# Information-Theoretic Multi-Robot Adaptive Exploration and Mapping of Environmental Hotspot Fields

Kian Hsiang Low[†], John M. Dolan[†§], and Pradeep Khosla[†§]
Department of Electrical and Computer Engineering[†], Robotics Institute[§]
Carnegie Mellon University
5000 Forbes Avenue Pittsburgh PA 15213 USA
{bryanlow, jmd}@cs.cmu.edu, pkk@ece.cmu.edu

## ABSTRACT

Recent research in robot exploration and mapping has focused on sampling hotspot fields. This exploration task is formalized by [3] in a decision-theoretic planning framework called MAXP. The time complexity of solving MAXP approximately depends on the map resolution, which limits its use in large-scale, high-resolution exploration and mapping. To alleviate this computational difficulty, this paper presents an information-theoretic approach to MAXP (*i*MAXP); by reformulating the cost-minimizing *i*MAXP as a reward-maximizing problem, its time complexity becomes independent of map resolution and is less sensitive to increasing robot team size. Using the reward-maximizing dual, we derive a novel adaptive variant of maximum entropy sampling, thus improving the induced policy performance. We also demonstrate the superior performance of exploration policies for sampling the log-Gaussian process to that of policies for the Gaussian process in mapping the hotspot field. Lastly, we provide sufficient conditions that, when met, guarantee adaptivity has no benefit under an assumed environment model.

## Categories and Subject Descriptors

G.1.6 [**Optimization**]: convex programming; G.3 [**Probability and Statistics**]: stochastic processes; I.2.8 [**Problem Solving, Control Methods, and Search**]: dynamic programming; I.2.9 [**Robotics**]: autonomous vehicles

## General Terms

Algorithms, Performance, Experimentation, Theory


*ESSA Workshop* '09, April 16, 2009, San Francisco, California, USA


## Keywords

active learning; adaptive sampling; non-myopic path planning; mobile sensor network; Gaussian process; log-Gaussian process; multi-robot exploration and mapping

## 1. INTRODUCTION

Recent research in multi-robot exploration and mapping [3, 8] has focused on sampling environmental fields, some of which typically feature a few small *hotspots* in a large region [9]. Such a *hotspot field* (e.g., plankton density and mineral distribution in Fig. 2) is characterized by continuous, positively skewed, spatially correlated measurements with the hotspots exhibiting extreme measurements and much higher spatial variability than the rest of the field. With limited (e.g., point-based) robot sensing range, a complete coverage becomes impractical in terms of resource costs. So, to accurately map the field, the hotspots have to be sampled at a higher resolution.

The hotspot field discourages static sensor placement because a large number of sensors has to be positioned to detect and refine the sampling of hotspots. If these static sensors are not placed in any hotspot initially, they cannot reposition by themselves to locate one. In contrast, a robot team is capable of performing high-resolution hotspot sampling due to its mobility. Hence, it is desirable to build a mobile robot team that can actively explore to map a hotspot field.

To learn a hotspot field map, the *exploration strategy* of the robot team has to plan resource-constrained observation paths that minimize the map uncertainty of a hotspot field. The recent work of [3] formalizes this exploration task in a decision-theoretic planning framework called the multi-robot adaptive exploration problem (MAXP). So, MAXP can be viewed as a generalization of active learning [1] due to its sequential nature. It unifies formulations of exploration problems along the entire adaptivity (see Def. 2.2) spectrum, thus allowing the performance advantage of a more adaptive exploration policy to be theoretically realized. The map

uncertainty is measured in terms of the mean-squared error criterion, which causes the time complexity of solving MAXP (approximately) to depend on the map resolution. This limits its practical use in large-scale, high-resolution exploration and mapping.

The principal contribution of this paper is to alleviate this computational difficulty through an information-theoretic approach to MAXP ($i$MAXP) (§2), which measures map uncertainty based on the entropy criterion. Unlike MAXP, reformulating the cost-minimizing $i$MAXP as a reward-maximizing problem causes its time complexity of being solved approximately to be independent of the map resolution and less sensitive to larger robot team size (§3 and §5). Additional contributions from this reward-maximizing formulation include:

- making the commonly-used non-adaptive maximum entropy sampling problem adaptive (§3), thus improving the performance of the induced exploration policy;
- given an assumed environment model (e.g., occupancy grid map), establishing sufficient conditions that, when met, guarantee adaptivity provides no benefit (§4); and
- explaining and demonstrating the superior performance of exploration policies for sampling the log-Gaussian process ($\ell$GP) to that of policies for the commonly-used Gaussian process (GP) in mapping the hotspot field (§4 and §6).

**Related Work.** Beyond its computational gain, $i$MAXP retains the beneficial properties of MAXP: it is novel in the class of model-based strategies to perform both wide-area coverage and hotspot sampling. The former considers sparsely sampled areas to be of high uncertainty and thus spreads the observations evenly across the environmental field. The latter expects areas of high uncertainty to contain highly-varying measurements and hence produces clustered observations. Since $i$MAXP builds upon the formal framework of MAXP, it uniquely covers the entire adaptivity spectrum; a more adaptive strategy can exploit clustering phenomena in a hotspot field to produce lower map uncertainty. In contrast, all other model-based strategies [4, 5, 8] are non-adaptive and achieve only wide-area coverage; they are observed to perform well only with smoothly-varying fields. Like MAXP, $i$MAXP can plan non-myopic multi-robot paths, which are more desirable than greedy or single-robot paths [4, 5, 8]. For a thorough discussion of existing exploration strategies, we refer the interested reader to the related work in [3].

## 2. COST-MINIMIZING PROBLEM FORMULATIONS

Using the methodology of constructing MAXP, we formalize here the information-theoretic exploration problems at the two extremes of the adaptivity spectrum. Exploration problems within the spectrum can be formalized in a similar manner. Not surprisingly, the resulting cost-minimizing formulations differ from that of MAXP by only the entropy criterion.

**Notation and Preliminaries**. Let $\mathcal{X}$ be the domain of the hotspot field corresponding to a finite, discretized set of grid cell locations. An observation taken (e.g., by a single robot) at stage $i$ comprises a pair of location $x_i \in \mathcal{X}$ and its measurement $z_{x_i}$. More generally, $k$ observations taken (e.g., by $k$ robots or 1 robot taking $k$ observations) at stage $i$ can be represented by a pair of vectors $\mathbf{x}_i$ and $\mathbf{z}_{\mathbf{x}_i}$, which, respectively, denote $k$ locations and their corresponding measurements.

Definition 2.1 (Posterior Data). *The posterior data $d_i$ at stage $i > 0$ comprises*

- *the prior data $d_0 = \langle \mathbf{x}_0, \mathbf{z}_{\mathbf{x}_0} \rangle$ available at stage 0, and*
- *a complete history of observations $\mathbf{x}_1, \mathbf{z}_{\mathbf{x}_1}, \ldots, \mathbf{x}_i, \mathbf{z}_{\mathbf{x}_i}$ induced by $k$ observations per stage over stages 1 to $i$.*

Let $\mathbf{x}_{0:i}$ and $\mathbf{z}_{\mathbf{x}_{0:i}}$ denote vectors comprising the location and measurement components of the data $d_i$ (i.e., concatenations of $\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_i$ and $\mathbf{z}_{\mathbf{x}_0}, \mathbf{z}_{\mathbf{x}_1}, \ldots, \mathbf{z}_{\mathbf{x}_i}$), respectively. Let $\overline{\mathbf{x}}_{0:i}$ denote the vector comprising locations of domain $\mathcal{X}$ not observed in $d_i$, and $\mathbf{z}_{\overline{\mathbf{x}}_{0:i}}$ be the vector comprising the corresponding measurements. Let $Z_{x_i}$, $\mathbf{Z}_{\mathbf{x}_i}$, $\mathbf{Z}_{\mathbf{x}_{0:i}}$, $\mathbf{Z}_{\overline{\mathbf{x}}_{0:i}}$ be the random counterparts of $z_{x_i}$, $\mathbf{z}_{\mathbf{x}_i}$, $\mathbf{z}_{\mathbf{x}_{0:i}}$, $\mathbf{z}_{\overline{\mathbf{x}}_{0:i}}$ respectively.

