Policy Search with Non-uniform State Representations for Environmental Sampling

Anonymous Authors[1]

Abstract

Surveying fragile ecosystems like coral reefs is important to monitor the effects of climate change. We present an adaptive sampling technique that generates efficient trajectories covering hotspots in the region of interest at a high rate. A key feature of our sampling algorithm is the ability to generate action plans for any new hotspot distribution using the parameters learned on other similar looking distributions.

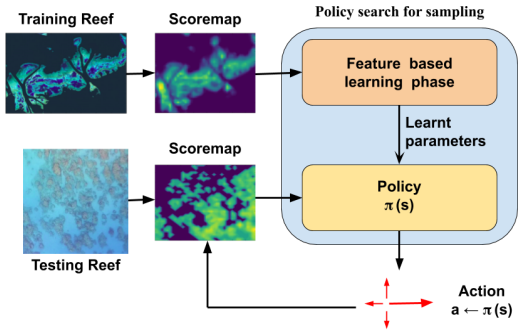
1. Introduction

In this paper, we consider monitoring the health of coral reefs by sampling visual data from the surface using an autonomous surface vehicle (ASV). Increase in the ocean temperatures has resulted in widespread coral bleaching at an ever-increasing rate (Hoegh-Guldberg, 1999) (Fig. 1a). Improved monitoring would enhance the currently poor understanding of the spatial and temporal dynamics of coral bleaching. Since we are sampling from the surface, higher information gain is provided in shallower regions where visibility is better.

We present an anytime (Zilberstein & Russell, 1993) adaptive sampling technique that generates paths to efficiently measure and then mathematically model a scalar field by performing non-uniform measurements in a given region of interest[1]. In particular, the class of scalar field we are interested is some physical or virtual parameter that varies with location, such as depth of the sea floor or algae blooms or suspended particles in air. As the measurements are collected at each sampling location, we can compute an estimate of the large-scale variation of the phenomenon of interest. We compute a sampling path that minimizes the expected time to accurately model the phenomenon of interest by visiting high information regions (hotspots) using non-myopic path generation based on reinforcement learning.

Exhaustively sampling each point of an unknown survey region (Xu et al., 2011; Choset & Pignon, 1998) can be tedious and impractical if the survey space is large and/or the phenomenon of interest has only a few regions with important information (*hotspots*). Also it has been observed that sampling rates far below the Nyquist rate can still be information preserving (Venkataramani & Bresler, 2000). This is the key guiding principle behind active and non-uniform sampling (Manjanna & Dudek, 2017; Low et al., 2008; Rahimi et al., 2005; Sadat et al., 2015).

(a) Effect of global warming on Corals (Vevers)

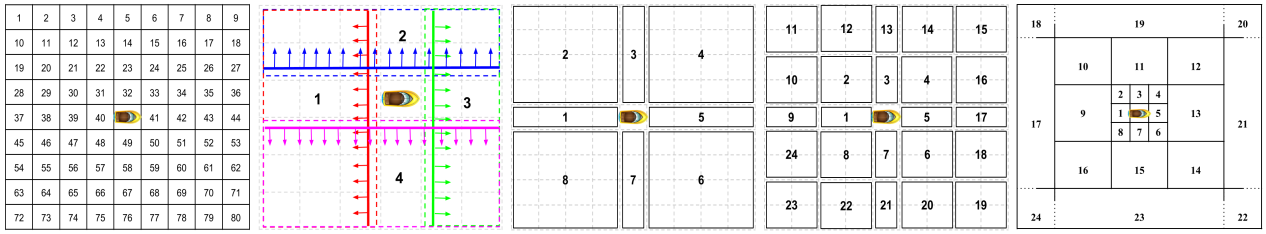


(b) System overview of our sampling approach.

Figure 1

In our approach (Fig.1b), a continuous two-dimensional sampling region is discretized into uniform grid-cells, such that the robot's position $\mathbf{x}$ can be represented by a pair of integers $\mathbf{x} \in \mathbb{Z}^2$. Each grid-cell (i, j) is assigned a score $\mathrm{q}(i, j)$ indicating the expected goodness of the visual data in that cell. The goal is to maximize the total accumulated score J over a trajectory τ within a fixed amount of time T. To specify the robot's behavior we use a parametrized policy $\pi_{\boldsymbol{\theta}}(\mathbf{s}, \mathbf{a}) = p(\mathbf{a}|\mathbf{s}; \boldsymbol{\theta})$ that maps the current state $\mathbf{s}$ of sampling to a distribution over possible *actions* $\mathbf{a}$. Our aim will be to automatically find good parameters $\boldsymbol{\theta}$, after which the policy can be deployed without additional training on new problems.

