal{C}^\circ_\delta(\mathcal{R}_{b_0})$ and $\mathcal{P}_\delta(\mathcal{R}_{b_0})$ are all well defined (i.e. finite) in this case. However, when $|\mathcal{S}| = \infty$ the finiteness of $\mathcal{C}_\delta(\mathcal{R}_{b_0})$ is not guaranteed in general. We therefore assert that $\mathcal{R}_{b_0}$ is totally bounded as per Assumption 1 in the main body of the paper.*

2 Some Useful Operators and Results

We will find it convenient to introduce some function operators for notational compactness. Recall that the operator $\tilde{\tau}_{B,\rho}$ was introduced in the main body of the paper.

Definition 3. *Let $\mathscr{V}$ and $\mathscr{Q}$ respectively denote the space of all functions $v : \mathcal{B} \to \mathbb{R}$ and $q : \mathcal{B} \times \mathcal{A} \to \mathbb{R}$ such that $\|v\|_\infty < \infty$ and $\|q\|_\infty < \infty$. For a given subset $B \subset \mathcal{B}$ and (stochastic) policy $\pi : \mathcal{B} \to \Delta(\mathcal{A})$ let*

$$\langle \pi, q \rangle(b) := \sum_{a \in \mathcal{A}} \pi(a \,|\, b) q(b, a) \quad (5)$$

$$[\mathcal{M}q](b) := \max_{a \in \mathcal{A}} q(b, a) \quad (6)$$

$$\langle P, v \rangle(b, a) := \sum_{o \in \mathcal{O}} P(o \,|\, b, a) v\big(\tau(b, a, o)\big) \quad (7)$$

$$\langle P, v \rangle_{B,\rho}(b, a) := \sum_{o \in \mathcal{O}} P(o \,|\, b, a) v\big(\tilde{\tau}_{B,\rho}(b, a, o)\big) \quad (8)$$

where $\tilde{\tau}_{B,\rho}$ was defined in Sect. 4.3 of the paper.

The next result is a simple consequence of Definition 3.

[1] In [Lee *et al.*, 2007], such a covering is called a *proper δ covering*. Our terminology seems more descriptive and is consistent with the broader mathematical literature.

Proposition 2. *For any $B \subset \mathcal{B}$, metric ρ on $\mathcal{B}$, $v_1, v_2 \in \mathscr{V}$ and $q_1, q_2 \in \mathscr{Q}$ and $\mu_1, \mu_2, \lambda_1, \lambda_2 \in \mathbb{R}$ we have*

$$\begin{aligned}\langle \pi, \mu_1 v_1 + \mu_2 v_2 + \lambda_1 q_1 + \lambda_2 q_2 \rangle \\ = \mu_1 v_1 + \mu_2 v_2 + \lambda_1 \langle \pi, q_1 \rangle + \lambda_2 \langle \pi, q_2 \rangle \end{aligned} \quad (9)$$

and

$$\begin{aligned}\langle P, \langle \pi, \mu_1 v_1 + \mu_2 v_2 + \lambda_1 q_1 + \lambda_2 q_2 \rangle \rangle_{B,\rho} \\ = \mu_1 \langle P, v_1 \rangle_{B,\rho} + \mu_2 \langle P, v_2 \rangle_{B,\rho} \\ + \lambda_1 \langle P, \langle \pi, q_1 \rangle \rangle_{B,\rho} + \lambda_2 \langle P, \langle \pi, q_2 \rangle \rangle_{B,\rho}. \end{aligned} \quad (10)$$

Moreover

$$\|\langle P, v_1 \rangle_{B,\rho}\|_\infty \leq \|v_1\|_{B,\infty}. \quad (11)$$

Proof. Equations (9) and (10) follow straightforwardly from the definitions. The inequality (11) follows from Jensen's inequality since

$$\begin{aligned}|\langle P, v_1 \rangle_{B,\rho}(b,a)| &= \Big| \sum_{o \in \mathcal{O}} P(o \,|\, b, a) v_1\big(\tilde{\tau}_{B,\rho}(b,a,o)\big) \Big| \\ &\leq \sum_{o \in \mathcal{O}} P(o \,|\, b, a) \big| v_1\big(\tilde{\tau}_{B,\rho}(b,a,o)\big) \big| \leq \|v_1\|_{B,\infty}\end{aligned}$$

for every $b \in \mathcal{B}$ and $a \in \mathcal{A}$. □

For a given stochastic policy $\pi \in \Pi$, let the self-mapping operators $\mathscr{T} : \mathscr{Q} \to \mathscr{Q}$, $\mathscr{T}^\pi : \mathscr{Q} \to \mathscr{Q}$ and $\mathscr{T}^\pi_{B,\rho} : \mathscr{Q} \to \mathscr{Q}$ be given by

$$[\mathscr{T} q](b,a) := R(b,a) + \gamma \langle P, [\mathcal{M} q] \rangle (b,a) \quad (12)$$

$$[\mathscr{T}^\pi q](b,a) := R(b,a) + \gamma \langle P, \langle \pi, q \rangle \rangle (b,a) \quad (13)$$

$$[\mathscr{T}^\pi_{B,\rho} q](b,a) := R(b,a) + \gamma \langle P, \langle \pi, q \rangle \rangle_{B,\rho} (b,a). \quad (14)$$

We have the following well-known result, which can be justified using a classical argument – see e.g. [Bertsekas, 2008] or [Ross, 1970].

