Information-Guided Robotic Maximum Seek-and-Sample in Partially Observable Continuous Environments

Genevieve Flaspohler*1,2 Victoria Preston*1,2 Anna P.M. Michel2 Yogesh Girdhar2 Nicholas Roy1

Abstract—We present PLUMES, a planner for localizing and collecting samples at the global maximum of an a priori unknown and partially observable continuous environment. This “maximum seek-and-sample” (MSS) problem is pervasive in the environmental and earth sciences. Experts want to collect scientifically valuable samples at an environmental maximum (e.g., an oil spill source), but do not have prior knowledge about the phenomenon’s distribution. We formulate the MSS problem as a partially-observable Markov decision process (POMDP) with continuous state and observation spaces, and a sparse reward signal. To solve the MSS POMDP, PLUMES uses an information-theoretic reward heuristic with continuous-observation Monte Carlo Tree Search to efficiently localize and sample from the global maximum. In simulation and field experiments, PLUMES collects more scientifically valuable samples than state-of-the-art planners in a diverse set of environments, with various platforms, sensors, and challenging real-world conditions.

Index Terms—Motion and Path Planning, Learning and Adaptive Systems, Reactive and Sensor-Based Planning

I. INTRODUCTION

In many environmental and earth science applications, experts want to collect scientifically valuable samples of a maximum (e.g., an oil spill source), but the distribution of the phenomenon is initially unknown. This maximum seek-and-sample (MSS) problem is pervasive. Canonically, samples are collected at predetermined locations by a technician or by a mobile platform following a uniform coverage trajectory. These non-adaptive strategies result in sample sparsity at the maximum and may be infeasible when the geometric structure of the environment is unknown (e.g., boulder fields) or changing (e.g., tidal zones). Increasing the number of valuable samples at the maximum requires adaptive online planning and execution. We present PLUMES — Plume Localization under Uncertainty using Maximum-ValuE information and Search — an adaptive algorithm that enables a mobile robot to efficiently localize and densely sample an environmental maximum, subject to practical challenges including dynamic constraints, unknown geometric map and obstacles, and noisy sensors with limited field-of-view. Fig. 1 shows a motivating application of coral head localization.

Informative Path Planning: The MSS problem is closely related to informative path planning (IPP) problems. Offline IPP techniques for pure information-gathering that optimize submodular coverage objectives can achieve near-optimal performance [1], [2]. However, in the MSS problem, the value of a sample depends on the unknown maximum location, requiring a robot to adaptively select actions that explore to localize the maximum and then seamlessly transition to selecting actions that exploitatively collect valuable samples there. Even for adaptive IPP methods, the MSS problem presents considerable challenges. The target environmental phenomenon is partially observable and most directly modeled as a continuous scalar function. Additionally, efficient maximum sampling with a mobile robot requires consideration of vehicle dynamics, travel cost, and a potentially unknown obstacle map. Handling these challenges in combination excludes adaptive IPP algorithms that use discrete state spaces [3], [4], known metric maps [5], [6], or unconstrained sensor placement [7].

The MSS POMDP: Partially-observable Markov decision processes (POMDPs) are general models for decision-making under uncertainty that allow the challenging aspects of the
MSS problem to be encoded. We define the MSS POMDP in which the partially observable state represents the continuous environmental phenomenon and a sparse reward function encodes the MSS scientific objective by giving reward only to samples sufficiently close to the global maximum. Solving a POMDP exactly is generally intractable, and the MSS POMDP is additionally complicated by both continuous state and observation spaces, and the sparse MSS reward function. This presents the two core challenges that PLUMES addresses: performing online search in a belief-space over continuous functions, and overcoming reward function sparsity.

