2014 IEEE International Conference on Robotics & Automation (ICRA)
Hong Kong Convention and Exhibition Center
May 31 - June 7, 2014. Hong Kong, China

Sequential Allocation of Sampling Budgets in Unknown Environments

P. Michael Furlong and David Wettergreen[1]

***Abstract*— This paper presents an algorithm based on ecological models of foraging and that uses uncertainty in scientific observations made by the robot to value future actions. It is the hypothesis of this work that the foraging strategy will be an improvement over strategies based on principles from the design of experiments literature for small budget sizes. The experiment in this paper shows that for small budget sizes the new algorithm performs at least as well as other methods. However the new algorithm does not exhaust its sampling budget by default and thus needs to be modified to outperform other approaches for large sample budgets, which provides opportunity for future work in this area.**

I. INTRODUCTION

Robot scientists seldom have sufficient sampling budget to thoroughly investigate all phenomena of interest. Time, energy, resources, or communication bandwidth always impose some limit. Consequently samples must be deployed in a way that maximizes information without knowledge of what opportunities lie ahead. This paper examines strategies and identifies a method to make good decisions to give up on available sampling opportunities in the hope of finding better ones.

Traditional strategies for selecting informative actions assume all sampling actions are possible at any time. However planetary exploration missions resemble a sequence of encounters where only a subset of sampling actions are possible and no knowledge of future encounters is available. This work proposes that *strategies based on ecological models of foraging will outperform strategies based on standard design of experiments approaches for small budget sizes*.

The mission in this paper is to characterize – i.e. learn the distributions of – different random variables. In the case of the Life in the Atacama Desert project the random variables are different classifications of soil and their distributions are the abundance of subsurface microbial life in that soil type. The robot scientists are trying to learn these distributions by sampling the different random variables.

Robots exploring over long distances in unfamiliar areas do not have the luxury of knowing how many future sampling opportunities remain, nor is it guaranteed that all random variables will be encountered as they continue to explore. To help the robot scientist make the decision to leave or not we employ a model of optimal foraging based on Mean Value Theorem [1],[2] to choose between available opportunities and unknown future opportunities.

This research was supported by NASA under grants NNX11-AJ87G, ASTEP program.

[1] The Robotics Institute, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA, USA, 15213 {furlong,dsw} at cmu.edu

This paper presents the results of simulating exploration along a transect – a path across terrain – characterizing the abundance of life in subsurface habitats, the objective of the mission of Zoe, shown in Figure 1. The experiment tests four different sampling strategies on multiple simulated transects and compares the average performance.

Fig. 1. Zoe in the Atacama desert, before beginning a transect.

The transect is simulated by repeatedly presenting sampling opportunities to the robot, the robot the has the choice to either sample or move on to the next sampling opportunity. Robots completing the traverse with remaining samples represent lost opportunities to improve their hypotheses. Similarly robots that are too liberal with their sampling may waste samples on already well characterised stimuli, yielding little additional information.

The four strategies' performance is measured by the average reconstruction error in the learned distributions of the random variables on the transect. We show empirically that for small sampling budgets strategies that use the ecologically inspired approach perform at least as well as result-agnostic strategies based on optimal experiment design.

II. BACKGROUND

Previous approaches to planetary scale science autonomy fall down in two respects. Firstly, these approaches model scientific exploration as a standard exploration/exploitation problem. A model that does not necessarily hold for planetary exploration. Secondly, they do not use the output of the scientific measurements to improve how the robots select between sampling actions. For stationary processes Bayesian experiment design dictates that the optimal set of experiments can be determined without ever knowing the results of those experiments [3]. However real world

quantities are not necessarily stationarity and they may not even obey a function.

