Niche Selection for Foraging Tasks in Multi-Robot Teams Using Reinforcement Learning

Patrick Ulam, Tucker Balch

1. College of Computing, Georgia Institute of Technology, Atlanta, Ga 30332, USA
Corresponding author: {pulam,tucker}@cc.gatech.edu

Abstract

We present a means in which individual members of a multi-robot team may allocate themselves into specialist and generalist niches in a multi-foraging task where there may exist a cost for generalist strategies. Through the use of reinforcement learning, we show that the members can allocate themselves into effective distributions consistent with those distributions predicted by optimal foraging theory. These distributions are established without prior knowledge of the environment, without direct communication between team members, and with minimal state.

Keywords: Multi-Robot Systems, Reinforcement Learning, Foraging, Optimal Foraging

1 Introduction and Motivation

Foraging tasks are a standard testbed for multi-robot research partly due to their strong biological analogs as well as their applicability in a large number of tasks ranging from sample collection to mine disposal. In multi-robot foraging tasks, the robots composing the team search for objects to collect (attractors) in an area and return the attractors found to a goal location. Multi-foraging is a variant of the typical foraging task in that instead of a single type of attractor, there exist multiple differing types. A significant amount of research has been conducted in the area of multi-robot foraging. This research includes but is not limited to the effects of communication on multi-robot foraging (Balch & Arkin, 1994), interference patterns in multi-robot foraging tasks (Goldberg & Matarić, 1997), as well as the dynamics of collective sorting in a foraging task (Denebourg *et al.*, 1990). Of particular interest is Balch's work concerning the diversity of multi-robot teams that learn to perform a multi-foraging task (Balch, 1998). In this work, Balch found that the team of robots did not learn to specialize in the foraging task even though multiple types of attractors existed. In fact, using the social entropy metric developed within his thesis, he found that team diversity and performance were negatively correlated. This paper focuses upon this result and attempts to answer why the robots that learned the foraging task did not specialize, and what could cause a multi-foraging team to specialize. To address these questions we look towards the optimal foraging theory literature to provide insight into models of natural organisms' foraging behavior and the parameters that may result in specialized foraging behavior as well as generalist foraging behavior.

1.1 Optimal Foraging Theory

Optimal foraging theory, which is used by behavioral ecologists to model the foraging behavior of organisms ranging from birds (Krebs *et al.*, 1997) and mantes (Charnov, 1976) to bees (Real, 1991), has looked extensively at the problem of finding the most efficient means in which an organism may forage. Optimal foraging theory operates under the assumption that evolution has adapted the foraging behavior of organisms to maximize certain factors while minimizing others as a means of increasing its reproductive fitness. The usual interpertation of these optimization factors include the maximization of caloric intake and the minimization of other factors such as predatory risk or energy expenditure.

Research in this area has produced numerous models to describe this behavior. Most of these models utilize some combination of four factors: a fitness set for the foraging activities, activity selection, negative density dependence, and variable environments (Wilson & Yoshimura, 1994). The fitness set captures the intuitive notion that specialized foragers are usually more effective than generalized foragers. Activity selection allows for the organism to change its foraging behavior between differing prey types or differing environments. Negative density dependence factors capture the notion that foraging decisions are made in the context of the number of other organisms already performing a particular foraging action. Lastly, temporally varying environments are used as a means of modeling seasonal variations or other changing factors in the environment that may cause different distributions of foragers.

MacArther and Pianka, in their seminal paper on optimal foraging theory (MacArthur & Pianka, 1966), developed a model to determine the most efficient means in which an organism can forage in a patchy environment. Of particular interest to this work, is the portion of the model that describe the conditions in which competing foraging species, one a specialist and one a generalist, will be overrun by the generalist forager. They describe the net intake of food for a specialist forager as kDH where k represents the foraging rate, D the density of food items, and H is the time spent foraging. For the generalist forager the net intake is equal to $k'DH'$ where $k' < k$ to represent the trade off between generalist and specialist strategies and $H < H'$ to represent the reduced search time incurred by the generalist. This defines the parameter ranges in which specialist foragers can be expected to intermingle with generalist foragers, namely while

$$\frac{H}{H'} > \frac{k'}{k}. \tag{1}$$

Another model of interest is that proposed by Wilson and Yoshimura concerning the coexistence of specialist and generalist foragers (Wilson & Yoshimura, 1994). In their model they define fitness levels for organisms across different environments though the use of a carrying capacity K. This carrying capacity is defined on a per species and per environment basis such that $K_{i,j}$ represents the carrying capacity of species i in habitat j. This carrying capacity is utilized to represent specialist and generalists though the use of constants a and b. Thus the carrying capacities of the different species in a particular environment can be expressed as