[1]This work has been accepted for publication at another venue (withheld to preserve anonymity). Submitting here as it is very relevant to this workshop.



(a) (b) (c) (d) (e)

Figure 2: Robot-centric feature space aggregations. (a) Uniform-grid feature aggregation. (b) 4-feature aggregation. (c) 8-feature aggregation. (d) 24-feature aggregation. (e) Multi-resolution feature aggregation.

2. Technical Approach

Our algorithm gets trained with a generic score-map (q) generated by the satellite data from areas that exemplify the target environments, for example images of coral reefs (Fig.1b). The system is trained to achieve paths that preferentially cover *hotspots* at the earlier stages of exploration. These learned parameters then define a *policy* π (in the sense of reinforcement learning) that is then used on the satellite image or any other sensor map of the target coral reef (Fig.1b) to generate an explicit action plan. Thus, the policy does not need to be re-trained for each new reef map. This property is a key feature of our approach.

In our approach, we formalize the sampling problem as a Markov Decision Process (MDP). We consider the state $\mathbf{s}$ to include the position of the robot $\mathbf{x}$ as well as the map q containing the per-location score for the visual data, $\mathbf{s} = (\mathbf{x}, \mathrm{q})$. The action space A consists of four actions (move North, East, South, or West) and the transitions are deterministic. Once the visual data at the current cell (i, j) is sampled, the score $\mathrm{q}(i, j)$ is reduced to 0. The discounted reward function is defined as $\gamma^t q(\mathbf{x})$, with a discount factor $0 \leq \gamma \leq 1$ encouraging the robot to sample cells with high scores in early time steps t.

2.1. Policy Gradient Method

In policy gradient methods, the gradient of the expected return ($\nabla_\theta J_\theta$) guides the direction of the parameter update ($\theta_{k+1} = \theta_k + \eta \nabla_\theta J_\theta$, where η is the learning rate). The likelihood ratio policy gradient (Williams, 1992) is given by $\nabla_\theta J_\theta = \int_\tau \nabla_\theta p_\theta(\tau) R(\tau) d\tau$.

We use the GPOMDP and REINFORCE algorithms (Baxter & Bartlett, 2001; Sutton et al., 2000; Deisenroth et al., 2013; Kober et al., 2013) for computing the policy gradient as they have fewer hyper-parameters, making it easier to deploy,

$$\nabla_\theta J_\theta = \frac{1}{m} \sum_{i=1}^{m} \sum_{t=0}^{H-1} \nabla_\theta \log \pi_\theta(a_t^{(i)} | s_t^{(i)}) \left(\sum_{j=t}^{H-1} r(s_j^{(i)}, a_j^{(i)}) - b(s_t^{(i)}) \right). \quad (1)$$

In this equation, the gradient is based on m sampled trajectories from the system, with $\mathbf{s}_j^{(i)}$ the state at the j^{th} time-step of the i^{th} sampled roll-outs. Furthermore, b is a variance-reducing baseline. In our experiments, we set the baseline to the observed average reward.

2.2. Feature Aggregation

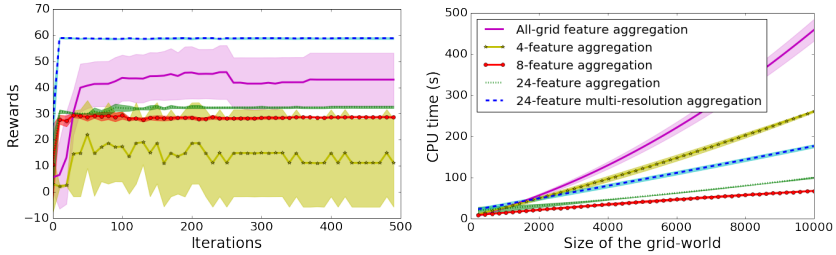
A popular method to define stochastic policies over a set of deterministic actions is the use of the Gibbs distribution as policy (also referred to as Boltzman exploration of softmax policy). We consider a commonly used linear Gibbs softmax policy parameterization (Sutton et al., 2000; Barto et al., 1991) given by,

$$\pi(\mathbf{s}, \mathbf{a}) = \frac{e^{\boldsymbol{\theta}^T \boldsymbol{\phi}_{\mathbf{s},\mathbf{a}}}}{\sum_{\mathbf{b}} e^{\boldsymbol{\theta}^T \boldsymbol{\phi}_{\mathbf{s},\mathbf{b}}}}, \quad \forall \mathbf{s} \in S; \mathbf{a}, \mathbf{b} \in A, \quad (2)$$

where $\boldsymbol{\phi}_{\mathbf{s},\mathbf{a}}$ is an l-dimensional feature vector characterizing state-action pair $(\mathbf{s}, \mathbf{a})$ and $\boldsymbol{\theta}$ is an l-dimensional parameter vector.