A. Foraging as Exploration

The exploration/exploitation problem asks the question: Is an agent rewarded better by exploiting already acquired knowledge or by exploring different options and improving that knowledge? The multi-armed bandit [4] was introduced to address the exploration/exploitation trade-off with a limited sampling budget. Multi-armed bandits model a fixed list of experiments as different slot machines each with their own random payout. An arm of a bandit is a metaphor for a random variable and the reward for playing that arm reveals information about that random variable. A shortcoming of the multi-armed bandit approach is that it assumes that at any given time all random variables are known and are available to conduct.

Active learning assumes an oracle and as such does not map well to exploration in unknown environments. In approaches like those of Robbins [4] or Balcan [5] the agent conducting experiments has at any time the opportunity to sample random variable they are characterizing. This is not the case in planetary exploration, we can only sample from those random variables that are present as robots follow their trajectories. The inaccuracy of the oracle model has been previously identified by Donmez and Carbonell [6].

Foraging theory provides an answer to the question of whether to stay or to go in the face of unknown future opportunities. This stands in contrast to the standard exploration/exploitation problem choosing from known sampling opportunities.

Optimal foraging strategies devised by Charnov [2] describe how predators hunt in different geographic regions with different levels of resources. Animals make the decision to forage by comparing the value of the options it has in front of it to the expected value of what it may obtain by searching for better options [1], less the cost of conducting a search. The distinction between exploration/exploitation and forage/engage is determined by two things: the recognition that there is not always a choice of what to explore and the realisation that the choice is between what is available and what may yet be encountered.

Kolling *et al* [1] found that humans make foraging decisions based on the arithmetic mean of the estimated values of the options they are presented with and the options that remain in the surrounding environment. From foraging literature we learn to compute the value of searching in an environment by taking the arithmetic mean of what is thought to be in that environment. The decision rule to stay or leave is a very simple comparison between the value of the current opportunity and the value of the environment.

Optimal foragers considering three things when choosing to leave a resource: Expected value of the current opportunity, the expected value of the rest of the environment, and the cost of searching for new opportunities [2],[1]. To adapt foraging to exploration we need to answer the question: What is the value of an option presented to the explorer? To answer that question we look to active learning.

B. Active learning

In active learning agents get to choose examples in order to resolve uncertainty or inaccuracy in models they are learning. An early version of active learning is the multi-armed bandit problem. The k-armed bandit was introduced in [4] as a means of sequentially selecting which experiments to conduct. In Robbins' work [4] selecting which experiment to conduct next is modelled on determining the payouts of one-armed bandit machines, where each machine represents a different experiment. The player has a fixed sampling budget and has to sequentially choose which machine to play, trading off exploiting the expected rewards for the different arms and exploring the different arms learning more accurately the payouts of those arms.

Lai *et al.* [8] introduced the Upper Confidence Bound (UCB) rule which values sampling opportunities with the sum of the expected reward for a sampling opportunity and a term that tries to balance the samples amongst all types of sampling opportunities.

$$Value = \mathbb{E}\left[R_i\right] + \sqrt{\frac{2\ln t_i}{T}}$$

Where R_i is the reward for sampling opportunity i, t_i is the number of times i has been sampled, and T is the total number of samples distributed. Work on proving the bounds of this algorithm has been continued by Agarawal [9] and Auer and Ortner[10].

Other approaches to the bandit problem use reward plus the uncertainty of that reward to indicate value. We see this in the work of Burnetas and Katehakis [11], and Auer [12]. This is a sentiment seen in other work, like the optimistic planners of Jurgen Schmidhuber's group [13], [14], [15], [16]. They choose actions that maximize the expected information gain with respect to some model they are learning. The most valuable actions are the ones that result in the greatest shift in the distribution the learner is building.

Balcan [5] presents a method for learning classifiers by requesting samples in the input space of the function where the classification error is the greatest. Classification error and uncertainty in function value are fungible quantities in this case. An analogy can be drawn between the classifiers used in [5] and the bandit arms used by Auer and Ortner[10].

