Learning To Learn

Nicholas J. Butko
Cognitive Science Department
University of California, San Diego
9500 Gilman Dr. #0515; *La Jolla, CA 92093 USA*
nbutko@cogsci.ucsd.edu

Javier R. Movellan
Institute for Neural Computation
University of California, San Diego
9500 Gilman Dr. #0523; *La Jolla, CA 92093 USA*
movellan@mplab.ucsd.edu

***Abstract*— Empirical evidence shows that infants 10 months of age can learn about 10 times faster than infants 2 months of age that a novel entity is socially contingent. This suggests that during the period from 2 to 10 months of age infants became better learners. One possible explanation for this change is that new brain structures grow, in a genetically predetermined manner, that support more efficient learning. An analogy for this point of view would be the increase in mastication efficiency due to the growth of teeth. An alternative hypothesis is that the increase in learning efficiency is itself the result of a learning process that operates on the time scale of months. Under this view, better learning is the consequence of learning itself. Here we explore the plausibility of the "learning to learn" hypothesis from a computational point of view. We show that with standard reinforcement learning algorithms using an internally generated reinforcement signal it is possible to develop agents that progressively learn to learn within a period of months. The results fit well at a qualitative level empirical evidence regarding the development of social contingency detection in infants. The learning techniques that we explored have potential application for robots that learn to learn on their own.**

***Index Terms*— Infomax Control, Infomax Reinforcement Learning (IRL), Social Contingency, Temporal Dynamics of Social Interaction, Probabilistic Functionalism, Developmental Robotics, Social Robotics, Probabilistic Robotics.**

I. The Role of Contingency in Social Development

In this paper we formalize and explore from a computational point of view the problem of learning to learn. For concreteness we focus on the development of contingency detection, a popular and fruitful experimental paradigm in the infancy learning literature. In contingency detection experiments a new contingency is created between a behavior of the infant and a sensory event (e.g, when the infant moves a leg, a sound is produced). Increments in the frequency of occurrence of the target behavior with respect to a control group are interpreted as evidence that the child has learned this contingency.

John Watson proposed that contingency detection plays a crucial role in the social and emotional development of infants. This view originated from an experiment in which 2-month-old infants learned to move their heads to activate a mobile above their cribs [1]. Infants in the experimental group were presented with a mobile that responded to movements of the infant's head. For the infants in the control group, the mobile activated at the same rate as in the experimental group but in a random, non-contingent manner. After four 10 minute daily sessions of exposure to this mobile, and an average of approximately 200 responses, there was evidence that the infants in the experimental group had detected the existence of a contingency: At about that time the experimental group exhibited significantly higher response rates than infants in the control group and displayed social responses that are typically directed towards caregivers (*e.g.*, cooing and social smiles).

Watson and Movellan [2], [3] conducted a similar experiment with 10-month-old infants. Infants were seated in front of a robot that did not look particularly human and were randomly assigned to an experimental group or a control group. In the experimental group the robot produced sounds contingent to the infant vocalizations. Each infant in the control group was matched to an infant in the experimental group and was presented the same temporal distribution of robot behaviors as was experienced by his/her matched participant. However, in the control group the robot was not responsive to the infant's behavior or to any other events in the room. Evidence was found that after 3.5 minutes of exposure to the robot, children in the experimental group learned that the robot was a contingent social agent: For example, they exhibited 5 times more vocalizations than infants in the control group. Moreover they followed the "line of regard" of the robot when it rotated, showing evidence for shared attention [2], [3].

Thus during the period from 2 to 10 months of age, infants became more efficient learners: While 2 month-old infants took 40 minutes to learn a new behavioral contingency, 10 month old infants only took 3.5 minutes.[1] During these 8 months infants became experts at learning new contingencies quickly and accurately. In fact [4] showed that some 10 month old infants learned contingencies in a manner that was very close to optimal given the statistics of social interaction. One possible explanation for this increase in learning efficiency is that during the period from 2 to 10 months new brain structures grow that support faster learning.

[1]These are the times at which significant differences were found between experimental and control group averages. Some individuals showed signs of having detected these contingencies much faster.