Lemma 1. *For arbitrary q and q' belonging to $\mathscr{Q}$, the operators $\mathscr{T}$, $\mathscr{T}^\pi$ and $\mathscr{T}^\pi_{B,\rho}$ satisfy*

$$\begin{aligned}\|\mathscr{T} q - \mathscr{T} q'\|_\infty &\leq \gamma \|q - q'\|_\infty \\ \|\mathscr{T}^\pi q - \mathscr{T}^\pi q'\|_\infty &\leq \gamma \|q - q'\|_\infty \\ \|\mathscr{T}^\pi_{B,\rho} q - \mathscr{T}^\pi_{B,\rho} q'\|_\infty &\leq \gamma \|q - q'\|_\infty. \end{aligned} \quad (15)$$

Thus, for $\gamma \in (0,1)$, the operators are contraction mappings with modulus of contraction γ and have unique fixed points (up to $\|\cdot\|_\infty$-equivalence). In particular, $Q^*, Q^\pi, Q^\pi_{B,\rho} \in \mathscr{Q}$ are the fixed points of $\mathscr{T}$, $\mathscr{T}^\pi$ and $\mathscr{T}^\pi_{B,\rho}$ respectively. Moreover, Q^* is the solution of the POMDP introduced in Sec. 2.1.

We can quantify the difference between the fixed points when B is a δ-covering $\mathcal{E}_\delta$ of $\mathcal{R}_{b_0}$ for the metric ρ_1. We will need two auxiliary results to prove this claim which we formalise in Proposition 3. The first is essentially Lemma 1 from [Lee *et al.*, 2007].

Lemma 2. *For any $\delta > 0$ and belief points b and b', we have*

$$|V^*(b) - V^*(b')| \leq Q_{\max} \|b - b'\|_1 \quad (16)$$

and, for fixed $a \in \mathcal{A}$,

$$|Q^*(b,a) - Q^*(b',a)| \leq Q_{\max} \|b - b'\|_1. \quad (17)$$

Proof. The inequality (16) comes directly from [Lee *et al.*, 2007] We sketch the proof for (17) since it follows in a similar way to Lemma 1 in [Lee *et al.*, 2007]. For a fixed $a \in \mathcal{A}$, the function $Q^* : \mathcal{B} \times \mathcal{A} \to \mathbb{R}$ can be approximated arbitrarily closely by a piecewise-linear function $Q^*(b,a) = \max_{\alpha \in \Gamma} (\alpha \cdot b)$ where $\Gamma \subset \mathbb{R}^{|\mathcal{S}|}$. For each $\alpha \in \Gamma$ the boundedness of the reward function ensures that the absolute values of the components of α are bounded by $Q_{\max}$. We can then argue in a similar way to [Lee *et al.*, 2007] to get the desired bound. □

Lemma 3. *Consider any stochastic policy $\pi \in \Pi$ and δ-covering $\mathcal{E}_\delta$ of $\mathcal{R}_{b_0}$ for some $\delta > 0$. Then*

$$\|\mathscr{T}^\pi Q^* - \mathscr{T}^\pi_{\mathcal{E}_\delta, \rho_1} Q^*\|_\infty \leq \gamma \delta Q_{\max}. \quad (18)$$

Proof. For any $\pi \in \Pi$, and $q \in \mathscr{Q}$ and fixed $(b,a) \in \mathcal{B} \times \mathcal{A}$, the operator definitions and Lemma 2 give us

$$\begin{aligned}&\big|\mathscr{T}^\pi Q^*(b,a) - \mathscr{T}^\pi_{\mathcal{E}_\delta, \rho_1} Q^*(b,a)\big| \\ &\leq \gamma \Big| \big\langle P, \langle \pi, Q^* \rangle \big\rangle (b,a) - \big\langle P, \langle \pi, Q^* \rangle \big\rangle_{\mathcal{E}_\delta, \rho_1} (b,a) \Big| \\ &\leq \gamma \sum_{o,a'} \theta^{o,a'}_{b,a} \big| Q^*(\tau(b,a,o), a') - Q^*(\tilde{\tau}_{\mathcal{E}_\delta, \rho_1}(b,a,o), a') \big| \\ &\leq \gamma \sum_{o,a'} \theta^{o,a'}_{b,a} Q_{\max} \big\| \tau(b,a,o) - \tilde{\tau}_{\mathcal{E}_\delta, \rho_1}(b,a,o) \, \big\|_1 \\ &\leq \gamma \delta Q_{\max}\end{aligned}$$

where $\theta^{o,a'}_{b,a} := P(o \,|\, b,a) \pi(a' \,|\, b)$ is a probability distribution over $\mathcal{O} \times \mathcal{A}$. The desired inequality follows since (b,a) was arbitrary in $\mathcal{B} \times \mathcal{A}$. □

Proposition 3. *For any $\delta > 0$*

$$\|Q^* - Q^{\pi^*}_{\mathcal{E}_\delta, \rho_1}\|_\infty \leq \frac{\gamma \delta Q_{\max}}{1 - \gamma}. \quad (19)$$

Proof. The policy was arbitrary in Lemma 3, so we can choose $\pi := \pi^*$ and we get

$$\begin{aligned}&\|Q^* - Q^{\pi^*}_{\mathcal{E}_\delta, \rho_1}\|_\infty \\ &\leq \|\mathscr{T}^{\pi^*} Q^* - \mathscr{T}^{\pi^*}_{\mathcal{E}_\delta, \rho_1} Q^* + \mathscr{T}^{\pi^*}_{\mathcal{E}_\delta, \rho_1} Q^* - \mathscr{T}^{\pi^*}_{\mathcal{E}_\delta, \rho_1} Q^{\pi^*}_{\mathcal{E}_\delta, \rho_1}\|_\infty \\ &\leq \gamma \delta Q_{\max} + \gamma \|Q^* - Q^{\pi^*}_{\mathcal{E}_\delta, \rho_1}\|_\infty\end{aligned}$$

where we have used the contraction property from Lemma 1. Rearranging the above gives us the desired result. □