Planning over Continuous Domains: In the MSS problem, the state of the environment can be modeled as a continuous function. PLUMES uses a Gaussian Process (GP) model to represent the belief over this continuous function, and must plan over the uncountable set of possible GP beliefs that arise from future continuous observations. To address planning in continuous spaces, state-of-the-art online POMDP solvers use deterministic discretization [8] or a combination of sampling techniques and particle filter belief representations [9]–[12]. Efficiently discretizing or maintaining a sufficiently rich particle set to represent the underlying continuous function in MSS applications is itself a challenging problem, and can lead to inaccurate inference of the maximum [13]. Other approaches have considered using the maximum-likelihood observation or any analytic form.

Rewards and Heuristics: In the MSS POMDP, the reward function is sparse and does not explicitly encode the value of exploration. Planning with sparse rewards requires long-horizon information gathering and is an open problem in robotics [18]. To alleviate this difficulty, less sparse heuristic reward functions can be optimized in place of the true reward, but these heuristics need to be selected carefully to ensure the planner performs well with respect to the true objective. In IPP, heuristics based on the value of information have been applied successfully [7], [14], [19], [20], primarily using the GP-UCB criteria [1], [21]. We demonstrate that within practical mission constraints, using UCB as the heuristic to the GP-UCB criteria [1], [21]. We demonstrate that within practical mission constraints, using UCB as the heuristic performs well with respect to the true objective. In IPP, heuristics based on the value of information have been applied successfully [7], [14], [19], [20], primarily using the GP-UCB criteria [1], [21]. We demonstrate that within practical mission constraints, using UCB as the heuristic performs well with respect to the true objective. In IPP, heuristics based on the value of information have been applied successfully [7], [14], [19], [20], primarily using the GP-UCB criteria [1], [21].

The contribution of this paper is the MSS POMDP formalism and the corresponding PLUMES planner, which by virtue of its belief model, information-theoretic reward heuristic, and search framework, enables efficient maximum sampling with asymptotic optimality guarantees for continuous environments. PLUMES extends the state-of-the-art in MSS planners by applying a BO heuristic reward function to MSS that alleviates the challenges of the true sparse MSS reward function, and integrating GP belief representations within continuous-observation MCTS. The utility of PLUMES for MSS applications is demonstrated in extensive simulation and field trials, showing a statistically significant performance improvement over state-of-the-art baselines.

II. Maximum Seek-and-Sample POMDP

We formalize the MSS problem by considering a target environment domain as a d-dimensional compact set \( \mathbb{X}_w \subset \mathbb{R}^d \). We allow \( \mathbb{X}_w \) to contain obstacles with arbitrary geometry and let \( \mathbb{X} \subset \mathbb{X}_w \) be the set of reachable points with respect to the robot’s initial pose. We assume there is an unknown underlying continuous function \( f : \mathbb{X}_w \rightarrow \mathbb{R} \) representing the value of a continuous phenomenon of interest. The objective is to find the unique global maximizer \( x^* = \arg\max_{x \in \mathbb{X}} f(x) \) by safely navigating while receiving noisy observations of this function \( f \). Because \( f \) is unknown, we cannot access derivative information or any analytic form.

We model the process of navigating and generating observations as the MSS POMDP: an 8-tuple \((\mathcal{S}, \mathcal{A}, \mathcal{Z}, T, O, R, \gamma, b_0)\):

- \( \mathcal{S} \): continuous state space of the robot and environment
- \( \mathcal{A} \): discrete set of action primitives
- \( \mathcal{Z} \): continuous space of possible observations
- \( T: \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{P}(\mathcal{S}) \), the transition function, i.e., \( \Pr(S_{t+1} = s' | S_t = s, A_t = a) \)
- \( O: \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{P}(\mathcal{Z}) \), the observation model, i.e., \( \Pr(Z_{t+1} = z | S_{t+1} = s, A_t = a) \)
- \( R: \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R} \), the reward of taking action \( a \) when robot’s state is \( s \), i.e., \( R(s, a) \)
- \( \gamma \): discount factor, \( 0 \leq \gamma \leq 1 \)
- \( b_0 \): initial belief state of the robot, \( b_0 \in \mathcal{P}(S_0) \)

where \( \mathcal{P}(\cdot) \) denotes the space of probability distributions over the argument.