An analogy for this point of view would be the increment in mastication efficiency due to the growth of teeth. An alternative explanation, which we explore in this document, is that better learning efficiency is itself a manifestation of the learning process, only at a longer time scale. This hypothesis does not deny the importance of brain development but it sees it as supporting the computational process of learning to learn.

Information theory provides a useful way to approach and formalize the problem of learning to learn. The key concept here is that behaviors have informational consequences, i.e., information value. In a way, behaviors can be seen as questions to the world. Good questions provide valuable information; bad questions do not. Thus, learning to learn can be seen as the process of learning to behave in a manner that causes the *most useful* information to become available as quickly as possible. With the most useful information, the infant can learn about the world more quickly. From this point of view "learning to learn" means learning to ask good questions, i.e., learning to produce behaviors that support the gathering of information that is useful for discovering facts about the world.

We explore the computational plausibility of the "learning to learn" hypothesis using the framework of Infomax Reinforcement Learning (IRL). The crux of IRL is that information gain can be used as an intrinsic reinforcement signal to progressively learn to become a better learner.

We show that using information gain as the reinforcement signal applied to an off-the-shelf reinforcement learning algorithm allows the algorithm to learn to learn in an optimal manner after the equivalent of 10 months of operation in the world. The point of this exercise is not to model precisely the developmental process in humans but to explore whether "learning to learn" is a computationally plausible explanation for some of the changes in learning efficiency observed in human infants.

II. Infomax Control

Movellan [4] formalized the problem of detecting social contingencies as an Infomax Control problem. The algorithm operates with a single binary sensor that encodes whether the sound level crosses a given threshold, and a binary actuator that either produces a vocalization or stays quiet. Under the model the sensor activations are caused by a background Poisson process. Humans, if present, respond to the controller's vocalization using a reaction time distribution collected from typical social interactions. Under these assumptions the problem of learning whether or not an object is socially responsive is formally equivalent to the problem of learning whether during the period following a vocalization there is a significant change in sensor activation rate with respect to the background level of activity. Under these conditions, it is possible to compute an optimal Infomax controller, *i.e.*, a controller that connects sensors and actuators in a manner that maximizes the long-term gathering of information about whether or not a social contingency is present (See Appendix A). It was found that this controller behaves in a manner remarkably similar to the way some 10 month old infants operate when discovering social contingencies [4].

A. Computational Complexity

Infomax Control is a specific instance of a general class of control problems known as Partially Observable Markov Decision Processes (POMDPs). In Infomax Control, information gain acts as an intrinsic reward mechanism, *i.e.*, the utility function optimized by the controller is the long term gathering of information about states of the world that are not directly observable. Typically POMDP controllers have a temporal horizon over which they are expected to maximize some reward function; in Infomax Control the controllers are evaluated in terms of how many bits of information they gathered after a finite period of time T. This time is called called the controller's horizon.

One reason why Infomax Control has not been pursued aggressively in the past is that the process of developing an optimal Infomax Controller can be computationally very expensive. Infomax controllers need to map the entire history of actions and observations into new actions on a moment-to-moment basis. This means that for the general case the number of possible states to keep track of grows exponentially with time. For example, if the system only has a binary sensor and a binary actuator, after T time steps there are a total of 4^T possible histories, each of which needs to be mapped to a current action. Fortunately in many problems of interest, sufficient statistics exist that condense all the past history of observation into a few numbers. For example, in Movellan's model, statistics representing the number of past vocalizations and the proportion of sensor activations after vocalizations and during silent periods are sufficient. Unfortunately the number of possible values of these statistics still grows rapidly with the temporal horizon, as T^4 [4] . Table I shows the growth in the minimum number of possible states needed to completely specify the Value as a function of the number of time steps in the controller's horizon. Note that a two-fold increase the number of time steps produced more than a twenty-fold increase in the number of possible state-action mappings that need to be represented.