Definition 4. *For any $\eta > 0$ and $q \in \mathscr{Q}$, define the operators $\mathcal{L}_\eta : \mathscr{Q} \to \mathscr{V}$ and $\mathcal{M}_\eta : \mathscr{Q} \to \mathscr{V}$ according to*

$$[\mathcal{L}_\eta q](b) := \frac{1}{\eta} \log \Big\{ \sum_{a \in \mathcal{A}} \exp[\eta q(b,a)] \Big\} \quad (20)$$

$$[\mathcal{M}_\eta q](b) := \frac{\sum_{a \in \mathcal{A}} \exp[\eta q(b,a)] q(b,a)}{\sum_{a' \in \mathcal{A}} \exp[\eta q(b,a')]} \quad (21)$$

for all $b \in \mathcal{B}$ so that $\mathcal{L}_\eta$ and $\mathcal{M}_\eta$ are the log-sum-exp *and* Boltzmann soft-max *operators respectively.*

We have the following useful properties.

Proposition 4. *For any $v \in \mathscr{V}$ and $q \in \mathscr{Q}$ we have*

$$[\mathcal{L}_\eta(v+q)] = v + [\mathcal{L}_\eta q]. \tag{22}$$

Moreover, if $\mathcal{A}$ is finite,

$$\Big|[\mathcal{L}_\eta q](b) - [\mathcal{M}_\eta q](b)\Big| \leq \frac{\log(|\mathcal{A}|)}{\eta} \tag{23}$$

and

$$0 \leq [\mathcal{L}_\eta q](b) - [\mathcal{M} q](b) \leq \frac{\log(|\mathcal{A}|)}{\eta} \tag{24}$$

for all $b \in \mathcal{B}$ and $\eta > 0$.

Proof. See [MacKay, 2003] for the proof of (23). We now prove (24). Fix a $b \in \mathcal{B}$. To show the lower bound, suppose $a^* \in \arg\max_{a\in\mathcal{A}} q(b,a)$ so that $q(b,a^*) = [\mathcal{M}q](b)$. Then

$$\begin{aligned}[\mathcal{L}_\eta q](b) &= \frac{1}{\eta}\log\Big\{\sum_{a\in\mathcal{A}}\exp[\eta q(b,a)]\Big\} \\ &\geq \frac{1}{\eta}\log\Big\{\exp[\eta q(b,a^*)]\Big\} = [\mathcal{M}q](b).\end{aligned}$$

Observe also that

$$\begin{aligned}[\mathcal{L}_\eta q](b) &\leq \frac{1}{\eta}\log\Big\{\sum_{a\in\mathcal{A}}\exp\big[\eta[\mathcal{M}q](b)\big]\Big\} \\ &\leq \frac{\log(|\mathcal{A}|)}{\eta} + [\mathcal{M}q](b)\end{aligned}$$

which is the desired upper bound. □

The following basic result will be useful.

Proposition 5. *Let $x \in X$ be an arbitrary element of a general set X and consider the real functions $f_1 : X \to \mathbb{R}$ and $f_2 : X \to \mathbb{R}$. Then*

$$\begin{aligned}&\sup_{x\in X} f_1(x) + \sup_{x\in X} f_2(x) \\ &\quad\leq \sup_{x\in X}\Big(f_1(x)+f_2(x)\Big) + 2\Big(\sup_{x\in X}|f_2(x)|\Big).\end{aligned} \tag{25}$$

Proof. Observe that

$$\begin{aligned}f_1(x_1)+f_2(x_2) &= f_1(x_1)+f_2(x_1)-f_2(x_1)+f_2(x_2) \\ &= \sup_{x\in X}\Big(f_1(x)+f_2(x)\Big) + |f_2(x_1)| + |f_2(x_2)| \\ &= \sup_{x\in X}\Big(f_1(x)+f_2(x)\Big) + 2\Big(\sup_{x\in X}|f_2(x)|\Big)\end{aligned}$$

for arbitrary $x_1, x_2 \in X$. □

3 Convergence of the Exact Scheme

We can now prove Theorem 1 in the main body of the paper. In fact the proof is slightly more general as it states the result for all $B \subset \mathcal{R}_{b_0}$ (NB: Theorem 1 is the special case when $B = \mathcal{R}_{b_0}$). To this end, let $\{Q_0, Q_1, Q_2, \ldots\} \subset \mathscr{Q}$ be an auxiliary sequence of action-value functions which is defined according to recursion

$$Q_0 := \hat{\Psi}_0 \tag{26}$$

$$Q_k := R + \frac{\gamma}{k}\langle P, \mathcal{L}_\eta[(k-1)Q_{k-1} + Q_0]\rangle_{B,\rho} + \frac{E_{k-1}}{k} \tag{27}$$

for $k \geq 1$. Our aim is to bound the quantity $\|Q^{\hat{\pi}_k}_{B,\rho} - Q_k\|_\infty$ which will help us ultimately bound $\|Q^{\pi^*}_{B,\rho} - Q^{\hat{\pi}_k}_{B,\rho}\|_\infty$.

