Exploring Multi-Armed Bandit Decision-Making Strategies in an Underwater Vehicle Testbed

Jonathan Valverde Lizano, Prof. Naomi E. Leonard

Department of Mechanical and Aerospace Engineering, Princeton University

ABSTRACT:

The problem of estimating a field in an unknown environment is approached as a Multi-Armed Bandit (MAB) problem, in which an agent must learn about an **unknown environment** while **maximizing expected reward**. The arms correspond to points in space with **spatial correlation**. The smoothness of the field is measured by λ^*, the length scale. **Upper-Confidence Limit (UCL)** [1], an algorithm for MAB problems with correlated rewards, is then applied to the problem. Knowledge of λ^* may give significant improvements in performance. Performance is explored for estimates that are equal, lower, and higher than the correct length scale. The search task is implemented with an underwater robot. In the experiments and simulations, best performance is obtained with **overestimates of the length scale**.

Applications include odor plume detection [2], mapping of forest fires [3], and ocean sampling for oil or dissolved oxygen concentrations.

MODELLING SMOOTHNESS OF THE FIELD

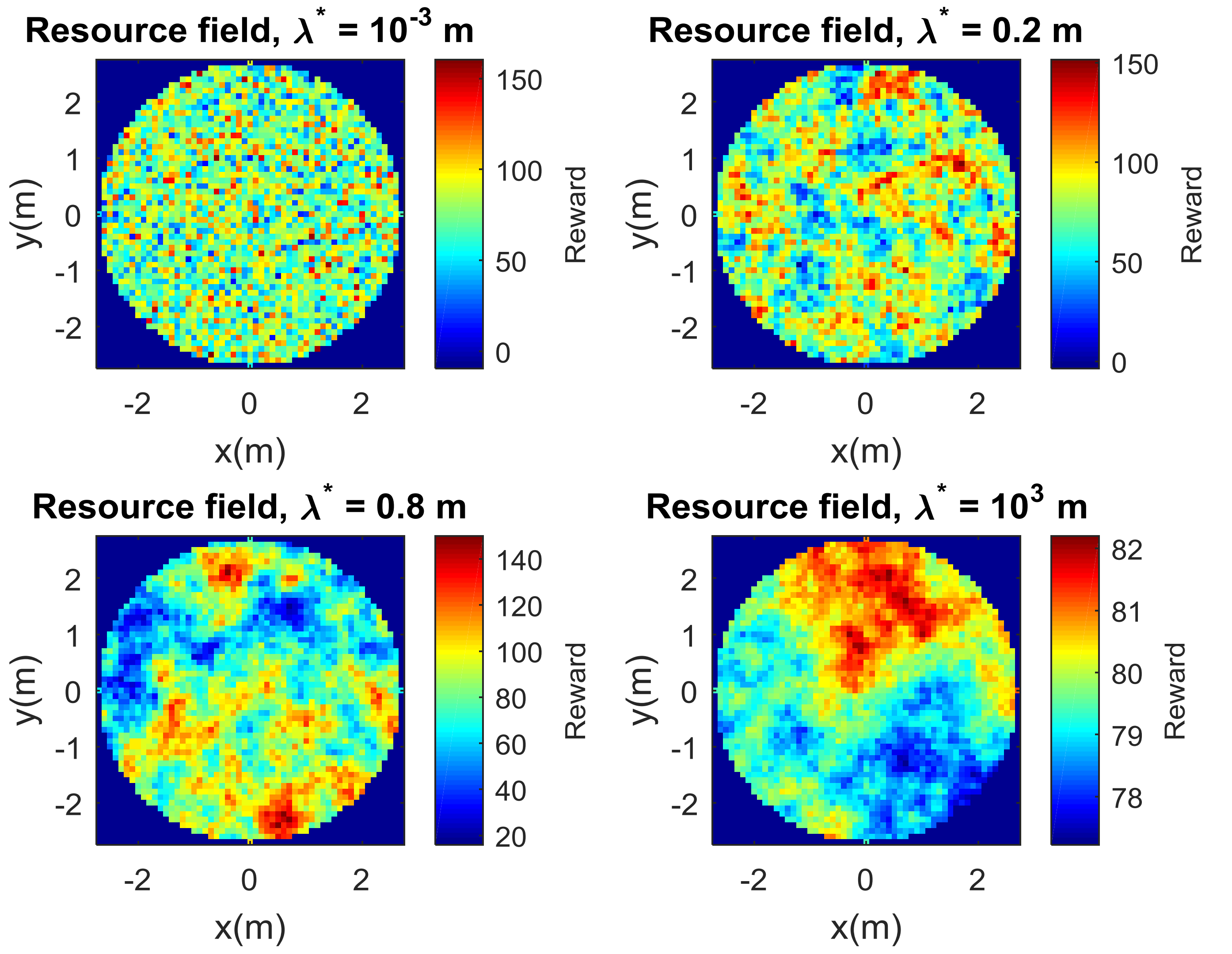
- The correlation between any pair of arms i and j.

