Budgeting Samples for Exploration in Unknown Environments

P. Michael Furlong*, David S. Wettergreen*

*The Robotics Institute, Carnegie Mellon University, USA
e-mail: furlong@cmu.edu,dsw@ri.cmu.edu

Abstract

This paper presents an algorithm that improves previous efforts to incorporate ecological models of foraging into automating exoplanetary exploration. The new budgeting strategy is an improvement over previous approaches in that it attempts to exhaust its sampling budget while exploring. Simulated experiments demonstrate that the budgeting algorithm is also a better approach for small budget sizes thank more traditional approaches based on principles from the design of experiments.

1 Introduction

As the scope of planetary exploration increases, so to does the demand on human scientists to direct exploring robots. Compounding this demand are limitations in bandwidth and delays in communication that hinder the ability to communicate with and control exploring robots. Time, energy, and resources impose limits on how many times robots can collect samples. Consequently samples must be deployed in a way that maximizes information without knowledge of what opportunities lie ahead.

The objective of this work is to provide a decision making algorithm for trading off currently available sampling opportunities against future sampling opportunities. With such an algorithm robots may choose whether to sample or not without human oversight. This research simulates a robot following a pre-defined path repeatedly making the decision to either take a sample or continue to the next sampling opportunity.

Determining where an explorer should distribute samples is a problem of sequentially selecting experiments, or actions, that increase the knowledge of the world. Choosing maximally informative next actions is a problem of design of experiments or active learning. More specifically robots face sequential experiment selection where after every action the value of possible actions can change and there is a limited sampling budget.

Traditional strategies for selecting informative actions assume all sampling actions are possible at any time. In contrast, 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. Planetary exploration resembles animals foraging, having to make the decision either to consume available resources or to continue searching for more valuable resources. Similarly robots exploring long distances in unfamiliar areas do not know how many sampling opportunities remain. Further there is no guarantee that any sampling opportunities will be repeated during exploration.

This work proposes that *budget aware sampling will improve upon the foraging strategy presented in* Furlong and Wettergreen[10].

This paper introduces a new algorithm that combines optimal foraging theory [6],[7] and multi-armed bandit literature [13],[12] to enable robots to make the decision between stopping and sampling or continuing to explore the environment. Literature on sequential experiment selection gives a means for valuing different sampling opportunities. Optimal foraging theory yields a mechanism for making the decision between currently available opportunities and uncertain future opportunities. Previous approaches to science autonomy have employed design of experiments techniques to set paths for exploration [19], they do not consider the results of sampling activities when planning future activities.

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, the robot used in the Life in the Atacama Desert project. The experiment tests three different sampling strategies on multiple simulated transects and compares the average performance. In these experiments we determined that the new budgeting strategy has improved performance over the foraging strategy of [10].

Reliably prioritizing sampling opportunities frees robots from human specified exploration mission priorities. Freeing exploring agents from the throughput of human decision makers, especially in increasingly remote locations, can improve the scientific yield of robotic exploration missions.

2 BACKGROUND

Previous approaches to planetary scale science autonomy fall down in two respects. Firstly, these approaches model scientific exploration as a standard explo-

ration/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 experiment design dictates that the optimal set of experiments can be determined without ever knowing the results of those experiments [17].

2.1 Sequential Action Selection

Sequential experiment selection, a type of active learning, is addressed in the multi-armed bandit literature. The multi-armed bandit was introduced in [13] as a means of sequentially selecting which experiments to conduct with a limited budget. In Robbins' work [13] selecting experiments is modelled on determining the payouts of one-armed bandit machines – 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.* [12] 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.

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

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 [1] and Auer and Ortner[3].

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 [5] and Auer [2]. This is a sentiment seen in other work, like the optimistic planners of Jurgen Schmidhuber's group [15, 14, 16, 18]. 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 [4] presents a method for learning classifiers by requesting samples from the input space with the greatest classification error. Classification error and uncertainty in function value are fungible quantities in this case. An analogy can be drawn between the classifiers used in [4] and the bandit arms used by Auer and Ortner[3].

Thompson and Wettergreen [19] 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.

Sequential experiment selection values actions by a combination of reward and uncertainty in that reward. Since the mission of exploration is learning the reward is the reduction in uncertainty by taking actions. Seeking uncertainty is a useful way to value options presented to a learning agent but it does not address the explorer's problem of either giving up on a sampling opportunity or searching for better opportunities. Further it is not guaranteed that sampling opportunities can be accessed at no cost, an assumption commonly made when querying an oracle.