Given the complexity of this problem is not clear *a priori* whether it is computationally plausible to learn realistic Infomax Controllers. Indeed some in the computational literature have proposed that Infomax Control with long time horizons is too difficult and advocate greedy approaches with one-step time horizons [5]. However, one-step solutions to information gathering often fail. For example, when a baby makes a sound it partially blocks the reception of auditory signals from the external world, temporarily reducing the gathering

of useful information. Thus a greedy infomax controller would prescribe to never vocalize, since it results in an immediate reduction of useful information. However, in the long run vocalizations are important to gather information as to whether a responsive human is present. Thus learning to vocalize as a way to gather information requires controllers with non-zero time horizons.

TABLE I
COMPLEXITY OF DYNAMIC PROGRAMMING SOLUTION

	Total Timesteps			
	12	25	50	100*
DP Runtime (Mins.)	.04	.40	10.6	275
Number States	1.5e4	4e5	1e7	2.3e8

* Estimated

B. Finite Horizon, Time Steps, and Social Interaction

It is not immediately clear how the notion of time steps applies to real life social interaction problems. The important notion is that social interaction happens on a certain time scale, and the abstract notion of "time step" should fit with that natural scale. For example, if the time step represented one hour, an infant would vocalize for an hour, and wait another few hours to see if somebody was responding. Such a strategy would be ineffectual because most people would respond while the baby was crying, but after an hour, they would probably give up and ignore it, and the baby would not notice any responses. We can develop a more reliable system by having very short time steps (*e.g.*, 1 msec steps). This would allow the system to be very responsive. However this would come at the cost of requiring a very large horizon (measured in number of time steps) leading to the combinatorial explosion problem described in the previous section.

In practice we have found that when the Infomax contingency detector is run on social robots, it operates well with time steps in the range from 1/4 of a second to 1 second. A time step of 1 second is also optimal for the Infomax control model to reproduce some of the observed behaviors in 10 month infants. As such hereafter we will proceed under the assumption that a time step of our discrete time computer simulations roughly corresponds to a second of an infant's life.

III. INFOMAX REINFORCEMENT LEARNING

Reinforcement learning is an area of machine learning and control whose goal is to develop approximately optimal controllers based on examples of state-action-reinforcement triplets. The reinforcement signals can be extrinsic, *e.g.*, water, food, or internally generated, *e.g.*, information gain. Infomax Reinforcement Learning (IRL) refers to Reinforcement Learning problems that use information gain as the basic reinforcement signal. The goal in these problems is to develop action strategies that maximize the long term gathering of information about targeted states of the world.

In this paper we will implement IRL using Temporal Difference Learning (TD), a popular reinforcement algorithm that has been shown to describe well the behavior of dopaminergic neurons in the basal ganglia [6]. The goal in TD learning is to learn state value functions that can then be used to choose the most valuable actions given each state. Such value function need to satisfy the Bellman Equation

$$V_t(x_t) = E[V_{t+1} \mid x_t] + \mathcal{R}_t \tag{1}$$

where, x_t denotes the state at time t of the sufficient statistics, E denotes the expected-value operator, and $\mathcal{R}_t$ is the reward signal (in our case information gain – see Appendix A for more information). TD learning is an iterative approach that starts with some initial estimates of the value function for all states and time steps, and progressively refines these estimates based on experience.

The goal of the computational experiments presented in this document is to explore whether it is feasible to learn an optimal controller using simple TD learning with information gain as the basic reinforcement signal. In particular our goal is to explore whether an optimal social contingency detector could be developed over a period of 10 months assuming no more than 200 vocalizations per day (a total of no more than 60,000 vocalizations).

A. Exact IRL Results

First we found that the required number of vocalizations needed for IRL to converge grew as a fifth power of the horizon (Table II). Convergence within 60,000 vocalizations was only achievable with horizons no larger than 12 time steps into the future.

TABLE II
TOTAL VOCALIZATIONS REQUIRED FOR EXACT TD(0) LEARNING

	Total Timesteps			
	8	12	16	20
# Vocalizations	7.5e3	5.7e4	2.3e5	7.0e5

B. IRL Approximation Results

Once we established that with current IRL techniques, it would be difficult to learn a controller with a time horizon longer than 12 time steps (i.e., approximately 12 seconds) we investigated the question of how 12 time step controllers compare to optimal controllers with longer time horizons. Given the statistics of social interaction, does it pay off to use time horizons longer than 12 seconds?