is modeled as decaying with distance based on a length scale λ^*:

$$e^{-\frac{d(i,j)}{\lambda^*}}$$

where $d(i,j)$ is the distance between arms i and j.

- This structure can then be used to generate random fields of varying smoothness by changing λ^*.



Randomly generated fields of varying smoothness

SEARCH ALGORITHM

- **Upper Confidence Bound (UCB)** is an algorithm for MAB search with no prior knowledge of the environment
- **UCL** incorporates priors. Input priors with mean μ_0 and correlation Σ_0
- Chooses arm with max:

$$\boldsymbol{Q_i^t = \mu_i^t + \sigma_i^t \Phi^{-1}\left(1 - \frac{1}{Kt}\right)}$$

- Whenever an arm is sampled, the belief state is updated:

$$q = \frac{r_t \phi_t}{\sigma_s^2} + \Lambda_{t-1}\mu_{t-1}$$

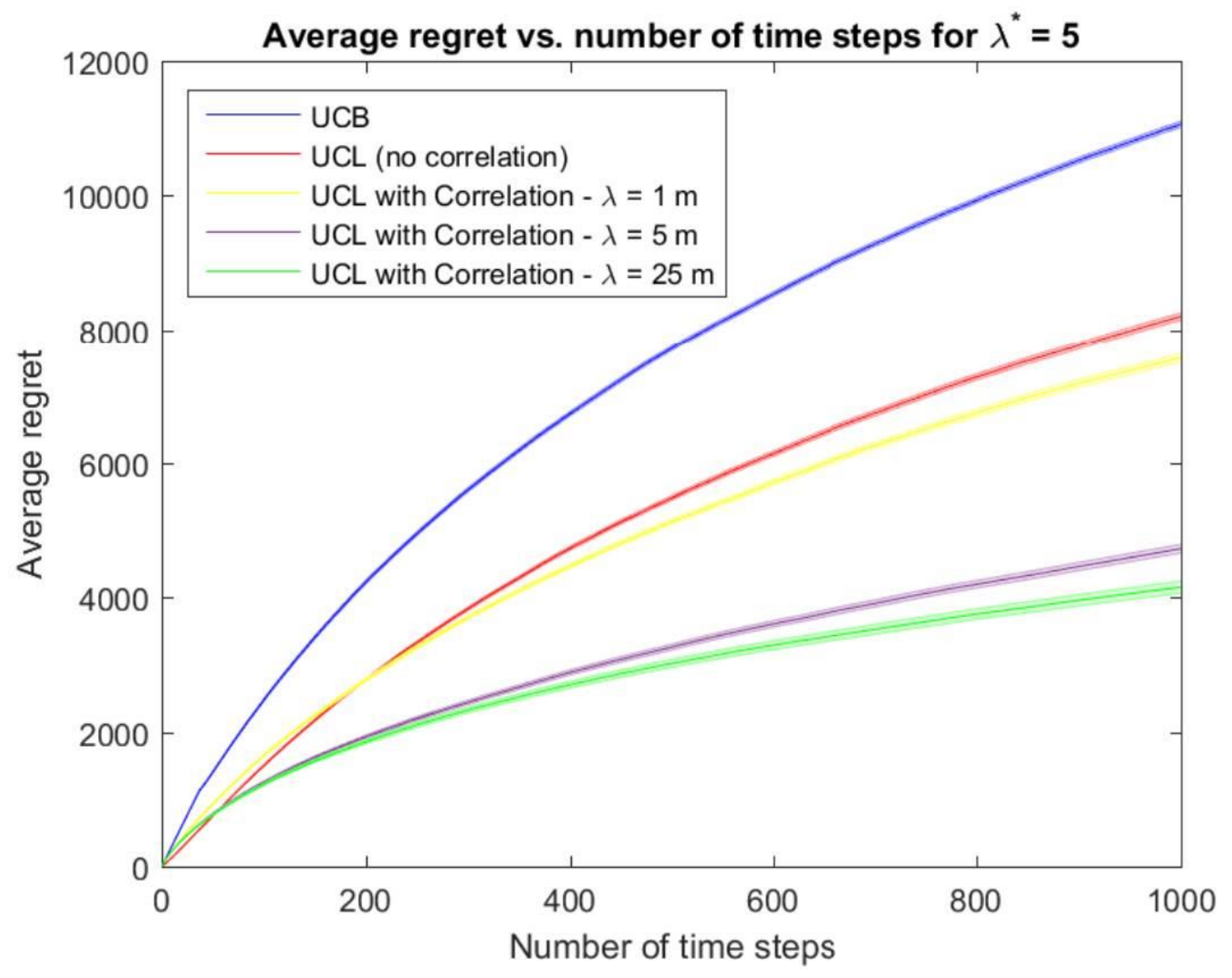
$$\Lambda_t = \frac{\phi \phi^T}{\sigma_s^2} + \Lambda_{t-1}, \Sigma_t = \Lambda_{t-1}$$

