Adaptive Sampling Site Selection for Robotic Exploration in Unknown Environments

Pranay Thangeda and Melkior Ornik

***Abstract*— Autonomously selecting the right sequence of locations to sample is critical during exploration missions in unknown environments, with constraints on the number of samples that can be collected, and a possibility of system failure. A key idea for decision-making in unknown environments is to exploit side information available to the agent, combined with the information gained from samples collected so far, to estimate the sampling values. In this paper, we pose the problem of sampling site selection as a problem of finding the optimal policy in a Markov decision process modeling the unknown sampling values and the outcomes associated with sampling attempts at different locations. Our solution exploits the fact that the partially unknown rewards of this Markov decision process are correlated to each other to devise a strategy that attempts to maximize the total sample value while also ensuring that the agent achieves its minimum mission requirement. We validate the utility of the proposed approach by evaluating the method against a baseline strategy that pursues collecting the samples that are estimated to be of the highest value. Our evaluations use a simulated sampling problem on Martian terrain and using OceanWATERS, a high-fidelity simulator of a future Europa lander mission.**

I. Introduction

Exploration of ocean worlds for signs of life and understanding the conditions for habitability is a critical component of the solar system exploration in the next decade [1], [2], [3]. Robotic planetary exploration missions, such as the Europa Lander mission concept [4] designed for studying Jupiter's moon Europa, are essential for exploration of potentially habitable worlds beyond Earth. One of the primary objective of such missions, including the Europa Lander, is to perform in situ analysis on surface and subsurface samples using onboard science instruments [4].

The surface operations in past and current missions such as the Mars exploration program [5] are traditionally designed as ground-in-the-loop systems with human experts on Earth making decisions based on the information obtained from the surface system. For example, in the Perseverance rover operations, the mission route planning and candidate sample site selection are handled remotely by the Perseverance mission team [6]. However, on ocean worlds such as Europa, remote operations might not be feasible due to several factors such as high communication latency and limited battery life, a consequence of the design choice to avoid radioisotope power sources due to planetary protection concerns and mission cost [4]. This constraint motivates the need for a highly autonomous exploration system that handles the entire sample collection process onboard with no human supervision. Further, autonomy has the potential to decrease the overall mission costs while simultaneously increasing the science return of the mission [7].

Traditionally, the robotic exploration missions are designed to last for several months with Earth-based teams interpreting data collected by the robotic explorer after each observation to iteratively guide future sampling site selection. However, in missions to ocean worlds, the explorer is highly constrained on the number of samples it can collect by the battery capacity [4]. Further, given the challenging and unknown conditions that the explorer operates in, there is always a possibility of total mission loss due to a critical system failure. Considering these factors, it is imperative that the explorer samples as quickly as possible while also ensuring that the selected locations maximize the science return of the mission.

Although performing sampling at a particular location is the only way to realize its true sampling value, one could get insights on the sampling value using indirect information. For example, the sample site selection for Martian exploration relies heavily on geological features in and around the candidate location [8]. These features, often interpreted from imagery captured by onboard cameras, are used to gauge for potential value of sampling at a location. Given the visual information of any two candidate locations, the extent of *visual similarity* between these two locations can be quantified and used as an indicator of the correlation between the sampling value of the two locations. This additional information, here-in-after referred to as *side information*, enables the agent to maintain an estimate of the value at unsampled locations based on the past samples from the sampled locations.

Given the side information, one approach to selecting the sampling locations is to maintain a point estimate of the sampling values and use an adaptive strategy that acts greedily with respect to the estimated samples values [9]. While this strategy guides the agent to locations with high estimated values, it has several pitfalls. For example, in the case where a sampling location is only slightly correlated with a previously observed location, the strategy would act on the current estimate that is unlikely to be true. Further, this approach also does not account for the chances of failure for each sampling location.

This work proposes an alternative approach where we model the unknown sampling values as a distribution of random variables which are updated by conditioning to observed

This work was supported by National Aeronautics and Space Administration under Award No. 80NSSC21K1030.

Authors are with the Department of Aerospace Engineering and the Coordinated Science Laboratory, University of Illinois Urbana-Champaign, Urbana, USA. {pranayt2,mornik}@illinois.edu

sampling values. Given the distributions of sampling values, which provide the decision-maker with a measure of the uncertainty in the estimation, sampling site selection problem can be naturally modeled as a Markov decision process (MDP) [10] with uncertain, time-varying reward. Inspired by the optimistic approaches in the area of active learning [11], we provide a heuristic to find a solution to the MDP. The proposed heuristic balances between exploring locations with uncertain sampling value estimates and exploiting locations with high-confidence sampling value estimates while ensuring that minimum mission requirements are met. The results from our simulations, including an example using a high-fidelity simulator for the Europa Lander concept, demonstrate utility of the proposed framework.