Fifty new simulations were performed, each with different starting points and with a time horizon of 12 time steps. On average, IRL converged after less than 60,000 vocalizations.

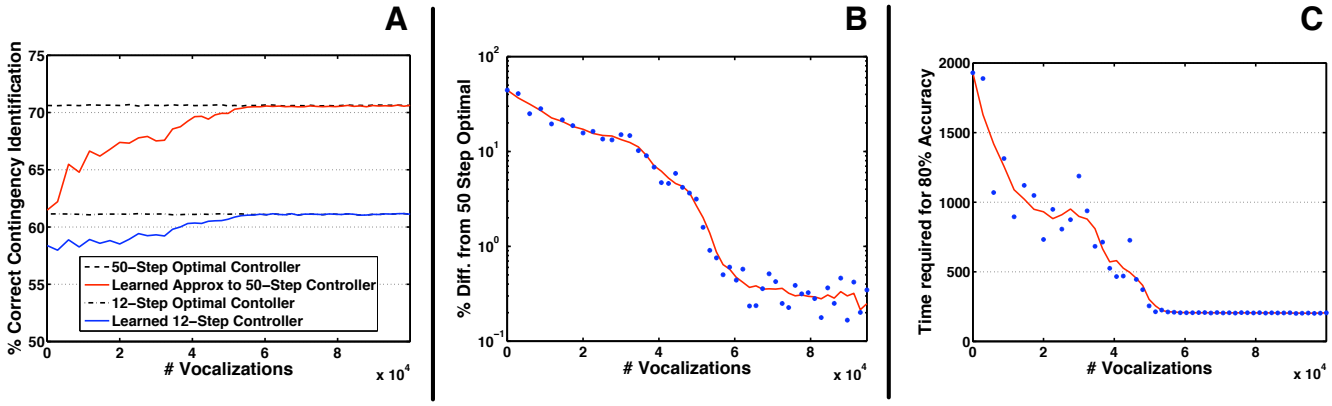

Fig. 2. **A:** Performance of TD(0) learner during learning in the exact, and in the approximate-continuing case, based on the total number of Vocalizations made since birth. **B:** When a 12-step controller is used to approximate a 50-step one, the final performance is very close to the best that could possibly be hoped from a perfectly efficient information gatherer (about 0.3% above optimal). **C:** Number of time steps spent acting, exploring, and listening to the world that are required to achieve 80% social agent identification accuracy.

We then used Dynamic Programming to compute optimal 12-step and 50-step controllers. Dynamic programming is a technique from the Theory of Stochastic Optimal Control. The advantage is that it allows finding exact optimal solutions to control problems. The disadvantage of exact dynamic programming solutions is that they tend to be more computationally expensive and less memory efficient than approximate methods like reinforcement learning. The performance of the optimal 12-step controllers found using dynamic programming (an exact method) was almost identical to the 12-Step controllers found using IRL (an approximate method) indicating that IRL actually converged to an optimal solution. Most importantly the performance of the 12-step IRL trained controllers was almost as good as the performance of the optimal 50-Step controllers: after 60,000 vocalizations, the average performance was better than 99.5% of the optimal performance, compared to chance. This indicates that given the uncertainty of real time social interaction, and for the purposes of detecting social contingency it is not worthwhile to attempt to "look–ahead" for more than about 12 seconds. These results are illustrated in Figure 2A&B.

Watson [1] found evidence that 2 month old infants could learn that a mobile was socially contingent within a 40 minute period. Movellan and Watson [2], [3] found that 10 month old infants could learn the same thing about a non-humanoid robot within a 3.5 minute period. Inspired by these results we tracked the average performance of the IRL trained controllers as a function of age (*i.e.*, number of vocalizations used for learning). Performance of the controllers was assessed in terms of how many time steps they required to learn whether they were being presented with a social agent to an accuracy level of 80%. The results are shown in Figure 2C. On average 10-month old controllers (trained with 60,000 vocalizations) were capable of detecting social contingencies in ≈ 200 time steps, the equivalent of 3.3 minutes. This was about 6 times faster than 2-month-old controllers (trained with 12,000 vocalizations), which required $\approx$ 1200 time steps, i.e., 20 minutes. This difference in performance is within range of the tenfold increase in performance observed empirically between 2 month and 10 month old infants.