The Bellman equation is used to recursively quantify the value of belief \( b_t = \mathcal{P}(S_t) \) over a finite horizon \( h \) under policy \( \pi: b_t \rightarrow a_t \) as:

\[
V^\pi_h(b_t) = \mathbb{E}[R(s_t, \pi(b_t))] + \gamma \int_{z \in \mathbb{Z}} V^\pi_{h-1}(b_{t+1}^z) \Pr(z | b_t, \pi(b_t)) \, dz,
\]

where the expectation is taken over the current belief and \( b_{t+1}^z \) is the updated belief after taking action \( \pi(b_t) \) and observing \( z \in \mathbb{Z} \). The optimal policy \( \pi^*_h \) over horizon-\( h \) is the maximizer of the value function over the space of possible policies \( \Pi: \pi^*_h = \arg\max_{\pi \in \Pi} V^\pi_h(b_t) \). However, Eq. 1 is intractable to compute in continuous state and observation spaces; the optimal policy must be approximated. PLUMES uses a receding-horizon, online POMDP planner and heuristic reward function to approximately solve the MSS POMDP in real-time on robotic systems.

III. THE PLUMES ALGORITHM

PLUMES is an online planning algorithm with a sequential decision-making structure:
1) Conditioned on $b_t$, approximate the optimal policy $\pi^*_h$ for finite horizon $h$ and execute the action $a = \hat{\pi}^*_h(b_t)$.

2) Collect observations $z \in Z$, according to $O$.

3) Update $b_t$ to incorporate this new observation; repeat.

In the following sections, we define the specific choice of belief model, search algorithm, and heuristic reward function PLUMES uses to solve the MSS POMDP.

**A. Gaussian Process Belief Model**

We assume the robot’s pose $x_t$ at planning iteration $t$ is fully observable, and the unknown environmental phenomenon $f$ is partially observable. The full belief-state is represented as a tuple $b_t$ of robot pose $x_t$ and environment belief $g_t = \mathcal{P}(f)$ at time $t$. Because $f$ is a continuous function, we cannot represent the belief $g_t$ as a distribution over discrete states, as is standard in POMDP literature [23], and must choose an alternate representation. PLUMES uses a Gaussian process (GP) [24] to represent $g_t$ conditioned on a history of past observations. This GP is parameterized by mean $\mu(x)$ and covariance function $\kappa(x, x')$.

As the robot traverses a location $x_t$, it gathers observations $z \in Z$ of $f$ subject to sensor noise $\sigma^2_z$, such that $z = f(x_t) + \epsilon$ with $\epsilon \sim N(0, \sigma^2_z)$. Given a history $D_t = \{x_i, z_i\}_{i=0}^D$ of $D$ observations and observation locations at planning iteration $t$, the posterior belief at a new location $x_t' \in \mathcal{X}$ is computed:

$$g_t(x_t') | D_t \sim \mathcal{N}(\mu_t(x_t'), \sigma^2_t(x_t'))$$

For $\mu_t(x_t') = \kappa_t(x_t', \mathcal{X})^T (K_t + \sigma^2_z I)^{-1} z_t$,

$$\sigma^2_t(x_t') = \kappa_t(x_t', x_t') - \kappa_t(x_t', \mathcal{X})^T (K_t + \sigma^2_z I)^{-1} \kappa_t(\mathcal{X}),$$

where $z_t = [z_0, \ldots, z_{D-1}]^T, K_t$ is the positive definite kernel matrix with $K_{[i, j]} = \kappa(x_i, x_j)$ for all $x_i, x_j \in D_t$, and $\kappa_t(x, \mathcal{X}) = [\kappa(x_0, x'), \ldots, \kappa(x_{D-1}, x')]^T$.

**B. Planning with Continuous-Observation MCTS**

PLUMES selects high-reward actions with receding-horizon search over possible belief states. This search requires a simulator that can sample observations and generate beliefs given a proposed action sequence. For PLUMES, this simulator is the GP model, which represents the belief over the continuous function $f$, and in turn generates continuous observations from action queries using Eq. 3 & 4.

