Bayesian Reinforcement Learning for Single-Episode Missions in Partially Unknown Environments Appendix

1 List of acronyms and abbreviations

AUV	Autonomous underwater vehicle
BAMCP	Bayes-adaptive Monte-Carlo planning
BAMDP	Bayes-adaptive Markov decision process
BO	Bayesian optimisation
BRL	Bayesian reinforcement learning
GNC	Guidance, navigation, control
GP	Gaussian process
GPDM	Gaussian process dynamical model
MCTS	Monte-Carlo tree search
POMDP	Partially observable Markov decision process
ROS	Robot operating system
SSP	Stochastic shortest path
U-MDP	MDP with unknown feature values

2 Proof of Proposition 2

Proof. With individual belief updates at every stage, the history density is (shortening $\mathcal{P}_{\pi}^{h_t}(h_{t+\tau})$ to $\mathcal{P}_{\pi}^{h_t}$ and $\tilde{\mathcal{P}}_{\pi}^{h_t}(h_{t+\tau})$ to $\tilde{\mathcal{P}}_{\pi}^{h_t}$):

$$\begin{aligned}
\mathcal{P}_{\pi}^{h_t} &= p(a_t s_{t+1} a_{t+1} ... s_{t+\tau} \mid h_t, \pi) \\
&= p(a_t \mid h_t, \pi) p(s_{t+1} \mid h_t, \pi, a_t) p(a_{t+1} \mid h_{t+1}, \pi) ... p(s_{t+\tau} \mid h_{t+\tau-1}, a_{t+\tau}, \pi) \qquad (5) \\
&= \prod_{t \le t' < t+\tau} \pi(h_{t'}, a_{t'}) \cdot \prod_{t < t' \le t+\tau} p(s_{t'} \mid h_{t'-1}, a_{t'-1}) \qquad (6) \\
&= \prod_{t \le t' < t+\tau} \pi(h_{t'}, a_{t'}) \ \cdot \prod_{t < t' \le t+\tau} \int_{T} T(s_{t'-1}, a_{t'-1}, s_{t'}) p(T \mid h_{t'-1}) dT, \qquad (7)
\end{aligned}$$

where $s_t = (s_k, s_e)$.

Given the definition of T^{+} in (2), and the fact that a history h_t uniquely specifies a GP $\mathcal{GP}_{\mathcal{D}_h}$ of observations up to time t:

$$\mathcal{P}_{\pi}^{h_t} = \prod_{t \le t' < t+\tau} \pi(h_{t'}, a_{t'}) \ \cdot \prod_{t < t' \le t+\tau} \left[T^{o}\big(s_{t'-1}, a, s_{k,t'}\big) p^{\mathcal{GP}}\big(s_{e,t'} \mid s_{k,t'}, \mathcal{D}_{t'-1}\big) \right]. \qquad (8)$$

The GP posterior $p^{\mathcal{GP}}\big(s_{e,t'} \mid s_{k,t'}, \mathcal{D}_{t'-1}\big)$ is a multivariate normal distribution (MVN). A GP belief update with a noise-free sampled observation is performed by conditioning the posterior MVN on the

6th Conference on Robot Learning (CoRL 2022), Auckland, New Zealand.

sampled value (for compactness we remove s_k from the MVN probability density function $p^{\mathcal{GP}}$):

$$\mathcal{P}_\pi^{h_t} = \prod_{t \le t' < t+\tau} \pi(h_{t'}, a_{t'}) \cdot \prod_{t < t' \le t+\tau} T^o\big(s_{t'-1}, a, s_{k,t'}\big) \cdot \\ \big[p^{\mathcal{GP}}\big(s_{e,t+1} \mid \mathcal{D}_t\big) \cdot \prod_{\substack{t+1<t' \\ \le t+\tau}} p\big(s_{e,t'} \mid s_{e,t'-1}, \dots, s_{e,t+1}\big)\big]. \tag{9}$$

The repeated belief update product in the square brackets in (9) can be recognised as being equivalent (via the chain rule for probability) as being equivalent to the joint distribution across all values of $s_{e,t'}$:

$$\big[\cdots\big] = p^{\mathcal{GP}}\big(s_{e,t+1}, \dots, s_{e,t+\tau} \mid \mathcal{D}_t\big). \tag{10}$$