IV. Conclusions

We explored the idea of development as a process of learning to learn. To this effect we focused on how infants learn to detect new contingencies between behaviors and their consequences. While it takes 2 month olds about 40 minutes to learn new contingencies, by 10 months it takes them less than 3.5 minutes. Such improvements in learning efficiency are well known in the developmental literature but are seldom modeled from a computational point of view. One popular explanation for these learning improvements is that they are due to the maturing brain structures that somehow are specially built for more efficient learning, just like teeth are built for more efficient chewing. The explanation that we explore in this document is that the better learning efficiency is itself a manifestation of the learning process, only at a longer time scale. This hypothesis does not deny the importance of brain growth but it does not see it in the same light as the role that growing teeth have on mastication.

In this paper we show evidence suggesting that Infomax Reinforcement Leaning (IRL) is a computationally reasonable approach that may help explain how infants improve on their capacity to learn. In 10 months of simulated experience, IRL agents show two properties: Given a fixed amount of time (50 seconds) to act and try to learn about the world, they perform 99.5% as well as could possibly be hoped. They also rapidly decrease the amount of time needed to exhibit a given level of learning.

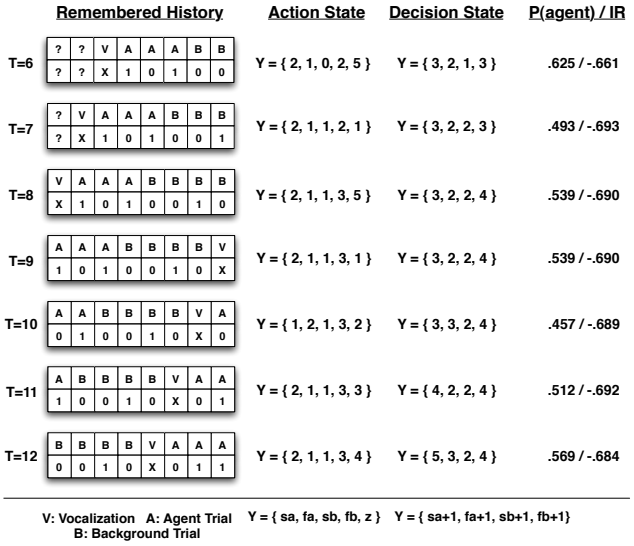

Fig. 1. Illustration of the method used to approximate a continuous controller. Eight recent events of history are used to make all decisions about how to act after the eighth time step. The state used for action is a 5-tuple consisting of s_a, f_a, s_b, f_b, z, where s/f are the successes and failures of agent and background trials, and z is a counter (*c.f.* Appendix A for more details). The decision state is used to make a judgment about the presence or absence of a contingent agent using a closed form equation (Equation 2), and summarizes all previous observations plus priors. $P(agent)$ is the current belief of the probability that an agent is present, which scores a high information reward for being close to 0 or 1, and a low information reward for being close to 0.5.

Under the IRL approach, the resultant improvement in learning speed and accuracy is due to a process of "learning to learn". This is a manifestation of a continuous learning process at the time scale of months. The same learning process manifests itself in the time scale of minutes as a process of detecting novel contingencies. An important aspect of IRL is that it uses an internally generated reinforcement signal: information gain. We showed that IRL is a computationally plausible explanation for the improvements in social contingency detection observed in human infants between 2 and 10 months of age. In addition the approach proposed here is well formalized, and computationally plausible opening new avenues for the development of robots that learn to learn on their own.

REFERENCES

[1] John S. Watson. Smiling, cooing, and "the game.". *Merrill-Palmer Quarterly*, 18:323–339, 1972.