PLUMES uses continuous-observation MCTS to overcome the challenges of planning in continuous state and observation spaces. There are three stages: selection, forward simulation, and back-propagation. Each node in the tree can be represented as the tuple of robot pose and GP belief, $b_t = \{x_t, g_t\}$. Additionally, we will refer to two types of nodes: belief nodes and belief-action nodes. The root of the tree is always a belief node, which represents the entire history of actions and observations up through the current planning iteration. Through selection and simulation, belief and belief-action nodes are alternately added to the tree (Fig. 2).

From the root, a rollout begins with the selection stage, in which a belief-action child is selected according to the Polynomial Upper Confidence Tree (PUCT) policy [17]. The PUCT value $Q^*_\text{aug}(b_t, a)$ is the sum of the average heuristic rewards (i.e., MVI) from all previous simulations and a term that favors less-simulated action sequences:

$$Q^*_\text{aug}(b_t, a) = \hat{Q}^*(b_t, a) + \sqrt{\frac{N(b_t)}{a}}$$

where $\hat{Q}^*(b_t, a)$ is the average heuristic reward of choosing action $a$ with belief $b_t$ in all previous rollouts, $N(b_t)$ is the number of times the node $b_t$ has been simulated, $N(b_t, a)$ is the number of times that particular action from node $b_t$ has been selected, and $\epsilon_d$ is a depth-dependent parameter.

Once a child belief-action node is selected, the action associated with the child is forward simulated using the generative observation model $O$, and a new belief node is generated $b_{t+1} = \{x_{t+1}, g_{t+1}\}$ as though the action were taken and samples observed. The simulated observations are drawn from the belief-action node’s GP model $g_t$, and the robot’s pose is updated deterministically based on the selected action. Since the observations in a GP are continuous, every sampled observation is unique with probability one. Progressive widening, with depth-dependent parameter $\alpha_d$, incrementally grows the tree by limiting the number of belief children of each belief-action node. When growing the tree, $b_{t+1}$ is either chosen to be the least visited node if $[N(b_t, a)]^{\alpha_d} = [(N(b_t) - 1)^{\alpha_d}]$, or otherwise is a new child with observations simulated from $b_t$. By limiting the width of the search tree and incrementally growing the number of explored children, progressive widening avoids search degeneracy in continuous environments.

Once a sequence of actions has been rolled out to a horizon $h$, the accumulated heuristic reward is propagated upward from the leaves to the tree root. The average accumulated heuristic reward and number of queries are updated for each node visited in the rollout. Rollouts continue until the computation budget is exhausted. The most visited belief-action child of the root node is executed.

Continuous-observation MCTS within PLUMES provides both practical and theoretical benefits. Practically, progressive-widening directly addresses search degeneracy by visiting belief nodes multiple times even in continuous observation spaces, allowing for a more representative estimate of their

*Refer to Table 1 of Auger et al. [17] for parameter settings.
value. Theoretically, PLUMES can be shown to select asymptotically optimal actions. We briefly describe how analysis in Auger et al. [17] for PUCT-MCTS with progressive widening in MDPs can be extended to PLUMES.

Using standard methods [23], we can reduce the MSS POMDP to an equivalent belief-state MDP. This belief-state MDP has a state space equal to the set of all possible beliefs, and a transition distribution that captures the effect of both the dynamics and the observation model after each action. Planning in this representation is often intractable as the state space is continuous and infinite-dimensional. However, PLUMES plans directly in the belief-state MDP by using its state space is continuous and infinite-dimensional. However, planning in this representation is often intractable as the state space is continuous and infinite-dimensional. However, PLUMES plans directly in the belief-state MDP by using its

Subsequently, Theorem 1 in Auger et al. [17] shows that for an MDP with a continuous state space, like the belief-state MDP representation suggested, the value function estimated by continuous-observation MCTS asymptotically converges to that of the optimal policy:

\[
\left| \hat{Q}_N^*(b_t, a) - Q_k^*(b_t, a) \right| \leq \frac{C}{N(b_t, a)^{\gamma_d}},
\]

with high probability [17], for constants \( C > 0 \) and \( \gamma_d^* \).