$$K_{1,1} = K_1, K_{2,1} = aK_1, K_{3,1} = bK_1. \tag{2}$$

These relationships between the carrying capacities of the different species are used to determine the individual fitness of a species as

$$W_{i,j} = e^{r(1-(N_{1,j}+N_{2,j}+N_{3,j})/K_{i,j})}, \tag{3}$$

where r is a constant rate of increase for all species and $N_{i,j}$ is the number of species i in habitat j. By iterating this fitness value along with the current population of a species in that habitat using a standard discrete-time population model,

$$N_{i,j,t+1} = N_{i,j,t} W_{i,j,t}, \tag{4}$$

they are able to predict the number of the three species that will be present in each habitat upon stabilization of the system.

1.2 Reinforcement Learning

This insight of treating foraging as an optimization process leads to the utilization in our investigation of a common optimization technique in robotics, namely reinforcement learning. Reinforcement learning is a machine learning technique in which an agent learns through trial and error to maximize rewards received for taking particular actions in particular states over an extended period of time. More precisely, given a set of environmental states $\mathcal{S}$, and a set of agent actions $\mathcal{A}$, the agent learns a policy, π, which maps the current state of the world $s \in \mathcal{S}$, to an action $a \in \mathcal{A}$, such that the sum of the reinforcement signals r are maximized over a period of time.

There are a number of techniques for maximizing this reinforcement signal including but not limited to such techniques as Q-learning and the adaptive heuristic critic algorithm (Kaelbling *et al.*, 1996). For our experiments, however, we chose a relatively simple method to calculate the value of taking a given action in a given state, namely by calculating the average reward over state, action pairs. This average can be calculated using

$$\mathcal{Q}(s,a) = \frac{\mathcal{N}(s,a)\mathcal{Q}(s,a) + r + \max_{a'} \mathcal{Q}^*(s,a')}{\mathcal{N}(s,a) + 2}, \tag{5}$$

where r is the reward received for taking the action, $\max_{a'} \mathcal{Q}^*(s,a')$ is the reward that would be received by taking the optimal action after that, and $\mathcal{N}(s,a)$ is the number of times the robot has taken action a in state s. By choosing the action with the highest Q-value, and allowing for the robot to choose a random action with a given probability, the robot can explore the state space and converge upon the action with the greatest average reward. For a more detailed discussion of reinforcement learning refer to (Sutton & Barto, 1998) and (Kaelbling *et al.*, 1996).

2 Related Work

A large body of research has looked at using reinforcement learning as a means of guiding multi-robot teams in foraging tasks. Matarić has analyzed the performance of such foraging robots using reinforcement learning (Matarić, 1997). Balch has measured the behavioral diversity of teams that have learned foraging tasks (Balch, 1998). A significant amount of research has also addressed the division of labor in foraging tasks, both in the context of multi-robot teams, as well as social insects. Jones and Matarić have looked at means of using limited sensory history to estimate the proper division of labor in a foraging task (Jones & Matarić, 2003). Martinson and Arkin investigated the utility of reinforcement learning as a means of guiding role switching in a military scenario involving foraging robots, soldier robots, and mechanics (Martinson & Arkin, 2003). Additional division of labour models have been proposed in the context of social insects. Bonabeau *et al.* developed a division of labor model for social insects utilizing response thresholds (Bonabeau *et al.*, 1996). Theraulaz *et al.* extended this model to allow for threshold adjustment via the use of a reinforcement process (Theraulaz *et al.*, 1998).

3 Method

Using the Teambots robot simulation environment, five different worlds were created with 5 percent random obstacle coverage and 40 randomly distributed attractors colored blue and red. Eleven variations of these worlds were generated by varying the proportion of red attractors present such that the number of red attractors ranged from 0 to 20. Four simulated Nomad 150 robots were placed into the world. Each robot can execute one of three foraging strategies: a specialist red attractor foraging behavior, a specialist blue attractor foraging behavior, and a generalist foraging behavior in which the robot would collect either type of attractors. Each robot's controller is designed using the Clay architecture (Balch, 1998) of Teambots which allows for the creation of motor schema based control systems. We use the reinforcement learning algorithm described previously to enable each robot to learn which of the three different foraging strategies to use. Each robot has only one state and in that state can choose from three actions corresponding to the three foraging behaviors described above.