Step 1. In the first step, we will inductively verify the relation:

$$\hat{\Psi}_k = kQ_k + Q_0 - \mathcal{L}_\eta[(k-1)Q_{k-1} + Q_0], \quad \forall k \geq 1. \tag{28}$$

For the base case when $k = 1$, notice that (27) gives $Q_1 = R + \gamma\langle P, \mathcal{L}_\eta Q_0\rangle_{B,\rho} + \epsilon_0$. Then, the RHS of (28) is

$$\begin{aligned}&R + \gamma\langle P, \mathcal{L}_\eta Q_0\rangle_{B,\rho} + Q_0 - \mathcal{L}_\eta Q_0 + \epsilon_0 \\ &\quad= R + \gamma\langle P, \mathcal{L}_\eta Q_0\rangle_{B,\rho} + \hat{\Psi}_0 - \mathcal{L}_\eta Q_0 + \epsilon_0 = \hat{\Psi}_1\end{aligned} \tag{29}$$

because of the synchronous scheme. For the induction step, suppose (28) holds up to some $k \geq 1$. Then, using Proposition 2 and our definitions, we get

$$\begin{aligned}\hat{\Psi}_{k+1} &= \hat{\Psi}_k - [\mathcal{L}_\eta\hat{\Psi}_k] + R + \gamma\langle P, \mathcal{L}_\eta\hat{\Psi}_k\rangle_{B,\rho} + \epsilon_k \\ &= kQ_k + Q_0 + R + \gamma\langle P, \mathcal{L}_\eta[kQ_k + Q_0] \\ &\quad - \mathcal{L}_\eta[(k-1)Q_{k-1} + Q_0]\rangle_{B,\rho} \\ &\quad - \mathcal{L}_\eta[kQ_k + Q_0] + \epsilon_k \\ &= kQ_k - kR - \gamma\langle P, \mathcal{L}_\eta[(k-1)Q_{k-1} + Q_0]\rangle_{B,\rho} \\ &\quad - E_{k-1} + (k+1)R + \gamma\langle P, \mathcal{L}[kQ_k + Q_0]\rangle_{B,\rho} \\ &\quad + E_k + Q_0 - \mathcal{L}_\eta[kQ_k + Q_0] \\ &= (k+1)Q_{k+1} + Q_0 - \mathcal{L}_\eta[kQ_k + Q_0]\end{aligned} \tag{30}$$

which proves the desired relation (28). Incidentally, since $\mathcal{L}_\eta q$ is independent of $a \in \mathcal{A}$, a trivial consequence of (28) is that

$$\hat{\pi}_k(a \,|\, b) := \frac{\exp[\eta\{Q_k(b,a) + Q_0(b,a)\}]}{\sum_{a'}\exp[\eta\{Q_k(b,a') + Q_0(b,a')\}]} \tag{31}$$

from which we conclude that

$$\begin{aligned}\mathcal{M}_\eta(kQ_k + Q_0) &= \langle\hat{\pi}_k, kQ_k + Q_0\rangle \\ &= k\langle\hat{\pi}_k, Q_k\rangle + \langle\hat{\pi}_k, Q_0\rangle.\end{aligned} \tag{32}$$

Step 2. Next, we try to explicitly bound $\|Q^{\pi^*}_{B,\rho} - Q_k\|_\infty$. We try to inductively prove the relation

$$\begin{aligned}\|Q^{\pi^*}_{B,\rho} - Q_k\|_\infty &\leq \frac{\gamma(4Q_{\max} + \log(|\mathcal{A}|)/\eta)}{(1-\gamma)k} \\ &\quad + \frac{1}{k}\sum_{j=1}^{k}\gamma^{k-j}\|E_{j-1}\|_\infty, \quad \forall k \geq 1.\end{aligned} \tag{33}$$

First we prove the base case for $k = 1$. Our hypothesis $\|Q_0\|_\infty = \|\hat{\Psi}_0\|_\infty \leq Q_{\max}$ together with the triangle inequality for norms yield

$$\|Q^{\pi^*}_{B,\rho} - Q_0\|_\infty \leq \|Q^{\pi^*}_{B,\rho}\|_\infty + \|Q_0\|_\infty \leq 2Q_{\max}. \tag{34}$$

Thus,

$$\begin{aligned}
&\|Q^{\pi^*}_{B,\rho} - Q_1\|_\infty \\
&\quad = \|\mathscr{T}^{\pi^*}_{B,\rho} Q^{\pi^*}_{B,\rho} - (R + \gamma\langle P, \mathcal{L}_\eta Q_0\rangle_{B,\rho} + E_0)\|_\infty \\
&\quad \le \|\mathscr{T}^{\pi^*}_{B,\rho} Q^{\pi^*}_{B,\rho} - \mathscr{T}^{\pi^*}_{B,\rho} Q_0\|_\infty + \|\mathscr{T}^{\pi^*}_{B,\rho} Q_0 \\
&\qquad - (R + \gamma\langle P, \mathcal{L}_\eta Q_0\rangle_{B,\rho})\|_\infty + \|E_0\|_\infty \\
&\quad \le \gamma\|Q^{\pi^*}_{B,\rho} - Q_0\|_\infty + \gamma\Big\|\big\langle P, \mathcal{M}Q_0 - \mathcal{L}_\eta Q_0\big\rangle_{B,\rho}\Big\|_\infty \\
&\qquad + \|E_0\|_\infty \\
&\quad \le \gamma\|Q^* - Q_0\|_\infty + \gamma\Big\|\mathcal{M}Q_0 - \mathcal{L}_\eta Q_0\Big\|_\infty + \|E_0\|_\infty \\
&\quad \le \gamma\Big[2Q_{\max} + \frac{\log(|\mathcal{A}|)}{\eta}\Big] + \|E_0\|_\infty
\end{aligned} \tag{35}$$

where we have made use of (11), Lemma 1 and Proposition 4. This validates the base case of (33).