2.2 Exploration as Foraging

Active learning assumes an oracle and as such does not map well to exploration in unknown environments. In approaches like those of Robbins [13] or Balcan [4] 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 [8].

Foraging theory provides a way to make the decision to stay or to go withouth knowledge of future opportunities. This stands in contrast to the standard exploration/exploitation problem choosing from known sampling opportunities.

Optimal foraging strategies devised by Charnov [7] 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 [11], less the cost of conducting a search.

Kolling *et al* [11] 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 comparison between the value of the current opportunity and the expected 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 [7],[11]. 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.

Previous work by the authors [10] addressed the problem of exploring along a transect by employing techniques from Foraging Theory. However that work did not address the fact that the sampler had a limited budget. Agents in that work did not expend all of their samples for large budgets, a problem this research addresses.

The strategy presented by Ferri *et al.* made a comparison between the perceived value of the available sampling opportunity and an arbitrary function of the remaining number of sampling opportunities [9]. The value of a sampling opportunity was determined by a fixed threshold and the proclivity for spending samples was likewise determined by an arbitrary constant. In contrast this work employs an information theoretic measure of opportunity value and a principled measure to determine when to expend a sample.

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.

3 METHOD

There are two experiments discussed in this paper. The first compares the efficacy of the new decision making algorithm against results from previous work. The second experiment examines the sensitivity of the new budgeting algorithm to incorrectly estimating the number of sampling opportunities on a transect.

The first experiment compares the proposed Budgeting strategy on a simulated transect against the algorithms tested in [10]. The execution of the transect is simulated by repeatedly making 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, so the robot cannot travel back to previous sampling opportunities. This scenario mirrors situations where foreknowledge of the area to be explored is not available, for example in underground, undersea, or planetary settings.

The budgeting algorithm depends on knowing how many sampling opportunities are available on a transect. In unknown environments the number of sampling opportunities must be estimated. The second experiment determines the sensitivity of the budgeting algorithm to error in the estimated number of sampling opportunities.

3.1 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 first 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 1. This distribution has one dominating random variable to model an environment with one dominating type of material to sample from. The dominating random variable represents the dominating material in the environment.

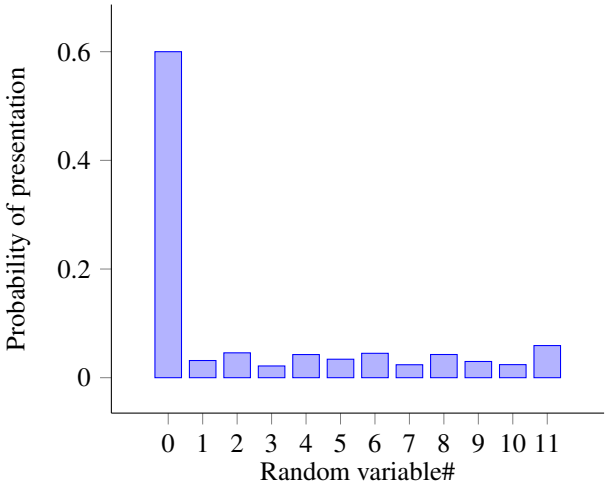

Figure 1. The probability of presentation is the probability that a sampling opportunity for a random variable will be presented to the robot scientist in an encounter. The probabilities here are for the second experiment condition.

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.

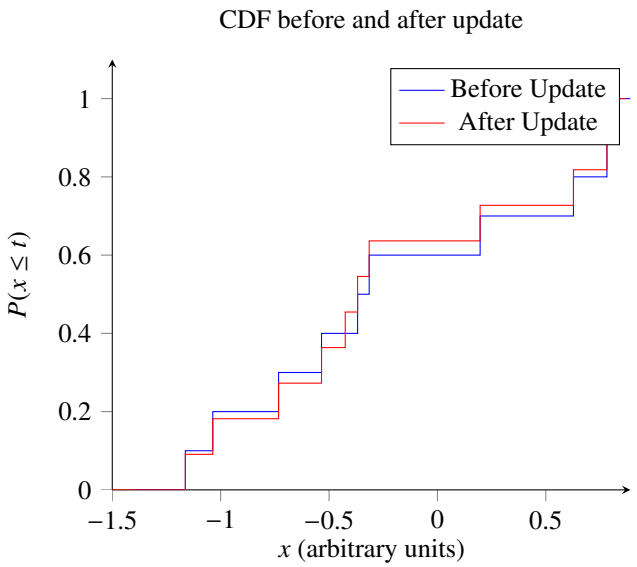

Figure 2. The value of a sampling action is the shift in the empirical distribution function due to adding a new new data point. Larger shifts in the distribution imply greater uncertainty in that distribution and thus more valuable opportunities.