Definition 2.2 (Characterizing Adaptivity). *Suppose prior data $d_0$ are available and $n$ new locations are to be explored. Then, an exploration strategy is*

- **adaptive** *if its policy to select each vector $\mathbf{x}_{i+1}$ of $k$ new locations depends only on the previously sampled data $d_i$ for $i = 0, \ldots, n/k - 1$. This strategy thus selects $k$ observations per stage over $n/k$ stages. When $k = 1$, this strategy is strictly adaptive. Increasing $k$ makes it less adaptive;*
- **non-adaptive** *if its policy to select each new location $x_{i+1}$ for $i = 0, \ldots, n-1$ is independent of the measurements $z_{x_1}, \ldots, z_{x_n}$. As a result, all $n$ new locations $x_1, \ldots, x_n$ can be selected prior to exploration. That is, this strategy selects all $n$ observations in a single stage.*

**Objective Function**. The exploration objective is to select observation paths that minimize the uncertainty of mapping a hotspot field. To achieve this, we use the entropy criterion to measure map uncertainty. Given the posterior data $d_n$, the *posterior map entropy* of domain $\mathcal{X}$ can be represented by the posterior entropy of the unobserved locations $\overline{\mathbf{x}}_{0:n}$:

$$\mathbb{H}[\mathbf{Z}_{\overline{\mathbf{x}}_{0:n}} | d_n] \triangleq -\int f(\mathbf{z}_{\overline{\mathbf{x}}_{0:n}} | d_n) \log f(\mathbf{z}_{\overline{\mathbf{x}}_{0:n}} | d_n) \, d\mathbf{z}_{\overline{\mathbf{x}}_{0:n}} \ . \quad (1)$$

**Value Function**. If only the prior data $d_0$ are available, an exploration strategy has to produce a policy for selecting observation paths that minimize the *expected* posterior map entropy instead. This policy must then collect the optimal observations $\mathbf{x}_1, \mathbf{z}_{\mathbf{x}_1}, \ldots, \mathbf{x}_n, \mathbf{z}_{\mathbf{x}_n}$ during exploration to form posterior data $d_n$. The value under an exploration policy $\pi$ is defined to be the expected posterior map entropy (i.e., expectation of (1)) when starting in $d_0$ and following $\pi$ thereafter:

$$\begin{aligned} V_0^{\pi}(d_0) &\stackrel{\triangle}{=} \mathbb{E}\{\mathbb{H}[\mathbf{Z}_{\overline{\mathbf{x}}_{0:n}}|d_n]|d_0, \pi\} \\ &= \int f(\mathbf{z}_{\mathbf{x}_{1:n}}|d_0, \pi)\ \mathbb{H}[\mathbf{Z}_{\overline{\mathbf{x}}_{0:n}}|d_n]\ \mathrm{d}\mathbf{z}_{\mathbf{x}_{1:n}}\ . \end{aligned} \tag{2}$$

The strategy of [8] has optimized a closely related *mutual information* criterion that measures the expected entropy reduction of unobserved locations $\overline{\mathbf{x}}_{0:n}$ by observing $\mathbf{x}_{1:n}$ (i.e., $\mathbb{H}[\mathbf{Z}_{\overline{\mathbf{x}}_{0:n}}|d_0] - \mathbb{E}\{\mathbb{H}[\mathbf{Z}_{\overline{\mathbf{x}}_{0:n}}|d_n]|d_0\}$). This is deficient for the exploration objective because mutual information may be maximized by a choice of $\mathbf{x}_{1:n}$ inducing a very large prior entropy $\mathbb{H}[\mathbf{Z}_{\overline{\mathbf{x}}_{0:n}}|d_0]$ but not necessarily the smallest expected posterior map entropy $\mathbb{E}\{\mathbb{H}[\mathbf{Z}_{\overline{\mathbf{x}}_{0:n}}|d_n]|d_0\}$. In the next two subsections, we will describe how the adaptive and non-adaptive exploration policies can be derived to minimize the expected posterior map entropy (2).

**Adaptive Exploration**. The adaptive policy $\pi$ for directing a team of $k$ robots is structured to collect $k$ observations per stage over a $n$-stage planning horizon. So, each robot observes 1 location per stage and is constrained to explore at most $n$ new locations over $n$ stages. Formally, $\pi \stackrel{\triangle}{=} \langle \pi_0(d_0), \ldots, \pi_{n-1}(d_{n-1}) \rangle$ where $\pi_i : d_i \rightarrow \mathbf{a}_i$ maps data $d_i$ to a vector of robots' actions $\mathbf{a}_i \in \mathcal{A}(\mathbf{x}_i)$ at stage $i$, and $\mathcal{A}(\mathbf{x}_i)$ is the joint action space of the robots given their current locations $\mathbf{x}_i$. We assume the transition function $\tau : \mathbf{x}_i \times \mathbf{a}_i \rightarrow \mathbf{x}_{i+1}$ *deterministically* moves the robots to their next locations $\mathbf{x}_{i+1}$ at stage $i+1$. Combining $\pi_i$ and $\tau$ gives $\mathbf{x}_{i+1} \leftarrow \tau(\mathbf{x}_i, \pi_i(d_i))$. We can observe from this assignment that the sequential (i.e., stagewise) selection of $k$ new locations $\mathbf{x}_{i+1}$ to be included in the observation paths depends only on the previously sampled data $d_i$ along the paths for stage $i = 0, \ldots, n-1$. Hence, policy $\pi$ is adaptive (Def. 2.2).

Solving the adaptive exploration problem $i$MAXP(1) means choosing $\pi$ to minimize $V_0^{\pi}(d_0)$ (2), which we call the *optimal adaptive policy* $\pi^1$ (i.e., $V_0^{\pi^1}(d_0) = \min_{\pi} V_0^{\pi}(d_0)$). Plugging $\pi^1$ into (2) gives the $n$-stage dynamic programming equations:

$$\begin{aligned} V_i^{\pi^1}(d_i) &= \int f(\mathbf{z}_{\mathbf{x}_{i+1}}|d_i, \pi_i^1)\ V_{i+1}^{\pi^1}(d_{i+1})\ \mathrm{d}\mathbf{z}_{\mathbf{x}_{i+1}} \\ &= \int f(\mathbf{z}_{\tau(\mathbf{x}_i, \pi_i^1(d_i))}|d_i)\ V_{i+1}^{\pi^1}(d_{i+1})\ \mathrm{d}\mathbf{z}_{\tau(\mathbf{x}_i, \pi_i^1(d_i))} \\ &= \min_{\mathbf{a}_i \in \mathcal{A}(\mathbf{x}_i)} \int f(\mathbf{z}_{\tau(\mathbf{x}_i, \mathbf{a}_i)}|d_i)\ V_{i+1}^{\pi^1}(d_{i+1})\ \mathrm{d}\mathbf{z}_{\tau(\mathbf{x}_i, \mathbf{a}_i)} \\ V_n^{\pi^1}(d_n) &= \mathbb{H}[\mathbf{Z}_{\overline{\mathbf{x}}_{0:n}}|d_n] \end{aligned} \tag{3}$$