C. Maximum-Value Information Reward

The true state-dependent reward function for the MSS POMDP would place value on collecting sample points \( x \) within an \( \epsilon \)-ball of the true global maximum \( x^* \):

\[
R(f, x) = I_{||x - x^*|| < \epsilon},
\]

where \( \epsilon \) is determined by the scientific application. Optimizing this sparse reward function directly is challenging, so PLUMES approximates the true MSS reward by using the maximum-value information (MVI) heuristic reward [22],

MVI initially encourages exploration behavior, but ultimately rewards exploitative sampling near the inferred maximum.

The belief-dependent MVI heuristic reward \( \tilde{R}(b_t, x) \) quantifies the expected value of having belief \( b_t \) and collecting a sample at location \( x \in \mathbb{X} \). MVI reward quantifies the mutual information between the random variable \( Z \), representing the observation at location \( x \), and \( Z^* \), the random variable representing the value of the function \( f \) at the global maximum:

\[
\tilde{R}(b_t, x) = I(\{x, Z\}; Z^* | b_t),
\]

where \( Z^* = \max_{x' \in \mathbb{X}} f(x') \). To compute the reward of collecting a random observation \( Z \) at location \( x \) under belief \( b_t \), we approximate the expectation over the unknown \( Z^* \) by sampling from the posterior distribution \( z^*_i \sim p(Z^* \mid b_t) \) and use Monte Carlo integration with \( M \) samples [22]:

\[
\tilde{R}(b_t, x) = H[\Pr(Z \mid x, b_t)] - \mathbb{E}_{z' \sim \Pr(Z^* \mid b_t)}[H[\Pr(Z \mid x, b_t, Z^* = z')]],
\]

\[
\approx H[\Pr(Z \mid x, b_t)] - \frac{1}{M} \sum_{i=0}^{M} H[\Pr(Z \mid x, b_t, Z^* = z^*_i)].
\]

Each entropy expression \( H[\cdot] \) can be respectively approximated as the entropy of a Gaussian random variable with mean and variance given by the GP equations (Eq. 3 & 4), and the entropy of a truncated Gaussian, with upper limit \( z^*_i \) and the same mean and variance.

To draw samples \( z^*_i \) from the posterior \( p(Z^* \mid b_t) \), we employ spectral sampling [25]. Spectral sampling draws a function \( f \), which has analytic form and is differentiable, from the posterior belief of a GP with stationary covariance function [22], [26]. To complete the evaluation of Eq. 10, \( z^*_i \sim p(Z^* \mid b_t) \) can be computed by applying standard efficient global optimization techniques (e.g., sequential least
shown to be sensitive with respect to $\beta$ MCTS [14]. The performance of UCB planners has been myopic planning with UCB reward (UCB-Myopic) [19], and environmental surveying: non-adaptive lawnmower-coverage breadth of MSS scenarios that feature convex and non-convex and UCB can be observed in Fig. 3. The difference in convergence characteristics between MVI in a geographically bounded, obstacle-free environment. In

In marine and atmospheric applications, MSS often occurs in a geographically bounded, obstacle-free environment. In

To evaluate the mission performance of all planners, we report accumulated MSS reward (Eq. 7), which directly corresponds to the number of scientifically valuable samples collected within an $\epsilon$-ball of the true maximum. This metric is reported for all trial scenarios in Table I. We additionally report several metrics commonly used in IPP to evaluate posterior model quality: overall environmental posterior root mean-squared error (RMSE) and error in posterior prediction of $x^*$ at the end of a mission ($x^*$ error). We use a Mann-Whitney U non-parametric significance test [29] to report statistical significance ($p = 0.05$ level) in performance between PLUMES and baseline algorithms.

