Learning from Experience:
Response of Q-Learning to the Saliency of Environmental Cues

Catie Meador
Stella Cho

May 11, 2012

Abstract

In the field of adaptive robotics, numerous studies have been concerned primarily with effectively simulating biological phenomena in order to take advantage of the many effective approaches to adaptation and development present in the world of biology. In this paper, we discuss an attempt to mimic the psychological phenomena of classical conditioning, used frequently in training animals, in robots. To do this, we trained a robot through Q-learning, a reinforcement learning algorithm that creates a Q-table with a row for every state and a column for every action, filling the table with the maximum reward possible when executing action a from state s. Our robot was trained for two tasks— a Light Training task in which the robot was expected to seek out a light in its environment in response to a reward it received for reaching the light, and a Sound Training task in which the robot used what it learned in Light Training in combination with sound cues that are predictive of the location of the light to reach the light, thus learning to respond to the sound cues in a similar way as the light cues. Overall, our experiments were unsuccessful, mainly due to the numerous limitations of the Q-learning algorithm, as the robot only managed to find the light with an average of 8.5% success during Light Training, and an average of approximately 12% success during Sound Training with an Expected Values light table, a table artificially filled with the values we expected to see.

1 Introduction

Much of the field of Adaptive Robotics is concerned with observing and simulating biological phenomena in order to make robots capable of developing on their own and adapting to a dynamic environment [1]. As Baloch et. al. writes:

> Though a great deal of work has been done in mobile robotics, most has been algorithmic in its orientation, requiring much hand programmed knowledge, and is usually unacceptably slow in performing. Our approach begins, instead, from a very different point of view: *The robot must be adaptive to its environment and learn from experience*. [2]

Modern-day robots are expected to operate in constantly changing environments, such as space exploration or military tasks, which require the robot to learn and adapt [2]. It would take too many resources and too much time to hard-code such a responsive, adaptive robot. Thus, adaptive systems are modeled after the mechanisms in biological organisms that exhibit such behavior.

In order to adapt to the environment, a robot must to be able to respond to environmental cues. Various experiments have successfully implemented systems that demonstrate robot response to environmental cues. For example, Sporns et. al. implement a modifiable value system modeled after the nervous system in order to teach a robot to avoid bad-"tasting" objects, an inherent value, based on visual cues that predict the taste of the object. The paper showed that,

“... initially neutral sensory stimuli that reliably precede other, innately salient sensory stimuli can ... become themselves capable of eliciting value system responses.” [3] This study demonstrated how robots could be induced to exhibit the responses expected from Pavlovian conditioning, a psychological model of biological organisms that illustrates the organism’s response to salient environmental cues.

In Pavlovian conditioning, also known as classical conditioning, there is an unconditioned, or natural, stimulus (US), which elicits an unconditioned, or natural, response (UR). A second, neutral stimulus (called the conditioned stimulus, CS) elicits no response at all [4]. In the Sporns et. al. experiment, the US, UR and CS, are represented by the taste of the object, the avoidant behavior, and the visual cues, respectively. When the two stimuli are coupled together, with the visual cue (CS) indicating the bad taste (US), the robot learns to use the visual cues as a predictor of the bad taste and avoids the bad-tasting object before it even tastes it.

Baloch et. al. demonstrate similar behavior in MAVIN, the Mobile Adaptive VIsual Navigator, via neural networks. MAVIN, which associates certain images with certain “feelings”, can be conditioned to feel fear or happiness in response to a previously neutral image (i.e. an image that elicited no emotional response) when the neutral image is coupled with an emotional one [2].

One of the challenges of conditioning robots, however, is the issue of delayed reward. As Balkenius and Morén state, “the goal of classical conditioning is to establish a temporal gradient that represents the distance to the goal event.” [4] In order for the robot to be able to appreciate the predictive value of an environmental cue, it must be able to associate something that happens in the past with the final reward.