The second experiment repeats the second condition of the first experiment but in this case only the budgeting strategy is used. The strategy is forced to have an inaccurate estimate of how many sampling opportunities exist on the transect. The performance is measured with $\pm 10\%, \pm 20\%, \pm 30\%$ error in the estimate of total sampling opportunities.

3.2 Option Value

The value of a random variable is the arithmetic mean of the reward history for sampling that variable. The sampling reward is the shift in the empirical distribution function caused by taking a sample as seen in Figure 2, 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 should decrease as it is sampled more.

3.3 Sampling Strategies

In this work we compare three different algorithms. The first strategy do not consider the result of the sampling action or the effect it has on distributions they are learning. The Uniform sampling strategy is a baseline for comparison to the Foraging and new Budgeting strategies. The details of these strategies are given below.

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

3.3.1 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 re-budgets 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 2 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

3.3.2 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 3 Foraging sampling strategy

function INIT_FORAGE_SAMPLING($sampling\_budget$)
  $RandomVars \leftarrow \emptyset$
  $Values \leftarrow \emptyset$
end function
function FORAGE_SAMPLE(s_t)
  if $s_t \notin RandomVars$ **then**
    $RandomVars \leftarrow RandomVars \cup s_t$
    return engage
  end if
  if $Values_{s_t} \geq \mathbb{E}_{RandomVars}[Values]$ **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. Unlike the work of [7] this algorithm does not incorporate the cost of travelling to the next sampling opportunity.

3.3.3 Budgeting

The budgeting algorithm makes the decision to sample if the encountered random variable is new to the exploring agent or if its rank is greater than or equal to $1 - \frac{remaining\_budget}{remaining\_opportunities}$. The rank of the random variable is where it is positioned in list of random variables sorted by their current value. The rank of the random variable can be considered the belief that this opportunity is worth sampling. The second term, $1 - \frac{remaining\_budget}{remaining\_opportunities}$, can be seen as the probability – provided $remaining\_budget \leq remaining\_opportunities$ – that the agent should take a sample. If the rank score is greater than the probability of sampling, then the agent engages with the sampling opportunity. Otherwise it continues along the transect.

Like the foraging strategy the budgeting strategy must sample any new random variable. This is to ensure it can compare the different random variables when making the decision to engage or forage.

Algorithm 4 Budgeting sampling strategy

function INIT_FORAGE_SAMPLING($sampling\_budget$, $estimated\_opportunities$)
  $RandomVars \leftarrow \emptyset$
  $Values \leftarrow \emptyset$
  $remaining\_samples \leftarrow sampling\_budget$
  $remaining\_opps \leftarrow estimated\_opportunities$
end function
function FORAGE_SAMPLE(s_t)
  $action \leftarrow$ continue
  if $rank(s_t, RandomVars, Values) \geq 1 - \frac{remaining\_samples}{remaining\_opportunities}$ $\wedge s_t \notin RandomVars$ **then**
    $remaining\_samples \leftarrow remaining\_samples - 1$
    $RandomVars \leftarrow RandomVars \cup s_t$
    $action \leftarrow$ engage
  end if
  $remaining\_opps \leftarrow remaining\_opps - 1$
  return $action$
end function
function RANK($s_t, RandomVars, Values$)
  $sorted\_vars \leftarrow ascending\_sort(RandomVars, key = Values)$
  return $index\_of(s_t, sorted\_vars) / len(sorted\_vars)$
end function

Since this algorithm is mindful its sampling budget we anticipate that it will out perform the foraging strategy. Like the foraging algorithm this approach does not explicitly take into account the cost of sampling or traversing between sampling sites.

4 Results

4.1 Experiment 1 Condition 1: Uniform Distribution of Sampling Opportunities

Initially the budgeting algorithm out performs the uniform and foraging strategies. However they quickly converge to similar performance for all budget sizes, within a 95% confidence interval of each other. Figure 3 shows that the budgeting algorithm does at least as good as the two competing algorithms for all budget sizes.