Therefore the rollout distribution is identical between individual belief updates and root sampling:

$$\mathcal{P}_\pi^{h_t} = \prod_{t \le t' < t+\tau} \pi(h_{t'}, a_{t'}) \cdot \prod_{t < t' \le t+\tau} T^o\big(s_{t'-1}, a, s_{k,t'}\big) \cdot p^{\mathcal{GP}}\big(s_{e,t+1}, \dots, s_{e,t+\tau} \mid \mathcal{D}_t\big) = \tilde{\mathcal{P}}_\pi^{h_t}. \tag{11}$$

Given the rollout distribution equivalence, search tree node statistics for both methods will converge to the same values in expectation. □

3 Radiation Domain Details

3.1 Radiation simulation

Radiation is simulated using $1/r^2$ "solid angle" radiation physics. Radiation sources $\{(s_i, \mathbf{x}_i), \dots\}_{i=0}^n$ have strength s_i and pose $\mathbf{x}_i$. Source strength is the exposure value at a distance of 1m. The radiation exposure $\lambda(\mathbf{x})$ at robot pose $\mathbf{x}$ from these radiation sources is then

$$\lambda(\mathbf{x}) = \sum_i^n \frac{s_i^{src}}{\|\mathbf{x} - \mathbf{x}_i^{src}\|^2}. \tag{12}$$

3.2 Domain Details

A problem instance consists of the following components: **a)** a randomly generated distribution of radiation sources in the environment (below), **b)** the start location in the grid map, sampled from a uniform distribution across the map, and **c)** 3 goal states, also uniformly sampled. Sampled problem instances are discarded when they result in trivial solutions (e.g. due to the start and goal locations being too close) or where one goal is significantly closer to the start location than the other sampled goals.

Random radiation fields are generated in both of the following ways:

1. Random point-source distribution: insert between 5 to 20 radiation sources (uniform random sampling), with randomly sampled z position values $z \in \{1.0, 1.5, 2.5\}$ and randomly sampled strengths $s \in \{1000, 2000, 5000, 10000\}$. x and y position values are sampled uniformly within the bounds of the map ± 2.0m.
2. Randomly generated Gaussian random field distribution: evenly cover the map with radiation sources at $z = 1.0$ and draw their log strengths from a Gaussian random field. The Gaussian random field is generated with a radial basis function kernel, using uniformly sampled lengthscale hyperparameter $l \in \{3.0, 5.0, 7.0\}$ and variance hyperparameter $\sigma \in \{60, 75, 90\}$.

3.3 Gazebo simulation

A visualisation of the reactor room world in Gazebo is shown in Figure 4. The simulated robot is a Clearpath Jackal[1], using standard Gazebo lidar and odometry sensor simulation and standard ROS components such as AMCL localisation[2].

[1] https://clearpathrobotics.com/jackal-small-unmanned-ground-vehicle/

[2] http://wiki.ros.org/amcl

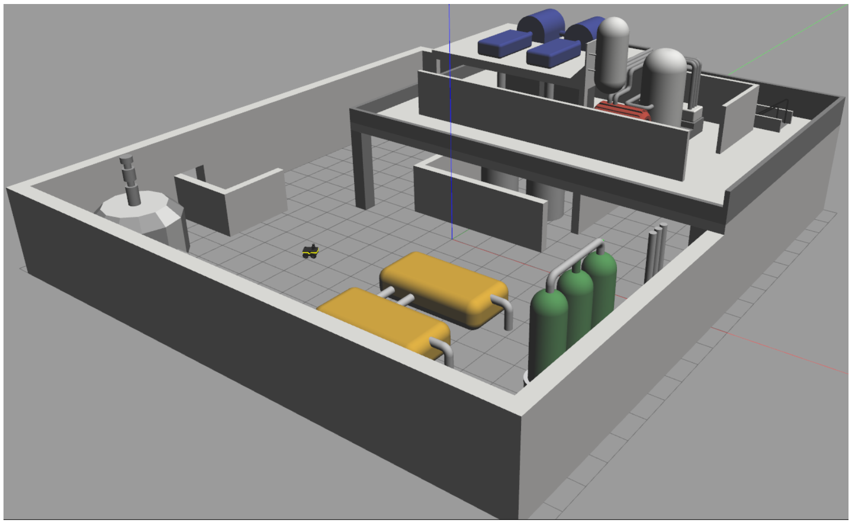

Figure 4: Visualisation of the reactor room Gazebo world in the middle of a simulated mission.