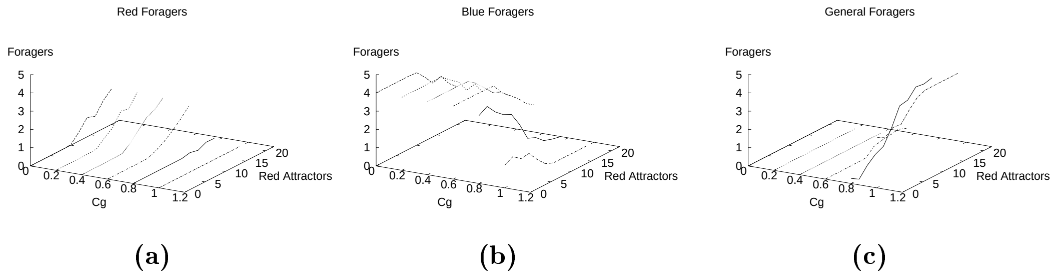
In order to capture the notion of the fitness set in optimal foraging theory we utilize a scalar c_g for the reward function of the general forager. This scalar can vary from 1, indicating there is no cost to being a general forager to, 0 which indicates that general foraging is ineffectual. This reward model is is further expanded to include the notion of search cost as MacArther and Pianka's model indicate that this may play an important role in specialization of natural organisms foraging behaviors. Thus for each timestep spent searching for attractors the robot receives a penalty of -1, indicating a significant search cost, or 0 indicating there is no significant cost to searching. Thus the reward function can be depicted as:

$$R(t) = \begin{cases} 1 & \text{if the attractor is returned using a specialist behavior at time t - 1;} \\ 1c_g & \text{if the attractor is returned using the generalist behavior at time t - 1;} \\ -1, 0 & \text{if the robot does not return an attractor at time t-1.} \end{cases}$$

The cost scalar was varied from 0 to 1 in increments of 0.2 and search cost was varied between 0 and -1. Three hundred trials were run in each configuration.

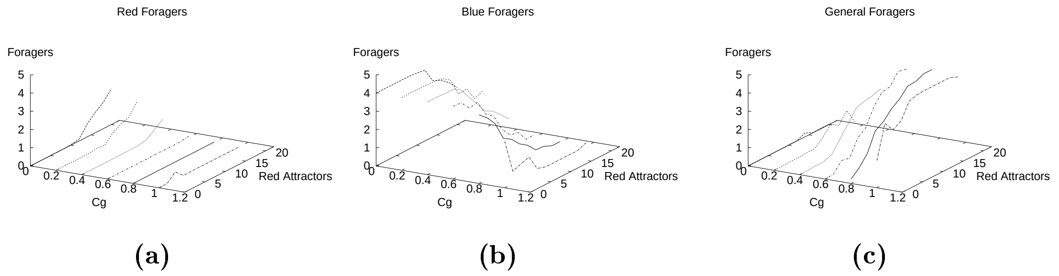
4 Results

The resulting behavior selection of the individual robots was measured. Figure 1 depicts the results of the experiments when the foraging robot does not take the cost of the actual foraging into account. Figures 1a, 1b, and 1c show the number of robots that perform the red foraging, blue foraging, and general foraging strategies respectively for a given configuration of general foraging reinforcement levels and attractor distribution. Figure 2 shows the same trials run with the addition of a penalty for each time step spent searching for attractors. Figures 2a, 2b, and 2c depict the resulting behavioral distribution for each of the three strategies with this additional negative reinforcement in place.



(a) (b) (c)

Figure 1: Distributions of four robots with no search cost as defined by the cost of generalization, c_g, and the number of red attractors out of 40



(a) (b) (c)

Figure 2: Distributions of four robots with significant search cost as defined by the cost of generalization, c_g, and the number of red attractors out of 40

5 Discussion

A baseline case occurs when there is no tradeoff between general and specialized foraging strategies. In the trials run, this parameterization occurs when the reward level for attractors returned using any strategy is 1. In the trials where there is a cost associated with searching as well as when there is no cost, a homogeneous team of general purpose foragers results as predicted by Balch's work on the diversity of multi-foraging teams. For the lower red attractor distribution, the general purpose foragers intermingle with blue foraging specialists as there are too few red attractors to mandate a fully homogeneous generalist team.