Thompson and Wettergreen [17] maximize diversity of collected samples by using mutual information sampling. This approach ensures diversity in the collected sample set, an act that reduces uncertainty in the input space of a function. Neither mutual information nor maximum entropy sampling methods, when used with stationary Gaussian processes, take into account the dependent variable when selecting samples.

The prior work described above assumes one is choosing among a number of options and want to choose the maximally informative one. While choosing the maximally informative option is a useful guiding principle when robot

explorers are presented with a number of sampling opportunities, it does not address the problem that explorers may have to give up a sampling opportunity in the hopes of finding better ones. Further it is not guaranteed that there is no cost associated with getting to sampling opportunities, an assumption commonly made when querying an oracle.

The prior work yields two observations. Firstly, foraging, a better model for planetary exploration, requires a measure of value of the sampling opportunities available to the exploring agent. Secondly, active learning uses uncertainty – in both input and output space of a function – to value potential exploration opportunities. What follows next is a method for exploring that reflects the limitations of a planetary setting and incorporates the result of sampling operations into decision making processes.

III. METHOD

This paper compares the proposed Foraging strategy on a simulated transect against three other strategies. The agent has to repeatedly make the decision between taking the current sampling opportunity and searching for more informative sampling opportunities. The robot is assumed to be travelling a predefined path without backtracking as it encounters different opportunities for sampling. This scenario is applicable to situations where preliminary scanning of the area to be explored is not necessarily available, for example in underground, undersea, or planetary settings.

A. Experiment Set Up

Transects are approximated by presenting the agent with 1000 sampling opportunities. Each sampling opportunity is presented as one of N random variables $s_t = i$ and a value $v_{i,t}$ that is revealed only if the agent chooses to sample the random variable. Where $i \in \{1, \ldots, N\}$ indicates the random variable being presented on the t-th presentation. In this experiment there were $N = 12$ random variables. If the random variables represent different classes of soil then $v_{i,t}$ would be the density of subsurface microbial life in soil type i on the t-th sampling opportunity.

The experiment has two conditions. In the first the probability of a random variable being presented is uniform across the number of random variables. In the second condition the probability of a random variable being presented follows the distribution in Figure 2. This distribution has one dominating random variable because in the field there is usually one dominating type of material to sample from. The dominating random variable represents the dominating material in the environment.

The success of the strategies on a transect is the error between the true and learned cumulative distribution functions (CDFs) averaged over all the random variables. The error function used is the sum of the absolute value of the difference between the empirical CDF learned by the strategy and the true CDF of the random variable. Each agent was tested on thirty different transects and their average performance is compared.

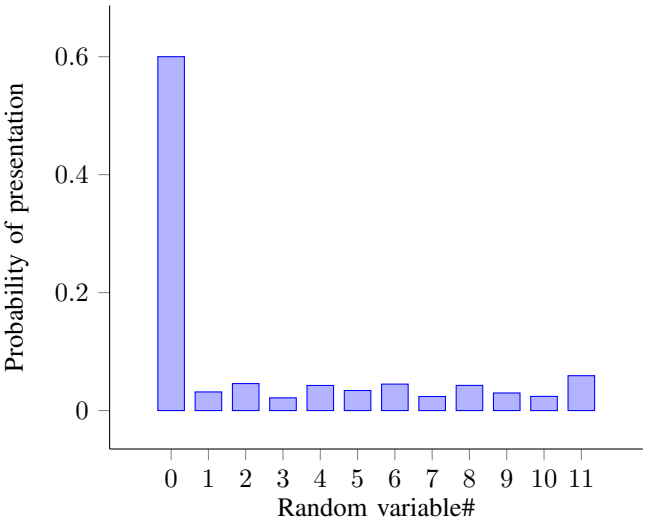

Fig. 2. The probability of presentation is the probability that a sampling opportunity for a random variable will be presented to the robot scientist in an encounter. The probabilities here are for the second experiment condition.