Now suppose (33) holds up to some $k \ge 1$. Then

$$\begin{aligned}
&\|Q^{\pi^*}_{B,\rho} - Q_{k+1}\|_\infty \\
&\quad = \Big\|\mathscr{T}^{\pi^*}_{B,\rho} Q^{\pi^*}_{B,\rho} \\
&\qquad - \Big(R + \frac{\gamma}{k+1}\langle P, \mathcal{L}_\eta[kQ_k + Q_0]\rangle_{B,\rho} + \frac{E_k}{k+1}\Big)\Big\|_\infty \\
&\quad = \frac{1}{k+1}\Big\|\mathscr{T}^{\pi^*}_{B,\rho} Q^{\pi^*}_{B,\rho} - \mathscr{T}^{\pi^*}_{B,\rho} Q_0 + \mathscr{T}^{\pi^*}_{B,\rho} Q_0 \\
&\qquad - \big[(k+1)R + \gamma\big\langle P, \mathcal{L}_\eta[kQ_k + Q_0]\big\rangle_{B,\rho}\big] \\
&\qquad\quad + k(\mathscr{T}^{\pi^*}_{B,\rho} Q^{\pi^*}_{B,\rho} - \mathscr{T}^{\pi^*}_{B,\rho} Q_k + \mathscr{T}^{\pi^*}_{B,\rho} Q_k) - E_k\Big\|_\infty \\
&\quad \le \frac{1}{k+1}\Big[\big\|\mathscr{T}^{\pi^*}_{B,\rho} Q^{\pi^*}_{B,\rho} - \mathscr{T}^{\pi^*}_{B,\rho} Q_0\big\|_\infty \\
&\qquad + k\big\|\mathscr{T}^{\pi^*}_{B,\rho} Q^{\pi^*}_{B,\rho} - \mathscr{T}^{\pi^*}_{B,\rho} Q_k\big\|_\infty \\
&\qquad\quad + \Big\|k\mathscr{T}^{\pi^*}_{B,\rho} Q_k + \mathscr{T}^{\pi^*}_{B,\rho} Q_0 \\
&\qquad - \big[(k+1)R + \gamma\big\langle P, \mathcal{L}_\eta[kQ_k + Q_0]\big\rangle_{B,\rho}\big]\Big\|_\infty + \|E_k\|_\infty\Big] \\
&\quad \le \frac{1}{k+1}\Big[\gamma\|Q^{\pi^*}_{B,\rho} - Q_0\|_\infty + \gamma k\|Q^{\pi^*}_{B,\rho} - Q_k\|_\infty \\
&\qquad\quad + \Big\|k\mathscr{T}^{\pi^*}_{B,\rho} Q_k + \mathscr{T}^{\pi^*}_{B,\rho} Q_0 \\
&\qquad - \big[(k+1)R + \gamma\big\langle P, \mathcal{L}_\eta[kQ_k + Q_0]\big\rangle_{B,\rho}\big]\Big\|_\infty + \|E_k\|_\infty\Big].
\end{aligned} \tag{36}$$

Now Proposition 5 and Proposition 4 yield

$$\begin{aligned}
&\big\|k\mathscr{T}^{\pi^*}_{B,\rho} Q_k + \mathscr{T}^{\pi^*}_{B,\rho} Q_0 - (k+1)R \\
&\quad - \gamma\big\langle P, \mathcal{L}_\eta[kQ_k + Q_0]\big\rangle_{B,\rho}\big\|_\infty \\
&\quad \le \big\|\gamma\big\langle P, \mathcal{M}(kQ_k) + \mathcal{M}Q_0 - \mathcal{L}_\eta[kQ_k + Q_0]\big\rangle_{B,\rho}\big\|_\infty \\
&\quad \le \gamma\big\|\mathcal{M}(kQ_k) + \mathcal{M}Q_0 - \mathcal{L}_\eta[kQ_k + Q_0]\big\|_\infty \\
&\quad \le \gamma\big\|\mathcal{M}(kQ_k + Q_0) + 2\mathcal{M}|Q_0| - \mathcal{L}_\eta[kQ_k + Q_0]\big\|_\infty \\
&\quad \le \gamma\Big[2\|Q_0\|_\infty + \frac{\log(|\mathcal{A}|)}{\eta}\Big].
\end{aligned} \tag{37}$$

Simple computations after substituting (37) into (36) then gives

$$\begin{aligned}
\|Q^{\pi^*}_{B,\rho} - Q_{k+1}\|_\infty \le \frac{1}{k+1}\Big[&\frac{\gamma(4Q_{\max} + \log(|\mathcal{A}|)/\eta)}{(1-\gamma)} \\
&+ \sum_{j=1}^{k+1}\gamma^{k+1-j}\|E_{j-1}\|_\infty\Big]
\end{aligned} \tag{38}$$

which verifies the desired relation.