II. Preliminaries and Background

In this section, we provide background on failure models for planetary exploration systems and Markov decision processes. We start by defining the notation used in the paper.

Given a finite set $\mathcal{A}$, $|\mathcal{A}|$ denotes its cardinality. $2^{\mathcal{A}}$ denotes the power set of $\mathcal{A}$, and $\Delta(\mathcal{A})$ denotes the set of all probability distributions over the set $\mathcal{A}$. $\Pr[\cdot]$ denotes the probability of an event and $\mathbb{E}[\cdot]$ denotes the expectation of a random variable. $\mathbb{N}$ denotes the set of natural numbers and $[N]$ for $N \in \mathbb{N}$ denotes the set $\{0, \ldots, N-1\}$.

A. Reliability Model

Ensuring system reliability in face of possible malfunctions and failures is essential for planning in mission-critical systems. In literature on reliability engineering [12], [13], reliability of a system at a given time is defined as the probability that the system will operate as intended. In our problem, we utilize the reliability models provided in [14] and [15] and assume that the subsystems in the planetary rovers and landers are independent in terms of their reliability. While operating within the service life, assuming a constant failure rate λ, the probability of survival for a subsystem while performing a task of duration Δt can be given [15] as $P_{\text{survival}} = e^{-\Delta t \lambda}$. Let N_{sub} be the number of mission-critical subsystems in our surface system. For a mission such as sampling, it is reasonable to assume that the failure of any subsystem will prevent the system from performing the mission. Therefore, the probability of survival for the entire system P^{sys} can be given by

$$P^{sys}_{\text{survival}} = \prod_{i=1}^{N_{\text{sub}}} P^{i}_{\text{survival}} = e^{-\Delta t \sum_{i=1}^{N_{\text{sub}}} \lambda_i} \tag{1}$$

where Δt is the duration of the task and $\lambda_i, i \in [N_{\text{sub}}]$ are the individual failure rates of the subsystems.

B. Markov Decision Process

A finite horizon Markov Decision Process (MDP) [16], [10] M is specified by the tuple $(\mathcal{S}, \mathcal{A}, P, R, H, s_0)$, where:

- $\mathcal{S}$ denotes a finite set of states.
- $\mathcal{A}$ denotes a finite set of actions with $A : \mathcal{S} \to 2^{\mathcal{A}}$ denoting the set of actions the agent is allowed to take in each state $s \in \mathcal{S}$.
- $P : \mathcal{S} \times A(s) \to \Delta(\mathcal{S})$ denotes the transition function.
- $R : \mathcal{S} \times A(s) \to [0, R_{\max}]$ denotes the reward function.
- H denotes the finite planning horizon for the problem.
- $s_0 \in \mathcal{S}$ denotes the initial state.

In the given MDP $M = (\mathcal{S}, \mathcal{A}, P, R, H, s_0)$, an agent interacts with the environment according to the following protocol: at each time step $t = 0, 1, 2, \ldots$, the agent takes an action $a_t \in A(s_t)$, obtains the immediate reward $r_t = R(s_t, a_t)$, and transitions to the next state s_{t+1} sampled according to $s_{t+1} \sim P(\cdot|s_t, a_t)$.

A *generalized non-stationary policy* $\pi : \mathcal{S} \times [H] \to \Delta(\mathcal{A})$, where $\Delta(\mathcal{A})$ is the space of probability distributions over $\mathcal{A}$, specifies a decision-making strategy in which an agents selects an action based on the current state and the time step. For a given policy π and a starting state s_0 at $t = 0$, the value function $V^{\pi} : \mathcal{S} \to \mathbb{R}$ is defined as

$$V^{\pi}(s) = \mathbb{E}\Big[\sum_{t=0}^{H-1} R(s_t, a_t) \mid a_t \sim \pi(s_t, t), s_0 = s \Big]. \tag{2}$$

Starting at a state s, the solution of the MDP M would be a policy π that maximizes the value function $V^{\pi}(s)$.