B. Option Value

The sampling reward is the shift in the empirical distribution function caused by taking a sample as seen in Figure 3, this is a measure of uncertainty in the learned distribution. Reward is computed by taking the sum of the absolute value of the error between the CDF before and after a sampling update, as described in Algorithm 1. The value of the random variable is the arithmetic mean of the rewards achieved from all samples of that random variable. This is a number that should decrease as a random variable is sampled.

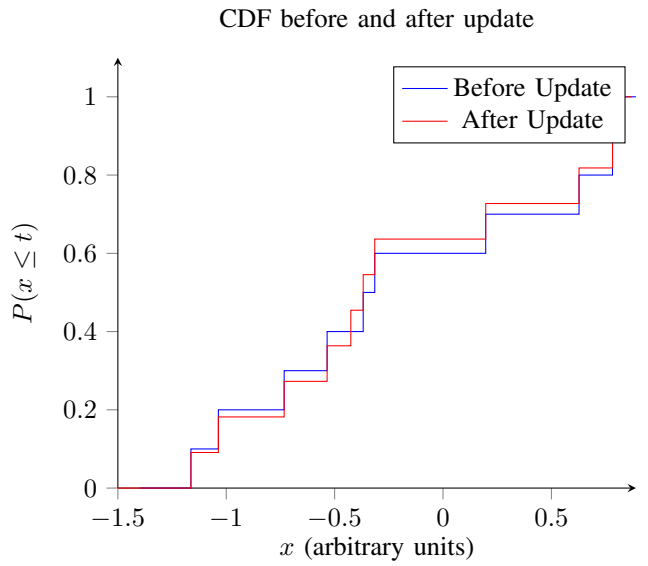

Fig. 3. The value of a sampling action is the difference between the empirical distribution function before and after a new data point is added. Larger shifts in the distribution are considered to imply the robot scientist was less certain in that hypothesis than one that yielded a smaller shift.

Algorithm 1 Option valuing.

function INIT_VALUE
  $RandomVars \leftarrow \emptyset$
  $Count \leftarrow ()$
  $Samples \leftarrow ()$
end function
function UPDATE_VALUE(s_t, v_t)
  if $s_t \notin RandomVars$ **then**
    $RandomVars \leftarrow RandomVars \cup s_t$
    $Count_{s_t} \leftarrow 0$
  end if
  $Count_{s_t} \leftarrow Count_{s_t} + 1$
  $Samples'_{s_t} \leftarrow (Samples_{s_t}, v_t)$
  $F_{old}(z) \leftarrow empirical\_dist(Samples_{s_t}, z)$
  $F_{new}(z) \leftarrow empirical\_dist(Samples'_{s_t}, z)$
  $Reward_{s_t, Count_{s_t}} \leftarrow \sum_{z \in D'_{s_t}} \| F_{old}(z) - F_{new}(z) \|$
  $Value_{s_t, Count_{s_t}} \leftarrow \frac{1}{Count_{s_t}} \sum_{i=0}^{Count_{s_t}} Reward_{s_t,i}$
  $Samples_{s_t} \leftarrow Samples'_{s_t}$
  return $Value_{s_t, Count_{s_t}}$
end function

C. Sampling Strategies

In this work we compare four different algorithms. The first three strategies do not consider the result of the sampling action or the effect it has on distributions they are learning. The Random, Always-Engage, and Uniform sampling strategies are baselines for comparison to the new Foraging strategy. The details of those strategies are given below.

1) Random Sampling: The random strategy engages with sampling opportunities with a probability q. q can be considered the desired sampling frequency. Ideally q would be tuned to the length of the transect being investigating. However it is not necessarily possible to know how many sampling opportunities are on a transect and consequently not possible to tune the parameter q *a priori*. In this experiment $q = 0.1$.