In this experiment, we attempt to reproduce such conditioned behavior in robots trained via reinforcement learning, specifically Q-learning, because it allows for the learning of events with delayed reward, or events that are temporally separated from the goal state. We study whether Q-learning allows the robot to be responsive to the environment; in other words, we are interested in the plasticity of the robot, or the robot’s ability to adapt to the saliency of environmental cues, that is afforded by Q-learning. This is accomplished in two parts: The first part, Light Training, uses reinforcement learning to condition the robot to seek a light. The robot does not know what its task is; it only knows that a series of actions either causes it to receive a reward -- when it reaches the light -- or to receive no reward. The experiment observes whether the robot learns with experience that heading towards the light, a salient environmental cue, gives it the highest reward.

In the second part, Sound Training, we compare the responses of a robot trained via reinforcement learning to the responses expected from Pavlovian conditioning. The aforementioned light serves as the US, and a sound cue, which predicts the location of the light, serves as the CS. If reinforcement learning allows the robot to respond to the saliency of outside cues, the robot will learn to react to and use the sound cue to faster reach the light, for which it is rewarded.

We will provide an overview of Q-learning, followed by a description of the experiments. Then we will present and explain the results and discuss future directions.

1.1 Q-learning

Q-learning is a form of reinforcement learning, a field of machine learning that operates under the principle that given a state, the robot should choose the action that leads to the highest-reward next state. Q-learning is an example of temporal difference learning, meaning that the reward is delayed, so after a trial is completed the resulting reward of that trial is backpropagated to all of the actions performed in that trial, rather than the robot receiving an explicit reward after each action.

Q-learning works by creating a Q-table with a row for every state and a column for every action. The value of Q(s,a) represents the maximum discounted cumulative reward possible that can be attained by executing action a as the first action from state s. This means the value of Q is the reward for executing action a from state s plus the value of performing all subsequent optimal actions, demonstrated by the following update rule:

$$Q(s, a) = Q(s, a) + \alpha(r + \gamma * \max Q(s', a') - Q(s,a)),$$

where r is the reward, α is the learning rate, and γ is the discount rate [6].

2 The Experimental Framework

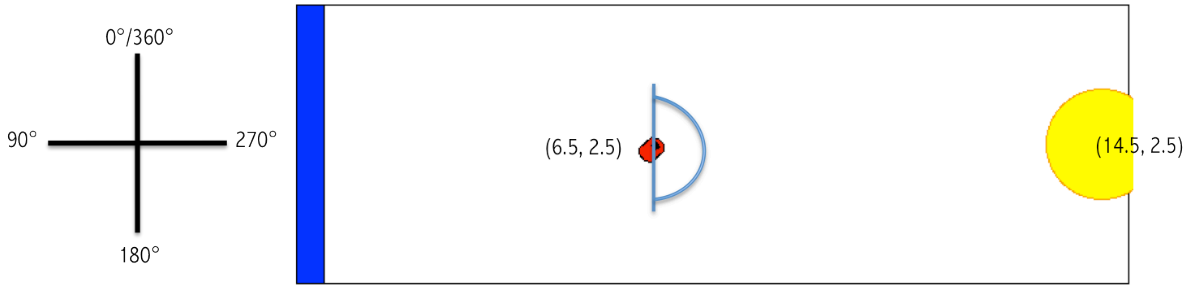

Figure 1: The training world, with a size of 10 by 5 units. The sound source is shown as a blue rectangle on the left wall, while the goal is represented as a circular light along the right wall. Shown to the left is the directional key used for all geometric calculations. For Sound Training, the robot was always placed at coordinates (6.5, 2.5) with a random heading in the range indicated by the blue semicircle.

2.1 Environment

To test our robot, we created a long, rectangular, 10 by 5 unit world in Pyrobot, a simulator developed by Doug Blank and Lisa Meeden that models robot interactions in a two dimensional world (see Figure 1) [5]. Along the left wall was a sound source, and near the center of the right wall was the light source, a circular light at position (14.5, 2.5) with a 1-unit radius. Depending on the trial, the robot was placed in different locations in the world, with a random heading either between 0 and 360 degrees for Light Training, or 180 and 360 degrees for Sound Training.