Further analysis is possible by looking at some of the model's predictions concerning the proper distribution of specialists and generalists. In particular, we can look MacArthur and Pianka's predictions concerning the the critical points in which generalist foragers will overrun specialist foragers. The parameters in our simulation can be mapped to the parameters in MacArthur and Pianka's model readily. The reinforcement levels for the generalist and specialist strategies times the reward r can be mapped to k' and k respectively. The parameters for the time spent foraging, H and H', can be mapped to the simulation by noting that the foraging time for the two strategies will be be proportional to the number of attractors available to be collected via a given strategy. Hence,

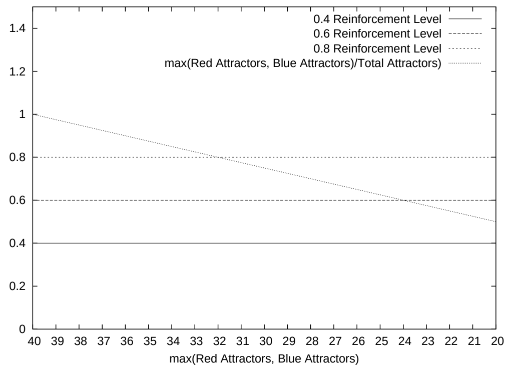
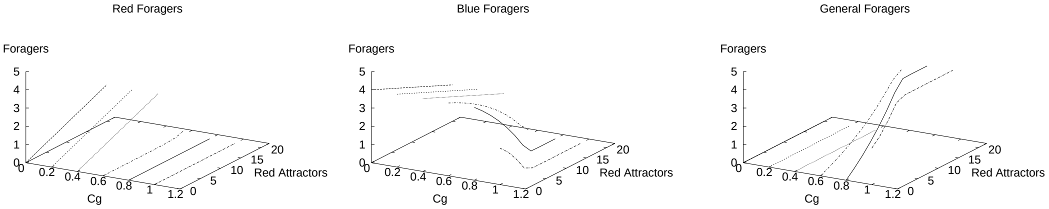


Figure 3: Critical points for the transition to homogeneous generalist teams via MacArther and Pianka's model.



(a) **(b)** **(c)**

Figure 4: Allocations for four robots using Wilson and Yoshimura's model

the time spent foraging for the specialist will be proportional to $\frac{max(A_{red},A_{blue})}{A_{red}+A_{blue}}$ and the time spent foraging for the generalist will be proportional to $\frac{A_{red}+A_{blue}}{A_{red}+A_{blue}}$ where A_{red} and A_{blue} represent the number of red and blue attractors respectively.

By placing these values in equation 1, the critical points in which the specialists are overrun with generalists can be calculated as:

$$\frac{h\frac{max(A_{red},A_{blue})}{A_{red}+A_{blue}}}{h\frac{A_{red}+A_{blue}}{A_{red}+A_{blue}}} = \frac{rc_g}{r}$$

$$\frac{max(A_{red},A_{blue})}{A_{red}+A_{blue}} = c_g, \tag{6}$$

where r is is the reward for returning an attractor under the specialist strategies and h represents some constant handling time for attractor collection.

Figure 3 shows a plot of $\frac{max(A_{red}, A_{blue})}{A_{red}+A_{blue}}$ and three different reward multiples for the generalist foraging strategies where the reward level was multiplied by 0.4, 0.6, and 0.8. The intersections between the reinforcement factor functions and the maximum proportion of the attractors available to the specialist depict the critical points at which the team should converge upon a homogeneous generalist strategy. The data presented in figure 2c shows that the team did in fact converge on the homogeneous strategy, but slightly later then predicted by the optimal foraging model. The trials in which $c_g = 0.6$ converged to a homogeneous team when at $A_{red} = 18$ as opposed to the predicted $A_{red} = 16$. At $c_g = 0.8$ the convergence did not occur until $A_{red} = 12$. In both the our simulation runs and in the MacArther and Pianka's model, a homogeneous team of foragers does not emerge for $c_g \leq 0.4$.

We can do a similar comparison to Wilson and Yoshimura's model described previously. We assign the carrying capacities of each species as

$$\begin{aligned} K_{red,red} = N_{red}, K_{blue,red} = aN_{red}, K_{gen,red} = bN_{red} \\ K_{blue,blue} = N_{blue}, K_{red,blue} = aN_{blue}, K_{gen,blue} = bN_{blue}, \end{aligned} \tag{7}$$

with $a = 0$, and $b = c_g$, utilize equations 3 and 4 to determine the stable configuration, and then normalize the results for four agents. The resulting configuration space is shown in figure 4. As can be readily seen, the results are strikingly similar to results of the simulations foraging runs made in which search cost was not considered as a significant portion of the reward function. The results from our reinforcement learning allocation produced slightly sharper curves at the data points with low values of c_g and low proportion of red attractors when compared to Wilson and Yoshimura's predictions. Also, the bifurcation that occurs at when $c_g = 0.6$ with the appearance of generalist foragers and the disappearance of red foragers is not as pronounced in our simulation results. The generalist foragers do begin to emerge and specialist red foragers begin to dissapear but not as drastically as their optimal foraging model would predict.