Algorithm 2 Random sampling strategy.

function INIT_RANDOM_SAMPLE($sampling\_frequency$)
  $q \leftarrow sampling\_frequency$
end function
function RANDOM_SAMPLE(s_t)
  $prob\_sample \sim U(0, 1)$
  if $prob\_sample \leq q$ **then**
    return engage
  end if
  return continue
end function

2) Always-Engage: The Always-Engage strategy engages with every sampling opportunity it encounters until its budget is exhausted. While this algorithm has no strategy in the limit of an infinite sampling budget it should have the least error. This algorithm will always use all of its sampling budget if the budget is less than the number of sampling opportunities.

Algorithm 3 Always-Engage sampling strategy.

function INIT_ALWAYS_ENGAGE_SAMPLE
  $num\_samples \leftarrow 0$
end function
function ALWAYS_ENGAGE_SAMPLE(s_t)
  if $num\_samples < sampling\_budget$ **then**
    return engage
  else
    return continue
  end if
end function

3) Uniform Sampling: Distributing samples uniformly between all the random variables is behaviour predicted by Bayesian optimal design of experiments. The Uniform sampling algorithm attempts to distribute the samples evenly among all random variables, changing the distribution as it discovers new random variables. Therefore the agent rebudgets its samples when new random variables are identified. Should any one random variable have already exceeded its new budget then it is never sampled again.

Algorithm 4 Uniform sampling strategy

function INIT_UNIFORM_SAMPLING($sampling\_budget$)
  $Budget \leftarrow sampling\_budget$
  $RandomVars \leftarrow \emptyset$
  $Count \leftarrow \emptyset$
end function
function UNIFORM_SAMPLE(s_t)
  if $|Count_{s_t}| < Budget$ **then**
    $Count_{s_t} \leftarrow Count_{s_t} + 1$
    return engage
  end if
  if $s_t \notin RandomVars$ **then**
    $RandomVars \leftarrow RandomVars \cup s_t$
    $Count_{s_t} \leftarrow Count_{s_t} + 1$
    $Budget \leftarrow sampling\_budget / \|RandomVars\|$
    return engage
  end if
  return continue
end function

4) Foraging: This algorithm compares the value of available random variable with the mean value of the known random variables – the environment value. If the mean value of all random variables is greater than the available random variable then the agent continues to search but if the current value is higher than the environment value then the agent will engage with the presented random variable.

Algorithm 5 Foraging sampling strategy

```
function INIT_FORAGE_SAMPLING(sampling_budget)
    RandomVars ← ∅
    Values ← ∅
end function
function FORAGE_SAMPLE(s_t)
    if s_t ∉ RandomVars then
        RandomVars ← RandomVars ∪ s_t
        return engage
    end if
    if Values_{s_t} ≥ E_{RandomVars}[Values] then
        return engage
    end if
    return continue
end function
```

The foraging strategy uses the uncertainty in the learned distributions for the random variables as the value for the different random variables. There is assumed a fix, unit cost for taking a sample. Since this cost is the same for all random variables it can be ignored. Finally, unlike the work of [2] this algorithm does not incorporate the cost of travelling to the next sampling opportunity.

IV. RESULTS

The four agents were tested on thirty trials of 1000 sampling opportunities. The sampling budget for the different trials were varied from 10 to 400 samples. Because there was no change in performance trends after about 200 samples Figures 4 and 5 only show results for up to 200 samples.

The graphs below report the average performance and a 95% confidence interval. When those intervals do not overlap the performance of the agents can be said to be statistically significantly different with 95% confidence under an unpaired Student's t-test.

A. Condition 1: Uniform Distribution of Sampling Opportunities

When the probability of a random variable being presented is uniform there is not a considerable difference in the performance for sampling budgets below 100. Figure 4 shows that for budgets greater than 100 the random sampling strategy plateaus. The plateau is because with more than 100 samples and a sampling probability of $\frac{1}{10}$ the agent will complete the transect before using all of its budget. That is not to reflect poorly on the random sampling algorithm – initially its performance is indistinguishable from the others – but to show how the algorithm must be tuned to the number of sampling opportunities on a transect.