2.2 Agent

We used Pyrobot's PyrobotRobot60000 robot, with two light sensors in the front that allowed the robot to determine whether it had reached the light, and two artificial sensors used by the robot to determine geometrically where the sound and light sources were relative to the robot's own position. These sensors informed the robot whether the light and sound sources were in front of, behind, to the left, or to the right of the robot.

2.3 Task

The experiment was divided into two main parts: whether the robot could learn the optimal behavior through environmental cues and rewards (Light Training), and whether the robot could learn to associate a salient environmental cue (CS) with another, inherently rewarded stimulus (US; Sound Training).

2.3.1 Light Training

In the first round of training, which we called Light Training, we intended for the robot to learn to seek the light through Q-learning. The robot created the Q-table based on the rewards it received when it accidentally hit upon the light. Over time, the robot would learn that certain actions (i.e. heading towards the light) would be rewarded and thus learn to seek the light. The robot was only able to sense the light source, and not the sound source for this trial.

2.3.2 Sound Training

In the second round, called Sound Training, the robot started training with a modified Q-table that we filled with expected values; it was already trained to respond to light. In other words, the light acted as the unconditioned, natural, stimulus (US). However, the light source did not appear until the robot had already performed a certain amount of steps (5 or 10, depending on the trial), meaning the robot had to rely on the sound source at the beginning of each trial. In this round of training, we expected the robot to adjust the Q-table to account for the new information it was receiving from the sound sensor so that it would still be able to find the goal quickly despite the initial absence of the light source.

2.4 Implementation

To implement Q-learning in our robot, we gave the robot a state that included the direction of the sound, the direction of light, and whether or not the robot was stalled. We simplified and discretized the states, using the integers 0-3 to represent whether the sound or light was in front, to the left, behind, or to the right of the robot, respectively, and 'T' or 'F' to represent whether the robot was stalled or not, respectively. Using this system, a typical state would look something like the string '23F', using a 2 to represent that the sound is behind the robot, a 3 to represent that the light is to the right of the robot, and an F to represent that the robot is not stalled. These states were placed in the Q-table with three possible actions to choose from: forward, forward-left, or forward-right. During execution, the robot chose from these states with decreasing randomness. If it was early on in the experiment, the robot would be more likely to choose a random action, in order to explore unknown states, whereas later on in the experiment, with more experience, the robot would be more likely to choose the best action of the three in order to more effectively find the light.

Furthermore, we noticed that the robot sometimes developed a circling behavior in order to avoid stalling, which avoided the negative reward of hitting a wall, but also prevented the robot from reaching the goal. To prevent this, we gave the robot a step limit of 175 steps, after which the trial would end and the robot would receive the same negative reward it received for stalling if it had failed to reach the light.

2.4.1 Implementation of Light Training

In addition to the simplification of the states and possible actions, we added many other adjustments to our Q-learning system in order to facilitate learning. After numerous unsuccessful trials, we began scaffolding the Q-learning in Light Training by starting the robot closer to the light so it would be more likely to reach the light, receive a reward, and learn the appropriate action. After every 1500 trials, we increased the distance the robot could start from the light, in the hopes that the robot would generalize the results from when it was close to the light to when it was further away. This distance increased 10 times by 1 unit increments.

During Light Training, sound direction values were not used. In order to make the Q-table created in Light Training usable with the addition of the sound sensor in Sound Training, we updated the Q-table in such a way that, given a specific light sensor and stall sensor value, the value in the Q-table would be the same regardless of the sound sensor value. This was implemented by giving the robot the sound direction value of 0 for all states. Given the state '0#F', with # representing the direction of the light, the Q-table's '0#F', '1#F', '2#F', and '3#F' rows were simultaneously updated with the same reward value.