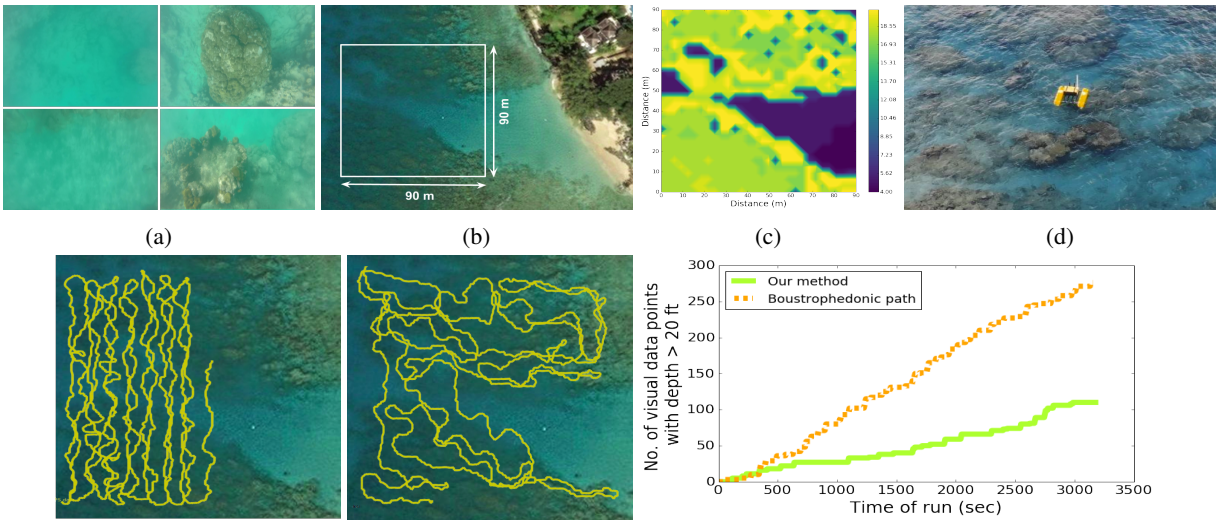
The final feature vector $\phi_{\mathbf{s},\mathbf{a}}$ is formed by concatenating a vector $\phi'_s \delta_{aa'}$ for every action $a' \in \{North, East, South, West\}$, where $\phi'_s \subset \mathbb{R}^k$ is a feature representation of the state space, and $\delta_{aa'}$ is the Kronecker delta. Thus, the final feature vector has $4 \times k$ entries, 75% of which corresponding to non-chosen actions will be 0 at any one time step. We consider five different types of robot-centric feature designs ($\phi'_\mathbf{s}$). The first one is to consider a vector with all the scores in the score-map q as presented in Fig.2a. This feature vector grows in length as the size of the sampling region increases resulting in higher computation times for bigger regions. The four other kinds of feature aggregations are illustrated in Fig.2b - 2e. These aggregations have a fixed number of features, corresponding to the average scores in the feature map in each of the indicated areas, irrespective of the size of the sampling region.

Fig.2e depicts a multi-resolution aggregation where the feature cells grow in size along with the distance from the robot. This results in high resolution features close to the robot and lower resolution features further from the robot's current position. The aggregated feature design is only used to achieve better policy search (Singh et al., 1995), but the robot action is still defined at the grid-cell level.



(a) Discounted total rewards (b) CPU-time vs. world size

Figure 3: Evaluation of feature aggregations. Shaded region indicates the SD over 5 trials on 3 different sized maps.



(a) (b) (c) (d)

(e) (f) (g)

Figure 4: Coral survey: (a) In first column, corals are too deep ($>$ 20ft), hence visual data is bad. Second column presents good quality visual samples of the coral-heads from shallower regions. (b) Survey region. (c) Scoremap generated. (d) ASV deployed in the field. (e) and (f) Paths of the boat for 40 mins. (g) Plot illustrating that the number of non-informative visual samples collected by a boustrophedonic sampler is almost three times the ones collected by our method.