A. Bounded Convex Environments

In marine and atmospheric applications, MSS often occurs in a geographically bounded, obstacle-free environment. In

![Fig. 4. Simulation Environments: The multimodal simulated $10m \times 10m$ environments. Yellow regions are high-valued; blue regions are low-valued. The global maximum is marked with a star. The left and center environments represent convex-worlds (Section IV-A), while the right environment is representative of a non-convex world (Section IV-B).](image)

50 simulated trials, we applied PLUMES and our baseline planners to a point robot in a $10m \times 10m$ multimodal environment drawn randomly from a GP prior with a squared-exponential covariance function and zero mean ($\ell = 1.0, \sigma^2 = 100.0, \sigma^2_n = 1.0 \{1\%\}$) (see Fig.4). The action set consisted of ten viable trajectories centered at the robot’s pose with path length 1.5 m, and samples are collected every 0.5 m of travel. Mission lengths were budgeted to be 200 m. Nonmyopic planners rolled out to a 5-action horizon and were allowed 250 rollouts per planning iteration. Summary simulation results are presented in Table I.

In these trials, PLUMES accumulated significantly (0.05-level) more reward than baselines. The distribution of accumulated reward (Fig. 5) shows that PLUMES has a single dominating mode near reward 200 and few low-performing missions (reward <50). In contrast, both UCB-based methods have distributions which are multimodal, with non-trivial modes in the low-performance region. Boustro. collected consistently few scientifically valuable samples. In addition to collecting many more samples at the maximum, PLUMES achieved statistically indistinguishable levels of posterior RMSE and $x^*$ error compared to baselines (Table I).

The corresponding field trial for convex-world maximum-search was performed in the Bellairs Fringing Reef, Barbados

![Fig. 5. Distribution of accumulated MSS reward in 50 convex-world simulations: Accumulated MSS reward is calculated for each trial and the distribution for each planner is plotted as a kernel density estimate (solid line). The dashed lines represent the median accumulated reward for each planner (reported in Table I). The gray area of the plot indicates a low performance region where the planner collected <50 samples near the maximum. PLUMES has a single mode near 200, whereas both UCB-based methods are multimodal, with modes in the low performance region.](image)
by a custom-built autonomous surface vehicle (ASV) with the objective of localizing the most exposed coral head. Coral head exposure is used to select vantage points for coral imaging [30] and in ultraviolet radiation studies on coral organisms [31]. Due to time and resource constraints, only one trial of two planners was feasible on the physical reef; we elected to demonstrate PLUMES and Boustro, one of the most canonical surveying strategies in marine sciences.

The ASV (1 m × 0.5 m) had holonomic dynamics and a downward-facing acoustic point altimeter (Tritech Micron Echosounder) with returns at 1 Hz. Ten dynamically-feasible 10 m straight paths radiating from the location of the ASV were used in the action set. The environment was bounded by a 50 m by 50 m geofence. Localization and control was provided by a PixHawk Autopilot with GPS and internal IMU; the fused state estimate was empirically suitable for the desired maximum localization accuracy (ε = 10 m). The budget for each mission was 1000 m, which took approx. 45 minutes to travel. The GP kernel was trained on altimeter data from a dense data collection deployment the day before (parameters $l = 2.01$, $\sigma^2 = 0.53$, $\sigma_n^2 = 0.02$ [26%]). Note the high noise in the inferred GP model, as well as the relatively small length-scale in the 2500 m$^2$ field site. The reconstructed bathymetry and vehicle are shown in Fig. 6.

PLUMES successfully identified the same coral head to be maximal as that inferred from the GP trained on prior dense data collection, as indicated by accumulated reward in Table I, overcoming the challenges of moving in ocean waves, noisy altimeter measurements, and highly multimodal environment. Additionally, the posterior prediction of $x^*$ had an error of only 1.78 m while Boustro reported 8.75 m error due to its non-adaptive sampling strategy.