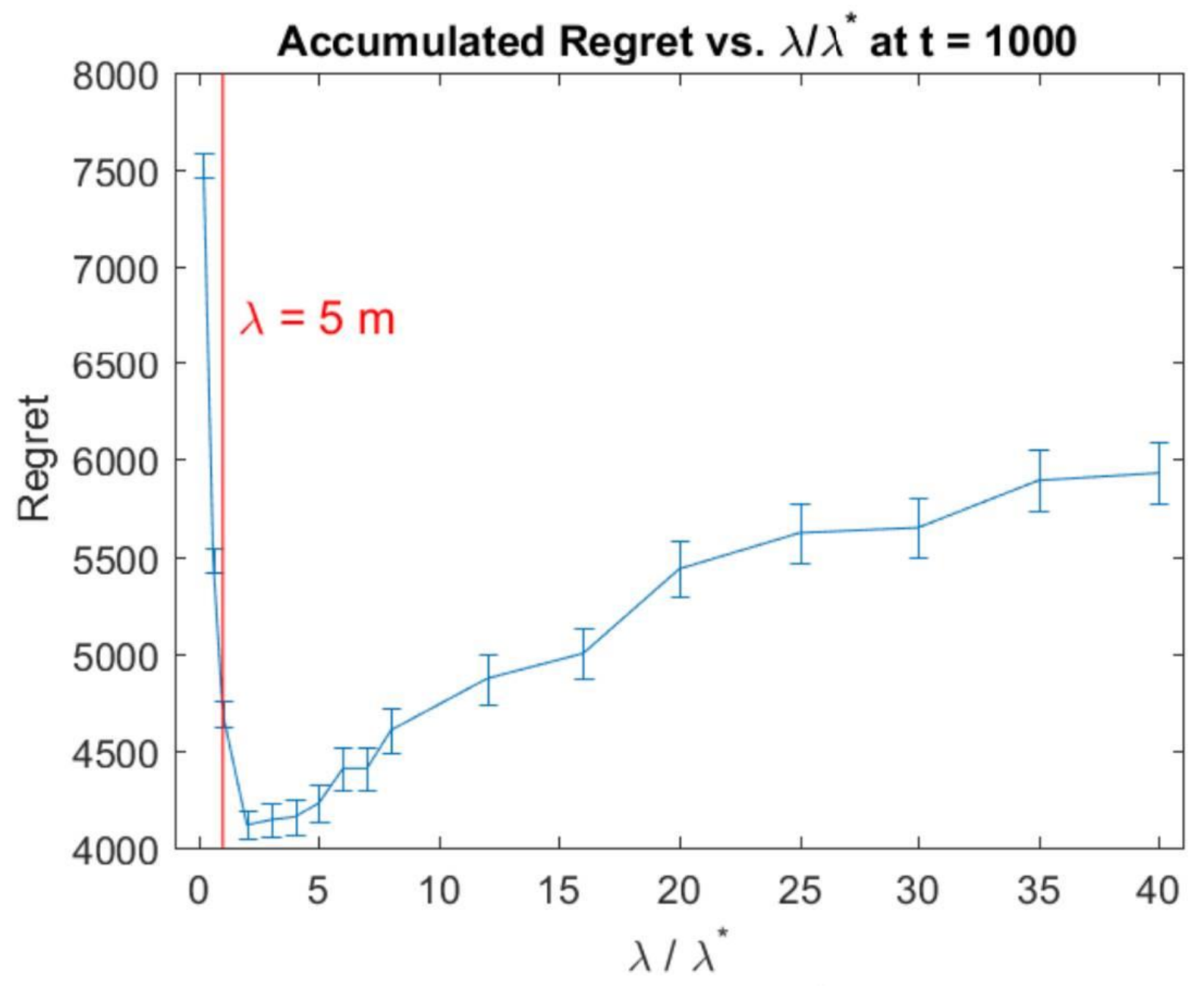
$$\mu_t = \Sigma_t q$$