2.4.2 Implementation of Sound Training

In Sound Training, the robot started at the same location in the middle of the environment for each trial, with heading randomized between 180 and 360 degrees, as shown in Figure 1.

We conducted two experiments for Sound Training. Information on the location of the sound source was made available from the beginning of the run. The light, however, was not turned on until 5 or 10 steps after the trial had begun, depending on the experiment. A new light value, 4, was introduced to indicate when no light source was available.

In addition, when we attempted Sound Training, we decreased the reward the robot received for reaching the light if it took more than 40 steps to do so, in order to encourage the robot to use the sound cues to reach the light faster, rather than just waiting for the light cue to appear to find the goal.

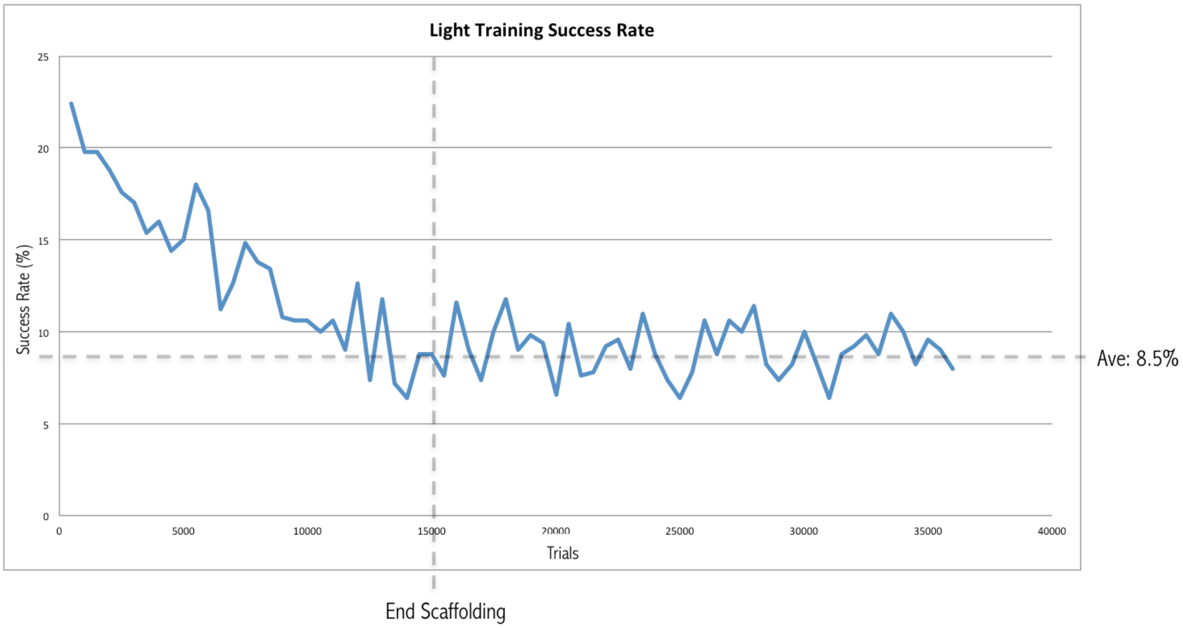

Figure 2: This figure shows the frequency of success, calculated every 500 trials, for Light Training. Success was determined by whether the robot reached the light. The initial high values and the downward slope can be attributed to scaffolding. Frequency of success was expected to increase with time as the Q-table learned that seeking the light correlated with a positive reward. As the results indicate, however, frequency of success did not increase with time.

3 Results and Discussion

3.1 Light Training

By 36000 trials, Light Training displayed an average frequency of success of 8.5%, as indicated by Figure 2. As expected, scaffolding allowed for an initially high rate of success and dropped as the robot's starting point was moved further and further away from the light source. However, the frequency of success did not rebound with time.