for stage $i = 0, \ldots, n-1$. The second equality follows from $\mathbf{x}_{i+1} \leftarrow \tau(\mathbf{x}_i, \pi_i^1(d_i))$ above. Policy $\pi^1 = \langle \pi_0^1(d_0), \ldots, \pi_{n-1}^1(d_{n-1}) \rangle$ can be determined in a stagewise manner by

$$\pi_i^1(d_i) = \arg\min_{\mathbf{a}_i \in \mathcal{A}(\mathbf{x}_i)} \int f(\mathbf{z}_{\tau(\mathbf{x}_i, \mathbf{a}_i)}|d_i)\ V_{i+1}^{\pi^1}(d_{i+1})\ \mathrm{d}\mathbf{z}_{\tau(\mathbf{x}_i, \mathbf{a}_i)}\ .$$

Note that the optimal action $\pi_0^1(d_0)$ at stage 0 can be determined prior to exploration using prior data $d_0$. However, each action rule $\pi_i^1(d_i)$ at stage $i = 1, \ldots, n-1$ defines the optimal action to take in response to $d_i$, part of which (i.e., $\mathbf{x}_1, \mathbf{z}_{\mathbf{x}_1}, \ldots, \mathbf{x}_i, \mathbf{z}_{\mathbf{x}_i}$) are only observed during exploration.

**Non-Adaptive Exploration**. The non-adaptive policy $\pi$ is structured to collect, in 1 stage, $n$ observations per robot with a team of $k$ robots. So, each robot is also constrained to explore at most $n$ new locations, but they have to do this within 1 stage. Formally, $\pi \stackrel{\triangle}{=} \pi_0(d_0)$ where $\pi_0 : d_0 \rightarrow \mathbf{a}_{0:n-1}$ maps prior data $d_0$ to a vector $\mathbf{a}_{0:n-1}$ of action components concatenating a sequence of robots' actions $\mathbf{a}_0, \ldots, \mathbf{a}_{n-1}$. Combining $\pi_0$ and $\tau$ gives $\mathbf{x}_{1:n} \leftarrow \tau(\mathbf{x}_{0:n-1}, \pi_0(d_0))$. We can observe from this assignment that the selection of $k \times n$ new locations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ to form the observation paths are independent of the measurements $\mathbf{z}_{\mathbf{x}_1}, \ldots, \mathbf{z}_{\mathbf{x}_n}$ obtained along the paths during exploration. Hence, policy $\pi$ is non-adaptive (Def. 2.2) and all new locations can be selected in a single stage prior to exploration.

Solving the non-adaptive exploration problem $i$MAXP($n$) involves choosing $\pi$ to minimize $V_0^{\pi}(d_0)$ (2), which we call the *optimal non-adaptive policy* $\pi^n$ (i.e., $V_0^{\pi^n}(d_0) = \min_{\pi} V_0^{\pi}(d_0)$). Plugging $\pi^n$ into (2) gives the 1-stage equation:

$$\begin{aligned} V_0^{\pi^n}(d_0) &= \int f(\mathbf{z}_{\mathbf{x}_{1:n}}|d_0, \pi_0^n)\ \mathbb{H}[\mathbf{Z}_{\overline{\mathbf{x}}_{0:n}}|d_n]\ \mathrm{d}\mathbf{z}_{\mathbf{x}_{1:n}} \\ &= \int f(\mathbf{z}_{\tau(\mathbf{x}_{0:n-1}, \pi_0^n(d_0))}|d_0)\ \mathbb{H}[\mathbf{Z}_{\overline{\mathbf{x}}_{0:n}}|d_n]\ \mathrm{d}\mathbf{z}_{\tau(\mathbf{x}_{0:n-1}, \pi_0^n(d_0))} \\ &= \min_{\mathbf{a}_{0:n-1}} \int f(\mathbf{z}_{\tau(\mathbf{x}_{0:n-1}, \mathbf{a}_{0:n-1})}|d_0)\ \mathbb{H}[\mathbf{Z}_{\overline{\mathbf{x}}_{0:n}}|d_n]\ \mathrm{d}\mathbf{z}_{\tau(\mathbf{x}_{0:n-1}, \mathbf{a}_{0:n-1})}. \end{aligned} \tag{4}$$

The second equality follows from $\mathbf{x}_{1:n} \leftarrow \tau(\mathbf{x}_{0:n-1}, \pi_0^n(d_0))$ above. Policy $\pi^n = \pi_0^n(d_0)$ can therefore be determined in a single stage by $\pi_0^n(d_0) =$

$$\arg\min_{\mathbf{a}_{0:n-1}} \int f(\mathbf{z}_{\tau(\mathbf{x}_{0:n-1}, \mathbf{a}_{0:n-1})}|d_0)\ \mathbb{H}[\mathbf{Z}_{\overline{\mathbf{x}}_{0:n}}|d_n]\ \mathrm{d}\mathbf{z}_{\tau(\mathbf{x}_{0:n-1}, \mathbf{a}_{0:n-1})}\ .$$

Note that the optimal sequence of robots' actions $\pi_0^n(d_0)$ (i.e., optimal observation paths) can be determined prior to exploration since the prior data $d_0$ are available.

## 3. REWARD-MAXIMIZING DUAL FORMULATIONS

In this section, we transform the cost-minimizing $i$MAXP(1) (3) and $i$MAXP($n$) (4) into reward-maximizing problems and show their equivalence. The reward-maximizing

$i$MAXP($n$) turns out to be the well-known *maximum entropy sampling* (MES) problem [7]:

$$U_0^{\pi^n}(d_0) = \max_{\mathbf{a}_{0:n-1}} \mathbb{H}[\mathbf{Z}_{\tau(\mathbf{x}_{0:n-1},\mathbf{a}_{0:n-1})}|d_0] \ , \qquad (5)$$

which is a 1-stage problem of selecting $k \times n$ new locations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ with maximum entropy to form the observation paths. This dual ensues from the equivalence result $V_0^{\pi^n}(d_0) = \mathbb{H}[\mathbf{Z}_{\overline{\mathbf{x}}_0}|d_0] - U_0^{\pi^n}(d_0)$ relating cost-minimizing and reward-maximizing $i$MAXP($n$)'s in the non-adaptive exploration setting, which follows from the entropy chain rule. This result says the original objective of minimizing expected posterior map entropy is equivalent to that of discharging, from prior map entropy, the largest entropy into the selected paths. So, their optimal non-adaptive policies coincide.