III. Problem Formulation

We consider the scenario of an autonomous robotic explorer operating in an extraterrestrial, unknown environment. The goal of the explorer is to collect surface and subsurface samples and perform in situ analysis for characterizing their potential for a scientific objective. Let $\mathcal{L}$ denote the set of candidate sampling locations in the workspace of the explorer with $|\mathcal{L}| = n_{\text{loc}}$ and $l_i \in \mathcal{L}$ representing each of the individual location. The limited onboard resources often constrain the agent in terms of the number of samples it can collect. This holds true in the case of our running example, the Europa Lander concept, where the agent is expected to collect a maximum of 5 samples using the onboard battery [4]. Keeping this constraint in mind, let $n_{\text{cap}} << n_{\text{loc}}$ represent the number of samples that the agent can collect and analyze. The actual number of samples collected during the mission can be less than n_{cap} due to the agent encountering other failure modes as explained in Section II. Finally, we assume that the agent can only sample at most once in a location, i.e., there is no value in sampling at a location that has been already sampled.

We define *sampling value* of a candidate sampling location as a measure of how valuable a sample collected at that location is in terms of a scientific objective of interest. Given the fact that the agent is acting autonomously in a completely unknown environment, it has no prior knowledge about the true sampling values of each sampling location. Let $\delta_i, i \in [n_{\text{loc}}]$ denote the true sampling values of different candidate locations. In the modeled interaction protocol, the sampling value of a location is realized only when the agent collects and analyzes a sample from that location. We assume that the agent has access to side information through onboard sensors that can be used quantify the sampling value similarity between different sampling locations. This similarity metric can be exploited to correlate the potential

sampling value at different locations, thereby helping us prioritize the locations that are likely to have a higher value based on past experience. In order to capture the fact that the sampling values are unknown a priori and correlated, we model them as correlated random variables $X_i, i \in [n_{\text{loc}}]$ jointly distributed as a multivariate normal distribution [17]. Formally, let $\boldsymbol{X} = (X_1, \ldots, X_{n_{\text{loc}}})^T$ be a random vector representing the sampling value at all the candidate locations following a multivariate normal distribution. Let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_{n_{\text{loc}}})^T = (\mathbb{E}[X_1], \ldots, \mathbb{E}[X_{n_{\text{loc}}}])^T$ denote the n_{loc}-dimensional mean vector and $\boldsymbol{\Sigma} : \Sigma(X_i, X_j) = \mathbb{E}[(X_i - u_i)(X_j - u_j)]$ denote the $n_{\text{loc}} \times n_{\text{loc}}$ dimensional covariance matrix. The mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$ characterize the distribution of sampling values at all locations.

We now formalize the steps involved in the sampling process. Let l_i denote the current location of the agent. A task T^{ij} of starting at location l_i and sampling at location l_j is defined as the process of performing the following steps in order: (i) analyze the sample collected from current location l_i to calculate the sampling value, (ii) proceed to a new location, and (iii) collect a sample at the location l_j. The time τ_{ij} taken to perform a task depends on the previous sampling location, the new sampling location, and the sample analysis time. We assume that the time τ_{ij} is deterministic and known a priori. Based on the discussion in Section II, the probability of agent failing during this task can be given as $1 - e^{-\tau_{ij}\lambda}$ where λ denotes the failure rate for the system.

A. Modeling as Markov Decision Process

Recall that the agent can perform a maximum of n_{cap} number of sampling actions during its mission. Let $(T_1^{i_0 i_1}, T_2^{i_1 i_2}, \ldots, T_{n_{\text{cap}}}^{i_{n_{\text{cap}}-2} i_{n_{\text{cap}}-1}})$ denote the sequence of n_{cap} sampling tasks that the agent can perform in the absence of any failures. The problem of finding the optimal sequence of sampling tasks where each task is associated with a non-negative probability of failure can be modeled as an MDP. Specifically, let $M_{\text{sampling}} = (\mathcal{S}, \mathcal{A}, P, R, H, s_o)$ be an MDP modeling the sampling process with the following components:

- $\mathcal{S} = \{l_0, l_1, \ldots, l_{n_{\text{loc}}-1}\} \cup \{s_o, s_f\}$ where s_o is the initial location and s_f is a state representing failure.
- $\mathcal{A} = \cup_{s \in \mathcal{S}} A(s)$ where $A(s_i) = \{a_{i0}, \ldots, a_{in_{\text{cap}}-1}, a_{if}\}$ for $i \in \{o, 0, 1, \ldots, n_{\text{cap}-1}\}$ and $A(s_f) = \{a_{ff}\}$. Action a_{ij} results in an attempt to perform the task T^{ij} that enables the transition from location l_i to l_j.
- $R(s_i) = X_i$ for $s_i \in \mathcal{L}$, $R(s_o) = 0$, and $R(s_f) = 0$ where X_i is the random variable representing the unknown sampling value at location l_i.