The average number of steps taken for successes also stabilized at around 34 steps. Neither average frequency or average number of steps taken appear to show any correlation with time. These results indicate that the robot was not developing a better method of reaching the light. Rather, they indicate that the robot developed a table that allowed it to reach the light 8.5% of the time. Had the Q-table enabled the robot to learn to reach the light, we would have expected a success frequency closer to the 64-73% range, in accordance with the results we obtained when we ran the simulator with a Q-table we filled with expected values (e.g. a high reward for turning left if the light is to the left and the sound is to the right). The table we created was used as the basis for Sound Training in order to maintain the unconditioned stimulus (US) premise.

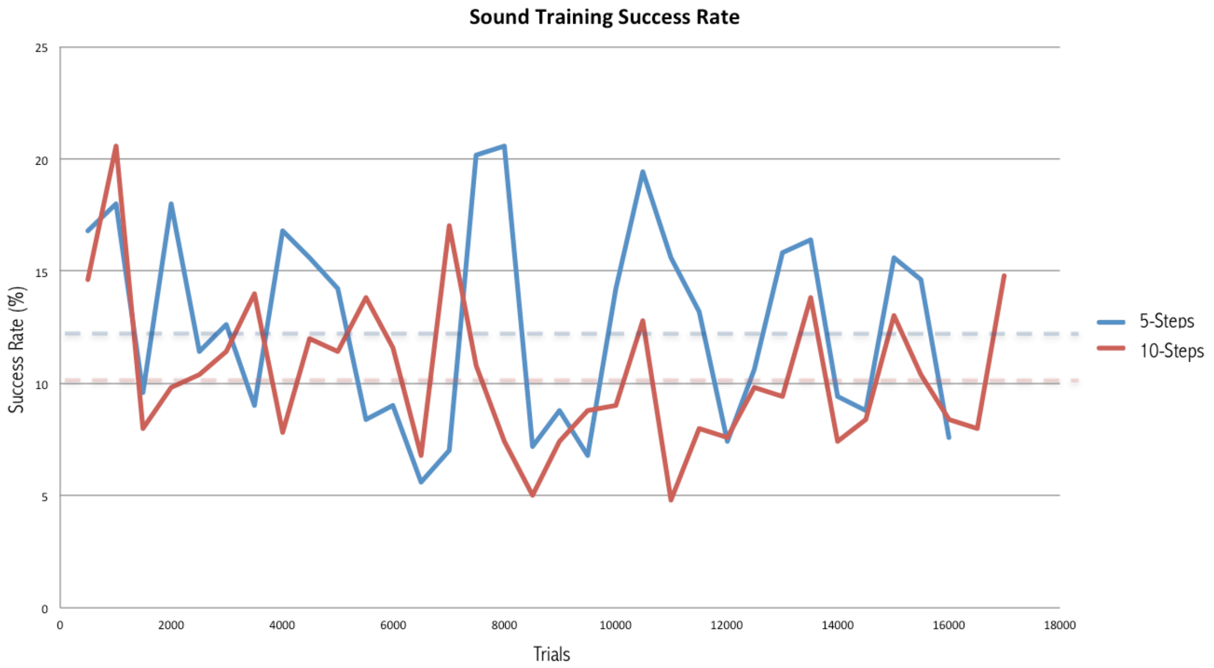

Figure 3. This figure shows the frequency of success, calculated every 500 trials, for the two Sound Training runs. The blue line indicates the run where the light went on after 5 time steps, and the red line indicates the run where the light went on after 10 time steps. Frequency of success was expected to start lower than the Light Training's success rate, but increase with time as the Q-table learned that moving away from the sound source allowed it to reach the light faster, and thus, receive higher rewards. As the results indicate, however, frequency of success did not increase with time.

3.2 Sound Training

We conducted 16000 trials in which the light appeared after 5 steps, and 17000 trials in which the light appeared after 10 steps. Overall, the results of the two experiments was very similar, with the 5-step trial averaging a success rate of approximately 12.5%, and the 10-step trial averaging a success rate of approximately 10%, as indicated by Figure 3. Based on these results, it appears that the robot did not learn to use the sound cues effectively, as it was trained on the light table that produced a 64-73% success rate, and yet with the absence of the light cue in the beginning of the trial, the success rate dropped by about 50%.