Our reward-maximizing $i$MAXP(1) is a novel adaptive variant of MES. Unlike the cost-minimizing $i$MAXP(1), it can be subject to convex analysis, which allows monotone-bounding approximations to be developed (§5). It comprises the following $n$-stage dynamic programming equations:

$$\begin{aligned} U_i^{\pi^1}(d_i) &= \max_{\mathbf{a}_i \in \mathcal{A}(\mathbf{x}_i)} \mathbb{H}[\mathbf{Z}_{\tau(\mathbf{x}_i,\mathbf{a}_i)}|d_i] + \\ &\qquad \int f(\mathbf{z}_{\tau(\mathbf{x}_i,\mathbf{a}_i)}|d_i)\ U_{i+1}^{\pi^1}(d_{i+1})\ \mathrm{d}\mathbf{z}_{\tau(\mathbf{x}_i,\mathbf{a}_i)} \\ U_t^{\pi^1}(d_t) &= \max_{\mathbf{a}_t \in \mathcal{A}(\mathbf{x}_t)} \mathbb{H}[\mathbf{Z}_{\tau(\mathbf{x}_t,\mathbf{a}_t)}|d_t] \end{aligned} \qquad (6)$$

for stage $i = 0, \ldots, t-1$ where $t = n-1$. Each stagewise reward reflects the entropy of $k$ new locations $\mathbf{x}_{i+1}$ to be potentially selected into the paths. By maximizing the sum of expected rewards over $n$ stages in (6), the reward-maximizing $i$MAXP(1) absorbs the largest expected entropy into the selected paths. In the adaptive exploration setting, the cost-minimizing and reward-maximizing $i$MAXP(1)'s are also equivalent (i.e., their optimal adaptive policies coincide):

THEOREM 3.1. $V_i^{\pi^1}(d_i) = \mathbb{H}[\mathbf{Z}_{\overline{\mathbf{x}}_{0:i}}|d_i] - U_i^{\pi^1}(d_i)$ *for stage* $i = 0, \ldots, n-1$.

In cost-minimizing $i$MAXP(1), the time complexity of evaluating the cost (i.e., posterior map entropy (1)) depends on the domain size $|\mathcal{X}|$ for the environment models in §4. By transforming into the dual, the time complexity of evaluating each stagewise reward becomes independent of $|\mathcal{X}|$ because it reflects only the uncertainty of the new locations to be potentially selected for observation. As a result, the runtime of the proposed approximation algorithm in §5 does not depend on the map resolution, which is clearly advantageous in large-scale, high-resolution exploration and mapping. In contrast, the reward-maximizing MAXP(1) [3] utilizing the mean-squared error criterion does not share this computational advantage, as the time needed to evaluate each stagewise reward still depends on $|\mathcal{X}|$. We will discuss this computational advantage further in §5.

# 4. LEARNING THE HOTSPOT FIELD MAP

Traditionally, a hotspot is defined as a location where its measurement exceeds a pre-defined extreme. But, hotspot locations do not usually occur in isolation but in clusters. So, it is useful to characterize hotspots with spatial properties. Accordingly, we define a hotspot field to vary as a realization of a random field $\{Y_x > 0\}_{x \in \mathcal{X}}$ with a positively skewed sampling distribution (e.g., Fig. 1b).

**Gaussian Process**. A widely-used random field to model environmental phenomena is the GP [4, 8]. The stationary covariance structure of GP is very sensitive to strong positive skewness of field measurements and can thus be destabilized by a few extreme ones [9]. So, if GP is used to model a hotspot field directly, it may not map well. To remedy this, a standard statistical practice is to take the log of the measurements (i.e., $Z_x = \log Y_x$) to remove skewness and extremity, and use GP to map the *log-measurements*. The entropy criterion (1) is therefore optimized in the transformed log-scale.

We will apply $i$MAXP(1) to sampling GP and determine if $\pi^1$ exhibits adaptive and hotspot sampling properties. Let $\{Z_x\}_{x \in \mathcal{X}}$ denote a GP, i.e., the joint distribution over any finite subset of $\{Z_x\}_{x \in \mathcal{X}}$ is Gaussian [6]. The GP can be completely specified by its mean $\mu_{Z_x} \triangleq \mathbb{E}[Z_x]$ and covariance $\sigma_{Z_x Z_u} \triangleq \mathrm{cov}[Z_x, Z_u]$ for $x, u \in \mathcal{X}$. We adopt a common assumption that the GP is second-order stationary, i.e., it has a constant mean and a stationary covariance structure (i.e., $\sigma_{Z_x Z_u}$ is a function of $x - u$ for all $x, u \in \mathcal{X}$). In this paper, we assume that the mean and covariance structure of $z_x$ are known. Given $d_n$, the distribution of $Z_x$ is Gaussian with posterior mean and covariance

$$\mu_{Z_x|d_n} = \mu_{Z_x} + \Sigma_{x\mathbf{x}_{0:n}} \Sigma_{\mathbf{x}_{0:n}\mathbf{x}_{0:n}}^{-1} \{\mathbf{z}_{\mathbf{x}_{0:n}} - \boldsymbol{\mu}_{\mathbf{Z}_{\mathbf{x}_{0:n}}}\}^\top \qquad (7)$$

$$\sigma_{Z_x Z_u|d_n} = \sigma_{Z_x Z_u} - \Sigma_{x\mathbf{x}_{0:n}} \Sigma_{\mathbf{x}_{0:n}\mathbf{x}_{0:n}}^{-1} \Sigma_{\mathbf{x}_{0:n}u} \qquad (8)$$

where, for the location components $v, w$ of $\mathbf{x}_{0:n}$, $\boldsymbol{\mu}_{\mathbf{Z}_{\mathbf{x}_{0:n}}}$ is a row vector with mean components $\mu_{Z_v}$, $\Sigma_{x\mathbf{x}_{0:n}}$ is a row vector with covariance components $\sigma_{Z_x Z_v}$, $\Sigma_{\mathbf{x}_{0:n}u}$ is a column vector with covariance components $\sigma_{Z_v Z_u}$, and $\Sigma_{\mathbf{x}_{0:n}\mathbf{x}_{0:n}}$ is a covariance matrix with components $\sigma_{Z_v Z_w}$. An important property of $\sigma_{Z_x Z_u|d_n}$ is its independence of $\mathbf{z}_{\mathbf{x}_{1:n}}$.

Policy $\pi^1$ can be reduced to be *non-adaptive*: observe that each stagewise reward is independent of the measurements

$$\mathbb{H}[\mathbf{Z}_{\tau(\mathbf{x}_i,\mathbf{a}_i)}|d_i] = \log \sqrt{(2\pi e)^k\ |\Sigma_{\mathbf{Z}_{\tau(\mathbf{x}_i,\mathbf{a}_i)}|d_i}|} \qquad (9)$$

where $\Sigma_{\mathbf{Z}_{\tau(\mathbf{x}_i,\mathbf{a}_i)}|d_i}$ is a covariance matrix with components $\sigma_{Z_x Z_u|d_i}$, $x, u$ of $\tau(\mathbf{x}_i, \mathbf{a}_i)$, that are independent of $\mathbf{z}_{\mathbf{x}_{1:n}}$. As a result, it follows from (6) that $U_i^{\pi^1}(d_i)$ and $\pi_i^1(d_i)$ are independent of $\mathbf{z}_{\mathbf{x}_{1:n}}$ for $i = 0, \ldots, n-$

1. The expectations in $i$MAXP(1) (6) can then be integrated out. As a result, $i$MAXP(1) for sampling GP can be reduced to a 1-stage deterministic problem $U_0^{\pi^1}(d_0) = \sum_{i=0}^{n-1} \max_{\mathbf{a}_i} \mathbb{H}[\mathbf{Z}_{\tau(\mathbf{x}_i,\mathbf{a}_i)}|d_i] = \max_{\mathbf{a}_0,\ldots,\mathbf{a}_{n-1}} \sum_{i=0}^{n-1} \mathbb{H}[\mathbf{Z}_{\tau(\mathbf{x}_i,\mathbf{a}_i)}|d_i] = \max_{\mathbf{a}_{0:n-1}} \mathbb{H}[\mathbf{Z}_{\tau(\mathbf{x}_{0:n-1},\mathbf{a}_{0:n-1})}|d_0] = U_0^{\pi^n}(d_0)$. This indicates the induced optimal values from solving $i$MAXP(1) and $i$MAXP($n$) are equal. So, $\pi^1$ offers no performance advantage over $\pi^n$.