B. Condition 2: Non-Uniform Distribution of Sampling Opportunities

When the distribution of random variables is non-uniform there is a difference in performance of the different algorithms. There are three important things to observe in Figure 5. First, the always-engage strategy quickly trails behind the uniform and foraging strategies, because always engaging

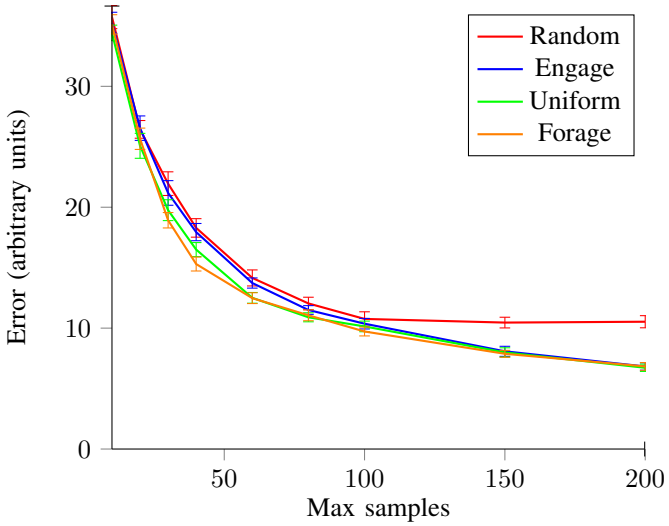

Fig. 4. There is no clear winner among the uniform, foraging, and engage strategies for sampling when all random variables are equally likely. Notice that the random algorithm plateaus after budget sizes of 100. With a sampling frequency of one in ten, the algorithm will reach the end of the transect before it has used all of its samples.

it spends approximately 60% of its samples on only one random variable. Over sampling one random variable is a result of the skewed arrival distribution in Figure 2.

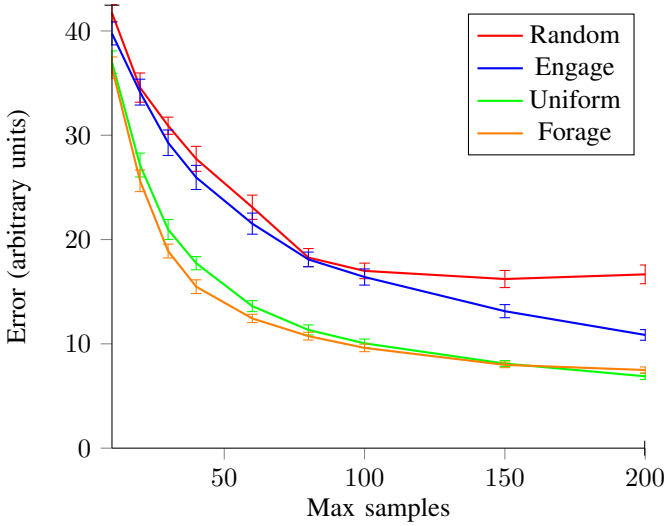

Fig. 5. For small sampling budgets the forage algorithm performs slightly, but statistically significantly better than uniform sampling. After 150 samples the foraging algorithm plateaus, simply because it is reserving its budget as nothing else meets the standard of being better than the average of the environment. The engage strategy converges towards the uniform strategy.

Second, Figure 5 shows that for sample sizes of 150 and below the foraging strategy is at least as good as the uniform strategy, both of which outperform the other two strategies for all budget sizes. For budgets of size 30, 40, and 60 the foraging strategy performs 9.8%, 12.7%, and 8.7% better than the uniform strategy, with $p < 0.05$. For budgets budget sizes uniform and foraging strategies are not statistically

significantly different.

Third, for budgets greater than 150 samples the foraging strategy does not use all its sampling budget before the transect is complete. To improve the foraging algorithm it needs to more readily dispose of its samples. Regardless, the foraging algorithm is an improvement over the uniform sampling strategy for smaller sample sizes. With a more aggressive decision rule we may also see continued improvement with larger budgets.