4.2 Experiment 1 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. The budgeting and foraging algorithms perfrom better than the uniform strategy for small budget sizes, as can be seen in Figure 5. For larger budget sizes the foraging strategy plateaus because it does not expend its entire sampling budget before the transect ends. The budgeting strategy, which does attempt to spend all of it sampling budget, improves its error with larger sampling

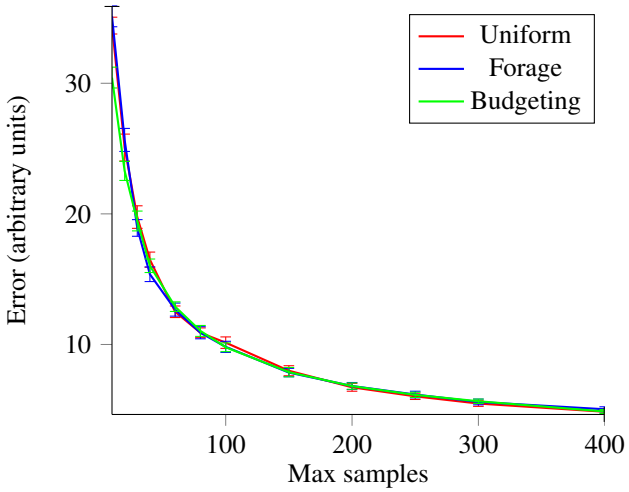

Figure 3. There is no clear winner among the uniform, foraging, and budgeting strategies for variables are equally likely, except for the smallest budget size.

budgets. However as Figure 4 shows uniform sampling still performs better for larger sampling budgets.

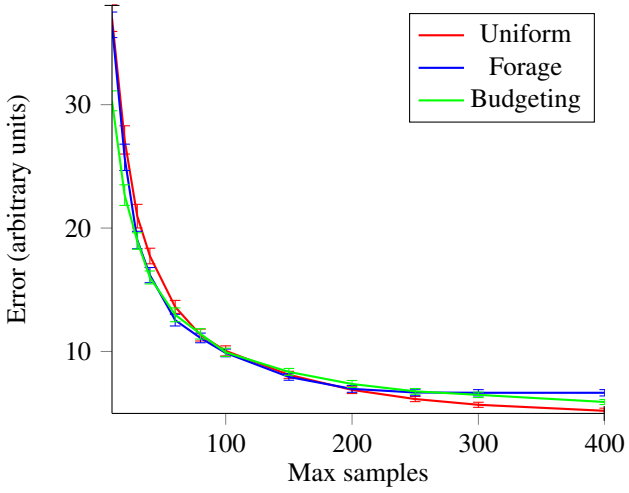

Figure 4. Initially the budgeting strategy performs better or as good as the competing strategies. However when the forgaging strategy plateaus the error of the budgeting strategy continues to improve. While the scores budgeting strategy for large sample budgets is lower-bounded by the uniform strategy.

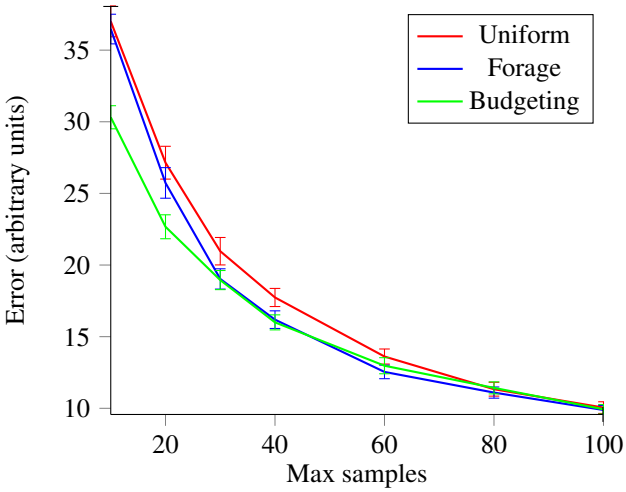

Figure 5. For small budget sizes the budgeting algorithm outperforms both the foraging and uniform strategies. As the sampling budget gets larger the performance of the algorithms converges to the other algorithms.

4.3 Experiment 2: Effect of Estimation Error on Budgeting

The second experiment determined the effect of incorrectly estimating the number of sampling opportunities. As can be seen in Figure 6 there is no significant penalty on the performance of the budgeting algorithm for under estimating the number of sampling opportunities. When the number of sampling opportunities is over estimated the performance of the budgeting algorithm decays.

5 Conclusions

The objective of the paper was develop a better algorithm than the foraging algorithm presented in previous work by the authors [10]. The budgeting algorithm does achieve that objective and provides a new rule for autonomous decision making while exploring in unknown environments. Employing these strategies should improve the performance of robot scientists.

The budgeting algorithm always performs at least as well as either of the foraging or uniform strategies. Unlike the the foraging algorithm the budgeting strategy always uses all of its sampling budget on a transect. Like the foraging strategy the budgeting strategy outperforms the uniform strategy for small budget sizes.