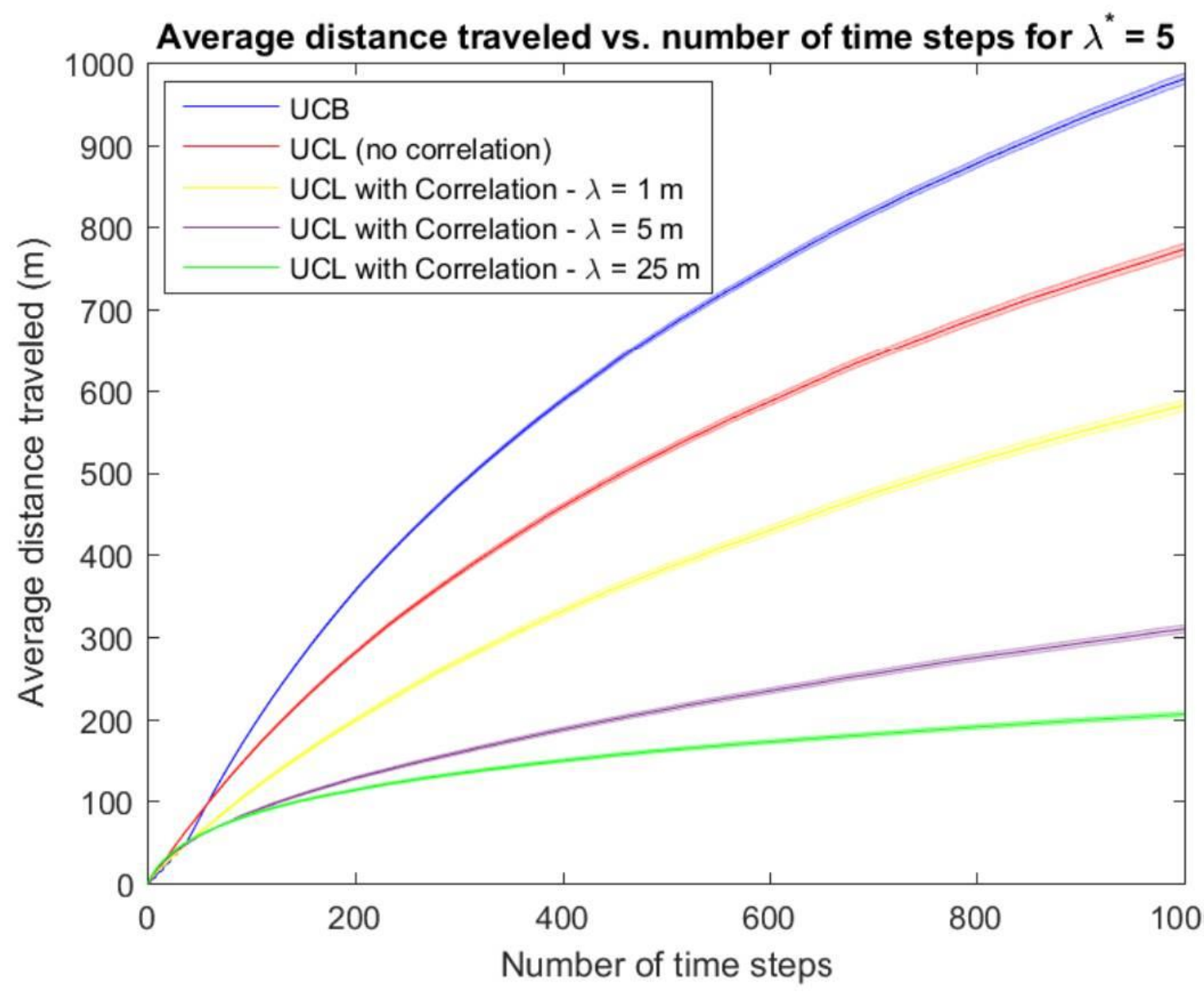
- Goal: minimize regret, defined by:

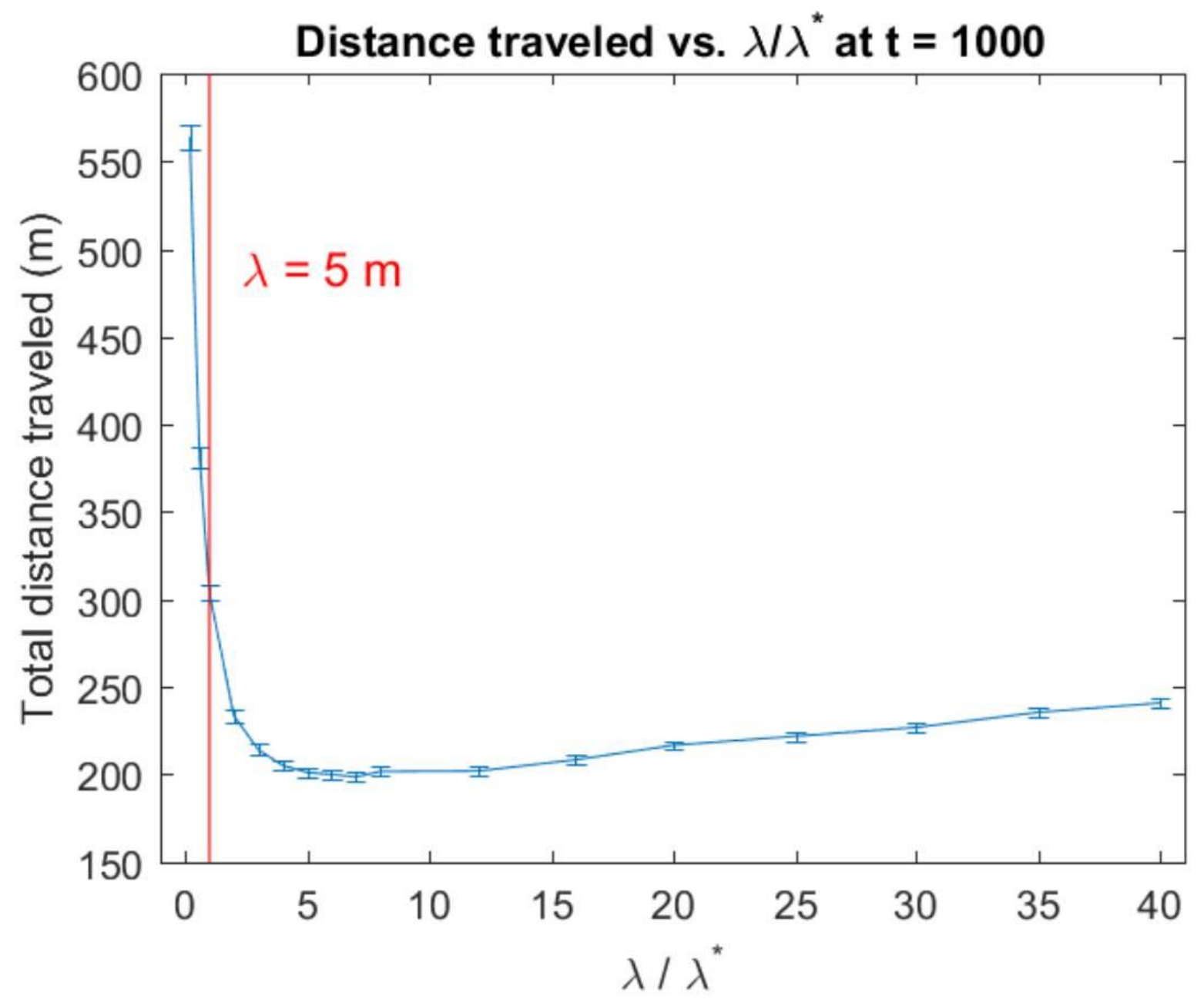
$$R = \Sigma_{t=1}^T \left(m^* - m_{i_t}\right)$$