[2] Javier R. Movellan and John S. Watson. Perception of directional attention. In *Infant Behavior and Development: Abstracts of the 6th International Conference on Infant Studies*, 1987.

[3] Javier R. Movellan and John S. Watson. The development of gaze following as a bayesian systems identification problem. In *Proceedings of the 2nd International Conference on Development and Learning (ICDL '02)*, volume 2. IEEE, 2002.

[4] Javier R. Movellan. An infomax controller for real time detection of social contingency. In *Proceedings of the 4th International Conference on Development and Learning (ICDL '02)*, pages 19–24, 2005.

[5] Sebastian Thrun, Wolfram Burgard, and Dieter Fox. *Probabilistic Robotics*. The MIT Press, 2005.

[6] Wolfram Schultz, Peter Dayan, and P. Read Montague. A neural substrate of prediction and reward. *Science*, 275:1593–1599, March 1997.

APPENDIX

A. Mathematics: Infomax Model of Social Contingency

In [4], Movellan modeled the problem of learning whether agents were socially contingent as an Infomax Control problem. We will refer to the socially contingent agent as "the agent," and the deciding agent as "the infant." All communication is through binary audio signals (sound-level in a time step is above threshold or not), and the agent and background generate audio events with certain unknown, different probabilities, each initially drawn uniformly from the range [0:1]. The agent may or may not be present with a probability 0.5, and the problem of contingency detection is ultimately the problem of deciding whether only a single background audio event rate is heard, or if two separate rates are observed for the background period and the agent period.

The infant has two choices for action: to vocalize or remain silent. If he vocalizes, he generates noise for a fixed number of time steps, called the self-period. Then, for the agent period, comprising another fixed number of time steps, the agent, if present, will respond at its appropriate rate. If the agent is not present, the background rate will be observed. After the agent's response period, the background rate is always observed. There is a timer z which is reset to 1 every time the agent vocalizes, and increments every subsequent time step, until the agent's process ends. This cap reflects that the steady state (background process) has been reached.

The problem is difficult because of tradeoffs enforced by the structure of social dynamics. When the infant makes a noise, he sacrifices observations because he is primarily hearing himself. During the time in which the agent is responding, the infant hears primarily the agent, and cannot get any information about the background. Thus the infant must choose carefully which distribution to sample at what time, subject to these constraints. To do this optimally, he should maximize the expected information he receives about whether an agent is present or absent.

Let the agent and background rates be r_a and r_b, and the number of observed audio "successes" and "failures" (audio events and no-audio events) during agent and background periods be s_a, s_b, f_a, and f_b. Sufficient statistics $\mathbf{y}$ for decision making are $\mathbf{y} = \{s_a, s_b, f_a, f_b, z\}$. The likelihoods of the infant's observations given the presence or absence of an agent are the binomial probabilities:

$$
\begin{aligned}
p(s_a, s_b, f_a, f_b|present) &= \binom{s_a + f_a}{s_a} r_a^{s_a}(1 - r_a)^{f_a} \\
&\quad \cdot \binom{s_b + f_b}{s_b} r_b^{s_b}(1 - r_b)^{f_b} \\
p(s_a, s_b, f_a, f_b|absent) &= \binom{s_a + s_b + f_a + f_b}{s_a + s_b} r_b^{s_a+s_b} \\
&\quad \cdot (1 - r_b)^{f_a+f_b}
\end{aligned}
$$

Integrating over all possible rates, it is easy to show that the probability of an agent given the infant's observations is simply:

$$
p(present|s_a, s_b, f_a, f_b) = \frac{1}{1 + \frac{\beta(s_a+s_b+1, f_a+f_b+1)}{\beta(s_a+1, f_a+1)\beta(s_b+1, f_b+1)}} \quad (2)
$$

where $\beta(x, y)$ is the β function. The probability that an agent is absent is $(1-$(Eqn. 2)$)$. From these probabilities, we can calculate the entropy (uncertainty) of the baby's estimate. Maximizing the Mutual Information between the observed data and this estimate is equivalent to minimizing the entropy (uncertainty), and so we take that negative entropy as a reward signal at each time step for a discrete time control problem.