Step 3. We are now ready to prove the main result. The triangle inequality and the contraction property of $\mathscr{T}^{\hat{\pi}_k}$ give

$$\begin{aligned}
&\|Q^{\pi^*}_{B,\rho} - Q^{\hat{\pi}_k}_{B,\rho}\|_\infty \\
&\quad \le \|Q^{\pi^*}_{B,\rho} - Q_{k+1}\|_\infty + \|Q_{k+1} - \mathscr{T}^{\hat{\pi}_k}_{B,\rho} Q^{\pi^*}_{B,\rho}\|_\infty \\
&\qquad + \|\mathscr{T}^{\hat{\pi}_k}_{B,\rho} Q^{\pi^*}_{B,\rho} - \mathscr{T}^{\hat{\pi}_k}_{B,\rho} Q^{\hat{\pi}_k}_{B,\rho}\|_\infty \\
&\quad \le \|Q^{\pi^*}_{B,\rho} - Q_{k+1}\|_\infty + \|Q_{k+1} - \mathscr{T}^{\hat{\pi}_k}_{B,\rho} \\
&\qquad + \gamma\|Q^{\pi^*}_{B,\rho} - Q^{\hat{\pi}_k}_{B,\rho}\|_\infty
\end{aligned} \tag{39}$$

Rearranging, we get

$$\begin{aligned}
(1-\gamma)\|Q^{\pi^*}_{B,\rho} - Q^{\hat{\pi}_k}_{B,\rho}\|_\infty \le \|Q^{\pi^*}_{B,\rho} - Q_{k+1}\|_\infty \\
+ \|Q_{k+1} - \mathscr{T}^{\hat{\pi}_k}_{B,\rho} Q^{\pi^*}_{B,\rho}\|_\infty.
\end{aligned} \tag{40}$$

Now, we can use (23), (32) and (33) to yield

$$\begin{aligned}
&\|Q_{k+1} - \mathscr{T}^{\hat{\pi}_k}_{B,\rho} Q^{\pi^*}_{B,\rho}\|_\infty \\
&\quad \le \Big\|R + \frac{\gamma}{k+1}\big\langle P, \mathcal{L}_\eta[kQ_k + Q_0]\big\rangle_{B,\rho} \\
&\qquad + \frac{E_k}{k+1} - \mathscr{T}^{\hat{\pi}_k}_{B,\rho} Q^{\pi^*}_{B,\rho}\Big\|_\infty \\
&\quad \le \Big\|\frac{\gamma}{k+1}\big\langle P, \mathcal{L}_\eta[kQ_k + Q_0] - \mathcal{M}_\eta[kQ_k + Q_0]\big\rangle_{B,\rho} \\
&\qquad + \frac{\gamma}{k+1}\big\langle P, \mathcal{M}_\eta[kQ_k + Q_0]\big\rangle_{B,\rho} + \frac{E_k}{k+1} \\
&\qquad - \big\langle P, \langle\hat{\pi}_k, Q^{\pi^*}_{B,\rho}\rangle\big\rangle_{B,\rho}\Big\|_\infty \\
&\quad \le \frac{\gamma}{k+1}\Big[\frac{\alpha}{1-\gamma} + \sum_{j=1}^{k}\gamma^{k-j}\|E_{j-1}\|_\infty\Big] + \frac{\|E_k\|_\infty}{k+1}
\end{aligned} \tag{41}$$

where

$$\alpha := 4Q_{\max} + \frac{\log(|\mathcal{A}|)}{\eta}. \tag{42}$$

Using the above bound and (33) in (40) and setting $B = \mathcal{R}_{b_0}$ then gives us

$$\begin{aligned}
(1-\gamma)\|Q^* - Q^{\hat{\pi}_k}\|_\infty \le \frac{2\gamma\alpha}{(1-\gamma)(k+1)} \\
+ \frac{2}{k+1}\sum_{j=0}^{k}\gamma^{k-j}\|E_j\|_\infty
\end{aligned} \tag{43}$$

which completes the proof. □

4 Convergence of the Approximate Scheme

We prove the high-probability loss bound presented in Theorem 2 of the main body. We will need the following result.

Lemma 4. *If* $\|\hat{\Psi}_0\|_\infty \leq Q_{\max}$ *then*

$$\|\epsilon_k\|_{B,\infty} \leq \frac{4\gamma \log(|\mathcal{A}|)}{\eta(1-\gamma)} + 2Q_{\max} =: U, \quad \forall k \geq 0 \tag{44}$$

for the sequence $(\hat{\Psi}_k)_{k\geq 0}$ *generated by the synchronous scheme.*

Proof. For any $(b,a) \in \mathcal{E}_\delta \times \mathcal{A}$, the error for the synchronous scheme is given by

$$\begin{aligned}\epsilon_k(b,a) = R(b,a) - \sum_{i=1}^{N_k(b,a)} \frac{R(s_i,a)}{N_k(b,a)} \\ + \frac{\gamma}{M_k} \sum_{j=1}^{M_k(b,a)} [\mathcal{L}_\eta \hat{\Psi}_k]\big(\tilde{\tau}_{\mathcal{E}_\delta,\rho_1}(b,a,o_k)\big) \\ - \gamma \langle P, [\mathcal{L}_\eta \hat{\Psi}_k]\rangle_{\mathcal{E}_\delta,\rho_1}(b,a).\end{aligned}$$