The reward values are unknown a priori and are represented as random variable to model the uncertainty in the estimate of the true sampling value at a location. The reward function $R(\cdot)$ changes at every time step as we enforce the constraint that no location is sampled more than once by assigning a reward of 0 to already visited states. The transition probability function of the MDP M_{sampling} will be:

$$P(s_j|s_i, a_{ij}) = \begin{cases} e^{-\tau_{ij}\lambda} & \text{if } \ s_i, s_j \in \mathcal{S} \setminus \{s_f\}, \\ 1 - e^{-\tau_{ij}\lambda} & \text{if } \ s_i \in \mathcal{S} \setminus \{s_f\}, \ s_j = s_f, \\ 1 & \text{if } \ a_{ij} = a_{ff}, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Finally, the horizon H of the MDP M_{sampling} is $n_{\text{cap}} + 1$.

B. Problem Statement

Consider an agent operating in a stochastic environment with unknown rewards as described above, modeled as an modeled as an MDP $M_{\text{sampling}} = (\mathcal{S}, \mathcal{A}, P, R, H, s_o)$. We assume that $\mathcal{S}, \mathcal{A}, P, H$ and s_0 are known to the agent a priori and satisfy $H < |\mathcal{S}|$. We study the following objective: find a policy $\pi(t)$ that ensures that the total obtained reward $\sum_{i=0}^{H-1} R(s_i)$ is maximized the agent starts from the state s_0 and follows the policy.

If the rewards are known a priori, a policy that maximizes the total collected sampling value can be synthesized using standard planning tools [10]. However, the rewards values in our case are unknown and are modeled as random variables. In the next section, we discuss approaches to find the best sequence of sampling tasks in our problem setting.

IV. Solution Approach

This section details our approach to online sampling site selection with the goal of maximizing the total collected sampling value when limited to a small number of sampling actions. We begin with a discussion of estimating the sampling values online using side information and the previous samples. Next, we present a baseline solution that selects the actions greedily with respect to the expected sampling values at different locations. Finally, we discuss an improved solution that actively explores potential high value sampling location while also exploiting current knowledge to maximize the collected sample value.

A. Sampling Value Estimation

At any point in the sample collection process, the sampling values observed thus far can be used to find the distribution of the sampling value at unobserved locations conditioned on the observed data. Formally, consider two subvectors $\boldsymbol{X}^a$ and $\boldsymbol{X}^b$ of the random vector $\boldsymbol{X} = (X_1, \ldots, X_{n_{\text{loc}}})^T$ where $\boldsymbol{X}^a$ consists of sampling locations with recently observed sampling values and $\boldsymbol{X}^b$ consists of the locations whose sampling values are yet to be observed. The mean vector and the covariance matrix can be divided into the components:

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}^a \\ \boldsymbol{\mu}^b \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix}. \tag{4}$$

Now, let $\boldsymbol{x}^a$ be the realized value of the random vector $\boldsymbol{X}^a$. The realized value for a sampling location l_i would be the true sampling value δ_i for that location. Given this, the expression for mean vector and covariance matrix of the conditional distribution corresponding to the travel time on the unobserved edges $f(\boldsymbol{X}^b|\boldsymbol{X}^a = \boldsymbol{x}^a) = \mathcal{N}(\overline{\boldsymbol{\mu}}, \overline{\boldsymbol{\Sigma}})$

can [17] be given as $\overline{\boldsymbol{\mu}} = \boldsymbol{\mu}^a + \Sigma_{ab}{\Sigma_{bb}}^{-1}(\boldsymbol{x}^a - \boldsymbol{\mu}^b)$ and $\overline{\boldsymbol{\Sigma}} = \Sigma_{aa} - \Sigma_{ab}{\Sigma_{bb}}^{-1}\Sigma_{ba}$.

The obtained posterior distribution at unobserved locations $\mathcal{N}(\overline{\boldsymbol{\mu}}, \overline{\boldsymbol{\Sigma}})$ reflects the information gained from the observations. The covariance matrix is designed based on the similarity of features at different sampling locations. Section V illustrates how the covariance matrix can be designed based on the features extracted from visual data.