Based on the above analysis, the following sufficient conditions, when met, guarantee that adaptivity has no benefit under an assumed environmental model:

THEOREM 4.1. *If $\mathbb{H}[\mathbf{Z}_{\tau(\mathbf{x}_i,\mathbf{a}_i)}|d_i]$ is independent of $\mathbf{z}_{\mathbf{x}_{1:n}}$ for stage $i = 0,\ldots,n-1$, $i$MAXP(1) and $\pi^1$ can be reduced to be single-staged and non-adaptive, respectively.*

For example, Theorem 4.1 also holds for the simple case of an *occupancy grid map* modeling an obstacle-ridden environment, which typically assumes $z_x$ for $x \in \mathcal{X}$ to be independent. As a result, $\mathbb{H}[\mathbf{Z}_{\tau(\mathbf{x}_i,\mathbf{a}_i)}|d_i]$ can be reduced to a sum of prior entropies over the unobserved locations $\tau(\mathbf{x}_i,\mathbf{a}_i)$, which are independent of $\mathbf{z}_{\mathbf{x}_{1:n}}$.

Policy $\pi^1$ performs *wide-area coverage* only: to maximize stagewise rewards (9), $\pi^1$ selects new locations with large posterior variance for observation. If we assume isotropic covariance structure (i.e., the covariance $\sigma_{Z_x Z_u}$ decreases monotonically with $||x-u||$) [6], the posterior data $d_i$ provide the least amount of information on unobserved locations that are far away from all observed locations. As a result, the posterior variance of unobserved locations in sparsely sampled regions are still largely unreduced by the posterior data $d_i$ from the observed locations. Hence, by exploring the sparsely sampled areas, a large expected entropy can be absorbed into the selected observation paths. But, the field of *original* measurements may not be mapped well because the under-sampled hotspots with extreme, highly-varying measurements contribute considerably to map entropy in the original scale, as discussed below.

**Log-Gaussian Process**. We will use another non-parametric probabilistic model called a $\ell$GP to map the original, rather than the log-, measurements directly, and hence optimize the entropy criterion (1) in the original scale. Let $\{Y_x\}_{x\in\mathcal{X}}$ denote a $\ell$GP: if $Z_x = \log Y_x$, $\{Z_x\}_{x\in\mathcal{X}}$ is a GP. A $\ell$GP can model a field with hotspots that exhibit much higher spatial variability than the rest of the field: Fig. 1 compares realizations of $\ell$GP and GP; the GP realization results from taking the log of the $\ell$GP measurements. This does not just dampen the extreme measurements, but also dampens and amplifies the difference between extreme and small measurements respectively, thus removing the positive skew. Compared to the GP realization, the $\ell$GP one thus exhibits higher spatial variability within hotspots but lower variability

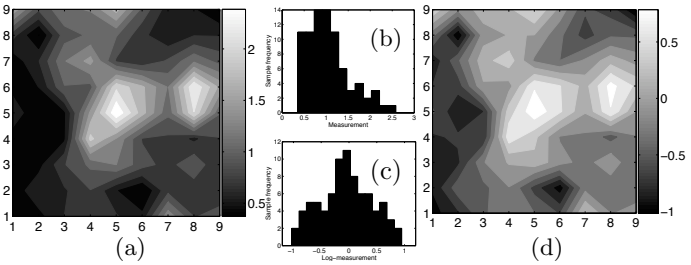


**Figure 1: Hotspot field simulation: (a-b) $\ell$GP and (c-d) GP.**

in the rest of the field. This intuitively explains why wide-area coverage suffices for GP but hotspot sampling is further needed for $\ell$GP.

Policy $\pi^1$ is *adaptive*: observe that each stagewise reward depends on the previously sampled data $d_i$:

$$\mathbb{H}[\mathbf{Y}_{\tau(\mathbf{x}_i,\mathbf{a}_i)}|d_i] = \log\sqrt{(2\pi e)^k|\Sigma_{\mathbf{Z}_{\tau(\mathbf{x}_i,\mathbf{a}_i)}|d_i}|} + \boldsymbol{\mu}_{\mathbf{Z}_{\tau(\mathbf{x}_i,\mathbf{a}_i)}|d_i}\mathbf{1}^\top \quad (10)$$

where $\boldsymbol{\mu}_{\mathbf{Z}_{\tau(\mathbf{x}_i,\mathbf{a}_i)}|d_i}$ is a mean vector with components $\mu_{Z_x|d_i}$ for $x$ of $\tau(\mathbf{x}_i,\mathbf{a}_i)$. Since $\mu_{Z_x|d_i}$ depends on $d_i$ by (7), $\mathbb{H}[\mathbf{Y}_{\tau(\mathbf{x}_i,\mathbf{a}_i)}|d_i]$ depends on $d_i$. Consequently, it follows from (6) that $U_i^{\pi^1}(d_i)$ and $\pi_i^1(d_i)$ depend on $d_i$ for $i = 0,\ldots,n-1$. Hence, $\pi^1$ is *adaptive*.

Policy $\pi^1$ performs both *hotspot sampling* and *wide-area coverage*: to maximize stagewise rewards (10), $\pi^1$ selects new locations with large Gaussian posterior variance and mean for observation. So, it directs exploration towards sparsely sampled areas and hotspots.

## 5. VALUE-FUNCTION APPROXIMATIONS

**Strictly Adaptive Exploration**. With a team of $k > 1$ robots, $\pi^1$ collects $k > 1$ observations per stage, thus becoming *partially adaptive*. We will now derive the optimal *strictly adaptive* policy (in particular, for sampling $\ell$GP), which, among policies of all adaptivity, selects paths with the largest expected entropy. By Def. 2.2, a strictly adaptive policy has to be structured to collect only 1 observation per stage. To achieve strict adaptivity, $i$MAXP(1) (6) can be revised as follows: (1) The space $\mathcal{A}(\mathbf{x}_i)$ of simultaneous joint actions is reduced to a constrained set $\mathcal{A}'(\mathbf{x}_i)$ of joint actions that allows one robot to move to observe a new location and the other robots stay put. This tradeoff for strict adaptivity allows $\mathcal{A}'(\mathbf{x}_i)$ to grow linearly, rather than exponentially, with the number of robots; (2) We constrain each robot to explore a path of at most $n$ new adjacent locations. The horizon then spans $k \times n$, rather than $n$, stages. This reflects the additional time of exploration incurred by strict adaptivity; (3) If $\mathbf{a}_i \in \mathcal{A}'(\mathbf{x}_i)$, the assignment $\mathbf{x}_{i+1} \leftarrow \tau(\mathbf{x}_i,\mathbf{a}_i)$ moves one chosen robot to a new location $x_{i+1}$ while the other unselected robots stay put at their current locations. Then, only one component of $\mathbf{x}_i$ is changed to $x_{i+1}$ to form $\mathbf{x}_{i+1}$; the other

components of $\mathbf{x}_{i+1}$ are unchanged from $\mathbf{x}_i$. Hence, there is only one unobserved component $Y_{x_{i+1}}$ in $\mathbf{Y}_{\mathbf{x}_{i+1}}$; the other components of $\mathbf{Y}_{\mathbf{x}_{i+1}}$ are already observed in the previous stages and can be found in $d_i$. As a result, the probability distribution of $\mathbf{Y}_{\mathbf{x}_{i+1}}$ can be simplified to a univariate $Y_{x_{i+1}}$.