This and the bound (11) give

$$\begin{aligned}\|\epsilon_k\|_\infty &\leq 2R_{\max} + \gamma \sup_{(b,a)\in\mathcal{E}_\delta\times\mathcal{A}} \Big|[\mathcal{L}_\eta \hat{\Psi}_k]\big(\tilde{\tau}_{\mathcal{E}_\delta,\rho_1}(b,a,o_k)\big)\Big| \\ &\quad + \gamma \big\|\langle P, [\mathcal{L}_\eta \hat{\Psi}_k]\rangle_{\mathcal{E}_\delta,\rho_1}(b,a)\big\|_{\mathcal{E}_\delta,\infty} \\ &\leq 2R_{\max} + 2\gamma \|\mathcal{L}_\eta \hat{\Psi}_k\|_{\mathcal{E}_\delta,\infty}\end{aligned}$$

so it suffices to bound $\|\mathcal{L}_\eta \hat{\Psi}_k\|_{\mathcal{E}_\delta,\infty}$ for all $k \geq 0$ which we can validate via induction. We claim that

$$\|\mathcal{L}_\eta \hat{\Psi}_k\|_{\mathcal{E}_\delta,\infty} \leq \frac{2\log(|\mathcal{A}|)}{\eta(1-\gamma)} + Q_{\max} \quad \forall k \geq 0. \tag{45}$$

The base case follows immediately from (24) so that, for any $b \in \mathcal{B}$,

$$\Big|[\mathcal{L}_\eta \hat{\Psi}_0](b)\Big| \leq \log(|\mathcal{A}|)/\eta + \|\hat{\Psi}_0\|_\infty \tag{46}$$

which satisfies the required bound since we hypothesised that $\|\hat{\Psi}_0\|_\infty \leq Q_{\max}$. For the induction step, it suffices to fix a $b \in \mathcal{E}_\delta$ and to observe that

$$\begin{aligned}&\Big|[\mathcal{L}_\eta \hat{\Psi}_{k+1}](b)\Big| \\ &= \Big|[\mathcal{L}_\eta \hat{\Psi}_{k+1}](b) - [\mathcal{M}\hat{\Psi}_0](b) + [\mathcal{M}\hat{\Psi}_{k+1}](b)\Big| \\ &\leq \Big|\big[\mathcal{M}\hat{\Psi}_k - [\mathcal{L}_\eta \hat{\Psi}_k] + R \\ &\quad + \gamma[\mathcal{L}_\eta \hat{\Psi}_k](\tilde{\tau}_{\mathcal{E}_\delta,\rho_1}(\cdot,\cdot,o_k))\big](b)\Big| + \frac{\log(|\mathcal{A}|)}{\eta} \\ &\leq \frac{\log(|\mathcal{A}|)}{\eta} + \Big|[\mathcal{M}\hat{\Psi}_k](b) - [\mathcal{L}_\eta \hat{\Psi}_k](b)\Big| + R_{\max} \\ &\quad + \gamma\Big|[\mathcal{M}[\mathcal{L}_\eta \hat{\Psi}_k](\tilde{\tau}_{\mathcal{E}_\delta,\rho_1}(\cdot,\cdot,o_k))](b)\Big| \\ &\leq \frac{2\log(|\mathcal{A}|)}{\eta} + R_{\max} + \gamma\|\mathcal{L}_\eta \hat{\Psi}_k\|_{\mathcal{E}_\delta,\infty} \\ &\leq \frac{2\log(|\mathcal{A}|)}{\eta} + R_{\max} + \frac{2\gamma\log(|\mathcal{A}|)}{\eta(1-\gamma)} + \gamma Q_{\max} \\ &= \frac{2\log(|\mathcal{A}|)}{\eta(1-\gamma)} + Q_{\max}\end{aligned}$$

and the result follows from the arbitrariness of $b \in \mathcal{E}_\delta$. □

Theorem 2 is valid for a *synchronous* backup. In other words, we sample the observations $o_1^{b,a}, \ldots, o_{N_k}^{b,a}$ from the distribution $P(\cdot \,|\, b,a)$ for *every* $(b,a) \in \mathcal{E}_\delta \times \mathcal{A}$ and then compute $\hat{\Psi}_{k+1}$ according to the synchronous scheme at each iteration k. Let $\mathbf{o}_k := [o_k^{b,a}]_{(b,a)\in\mathcal{E}_\delta\times\mathcal{A}}$ represent the collective sampled random variable after one synchronous iteration.

TODO: Tidy up the discussion about filtrations here. Now, let $(\mathcal{F}_k)_{k\geq 0}$ be the filtration generated by the random variables $(\mathbf{o}_i)_{0\leq i\leq k}$. Intuitively, each $\mathcal{F}_k$ can be seen as the set of events that can be distinguished as true or false after having observed $(\mathbf{o}_i)_{0\leq i\leq k}$. By our definition of the approximate sequence, it is clear that

$$\mathbb{E}[\epsilon_k(b,a) \,|\, \mathcal{F}_{k-1}] = 0 \quad \forall b, \forall a, \forall k \geq 1 \tag{47}$$

from which we can conclude that $E_k(b,a)$ is a martingale with respect to $(\mathcal{F}_k)_{k\geq 0}$ satisfying $E_0(b,a) = 0$. Hence, we can apply Theorem 1 with the uniform bound from Lemma 4 to conclude that, for any $\beta > 0$,