B. Baseline Method

One approach to handle the uncertainty in reward values, which as defined in the sampling MDP M_{sampling} are the same as the sampling values, is to utilize the expected values of the rewards. At each time instant t, we assume that the current expected value of rewards is the true reward function. Formally, the reward function for every unsampled location at time t is given as $R(s_i) = \mu_i$ for all $s_i \in \mathcal{S} \setminus \{s_0, \ldots, s_t\}$ where μ_i is the expected value of the random variable modeling the sampling value at the corresponding location. Under this assumption, the MDP M_{sampling} is completely known and the optimal policy π_t^* can be synthesized by algorithms such as value iteration [10]. The next sampling location would therefore be the state s_j reached by taking the action $\pi_t^*(s_t) = a_{s_t s_j}$ recommended by the optimal policy.

After reaching the next state by taking the action suggested by the synthesized policy, we update the reward distributions based on the realized sample value and then repeat the process for H steps or until a failure occurs.

Using the optimal policy generated by the greedy baseline approach may not maximize the overall collected sampling value as there is a possibility that the unexplored locations may have a higher true reward than the locations that currently have a high expected reward.

C. Proposed Approach

Given that the agent makes decisions based on the distribution of reward values at different states, the key to maximizing the total reward in our scenario would be to consider both of the following objectives: (i) explore potentially valuable locations that either likely to have a high true expected value or provide us with information about the sampling value at many other similar locations, and (ii) exploit the current estimates by sampling at locations where, with high confidence, we have a higher expected value than other uncertain locations. The first objective ensures that we maximize the long term total collected reward where as the second objective ensures that we maximize the collected sampling value in the short term and meet the minimum mission requirements before a potential failure.

As discussed in the previous section, selecting the sampling locations solely on the basis of expected reward values is highly influenced by the informative prior selected for the expected values and may lead to sub-optimal policies. Given that the agent has access to the entire probability distribution of the reward values, it can exploit the uncertainty in the distribution to direct the exploration process. Specifically, at each decision step $k, k \in [H]$, we use the following metric as the reward of each sampling location:

$$\tilde{R}(s_i) = \mu_{s_i} + \alpha \left(\frac{k}{H}\right) \sigma_{s_i} \tag{5}$$

where H is the horizon of the sampling MDP M_{sampling}, $\alpha > 0$ is a tunable parameter, μ_i and $\sigma_i = \Sigma(i,i)$ are the expected value and the standard deviation of the reward distribution at location corresponding to the state s_i. Intuitively, the above reward function ensures that the agent samples conservatively in the early stages of the sampling process by sampling at locations with high expected sampling value. With time, as the agent collects more samples thereby ensuring that minimum mission requirements are met, it is incentivized to explore remaining uncertain location by providing optimistic estimates of the reward at those locations. We summarize the proposed approach in Algorithm 1.

Algorithm 1 *EndOptimism*: Eventual Optimistic Heuristic

input: Markov decision process $(\mathcal{S}, \mathcal{A}, P, R, H, s_o)$
initialize: $t = 0, s_t = s_o, \tilde{V}_H^*(\cdot) = 0$
repeat
  $\tilde{R}(s^i) = \mu_i + \alpha \left(\frac{t}{H}\right) \sigma_i \;\forall\; s^i \in \mathcal{S}$
  for all $i = H-1, H-2, \ldots, t$ **do**
    for all $s \in \mathcal{S}$ **do**
      $\tilde{V}_i^*(s) = \max_a \left[\tilde{R}(s) + \mathbb{E}_{s'}(\tilde{V}_{i+1}^*(s'))\right]$
      $\tilde{\pi}_i^*(s) = \arg\max_a \left[\tilde{R}(s) + \mathbb{E}_{s'}(\tilde{V}_{i+1}^*(s'))\right]$
    end for
  end for
  $f(\boldsymbol{X}|X_{s_t} = \delta_{s_t}) = \mathcal{N}(\overline{\boldsymbol{\mu}}, \overline{\boldsymbol{\Sigma}})$
  $R(s_t) = 0$
  $s_{t+1} \sim P(\cdot|s_t, \tilde{\pi}_t^*(s_t))$
  $t = t + 1$
until $t = H - 1$

Now that we proposed the heuristic for selecting the sequence of sampling locations, we proceed to demonstrate its utility using two case studies.

V. Evaluation

This section presents two examples that serve the purpose of demonstrating our approach and validating the utility of the proposed heuristic. We investigate the performance of the heuristic using two case studies: (i) a grid-world problem modeling a rover operating and collecting samples on Martian surface that demonstrates the advantage of the heuristic over baseline solution, and (ii) the problem of collecting samples on Europa surface using a high-fidelity simulator of the Europa Lander that shows the utility of our approach for solving real-world problems.