These revisions of $i$MAXP(1) yield the strictly adaptive exploration problem called $i$MAXP($\frac{1}{k}$):

$$\begin{aligned} U_i(d_i) &= \max_{\mathbf{a}_i \in \mathcal{A}'(\mathbf{x}_i)} \mathbb{H}[Y_{x_{i+1}}|d_i] + \\ &\quad \int f(y_{x_{i+1}}|d_i)\ U_{i+1}(d_{i+1})\ \mathrm{d}y_{x_{i+1}} \\ &= \max_{\mathbf{a}_i \in \mathcal{A}'(\mathbf{x}_i)} \mathbb{H}[Y_{x_{i+1}}|d_i] + \mathbb{E}[U_{i+1}(d_i, x_{i+1}, Y_{x_{i+1}})|d_i] \\ U_t(d_t) &= \max_{\mathbf{a}_t \in \mathcal{A}'(\mathbf{x}_t)} \mathbb{H}[Y_{x_{t+1}}|d_t] \end{aligned} \tag{11}$$

for stage $i = 0, \ldots, t-1$ where $t = kn - 1$. Without ambiguity, we omit the superscript $\pi^{\frac{1}{k}}$ (i.e., the optimal strictly adaptive policy) from the optimal value functions above.

Since $Y_{x_{i+1}}$ is continuous, it entails infinite state transitions. So, unless the expectation $\mathbb{E}[U_{i+1}(d_i, x_{i+1}, Y_{x_{i+1}})|d_i]$ can be evaluated in closed form, $i$MAXP($\frac{1}{k}$) cannot be solved exactly and needs to be approximated. For ease of exposition, we will revert to using $Z_{x_{i+1}}$ (i.e., $= \log Y_{x_{i+1}}$) for $\ell$GP in the rest of this paper.

**Approximately Optimal Exploration**. To approximate $i$MAXP($\frac{1}{k}$), we employ the proposed method in [3] of first approximating the expectation from below using generalized Jensen bound. To do this, we need the following convexity result for $i$MAXP($\frac{1}{k}$) (11):

LEMMA 5.1. *$U_i(d_i)$ is convex in $\mathbf{z}_{\mathbf{x}_{0:i}}$ for $i = 0, \ldots, t$.*

Let the support of $Z_{x_{i+1}}$ given $d_i$ be partitioned into $\nu$ disjoint intervals $\mathcal{Z}^{[j]}_{x_{i+1}}$ for $j = 1, \ldots, \nu$. Then,

$$\sum_{j=1}^{\nu} p^{[j]}_{x_{i+1}} U_{i+1}(d_i, x_{i+1}, z^{[j]}_{x_{i+1}}) \leq \mathbb{E}[U_{i+1}(d_i, x_{i+1}, Z_{x_{i+1}})|d_i] \tag{12}$$

where

$$p^{[j]}_{x_{i+1}} \triangleq P(z_{x_{i+1}} \in \mathcal{Z}^{[j]}_{x_{i+1}}|d_i),\ z^{[j]}_{x_{i+1}} \triangleq \mathbb{E}[Z_{x_{i+1}}|d_i, \mathcal{Z}^{[j]}_{x_{i+1}}]\,.$$

By increasing $\nu$ to refine the partition, the bound can be improved. The approximate problem $\underline{i}$MAXP($\frac{1}{k}$) is constructed by replacing the expectation in $i$MAXP($\frac{1}{k}$) with the lower Jensen bound (12) to yield the optimal value functions $\underline{U}^{\nu}_i(d_i)$ for $i = 0, \ldots, t$ and optimal policy $\underline{\pi}^{\frac{1}{k}}$. The previous results of [3] guarantee that $\underline{U}^{\nu}_0(d_0)$ is a pessimistic estimate of largest expected entropy achieved by $\pi^{\frac{1}{k}}$, and $\underline{\pi}^{\frac{1}{k}}$ can achieve an expected entropy not worse than $\underline{U}^{\nu}_0(d_0)$.

**Real-Time Dynamic Programming**. For our bounding approximation scheme, the state size grows exponentially with the number of stages. This is due to the nature of dynamic programming problems (e.g., $\underline{i}$MAXP($\frac{1}{k}$)), which takes into account all possible states. To alleviate this computational difficulty, we modify the anytime algorithm URTDP in [3] based on $\underline{i}$MAXP($\frac{1}{k}$), which can guarantee its policy performance in real time. It simulates greedy exploration paths through a large state space, resulting in desirable properties of focused search and good anytime behavior. The greedy exploration is guided by computationally efficient, informed initial heuristic bounds independent of state size.

URTDP($d_0, t$):
  **while** $\overline{U}_0(d_0) - \underline{U}_0(d_0) > \alpha$ **do**
    SIMULATED-PATH($d_0, t$)

SIMULATED-PATH($d_0, t$):
1: $i \leftarrow 0$
2: **while** $i < t$ **do**
3: $\quad \mathbf{a}^*_i \leftarrow \arg\max_{\mathbf{a}_i} \overline{Q}_i(\mathbf{a}_i, d_i)$
4: $\quad \forall j, \quad \Xi_j \leftarrow p^{[j]}_{x^*_{i+1}} \{\overline{U}_{i+1}(d_i, x^*_{i+1}, z^{[j]}_{x^*_{i+1}}) - \underline{U}_{i+1}(d_i, x^*_{i+1}, z^{[j]}_{x^*_{i+1}})\}$
5: $\quad z \leftarrow$ sample from distribution at points $z^{[j]}_{x^*_{i+1}}$ of probability $\Xi_j / \sum_k \Xi_k$
6: $\quad d_{i+1} \leftarrow d_i, x^*_{i+1}, z$
7: $\quad i \leftarrow i + 1$
8: $\overline{U}_i(d_i) \leftarrow \max_{\mathbf{a}_i} \mathbb{H}[Y_{x_{i+1}}|d_i],\ \underline{U}_i(d_i) \leftarrow \overline{U}_i(d_i)$
9: **while** $i > 0$ **do**
10: $\quad i \leftarrow i - 1$
11: $\quad \overline{U}_i(d_i) \leftarrow \max_{\mathbf{a}_i} \overline{Q}_i(\mathbf{a}_i, d_i)$
12: $\quad \underline{U}_i(d_i) \leftarrow \max_{\mathbf{a}_i} \underline{Q}_i(\mathbf{a}_i, d_i)$