3.4 Algorithm parameters

The UCT exploration constant was dynamically set to equal the value of the decision node multiplied by 1.414. The guided rollout policy attempts to minimise the L2 distance from the goal state at the next state:

$$\pi_{rollout}(s_k) = \underset{a \in A(s_k)}{\arg\min}\, \text{dist}\,(s'_k, s_g)\, \forall s_g \in G, \tag{13}$$

where $A(s_k)$ are the enabled actions at a state and dist is a function that returns the shortest grid map path distance between two states in the grid map. The value of the ϵ parameter was 1.0 in the log GP space, meaning that MCTS search nodes covers increasingly large ranges of continuous value with increasing radiation level.

3.5 U-MDP Cost and Transition Structures

Transitions between grid map states are assumed to be deterministic as robot navigation does not fail in this easy-to-localise environment.

The state space $S^o = S_k \times S_e$ where $\{x, y\} \subseteq S_k$ and $rad\_exp \in S_e$. The set of actions $A^o = \{\text{left, left-up, left, up, right-up, right, right-down, down, left-down}\}$.

The cost structure is

$$C^o\big((s_k, s_e), a\big) = rad\_exp \cdot (\text{mcts-time} + d_a/v_{robot})\,, \tag{14}$$

where mcts-time is the MCTS time budget allocated to the algorithm, d_a is the travel distance associated with the action (diagonal actions are 1.414 times further), and v_{robot} is the manually estimated average velocity of the simulated robot, taken to be $= 0.3\text{ms}^{-1}$.

3.6 GP model

The GP is a single-output GP that models $o : \mathbb{R}^2 \to \text{Dist}(\mathbb{R})$ where the GP input is $\{x, y\} \subseteq S_k$. The GP is trained on log radiation measurements and predicts log radiation levels.

The GP kernel is a combination of a bias kernel and a radial basis function kernel:

$$k(s_k, s'_k) = \sigma^2 \exp\left(-\frac{\|s_k - s'_k\|^2}{2l^2}\right). \tag{15}$$

The radial basis function kernel hyperparameters are assigned the following uninformative Gamma distribution priors:

GP hyperparameter	Prior
Lengthscale l	$\Gamma(1, 0.5)$
Variance σ	$\Gamma(1, 4)$

This corresponds roughly to an expected lengthscale of 2m (standard deviation 2m) and expected log magnitude variance of 0.5 (standard deviation 0.5). At the beginning of each experiment the agent is provided with observations at the start state and two immediately neighbouring states as prior knowledge.

4 Underwater Currents Experiment

4.1 Domain Details

A problem instance consists of the following components: **a)** a 10km × 10km current field drawn from the real-world currents dataset in the same manner as [12], sampled from a fixed set of 12 fields where there is some variation in current across the field (rather than e.g. consistent current in one direction), **b)** the AUV start location, sampled from a uniform distribution across the field, and **c)** between 1 and 3 (with uniform probability) goal states, also uniformly sampled. Sampled problem instances are discarded when they result in trivial solutions (e.g. due to the start and goal locations being too close) or where one goal is significantly closer to the start location than the other sampled goals.

This experiment makes use of E.U. Copernicus Marine Service Information[3]. This is a dataset of north/east ocean current vectors on a 1500m spaced grid. To allow sampling of ground truth values at locations other than the grid locations, interpolation of the dataset using a spatio-temporal GP is carried out in the same manner as [12]. The currents experiment, the dataset used covers the region from approximately 47 to 62 latitude and -12 to 5 longitude. The dataset was originally collected on May 1 2020.

4.2 Kinematic GNC simulation

The vehicle's movement is determined by a kinematic calculation given of the vehicle's yaw, pitch and velocity control demands, process noise and the currents acting on the vehicle. Currents are drawn from the currents dataset at the vehicle's simulated true position. Underwater localisation is via an underwater acoustic beacon-aided extended Kalman filter. The vehicle uses this acoustic time-of-flight position feedback to navigate to the target location of the action selected by the MCTS planner. An example detailed run through the kinematic GNC simulator is shown in Figure 5.

4.3 Algorithm parameters

The UCT exploration constant was dynamically set to equal the value of the decision node multiplied by 1.414. The guided rollout policy attempts to minimise the L2 distance from the goal state at the next state:

$$\pi_{rollout}(s_k) = \underset{a \in A(s_k)}{\arg\min} \|s'_k - s_g\|_2 \forall s_g \in G, \tag{16}$$

where $A(s_k)$ are the enabled actions at a state. The value of the ϵ parameter was 0.1.