$$\begin{aligned}&\mathbb{P}\Big(\sup_{0\leq j\leq k} \|E_j\|_{\mathcal{E}_\delta,\infty} \geq \beta\Big) \\ &= \mathbb{P}\Big(\sup_{(b,a)\in\mathcal{E}_\delta\times\mathcal{A}} \sup_{0\leq j\leq k} |E_j(b,a)| \geq \beta\Big) \\ &= \mathbb{P}\Big(\bigcup_{(b,a)\in\mathcal{E}_\delta\times\mathcal{A}} \Big\{\sup_{0\leq j\leq k} |E_j(b,a)| \geq \beta\Big\}\Big) \\ &\leq \sum_{(b,a)\in\mathcal{E}_\delta\times\mathcal{A}} \mathbb{P}\Big(\sup_{0\leq j\leq k} |E_j(b,a)| \geq \beta\Big) \\ &= 2|\mathcal{E}_\delta||\mathcal{A}| \exp\Big(-\frac{2\beta^2}{(k+1)U^2}\Big).\end{aligned}$$

where U is the uniform error bound obtained in (44). Hence

$$\begin{aligned}\mathbb{P}\Big(\sup_{0\leq j\leq k} \|E_j\|_{\mathcal{E}_\delta,\infty} < \beta\Big) &\geq 1 - 2|\mathcal{E}_\delta||\mathcal{A}| \exp\Big[-\frac{2\beta^2}{(k+1)U^2}\Big] \\ &=: 1-\alpha\end{aligned}$$

and with probability at least $1-\alpha$ we have

$$\begin{aligned}\sum_{j=0}^{k} \gamma^{k-j} \|E_j\|_{\mathcal{E}_\delta,\infty} &\leq \sum_{j=0}^{k} \gamma^{k-j} \sup_{0\leq j\leq k} \|E_j\|_{\mathcal{E}_\delta,\infty} \\ \leq (1-\gamma)^{-1} \sup_{0\leq j\leq k} \|E_j\|_{\mathcal{E}_\delta,\infty} &\leq \frac{U}{1-\gamma}\sqrt{\frac{k+1}{2}\log\Big[\frac{2|\mathcal{E}_\delta||\mathcal{A}|}{\alpha}\Big]} \\ &\leq \frac{4\gamma\alpha}{1-\gamma}\sqrt{\frac{k+1}{2}\log\Big[\frac{2|\mathcal{E}_\delta||\mathcal{A}|}{\alpha}\Big]}.\end{aligned}$$

where $\alpha := 4Q_{\max} + \log(|\mathcal{A}|)/\eta$. Finally, we conclude from (**??**) that

$$\|Q^* - Q^{\hat{\pi}}_{\mathcal{E}_\delta,\rho_1}\|_\infty$$
$$\leq \frac{2}{(1-\gamma)(k+1)}\Big[\frac{\gamma\alpha}{1-\gamma} + \sum_{j=0}^{k}\gamma^{k-j}\|E_j\|_{\mathcal{E}_\delta,\infty}\Big]$$
$$\leq \frac{2\gamma B}{(1-\gamma)^2}\left[\frac{1}{k+1} + \frac{1}{1-\gamma}\sqrt{8\log\Big[\frac{2|\mathcal{E}_\delta||\mathcal{A}|}{\alpha}\Big]}\frac{1}{\sqrt{k+1}}\right] + \frac{\gamma\delta Q_{\max}}{(1-\gamma)}.$$

which concludes the proof.

4.1 A Maximal Azuma-Hoeffding Inequality

We employ a maximal version of the Azuma-Hoeffding inequality (see e.g. [Cesa-Bianchi and Lugosi, 2006]). It follows by replacing Markov's inequality with the Doob's maximal inequality for sub- or supermartingales (see [Doob, 1953] p. 314) in the proof of the standard (i.e non-maximal) version of the inequality (see e.g. [Hoeffding, 1963]). Intuitively, it bounds the likelihood of a martingale (or, more generally, a submartingale) having ever exceeded a given distance from its starting point, where the bound increases to one with the number of steps. As such, it can be seen as a concentration bound.

Theorem 1 (Maximal Azuma-Hoeffding Inequality)**.** *Let $(M_t)_{t\geq 0}$ be a discrete-time martingale with respect to a given filtration $\mathbb{F} = (\mathcal{F}_t)_{t\geq 0}$ on an arbitrary probability space $(\Omega, \mathcal{F}_\infty, \mathbb{P})$. Assume that there are $\mathbb{F}$-predictable processes processes $(A_t)_{t\geq 0}$ and $(B_t)_{t\geq 0}$ and constants $0 < c_t < +\infty$ such that:*

$$A_t \leq M_t - M_{t-1} \leq B_t \quad \textit{and} \quad B_t - A_t \leq c_t \quad P\textit{-a.s.}. \tag{48}$$

Then for all $\beta > 0$

$$\mathbb{P}\left[\sup_{0\leq s\leq t}(M_s - M_0) \geq \beta\right] \leq \exp\left(-\frac{2\beta^2}{\sum_{0\leq s\leq t} c_s^2}\right). \tag{49}$$

References

[Bertsekas, 2008] D. P. Bertsekas. *Dynamic Programming and Optimal Control*, volume II. Athena Scientific, 3rd edition, 2008.

[Cesa-Bianchi and Lugosi, 2006] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[Doob, 1953] J. L. Doob. *Stochastic Processes*. Wiley, 1953.

[Hoeffding, 1963] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[Kolmogorov and Fomin, 1970] A. N. Kolmogorov and S. V. Fomin. *Introductory Real Analysis*. Dover, 1970.

[Lee *et al.*, 2007] Wee Lee, Nan Rong, and David Hsu. What makes some POMDP problems easy to approximate? In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007.

[MacKay, 2003] D. J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

[Ross, 1970] S. M. Ross. *Applied Probability Models with Optimization Applications*. Dover, 1970.