**Algorithm 1:** URTDP ($\alpha$ is user-specified bound).

In URTDP (Algorithm 1), each simulated path involves an alternating selection of actions and their corresponding outcomes till the last stage. Each action is selected based on the upper bound (line 3). For each encountered state, the algorithm maintains both lower and upper bounds, which are used to derive the uncertainty of its corresponding optimal value function. It exploits them to guide future searches in an informed manner; it explores the next state/outcome with the greatest amount of uncertainty (lines 4-5). Then, the algorithm backtracks up the path to update the upper heuristic bounds using $\max_{\mathbf{a}_i} \overline{Q}_i(\mathbf{a}_i, d_i)$ (line 11) where

$$\overline{Q}_i(\mathbf{a}_i, d_i) \triangleq \mathbb{H}[Y_{x_{i+1}}|d_i] + \sum_{j=1}^{\nu} p^{[j]}_{x_{i+1}} \overline{U}_{i+1}(d_i, x_{i+1}, z^{[j]}_{x_{i+1}})$$

and the lower bounds via $\max_{\mathbf{a}_i} \underline{Q}_i(\mathbf{a}_i, d_i)$ (line 12) where

$$\underline{Q}_i(\mathbf{a}_i, d_i) \triangleq \mathbb{H}[Y_{x_{i+1}}|d_i] + \sum_{j=1}^{\nu} p^{[j]}_{x_{i+1}} \underline{U}_{i+1}(d_i, x_{i+1}, z^{[j]}_{x_{i+1}})\,.$$

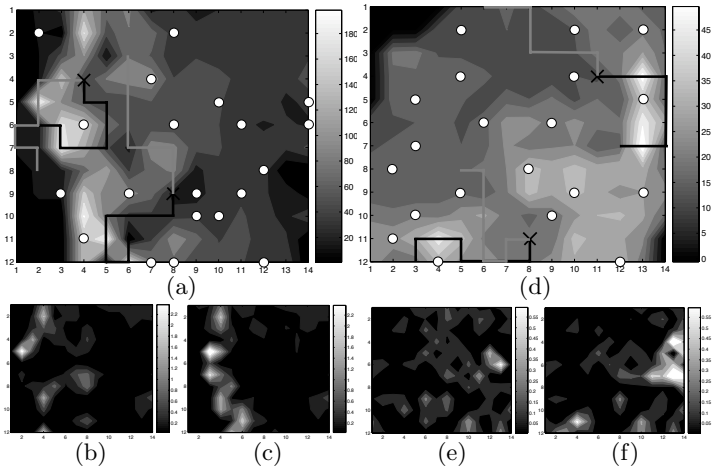


**Figure 2: (a) chl-a field with prediction error maps for (b) strictly adaptive $\underline{\pi}^{1/k}$ and (c) non-adaptive $\pi^n$: 20 units (white circles) are randomly selected as prior data. The robots start at locations marked by '×'s. The black and gray robot paths are produced by $\underline{\pi}^{1/k}$ and $\pi^n$ respectively. (d-f) K field with error maps for $\underline{\pi}^{1/k}$ and $\pi^n$.**

When an exploration policy is requested, we provide the greedy policy induced by the lower bound. The policy performance has a similar guarantee to that of $\underline{\pi}^{\frac{1}{k}}$.

We will show that the time complexity of SIMULATED-PATH$(d_0, t)$ is independent of map resolution but the same procedure in [3] is not. It is also less sensitive to increasing robot team size. Assuming no prior data and $|\mathcal{A}'(\mathbf{x}_i)| = \Delta$, the time needed to evaluate the stagewise rewards $\mathbb{H}[Y_{x_{i+1}}|d_i]$ for all $\Delta$ new locations $x_{i+1}$ (i.e., using Cholesky factorization) is $\mathcal{O}(t^3 + \Delta t^2)$, which is independent of $|\mathcal{X}|$ and results in $\mathcal{O}(t(t^3 + \Delta(t^2 + \nu)))$ time to run SIMULATED-PATH$(d_0, t)$. In contrast, the time needed to evaluate the stagewise rewards in [3] is $\mathcal{O}(t^3 + \Delta(t^2 + |\mathcal{X}|t) + |\mathcal{X}|t^2)$, which depends on $|\mathcal{X}|$ and entails $\mathcal{O}(t(t^3 + \Delta(t^2 + |\mathcal{X}|t + \nu) + |\mathcal{X}|t^2))$ time to run the same procedure. When the joint action set size $\Delta$ increases with larger robot team size, the time to run the procedure in [3] increases faster than that of ours due to the gradient factor $|\mathcal{X}|t$ involving large domain size. In §6, we will report the time taken to run this procedure empirically.

## 6. EXPERIMENTS AND DISCUSSION

This section evaluates, empirically, approximately optimal strictly adaptive policy $\underline{\pi}^{\frac{1}{k}}$ on 2 real-world datasets exhibiting positive skew: (a) June 2006 plankton density data (Fig. 2a) of Chesapeake Bay bounded within lat. $38.481 - 38.591$N and lon. $76.487 - 76.335$W, and (b) potassium distribution data (Fig. 2d) of Broom's Barn farm spanning 520m by 440m. Each region is discretized into a $14 \times 12$ grid of sampling units. Each unit $x$ is, respectively, associated with (a) plankton density $y_x$ (chl-a) in mg m$^{-3}$, and (b) potassium level $y_x$ (K) in mg l$^{-1}$. Each region comprises, respectively, (a) $|\mathcal{X}| = 148$ and (b) $|\mathcal{X}| = 156$ such units. Using a team of 2 robots, each robot is tasked to explore 9 adjacent units in its path including its starting unit. If only 1 robot is used, it is placed, respectively, in (a) top and (b) bottom starting unit, and samples all 18 units. Each robot's actions are restricted to move to the front, left, or right unit. We use the data of 20 randomly selected units to learn the hyperparameters (i.e., mean and covariance structure) of GP and $\ell$GP through maximum likelihood estimation [6]. So, prior data $d_0$ comprise the randomly selected and robot starting units.

The performance of $\underline{\pi}^{\frac{1}{k}}$ is compared to the policies produced by state-of-the-art exploration strategies, namely, the greedy and optimal non-adaptive policies. The greedy strategies are applied to sampling GP and $\ell$GP; a greedy policy repeatedly chooses a reward-maximizing action (i.e., by repeatedly solving $i$MAXP$(\frac{1}{k})$ with $t = 0$ in (11)) to form the paths. The optimal non-adaptive policy $\pi^n$ for GP is produced by solving $i$MAXP$(n)$ (5). Although $\underline{i}$MAXP$(\frac{1}{k})$ and $i$MAXP$(n)$ can be solved exactly, their state size grows exponentially with the number of stages. To alleviate this computational difficulty, we use anytime heuristic search algorithms URTDP (Algorithm 1) and Learning Real-Time A$^*$ [2] to, respectively, solve $\underline{i}$MAXP$(\frac{1}{k})$ and $i$MAXP$(n)$ approximately.

Two performance metrics are used to evaluate the above policies: (a) *Posterior map entropy* (ENT) $\mathbb{H}[\mathbf{Y}_{\overline{\mathbf{x}}_{0:t}}|d_t]$ of the unobserved locations $\overline{\mathbf{x}}_{0:t}$ is measured in the original scale where $t = 16$ (17) for the case of 2 (1) robots. A smaller ENT implies lower map uncertainty; (b) *Mean-squared relative error* (ERR) $|\mathcal{X}|^{-1} \sum_{x \in \mathcal{X}} \{(y_x - \mu_{Y_x|d_t})/\bar{\mu}\}^2$ measures the posterior map error by using the best unbiased predictor $\mu_{Y_x|d_t}$ (i.e., $\ell$GP posterior mean) [3] of the measurement $y_x$ to predict the hotspot field where $\bar{\mu} = |\mathcal{X}|^{-1} \sum_{x \in \mathcal{X}} y_x$. Although this criterion is not the one being optimized, it allows the use of ground truth measurements to evaluate if the field is being mapped accurately. A smaller ERR implies lower map error.