The budgeting strategy requires an estimate of the number of sampling opportunities on a transect. It is more harmful to overestimate the number of available sampling opportunities. This teaches us that it is best to err in favour

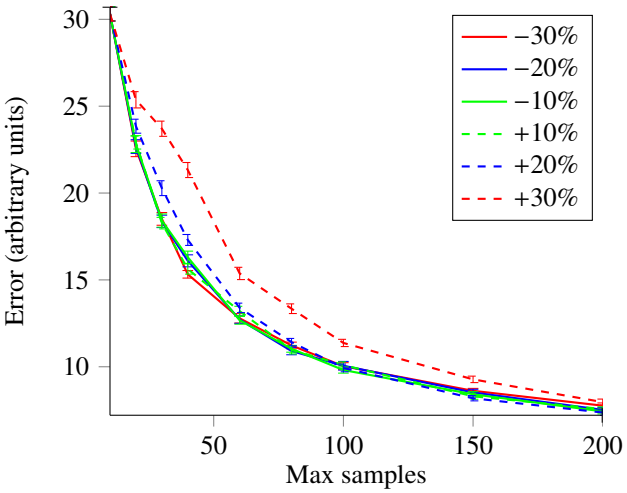

Figure 6. The budgeting strategy is harmed more by overestimating the number of samples on a transect that underestimating that number. The increased error is more pronounced for small sample sizes than large ones.

of conservative estimates in the number of opportunities to sample on a transect.

In the future this work will be expanded to incorporate estimated energy costs of traverse and the risk of vehicle failure associated with different path choices. This work will also be deployed on future missions to the Atacama Desert as part of the Life in the Atacama Desert project.

Acknowledgement

The authors would like to thank Drs. Jeff Schneider, Reid Simmons, and Stephane Ross for their advice on this work. This work is supported by the Life in the Atacama project by NASA Astrobiology Science and Technology for Exploring Planets (ASTEP) Grant NNX11AJ87G, with program executive Mary Voytek.

References

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

[2] Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *The Journal of Machine Learning Research*, 3:397–422, 2003.

[3] Peter Auer and Ronald Ortner. UCB revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 61(1-2):55–65, 2010.

[4] Maria-Florina Balcan, Alina Beygelzimer, and John Langford. Agnostic active learning. In *Proceedings of the 23rd international conference on Machine learning*, pages 65–72. ACM, 2006.

[5] Apostolos N Burnetas and Michael N Katehakis. Optimal adaptive policies for markov decision processes. *Mathematics of Operations Research*, 22(1):222–255, 1997.

[6] Eric Charnov and Gordon H Orians. *Optimal foraging: some theoretical explorations*. PhD thesis, University of Washington, 1973.

[7] Eric L Charnov. Optimal foraging, the marginal value theorem. *Theoretical population biology*, 9(2):129–136, 1976.

[8] Pinar Donmez and Jaime G Carbonell. Proactive learning: cost-sensitive active learning with multiple imperfect oracles. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 619–628. ACM, 2008.

[9] Gabriele Ferri, Michael V Jakuba, and Dana R Yoerger. A novel trigger-based method for hydrothermal vents prospecting using an autonomous underwater robot. *Autonomous Robots*, 29(1):67–83, 2010.

[10] P Michael Furlong and David Wettergreen. Sequential allocation of sampling budgets in unknown environments. In *Robotics and Automation (ICRA), 2014 IEEE International Conference on*. IEEE, 2014.

[11] Nils Kolling, Timothy EJ Behrens, Rogier B Mars, and Matthew FS Rushworth. Neural mechanisms of foraging. *Science*, 336(6077):95–98, 2012.

[12] Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[13] Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.

[14] Juergen Schmidhuber. Exploring the predictable. In *Advances in evolutionary computing*, pages 579–612. Springer, 2003.

[15] Jürgen Schmidhuber. What's interesting? 1997.

[16] Jurgen Schmidhuber. Simple algorithmic theory of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes. *Journal of SICE*, 48(1), 2009.

[17] Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*, 2009.

[18] Yi Sun, Faustino Gomez, and Jürgen Schmidhuber. Planning to be surprised: Optimal bayesian exploration in dynamic environments. In *Artificial General Intelligence*, pages 41–51. Springer, 2011.

[19] David R Thompson and David Wettergreen. Intelligent maps for autonomous kilometer-scale science survey. In *Proc. i-SAIRAS*, 2008.