6 Conclusions and Future Work

Reinforcement learning appears to be an effective means for individual robots to learn foraging strategies in environments where multiple types of attractors exist and the effectiveness between generalized foraging strategies and specialized strategies may be variable. Using the method described in this paper, the robots were able to achieve effective distributions in unknown environmens without the use of direct communication and with the use of minimal state. By modeling the trade-off between the effectiveness of general and specialized foraging strategies via a reward function, the individual robots were able to learn strategies resulting in team composition that is consistent with the foraging distributions predicted by Wilson and Yoshimura's model. When search cost became the defining factor in the foraging behavior the distributions closely converged to homogeneous generalist teams at the points predicted by MacArther and Pianka. Balch's work concerning the diversity in multi-robot teams that learn the foraging task have been shown to be consitant with both the optimal foraging models as well as the simulation trials described in this paper in which there existed no cost to performing generalist foraging. While we have looked at the extreme parameterizations of search cost in our simulations, it may prove fruitful to investigate the effect of more moderate search cost on niche selection for a foraging task. Additional investigation into the scalability of the method described in this paper over additional foraging behaviors as well as attractor types may also prove interesting.

References

Balch, T. 1998. *Behavioral diversity in learning robot teams.* Ph.D. thesis, College of Computing, Georgia Institute of Technology.

Balch, T. & Arkin, R. 1994. Communication in reactive multiagent robotic systems. *Autonomous Robots* 1(1).

Bonabeau, E. Thérauluz, G. & Deneubourg, J.L. 1996. *Quantitative study of the fixed threshold model for the regulation of division of labour in insect societies.* in: *Proceedings Roy. Soc. London B.* **263**: 1565–1569.

Charnov, E. 1976. Optimal foraging: Attack strategy of a mantid. *The American Naturalist.* **110**:141-151

Denebourg, J.L. Goss, S. Franks, N. SendovaFranks, A. Detrain, C. & Chretien, L. 1990. *The dynamics of collective sorting robot-like ants and ant-like robots.* in: Meyers, J.A. & Wilson, S.W. eds. (*SAB96, From Animals to Animats 4 : Proceedings of the 4th International Conference on Simulation of Adaptive Behavior*, Cambridge, MA, MIT Press. pp. 356–365.

Goldberg, D. & Matarić, M. 1997. *Interference as a tool for designing and evaluating multi-robot controllers.* in: *AAAI/IAAI*, Providence, RI, pp. 637–642.

Jones, C. & Matarić, M. 2003. *Adaptive division of labor in large-scale minimalist multi-robot systems.* in: *IEEE/RSJ International Conference on Robotics and Intelligent Systems (IROS)*, Las Vegas, Nevada.

Kaelbling, L.P., Littman, M.L., Moore, A.P. 1996. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research* **4**:247–285.

Krebs, J. Erichsen, J. Webber, J. & Charnov, E. 1997. Optimal Prey Selection in the Great Tit (*Parus Major*). *Animal Behaviour* **4**:30–38

MacArthur, R. & Pianka, E. 1966. On Optimal Use of a Patchy Environment. *The American Naturalist* **100**: 603–609.

Martinson, E. & Arkin, R. 2003. *Learning to role-switch in multi-robot systems.* in: *IEEE International Conference on Robotics and Automation (ICRA).* Taipei, Taiwan.

Matarić, M. 1997. Reinforcement learning in the multi-robot domain. *Autonomous Robots* **4**: 73–83.

Real, L. 1991. Animal choice behavior and the evolution of cognitive architecture. *Science* **243**: 980–986.

Sutton, R.S. & Barto, A.G. 1998. *Reinforcement learning: An introduction.* MIT Press, Cambridge, Ma.

Thérauluz, G., Bonabeau, E. & Deneubourg, J.L. 1998. *Threshold reinforcement and the regulation of division of labour in insect societies.* in: *Proceedings Roy. Soc. London B.* **265**: 327-335.

Wilson, D.S. and Yoshimura, J. 1994. On the coexistance of specialists and generalists. *The American Naturalist* **144**: 607–707.