3. Results

Comparing different feature aggregations shows that multi-resolution aggregated features achieve the highest discounted total rewards (Fig. 3a). Also for the uniform grid aggregation, the computation increases quadratically with the size of the area to be mapped (Fig.3b). These results further strengthen our observation that immediate actions are influenced by nearby rewards and the farther low-resolution features enhance non-myopic planning of the complete trajectory.

In the field, we did a visual survey of the reef with our sampling method with multi-resolution feature representation and evaluated it using the bathymetric data as a measure for shallowness of the region covered. Fig.4a presents the images captured at different locations of the reef region with varying depths. These images strengthen our hypothesis of covering shallower reefs to achieve high quality visual survey of corals. We compare the coverage performed by our method with a traditional exhaustive coverage technique using boustrophedonic path. Fig. 4e and Fig.4f illustrate both the sampling paths of the robotic boat for the first 40 minutes. Fig.4g presents the total number of visual data points collected from regions which are deeper than 20 feet (i.e. visual data samples that are not useful to monitor the health of the corals) plotted against the time spent surveying the region. The comparison plot illustrates that the number of non-informative visual samples collected by a boustrophedonic sampler is more than twice the ones collected by our method.

4. Conclusions

The results from exhaustive experiments suggest that multi-resolution non-uniform state aggregation can have a major impact on the efficiency of state exploration and modeling. We further validated this expectation in our field deployments. The proposed incremental sampling-and-modeling paradigm can be applied to many domains where the benefits of efficient sample acquisition should accrue. Such efficient sampling techniques facilitate a faster understanding of the effects of climate change on our environment.

References

Barto, A. G., Bradtke, S. J., and Singh, S. P. *Real-time learning and control using asynchronous dynamic programming*. University of Massachusetts at Amherst, Department of Computer and Information Science, 1991.

Baxter, J. and Bartlett, P. L. Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 15:319–350, 2001.

Choset, H. and Pignon, P. Coverage path planning: The boustrophedon cellular decomposition. In *Field and service robotics*, pp. 203–209. Springer, 1998.

Deisenroth, M. P., Neumann, G., and Peters, J. A survey on policy search for robotics. *Foundations and Trends® in Robotics*, 2(1–2):1–142, 2013.

Hoegh-Guldberg, O. Climate change, coral bleaching and the future of the world's coral reefs. *Marine and freshwater research*, 50(8):839–866, 1999.

Kober, J., Bagnell, J. A., and Peters, J. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.

Low, K. H., Dolan, J. M., and Khosla, P. Adaptive multi-robot wide-area exploration and mapping. In *Proceedings of the 7th International Joint Conference on Autonomous agents and Multiagent systems-Volume 1*, pp. 23–30. International Foundation for Autonomous Agents and Multiagent Systems, 2008.

Manjanna, S. and Dudek, G. Data-driven selective sampling for marine vehicles using multi-scale paths. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 6111–6117. IEEE, 2017.

Rahimi, M., Hansen, M., Kaiser, W. J., Sukhatme, G. S., and Estrin, D. Adaptive sampling for environmental field estimation using robotic sensors. In *Int. Conf. Intelligent Robots and Systems (IROS)*, pp. 3692–3698. IEEE, 2005.

Sadat, S. A., Wawerla, J., and Vaughan, R. Fractal trajectories for online non-uniform aerial coverage. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2971–2976, 2015.

Singh, S. P., Jaakkola, T., and Jordan, M. I. Reinforcement learning with soft state aggregation. In *Advances in neural information processing systems*, pp. 361–368, 1995.

Sutton, R. S., McAllester, D. A., Singh, S. P., and Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, pp. 1057–1063, 2000.

Venkataramani, R. and Bresler, Y. Perfect reconstruction formulas and bounds on aliasing error in sub-nyquist nonuniform sampling of multiband signals. *IEEE Transactions on Information Theory*, 46(6):2173–2183, 2000.

Vevers, R. Xl catlin seaview survey.

Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. In *Reinforcement Learning*, pp. 5–32. Springer, 1992.

Xu, A., Viriyasuthee, C., and Rekleitis, I. Optimal complete terrain coverage using an unmanned aerial vehicle. In *IEEE Int. Conf. Robotics and Automation (ICRA)*, pp. 2513–2519, 2011.

Zilberstein, S. and Russell, S. J. Anytime sensing, planning and action: A practical model for robot control. In *IJCAI*, volume 93, pp. 1402–1407, 1993.