Table 1 shows the results of various policies with different assumed models and robot team sizes for chl-a and K fields. For $\underline{i}$MAXP$(\frac{1}{k})$ and $i$MAXP$(n)$, the results are obtained using the policies provided by the anytime algorithms after running 120000 simulated paths.

**Plankton density data**. The results show policies for $\ell$GP achieve lower ENT and ERR than that of GP. The strictly adaptive $\underline{\pi}^{\frac{1}{k}}$ achieves lowest ENT and ERR as compared to non-adaptive and greedy policies. From Fig. 2a, $\underline{\pi}^{\frac{1}{k}}$ moves the robots to sample the hotspots showing higher spatial variability whereas $\pi^n$ moves them to sparsely sampled areas. Figs. 2b and 2c show, re-

spectively, the prediction error maps resulting from $\underline{\pi}^{\frac{1}{k}}$ and $\pi^n$; the prediction error at each location $x$ is measured using $|y_x - \mu_{Y_x|d_t}|/\bar{\mu}$. Locations with large errors are mostly concentrated in the left region where the field is highly-varying and contains higher measurements. Compared to $\underline{\pi}^{\frac{1}{k}}$, $\pi^n$ incurs large errors at more locations in or close to hotspots, thus resulting in higher ERR.

We also compare the time needed to run the first 10000 SIMULATED-PATH$(d_0, t)$'s of our URTDP algorithm to that of [3], which are 115s and 10340s respectively for 2 robots (i.e., 90× faster). They, respectively, take 66s and 2835s for 1 robot (i.e., 43× faster). So, scaling to 2 robots incurs 1.73× and 3.65× more time for the respective algorithms. Policy $\underline{\pi}^{\frac{1}{k}}$ can already achieve the performance reported in Table 1 for 2 robots, and ENT of 389.23 and ERR of 0.231 for 1 robot. In contrast, the policy of [3] only improves to ENT of 377.82 (391.85) and ERR of 0.233 (0.252) for 2 (1) robots, which are slightly worse off.

**Potassium distribution data**. The results show $\underline{\pi}^{\frac{1}{k}}$ achieves lowest ENT and ERR. From Fig. 2d, $\underline{\pi}^{\frac{1}{k}}$ again moves the robots to sample the hotspots showing higher spatial variability whereas $\pi^n$ moves them to sparsely sampled areas. Compared to $\underline{\pi}^{\frac{1}{k}}$, $\pi^n$ incurs large errors at a greater number of locations in or close to hotspots as shown in Figs. 2e and 2f, thus resulting in higher ERR.

To run 10000 SIMULATED-PATH$(d_0, t)$'s, our URTDP algorithm is 84× (48×) faster than that of [3] for 2 (1) robots. Scaling to 2 robots incurs 1.93× and 3.37× more time for the respective algorithms. Policy $\underline{\pi}^{\frac{1}{k}}$ can already achieve the performance reported in Table 1 for 1 and 2 robots. In contrast, the policy of [3] achieves worse ENT of 67.132 (55.015) for 2 (1) robots. It achieves worse ERR of 0.032 for 2 robots but better ERR of 0.025 for 1 robot.

To summarize, the above results show that $\underline{\pi}^{\frac{1}{k}}$ can learn the highest-quality hotspot field map (i.e., lowest ENT and ERR) among greedy and non-adaptive strategies. After evaluating whether MAXP vs. $i$MAXP planners are time-efficient for real-time deployment, we observe $\underline{\pi}^{\frac{1}{k}}$ can achieve mapping performance comparable to the policy of [3] using significantly less time, and the incurred planning time is also less sensitive to larger robot team size.

## 7. CONCLUSION

This paper describes an information-theoretic adaptive path planner, $i$MAXP, for actively exploring a hotspot field map. Like MAXP, it performs both hotspot sampling and wide-area coverage to minimize map uncertainty (§4). In contrast to MAXP, the time complexity of solving (reward-maximizing) $i$MAXP approximately is independent of map resolution, which is clearly advantageous in large-scale exploration and mapping. It is also less sensitive to increasing robot team size. For our future work, we will test the $i$MAXP-based planner on the robotic sensor boats in our Telesupervised Adaptive Ocean Sensor Fleet (TAOSF) project for mapping harmful algal blooms.

**Table 1: Performance comparison of exploration policies for (a) chl-a and (b) K fields: 1R (2R) denotes 1 (2) robots.**

| (a) chl-a field | | ENT | | ERR | |
|---|---|---|---|---|---|
| Exploration policy | Model | 1R | 2R | 1R | 2R |
| Strictly adaptive $\underline{\pi}^{1/k}$ | $\ell$GP | 381.37 | 376.19 | 0.183 | 0.232 |
| Greedy | $\ell$GP | 382.97 | 383.55 | 0.292 | 0.258 |
| Non-adaptive $\pi^n$ | GP | 390.62 | 399.63 | 0.415 | 0.320 |
| Greedy | GP | 392.35 | 392.51 | 0.300 | 0.336 |
| (b) K field | | ENT | | ERR | |
| Exploration policy | Model | 1R | 2R | 1R | 2R |
| Strictly adaptive $\underline{\pi}^{1/k}$ | $\ell$GP | 47.330 | 48.287 | 0.029 | 0.021 |
| Greedy | $\ell$GP | 61.080 | 56.181 | 0.046 | 0.030 |
| Non-adaptive $\pi^n$ | GP | 67.084 | 59.318 | 0.043 | 0.036 |
| Greedy | GP | 58.704 | 64.186 | 0.043 | 0.033 |


## Acknowledgements

We would like to thank Dr R. Webster from Rothamsted Research for providing the Broom's Barn Farm data.


## 8. REFERENCES


[1] D. A. Cohn, Z. Ghahramani, and M. I. Jordan. Active learning with statistical models. *J. Artif. Intell. Res.*, 4:129–145, 1996.

[2] R. Korf. Real-time heuristic search. *Artif. Intell.*, 42(2-3):189–211, 1990.

[3] K. H. Low, J. M. Dolan, and P. Khosla. Adaptive multi-robot wide-area exploration and mapping. In *Proc. AAMAS*, pages 23–30, 2008.

[4] A. Meliou, A. Krause, C. Guestrin, W. Kaiser, and J. M. Hellerstein. Nonmyopic informative path planning in spatio-temporal models. In *Proc. AAAI*, pages 602–607, 2007.

[5] D. O. Popa, M. F. Mysorewala, and F. L. Lewis. EKF-based adaptive sampling with mobile robotic sensor nodes. In *Proc. IROS*, pages 2451–2456, 2006.

[6] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006.

[7] M. C. Shewry and H. P. Wynn. Maximum entropy sampling. *J. Applied Stat.*, 14(2):165–170, 1987.

[8] A. Singh, A. Krause, C. Guestrin, W. Kaiser, and M. Batalin. Efficient planning of informative paths for multiple robots. In *Proc. IJCAI*, pages 2204–2211, 2007.

[9] R. Webster and M. Oliver. *Geostatistics for Environmental Scientists*. John Wiley & Sons, 2nd edition, 2007.