V. CONCLUSIONS

The hypothesis of this paper is that the foraging-style algorithm will produce improvements in performance over the supplied baseline algorithms for small budget sizes. We have empirically confirmed that hypothesis for the simulation of the transect described in Section III. Incorporating the Foraging strategy should improve the performance of robot scientists.

The Foraging algorithm does not perform as well with larger sample budgets as the Uniform algorithm. Foraging reaches the end of the transect with unused samples. Making the Foraging algorithm spend its entire sample budget is the subject of future work.

The proposed algorithm not does consider all the constraints of an exploratory mission. The cost of traversing between sampling opportunities is not considered, contextual information does not exist in this model, and it assumes a uniform distribution of random variables in the environment, even when that is not the case. We expect to gain further improvements on the algorithm by incorporating these factors.

ACKNOWLEDGMENT

The authors would like to thank Drs. Jeff Schneider, Reid Simmons, and Stephane Ross for their invaluable critiques of this work. The authors would also like to thank Greydon Foil for the picture in Figure 1.

REFERENCES

[1] N. Kolling, T. E. Behrens, R. B. Mars, and M. F. Rushworth, "Neural mechanisms of foraging," *Science*, vol. 336, no. 6077, pp. 95–98, 2012.

[2] E. L. Charnov, "Optimal foraging, the marginal value theorem," *Theoretical population biology*, vol. 9, no. 2, pp. 129–136, 1976.

[3] N. Srinivas, A. Krause, S. M. Kakade, and M. Seeger, "Gaussian process optimization in the bandit setting: No regret and experimental design," *arXiv preprint arXiv:0912.3995*, 2009.

[4] H. Robbins, "Some aspects of the sequential design of experiments," *Bulletin of the American Mathematical Society*, vol. 58, no. 5, pp. 527–535, 1952.

[5] M.-F. Balcan, A. Beygelzimer, and J. Langford, "Agnostic active learning," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 65–72.

[6] P. Donmez and J. G. Carbonell, "Proactive learning: cost-sensitive active learning with multiple imperfect oracles," in *Proceedings of the 17th ACM conference on Information and knowledge management*. ACM, 2008, pp. 619–628.

[7] E. Charnov and G. H. Orians, "Optimal foraging: some theoretical explorations," Ph.D. dissertation, University of Washington, 1973.

[8] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in applied mathematics*, vol. 6, no. 1, pp. 4–22, 1985.

[9] R. Agrawal, "Sample mean based index policies with o (log n) regret for the multi-armed bandit problem," *Advances in Applied Probability*, pp. 1054–1078, 1995.

[10] P. Auer and R. Ortner, "UCB revisited: Improved regret bounds for the stochastic multi-armed bandit problem," *Periodica Mathematica Hungarica*, vol. 61, no. 1-2, pp. 55–65, 2010.

[11] A. N. Burnetas and M. N. Katehakis, "Optimal adaptive policies for markov decision processes," *Mathematics of Operations Research*, vol. 22, no. 1, pp. 222–255, 1997.

[12] P. Auer, "Using confidence bounds for exploitation-exploration tradeoffs," *The Journal of Machine Learning Research*, vol. 3, pp. 397–422, 2003.

[13] J. Schmidhuber, "What's interesting?" 1997.

[14] ——, "Exploring the predictable," in *Advances in evolutionary computing*. Springer, 2003, pp. 579–612.

[15] ——, "Simple algorithmic theory of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes," *Journal of SICE*, vol. 48, no. 1, 2009.

[16] Y. Sun, F. Gomez, and J. Schmidhuber, "Planning to be surprised: Optimal bayesian exploration in dynamic environments," in *Artificial General Intelligence*. Springer, 2011, pp. 41–51.

[17] D. R. Thompson and D. Wettergreen, "Intelligent maps for autonomous kilometer-scale science survey," in *Proc. i-SAIRAS*, 2008.