A. Grid-World Environment

We consider a rover operating on Martian surface modeled as a grid-world environment. The environment is adopted from [18] and models a section of Jezero crater, the landing site of Mars Perseverance rover mission. A map of the crater

used for generating the grid-world is shown in Fig. 1. Each state in the grid-world belongs to one of the three terrain types: *benign*, *rough*, and *rippled*. We assume that the agent has prior knowledge of the terrain type at each location.

For our simulations, we randomly select 50 locations in the environment that are, as shown in Fig. 1, uniformly distributed over all three terrain types as the candidate sampling locations. The sampling capacity of the agent n_{cap} is also set to 50. We assume that the sampling values at locations of same terrain type are positively correlated, while the sampling locations of different terrain types have no correlation between their sampling values. Each sampling task is defined as the rover sampling at the location, analyzing the sample, and heading to the next target location. The time taken for a sampling task is therefore proportional to the analysis time and the length of the optimal path between the initial location and the target location. Finally, we assume that the true sampling value at each location is in $[0, 1]$.

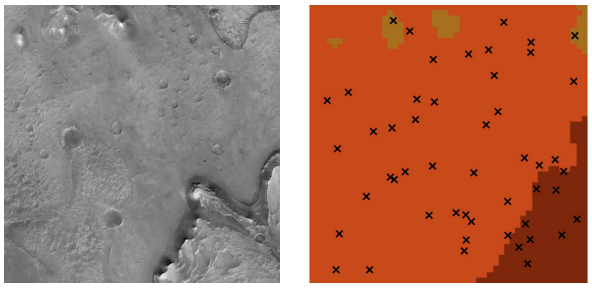

Fig. 1: (left) Original image of the area of interest in the Jezero crater, (right) discretized image with $\times$ markers indicating the candidate sample collection locations. The colors light brown, brown, and dark brown represent three terrain types benigh, rough, and rippled respectively.

We perform 10000 simulations where the values of true sampling value at each location and the correlation of sampling value between different locations within the same class is randomly generated for every simulation. We repeat the experiments twice, once with a failure rate $\lambda = 0.003$ and another time with a zero failure rate.

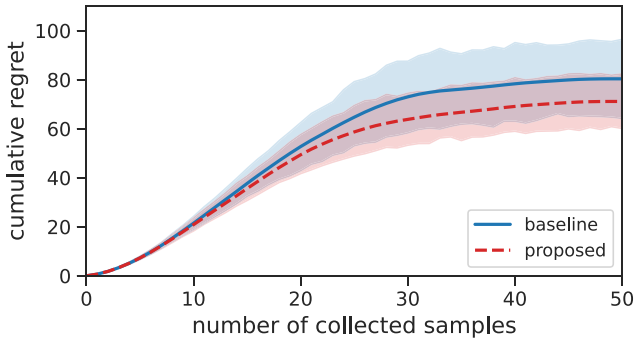

Fig. 2: The average cumulative regret of the baseline approach and the proposed approach in the grid-world environment with 50 sample locations, a sampling capacity of 50, and a 0 failure rate (lower is better).

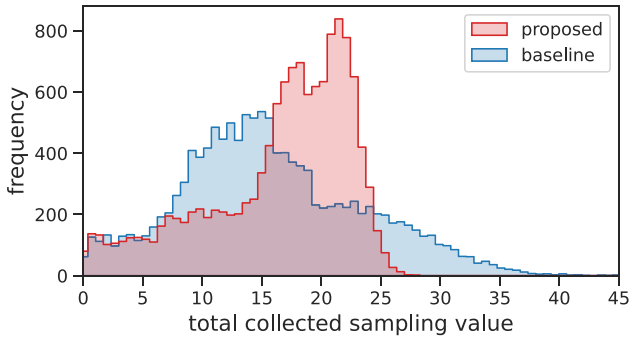

Fig. 3: Distribution of total sampling value obtained by running 10000 simulations of the grid-world environment with 50 candidate locations and a failure rate of 0.003.

Fig. 2 and 3 summarize the results from the experiments where Fig. 2 shows the average cumulative regret. Let $\mathcal{D}_n$ denote the set of n largest true sampling values. Then, the cumulative regret at a time step t is defined as:

$$\sum_{i=0}^{t} \left(\sum_{\delta_j \in \mathcal{D}_i} \delta_j - \sum_{j=0}^{i} R(s_j) \right). \tag{6}$$

The cumulative regret measures how well the agent could have performed over time and hence a lower value indicates an agent taking better sampling decisions through out the sampling site selection process. From Fig 2, we can see that the baseline approach has a higher average cumulative regret than the proposed approach.