In the Bellairs Fringing Reef trials, the environment was assumed to be static. However, in many marine domains the impact of sediment transport, waves, and tides could physically change the location of a maximum over the course of a mission. PLUMES can be extended to dynamic environments by employing a spatiotemporal kernel in the GP model, which allows for the predictive mean and variance to change temporally [32]. If the dynamics of an environment can be encoded in the kernel function, no other changes to PLUMES are necessary; MVI will be distributed according to the time dynamic. Fig. 7 demonstrates the properties of PLUMES with a squared-exponential kernel over space ($l = 1.5$, $\sigma^2 = 100$, $\sigma_n^2 = 0.5$) and time ($l = 100$, $\sigma^2 = 100$, $\sigma_n^2 = 0.5$). In this illustrative scenario, the global maximum moved between planning iteration $T = 230$ and $T = 250$. PLUMES with a spatiotemporal kernel maintained multiple hypotheses about the maximum’s location given the random-walk dynamic of the environment, resulting in MVI reward being re-distributed between the two maxima over time.

### B. Non-Convex Environments

We next consider non-convex environments with potentially unknown obstacles, a situation that occurs frequently in practical MSS applications with geographical no-go zones for rover or ASV missions, and in indoor or urban settings. We evaluated PLUMES, UCB-Myopic, and UCB-MCTS planners.

---

**TABLE I**

<table>
<thead>
<tr>
<th>Convex Simulation Trials</th>
<th>ASV Trial $\epsilon = 1.5$ m, 50 trials</th>
<th>Non-convex Simulation Trials</th>
<th>Dubins Car Trials</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLUMES</td>
<td>MSS Reward: 199 (89) RMSE: 3.8 (0.2) $x^*$ Error: 0.21 (0.23)</td>
<td>MSS Reward: 524</td>
<td>MSS Reward: 159 (74)</td>
</tr>
<tr>
<td>UCB-MCTS</td>
<td>171 (179)* MSS Reward: 3.7 (0.6) RMSE: 0.24 (0.29) $x^*$ Error: 0.25 (0.56)</td>
<td>115 (184)* MSS Reward: 3.6 (1.5) RMSE: 0.27 (1.18) $x^*$ Error: 0.23 (0.34)</td>
<td>52 (17)</td>
</tr>
<tr>
<td>UCB-Myopic</td>
<td>148 (199)* MSS Reward: 3.6 (0.2) RMSE: 0.33 (2.35) $x^*$ Error: -</td>
<td>86 (102)* MSS Reward: 3.4 (1.0) RMSE: 0.23 (0.34) $x^*$ Error: -</td>
<td>42 (66)</td>
</tr>
<tr>
<td>Boustro.</td>
<td>27 (3)* MSS Reward: 2.7 (10.4) RMSE: 0.26 (0.46) $x^*$ Error: 2.7 (10.4)</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

*Fig. 6. Coral head map and ASV: (A) The ground truth bathymetric map inferred from all collected data, mean corrected in depth. Yellow represents shallower depths, and blue is deeper. The global maximum is marked with a black star. (B) The custom ASV used to traverse the 2500 m$^2$ region.

*Fig. 7. Extending PLUMES for Spatiotemporal Monitoring: (A) The ground truth map at two planning iterations for a dynamic environment. The maximum is marked with a black star, and migrates from the top left to the top right of the world. (B) MVI reward is redistributed by using a spatiotemporal kernel within PLUMES that captures the environment’s dynamics.*
in 50 simulated trials with the same environments, vehicle, and actions as described in Section IV-A, with the inclusion of 12 block obstacles placed uniformly around the world in known locations (see Fig.4). Boustro. was not used as a baseline because of non-generality of the offline approach to unknown obstacle maps.