As indicated by Figure 4, the average number of steps taken in each experiment is also similar, with both averaging close to 46-50 steps per successful trial. Clearly this is much higher than the 36 steps average reached in Light Training using the Expected Values Table we created, which is the same Q-table that both Sound Training experiments were initially given. This again shows that the robot failed to adequately learn to use the sound cues, as it appears that it made little or no progress toward the goal in the 5 or 10 steps for which the light was absent. However, as shown in the graph in Figure 4, the 10-step experiment appears to be consistently faster by a few steps when it does reach the light, so it is possible that the robot learned to use the sound cues a little in the 10-step experiment to allow it to do better than the 5-step experiment. This can be accounted for by the fact that the 10-step experiment has longer periods of no-light states, giving Q-learning more experience with the no-light states and therefore more opportunity to learn.

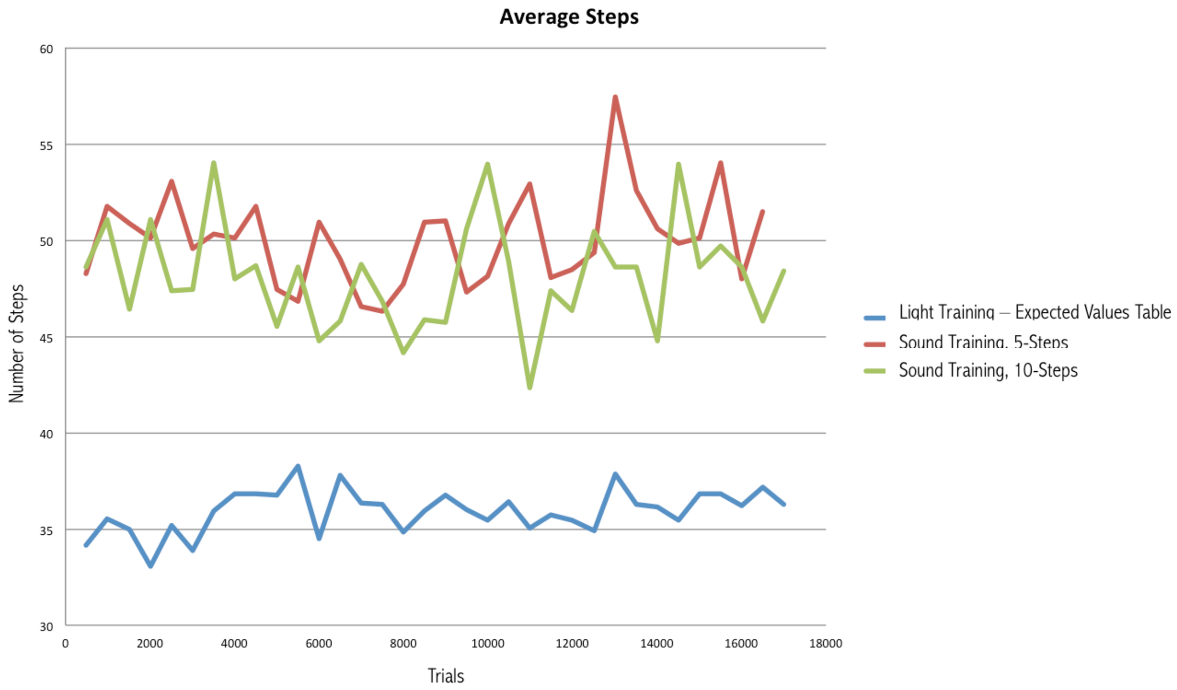

Figure 4. This figure shows the average number of steps, calculated every 500 trials, for successes only. The blue line, placed for comparison, indicates the average number of steps taken when using a table filled with expected values. The average number of steps were expected to trend towards the blue line as Q-learning learned to minimize the number of steps it takes to reach the light by using the sound cue. As the results indicate, however, average number of steps did not decrease with time.