Fig. 3 shows the average value of the total sampling value collected in each run with a non-zero probability of failure during the run. The proposed approach, selecting the samples conservatively in the beginning, performs similarly to the baseline approach in the low total sampling value runs. However, as shown in Fig. 3, the proposed approach on average performs better than the baseline approach and also has a higher probability of achieving any minimum mission requirement.

B. Ocean World Environment

This section provides a case study using the Ocean Worlds Autonomy Testbed for Exploration (OceanWATERS) [19] simulation testbed. The OceanWATERS simulator is an open-source simulator designed to aid in the development of autonomy software for robotic exploration of ocean worlds such as Europa. The simulator is built using the Robot Operating System (ROS) middleware [20] and Gazebo [21] and takes the Europa Lander mission as a reference. The lander, equipped with a robotic arm and a stereo camera, rests on a photo-realistic terrain model simulating the surface of Europa. A complete description of the simulator can be found in [19].

For our experiment, we consider 20 sampling locations in the workspace of the robot, selected to uniformly cover the workspace and all possible terrain profiles. Based on the specifications of the Europa Lander concept [4], we

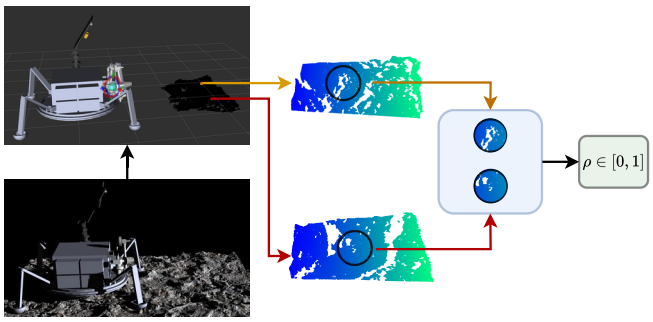

Fig. 4: Illustration of correlation estimation using visual similarity. The figure in the bottom left displays the Gazebo simulation environment and top left figure shows the point cloud captured using onboard stereo camera. The two point cloud patches represent the area of a pair of sampling locations and are used to calculate their similarity score.

assume that the lander can collect a maximum of 5 samples. We assign sampling values in the range $[0, 1]$ to different locations while ensuring that the values are consistent with their similarity. For the sampling value estimator, we design the prior mean at all locations as 0.5 and the prior covariance matrix using standard deviations of 0.25 at each location and correlations generated by comparing the features around each sampling location using the point cloud structural similarity metric [22]. An overview of computing the similarity between two sampling locations is provided in Fig. 4.

Table I shows the results from the experiment comparing the baseline approach to the proposed approach for different failure rates. While the performance of the baseline approach and the proposed approach are comparable for the lower failure rate scenario, the proposed approach performs significantly better for the higher failure rate scenario which is consistent with the results of the previous experiment.

TABLE I: Average total collected sampling value comparison

Method	$\lambda = 0.001$	$\lambda = 0.003$	$\lambda = 0.005$
Baseline	8.2	6.6	3.5
Proposed	8.6	7.8	6.4

VI. Conclusion and Future Work

This paper presents a heuristic for the problem of selecting a sequence of few most valuable sampling locations, one by one, out of many possible locations when the value of any location is only realized after sampling at that location. We formulated the problem as a Markov decision process where the sequential decision-making approach and stochastic dynamics naturally models the replanning involved in sample site selection and the possibility of failure. Using simulations on two experiments, including a high-fidelity simulator of the Europa Lander mission, we showed that the proposed approach outperforms a greedy baseline approach, while also ensuring that the agent, on average, is more likely to meet the minimum mission requirements – a factor to consider in highly expensive space exploration missions.

While this paper develops the initial framework for autonomous sampling in unknown environments, much work remains to be done in incorporating mission-specific priorities and side information. Working on the multi-agent version of the problem where we decide the sampling strategy for a cooperative team of robotic explorers in another possible direction for future work.

References

[1] A. R. Hendrix, T. A. Hurford, L. M. Barge, M. T. Bland, J. S. Bowman, W. Brinckerhoff, B. J. Buratti, M. L. Cable, J. Castillo-Rogez, G. C. Collins, *et al.*, "The NASA roadmap to ocean worlds," *Astrobiology*, vol. 19, no. 1, pp. 1–27, 2019.