4.4 U-MDP Cost and Transition Structures

Transitions between grid map states are assumed to be deterministic as the AUV will always eventually reach a specified goal. The state space $S^o = S_k \times S_e$ where $\{x, y\} \subseteq S_k$ and $\{v_x, v_y\} \in S_e$. The set of actions $A^o = \{0°, 60°, 120°, 180°, 240°, 300°, \}$ corresponds to the direction the vehicle wishes to travel to the hex grid state in that direction. The vehicle travels against the current such that its net movement is in the action-specified direction.

The cost is the time taken to carry out the transition, given a constant vehicle speed $v = 0.6$m s^{-1}, the currents acting on the vehicle, and the requirement from the navigation controller that the net

[3] Available: https://resources.marine.copernicus.eu/?option=com_csw&view=details&product_id=NORTHWESTSHELF_ANALYSIS_FORECAST_PHY_004_013. . Accessed 2021-08.

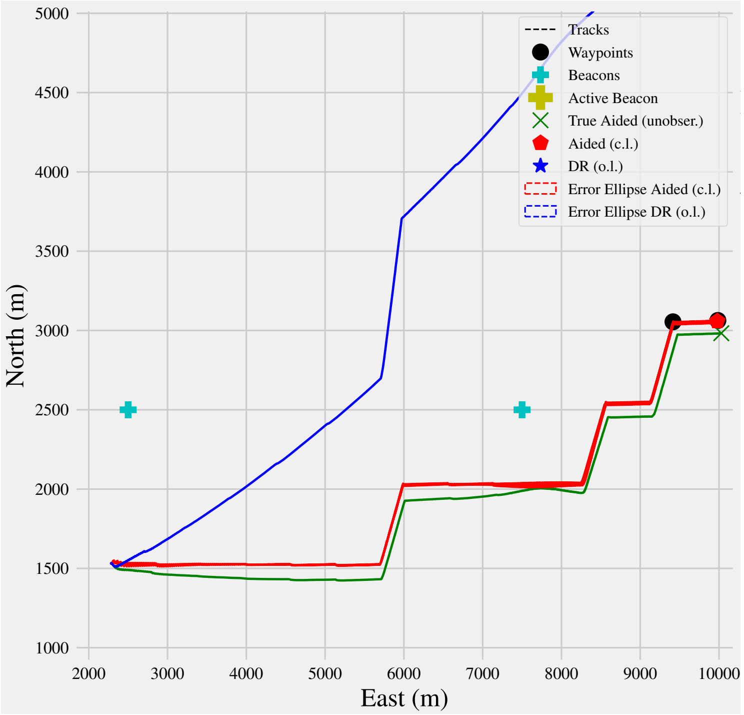

Figure 5: Example run through kinematics / GNC simulator. The green line is the vehicle's true path, the red line is the demanded path the vehicle has attempted to follow (where the demanded path is a result of an MCTS action selection at each grid MDP state), and the blue line is the "dead-reckoned/odometry-only" position estimate for the AUV. This is where the vehicle would localise to without any external EKF acoustic time-of-flight feedback.

direction of travel is in the action-specified direction. Given the chosen action this is solved to find the net velocity v_a^{eff} and the direction the vehicle must steer against the current at the state.

$$C^o\big((s_k, s_e), a\big) = \frac{d_s}{v_a^{eff}}, \tag{17}$$

where d_s is the distance between states in the hexagonal grid.

4.5 GP model

The GP is a vector-output coregionalised GP that models $o : \mathbb{R}^2 \to \text{Dist}(\mathbb{R}^2)$ where the GP input is $\{x, y\} \subseteq S_k$.

The GP kernel is the sum of a bias kernel and a radial basis function kernel. The RBF hyperparameters are assigned the following relatively broad Gamma distribution priors based on sensible current values seen in the whole currents dataset, and distribution of optimised lengthscale parameters seen when GPs are trained on a small number of random subsets of the dataset:

GP hyperparameter	Prior
Lengthscale l	$\Gamma(a = 49.0, b = 0.014)$
Variance σ	$\Gamma(a = 1.0, b = 4.0)$

This corresponds roughly to an expected lengthscale of 3500m (standard deviation 500m) and expected current magnitude variance of 0.25 (standard deviation 0.0625). At the beginning of each experiment the agent is provided with observations at the start state and two immediately neighbouring states as prior knowledge.