3.3 Limitations of Q-Learning

As demonstrated by our results, we were largely unsuccessful in implementing classical conditioning in our robot. There are many possible reasons for this, many having to do with the limitations of Q-learning. In general, Q-learning is an effective learning algorithm in situations with a small amount of clearly defined, discrete states. However, in robotics, states are often continuous, noisy, and numerous, and so adapting a robotics problem to a Q-learning algorithm is generally not an easy, or effective, task.

Our main problems with Q-learning arose from the complicated nature of the task we were trying to implement. Initially, we attempted to give the robot states that included discretized values for each of the two light sensors, but this proved to be far too many states, and so we were forced to instead give the robot a state that simply indicated the direction of the light. Even with the 32 states we used in our final experiments, it still seemed the robot had a difficult time learning a reliable strategy for each state.

In addition, using a temporal difference algorithm such as Q-learning for this task is difficult, because for each trial, the reward for all steps in the trial is based on whether or not the robot succeeded. Even if the robot made all the right moves up until the end of the trial, if it fails to reach the light in the end all of those moves are negatively reinforced, and so it is less likely to perform the correct moves in future trials. In addition, we noticed that in many trials, the robot would approach the light and then turn away from it right before the robot reached the light, when our intention was for the robot to learn to continue straight to the light. However, because it was so close to to the light, the action of turning still caused the robot to enter the goal area, and so the incorrect action of turning away from the light received a reward for reaching the goal state anyways. This is not a problem in situations where the robot is close to the light, but when the robot is further away and has learned to turn away from the direction of the light, its chances of success are small.

3.4 Future Work

Overall, it is clear that Q-learning is not the best algorithm for this task; the large amount of continuous states are not ideal for this algorithm, and the delayed reward serves to inhibit learning in many situations. For future experiments, we would like to try different methods of learning to see if this task is possible with more appropriate methods. One option would be to continue using Q-learning, but with the inclusion of a neural net to determine the states, so the continuous nature of the states can be dealt with more adequately. In addition, we could look into using the Modifiable Value System described in Sporns et al. [3], or using other types of algorithms, such as categorization, evolution, or a reinforcement algorithm that provides immediate reward (perhaps based on the distance from the goal) after each action.

Furthermore, upon success, we would like to perform more experiments to test how useful these results are. As an extension of Sound Training, we would like to perform an experiment in which the sound cue is no longer salient, to see if the robot is capable of learning to ignore this cue and only use the light cue to find the goal. In addition, we would be interested in experimenting with our sound- and light- trained robot in other worlds to see if what it learned in the environment we created can be used in other situations. This would especially be interesting in a maze-type environment, to see if we could train the robot to get around obstacles with strategically placed sound and light sources.

Acknowledgements

This experiment would not have been possible without the support of Lisa Meeden, our professor; tarragon, our friendly CS department computer; and the CS department. We'd also like to acknowledge Kristen Allen and Peter Ballen for lending us the code for determining the direction of sound and light geometrically.

References

[1] M. Lungarella, G. Metta, R. Pfeifer, and G. Sandini. Developmental robotics: a survey. *Connection Science*, 2003.

[2] A. A. Baloch and A. M. Waxman. Visual Learning, Adaptive Expectations, and Behavioral Conditioning of the Mobile Robot MAVIN. *Neural Networks*, 1991

[3] O. Sporns, N. Almássy, and G. M. Edelman. Plasticity in Value Systems and Its Role in Adaptive Behavior. *Adaptive Behavior*, 2000

[4] C. Balkenius and J. Morén. Dynamics of a Classical Conditioning Model. *Autonomous Robots*, 1999

[5] L. Meeden and D. Blank. www.pyrobotics.org

[6] T. Mitchell. Machine Learning. McGraw Hill, 1997.