[2] S. Howell, W. C. Stone, K. Craft, C. German, A. Murray, A. Rhoden, and K. Arrigo, "Ocean worlds exploration and the search for life," *Bulletin of the American Astronomical Society*, vol. 53, no. 4, p. 191, 2021.

[3] B. Sherwood, J. Lunine, C. Sotin, T. Cwik, and F. Naderi, "Program options to explore ocean worlds," *Acta Astronautica*, vol. 143, pp. 285–296, 2018.

[4] K. Hand, A. Murray, J. Garvin, W. Brinckerhoff, B. Christner, K. Edgett, and T. Hoehler, "Report of the Europa lander science definition team," NASA, Tech. Rep., 2017.

[5] D. Shirley and D. McCleese, "Mars exploration program strategy-1995-2020," in *Aerospace Sciences Meeting and Exhibit*, 1996, p. 333.

[6] M. Ehrenfried *et al.*, "Surface operations and science," in *Perseverance and the Mars 2020 Mission*. Springer, 2022, pp. 91–110.

[7] G. Reeves, B. A. Kennedy, G. H. Tan-Wang, P. G. Backes, S. A. Chien, V. Verma, K. P. Hand, and C. B. Phillips, "Development of autonomous actions to enable the next decade of ocean world exploration," *Bulletin of the American Astronomical Society*, vol. 53, no. 4, p. 328, 2021.

[8] R. Francis, T. Estlin, D. Gaines, B. Bornstein, S. Schaffer, V. Verma, R. Anderson, M. Burl, S. Chu, R. Castaño, *et al.*, "AEGIS autonomous targeting for the Curiosity rover's ChemCam instrument," in *IEEE Applied Imagery Pattern Recognition Workshop*, 2015, pp. 1–5.

[9] M. Ghaffari Jadidi, J. Valls Miro, and G. Dissanayake, "Gaussian processes autonomous mapping and exploration for range-sensing mobile robots," *Autonomous Robots*, vol. 42, no. 2, pp. 273–290, 2018.

[10] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., 1994.

[11] B. Settles, "Active learning literature survey," University of Wisconsin–Madison," Computer Sciences Technical Report, 2009.

[12] M. Ram, "On system reliability approaches: a brief survey," *International Journal of System Assurance Engineering and Management*, vol. 4, no. 2, pp. 101–117, 2013.

[13] V. Volovoi, "System reliability at the crossroads," *International Scholarly Research Notices*, vol. 2012, 2012.

[14] S. Stancliff, J. Dolan, and A. Trebi-Ollennu, "Towards a predictive model of mobile robot reliability," *Technical Report CMU-RI-TR-05-38*, 2005.

[15] S. Stancliff, J. Dolan, and A. Trebi Ollennu, "Planning to fail: reliability as a design parameter for planetary rover missions," in *Workshop on Performance Metrics for Intelligent Systems*, 2007, pp. 204–208.

[16] R. Bellman, "The theory of dynamic programming," *Bulletin of the American Mathematical Society*, vol. 60, no. 6, pp. 503–515, 1954.

[17] Y. L. Tong, *The Multivariate Normal Distribution*. Springer Science & Business Media, 1990.

[18] M. Ornik, J. Fu, N. T. Lauffer, W. Perera, M. Alshiekh, M. Ono, and U. Topcu, "Expedited learning in MDPs with side information," in *IEEE Conference on Decision and Control*. IEEE, 2018, pp. 1941–1948.

[19] D. Catanoso, A. Chakrabarty, J. Fugate, U. Naal, T. M. Welsh, and L. J. Edwards, "OceanWATERS lander robotic arm operation," in *IEEE Aerospace Conference*, 2021, pp. 1–11.

[20] M. Quigley, K. Conley, B. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, A. Y. Ng, *et al.*, "ROS: an open-source robot operating system," in *ICRA workshop on open source software*, vol. 3, no. 3.2, 2009, p. 5.

[21] N. Koenig and A. Howard, "Design and use paradigms for Gazebo, an open-source multi-robot simulator," in *International Conference on Intelligent Robots and Systems*, vol. 3, 2004, pp. 2149–2154.

[22] E. Alexiou and T. Ebrahimi, "Towards a point cloud structural similarity metric," in *IEEE International Conference on Multimedia and Expo Workshops*, 2020, pp. 1–6.