As indicated in Table I, PLUMES accumulated significantly more MSS reward than UCB-MCTS and UCB-Myopic, at the 0.05-level. The distribution of reward across the trials is visualized in Fig. 8. Like in the convex-world, the PLUMES has a primary mode between reward 200-250, while the UCB-based planners have a primary mode in the low-performance region (reward <50). There was no significant difference between planners with respect to RMSE or $\times^2$ error. The fact that MVI maximized the true MSS reward while achieving statistically indistinguishable error highlights the difference in exploitation efficiency between MVI and UCB-based methods.

The simulation experiments assume that a geometric map is known a priori. However in practical applications, like indoor gas leak detection, access to a map may be limited or unavailable. We simulate the scenario in which a nonholonomic car equipped with a laser range-finder must build a map online as it seeks the maximum in a cluttered indoor environment (Fig. 9). We generate a simulated chemical phenomenon from a GP ($l = 0.8$, $\sigma^2 = 100.0$, $\sigma_n^2 = 2.0$ [2%]), and simulate observations at 1 Hz. The action set for the vehicle consists of eleven 1.5 m Dubins curves projected in front of the vehicle, one straight path behind the vehicle, and a “stay in place” action. Results for five trials are shown in Table I and illustrate that PLUMES accumulates more MSS reward than baselines, indicating robust performance.

These simulation and robot trials demonstrate the utility of PLUMES compared to canonical and state-of-the-art baselines in a diverse set of environments with challenging practical conditions. For high-stakes scientific deployments, the consistent convergence and sampling performance of PLUMES is critical and beneficial.

V. DISCUSSION AND FUTURE WORK

Online planning methods for robotic maximum seek-and-sample are critical in a variety of contexts, including general environmental monitoring (scientific inquiry, reconnaissance) and disaster response (oil spill, gas leak, radiation). For partially observable environments that can be modelled using a GP, PLUMES is a novel approach for global maximum seek-and-sample that provides several key insights.

This work presents MVI as an empirically suitable alternative to the canonical GP-UCB heuristic in MSS solvers, which is both naturally adaptive and avoids a hand-tuned parameter to balance exploration and exploitation. MVI samples potential global maxima from the robot’s full belief state to manage exploration and exploitation. In contrast, heuristic functions like UCB place reward on all high-valued or highly uncertain regions, leading to unnecessary exploration and limiting the time available to exploit knowledge of the true maximum. Ultimately, the MVI heuristic allows PLUMES to collect exploitative samples, while still achieving the same overall level of posterior model accuracy (shown by RMSE) as UCB-based planners. Additionally, continuous-observation MCTS allows PLUMES to search over belief-spaces on continuous functions without discretization or maximum-likelihood assumptions.

One important area of future work for PLUMES is online hyperparameter learning [33], which is important when only one mission is possible and there is insufficient prior knowledge for hyperparameter selection. Another avenue of future work could examine the proprieties of the maxima sampled by MVI, to be used as a heuristic for meta-behavior transitions (e.g., action model switching, dynamic horizon setting) or
mission termination. Finally, the performance of PLUMES in non-convex environments is impacted by the discrete action set encoded. Extending PLUMES to continuous actions spaces, following the approach of Morere et al. [34], would allow flexibility in these environments.

VI. CONCLUSION

This paper formalizes the maximum-seek-and-sample POMDP and presents PLUMES, an adaptive planning algorithm that employs continuous observation MCTS and maximum-value information reward to perform efficient maximum-seeking in partially observable, continuous environments. PLUMES outperforms canonical coverage and UCB-maximum-seeking in partially observable, continuous environments (e.g. multiple local maxima, unknown obstacles, sensor noise). Maximum seek-and-sample is a critical task in environmental monitoring for which PLUMES, with theoretical convergence guarantees, strong empirical performance, and robustness under real-world conditions, is well-suited.

ACKNOWLEDGMENT

We would like to thank our reviewers for their feedback on this manuscript. Additionally, we thank the SLI group, RRG and WARPlab for their insight and support.

REFERENCES


[29] H. B. Mann and D. R. Whitney, “On a test of whether one of two random variables is stochastically larger than the other,” The annals of mathematical statistics, pp. 50–60, 1947.