SIMULATIONS



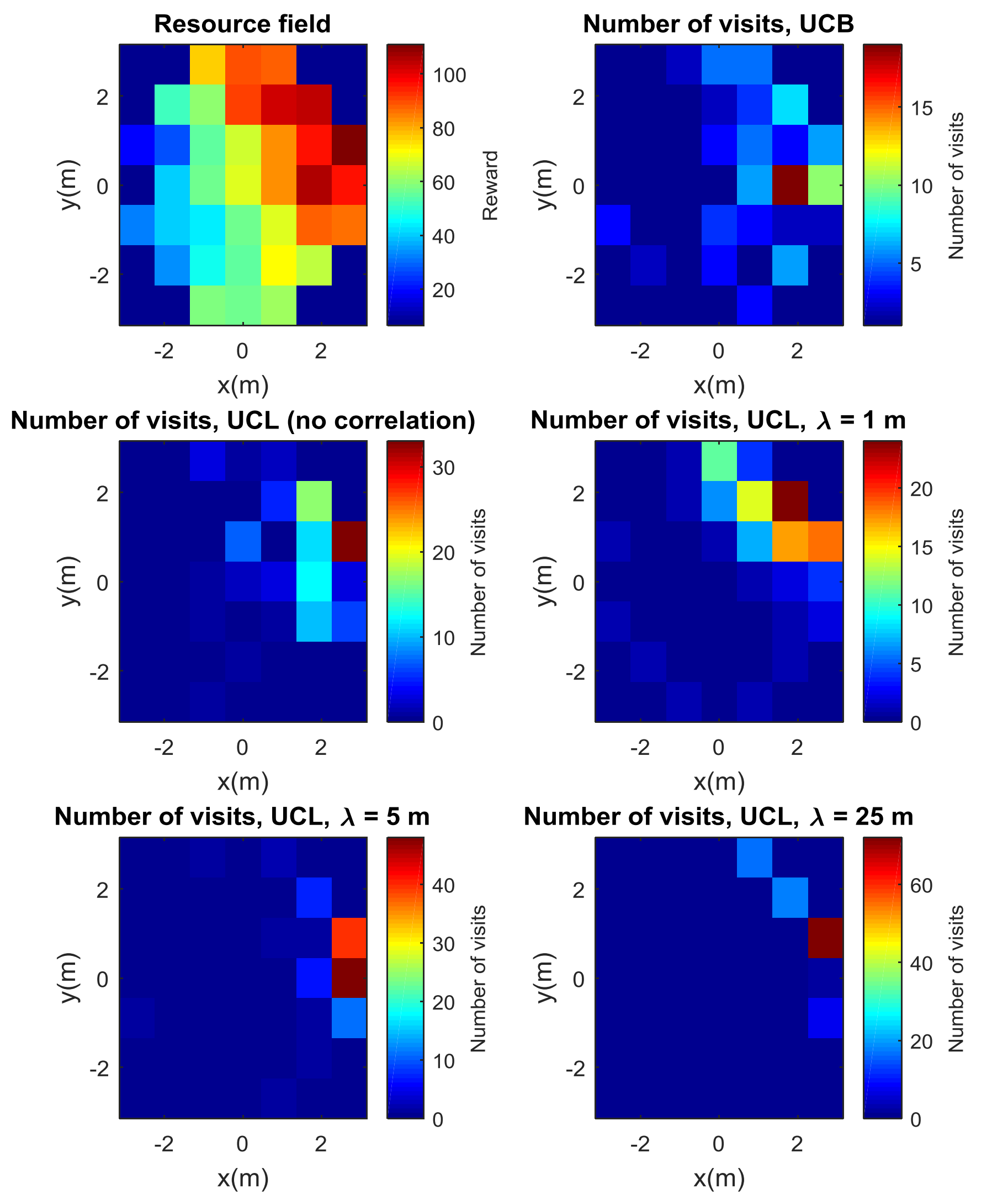






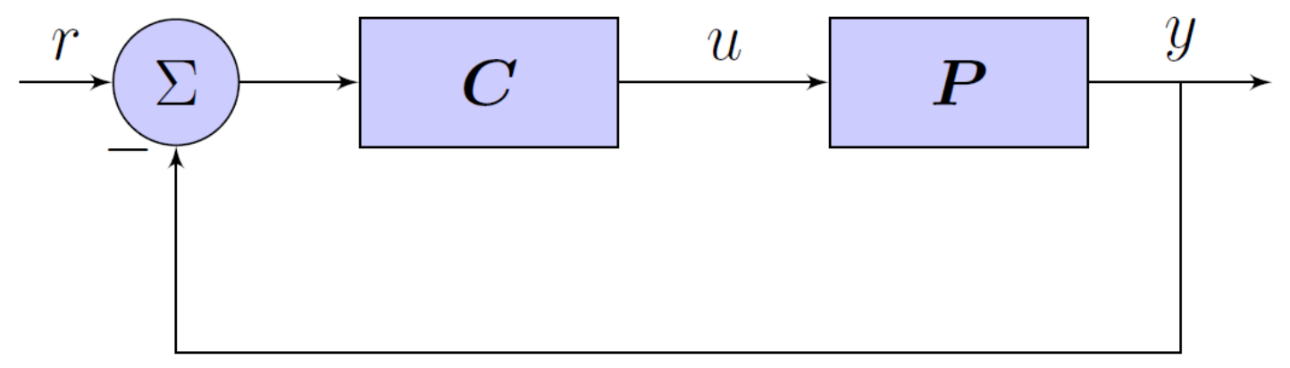
- More reliable performance obtained for UCL with correlation when μ_0 was the mean of the field for all arms instead of an estimate of each reward.
- Best performance both in distance traveled and regret was obtained at an overestimate of the true length scale.
- However, gross overestimation of the length scale can lead to linear regret.

EXPERIMENTAL IMPLEMENTATION



Underwater robot. Credit: Peter Landgren

Testbed with tank and underwater vehicles. Credit: Peter Landgren



Feedback loop used to track a reference height

- Use of PI control to address the variable pull of the robot's tether.
- UCL reduced the amount of exploration needed, with respect to UCB.
- Less exploration with higher length scale estimates because a sample gives more information about the field than with lower scale estimates. Faster convergence.
- Variations in distance traveled also due to the dynamics of the tether.

CONCLUSIONS

- An agent running a field estimation task using this MAB based approach should **err on the side of overestimating the length scale of the field**, within a certain threshold.
- Future work should focus on replicating these experiments with varying smoothness, area and shape of the region, number of dimensions, and number of discretization, for calculation of the optimal length scale estimate.
- Desirable to find a way to calculate the accuracy necessary in the length scale estimate for more general configurations.

REFERENCES: [1] P. B. Reverdy, V. Srivastava, and N. E. Leonard. Modeling human decision making in generalized gaussian multiarmed bandits. Proceedings of the IEEE, 02(4):544-571, 2014. [2] A. Marjovi and L. Marques. Optimal spatial formation of swarm robotic gas sensors in odor plume finding. Autonomous Robots, 35(2):93-109, 2013. [3] M. F. Mysorewala, D. O. Popa, and F. L. Lewis. Multi-scale adaptive sampling with mobile agents for mapping of forest fires. Journal of Intelligent and Robotic Systems, 54(4), 2008.