Proceedings of the Second International Conference on Development and Learning, Cambridge, MA, June 12–15, 2002.

Combining Configural and TD Learning on a Robot

David S. Touretzky, Nathaniel D. Daw, and Ethan J. Tira-Thompson
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

We combine configural and temporal difference learning in a classical conditioning model. The model is able to solve the negative patterning problem, discriminate sequences of stimuli, and exhibit second order conditioning. We have implemented the algorithm on the Sony AIBO entertainment robot, allowing us to interact with the conditioning model in real time.

1 Introduction

To gain insight into the workability of real-time animal learning models, we implemented a classical conditioning model based on temporal difference (TD) learning on the Sony AIBO entertainment robot (Figure 1.) We also created a configural learning component whose output served to enrich the state representation available to the conditioning model. In this paper we review existing models of conditioning and describe our configural learning mechanism in detail. We found the exercise of implementing a learning model as a real-time system in the physical world leads to a better understanding of its limitations.

2 Temporal Difference Learning

Temporal difference learning [1] arose as a real-time generalization of the Rescorla-Wagner model [2] of classical conditioning. Rescorla-Wagner was a trial-level model, and although it could account for many important effects, such as additivity, blocking, overshadowing, and conditioned inhibition, it could not represent effects of stimulus timing or account for second-order conditioning. The Rescorla-Wagner learning rule is the same as the Widrow-Hoff or LMS (least-mean-square) rule [3] used to train neural networks with a single layer of weights. Hence, Rescorla-Wagner also suffers from the usual limitations of linear models, such as the inability to learn negative patterning (exclusive-or) tasks.

Barto and Sutton extended the Rescorla-Wagner model into a real-time model by suggesting that stimulus onset and offset events produce stimulus *traces* that rise rapidly and decay gradually over time. Their

Figure 1: The Sony AIBO entertainment robot, model ERS-210.

"y-dot" theory can thus account for some of the effects of inter-stimulus interval (ISI) on learning rate. They went on to posit a memory representation where time is converted to space by using an array of units to reflect the temporal structure of a trial. This allows the model to represent temporal relationships between stimuli and rewards, so that it can generate an appropriately-timed CR (conditioned response.) This "complete-serial-compound" representation [1] can be implemented by a set of shift registers. Memory is organized as a fixed array of time slots, and new stimuli enter the memory at one end and shift to the next slot (while decaying in amplitude) with each successive clock tick.

Barto and Sutton further extended y-dot theory

into temporal difference learning [4] by requiring that the model compute a value function $V(t)$ based on total expected future reward, rather than merely predicting reward at the next time step. A typical TD implementation uses exponential discounting of rewards in order to ensure that the expected total discounted future reward converges:

$$V(t) = E\left[\sum_{\tau=0}^{\infty} \gamma^{\tau} r(t+\tau)\right]$$

where $r(t)$ is the reward received at time t, and γ is a discount factor less than one. Exponential discounting is favored because it leads to a simple recursive formulation of the value function:

$$V(t) = r(t) + \gamma V(t+1)$$

The reward prediction error $\delta(t)$ used to train a network to predict $V(t)$ is simply the difference between the left and right sides of the above equation:

$$\delta(t) = \gamma V(t+1) - V(t) + r(t)$$

The error signal $\delta(t)$ has attracted considerable interest because it appears to be a good model of the primate dopamine system's response to reward prediction error [5, 6, 7]. This led to the suggestion that $V(t)$ may be computed in the striatum [5].

Daw and Touretzky [8] argue that an average reward version of TD suggested by Tsitisklis [9] is preferable to the exponential discounting formulation because it connects more directly with psychological theories of animal choice. Animals appear to discount rewards hyperbolically [10], and this behavior can be obtained more naturally with an average reward model. (The same result can be achieved with exponential discounting by making γ very close to 1 and running the trials as a continuous sequence, rather than starting each trial with an empty memory buffer as most simulations do.) In rate-based TD, the mean reward rate is subtracted from $r(t)$ to prevent $V(t)$ from diverging:

$$\delta(t) = V(t+1) - V(t) + [r(t) - \bar{r}(t)]$$

$\bar{r}(t)$ is the exponentially weighted average reward received up through time t. The average reward model is able to account for changes in the tonic firing rates of dopamine cells in terms of a change in $\bar{r}(t)$ [8].

3 Memory Representation

One problem with the complete serial compound representation is that it does not generalize well. If a model is trained with a constant ISI of 2 seconds, it will be completely surprised by a reward that comes a little too early or late. Varying the ISI across trials slows learning because the weight for each slot in the memory buffer is being trained separately. Another problem that arises specifically when this representation is used for TD is that the finer the temporal resolution of the memory, the greater the number of slots, and hence the greater the number of trials required for $V(t)$ to propagate backwards from the time of reward to a memory state t seconds earlier.

The spectral timing model proposed by Grossberg and Schmajuk [11] is a continuous-time variant of the complete serial compound representation. Spectral timing utilizes an array of stimulus detectors with various response latencies. The breadth of the tuning curve increases with latency. As stimuli age, activity smoothly shifts to units tuned to longer latencies. Spectral timing models are less brittle because their units have overlapping tuning curves with smooth falloffs, so information learned at time t will naturally generalize to nearby times. In our simulation, the detectors are gaussian functions with preferred latency μ and standard deviation $\sigma = \mu/3$.

Although feature detectors with broad variance aid generalization, the cost is a loss of temporal precision in the response. When the model is trained with a Rescorla-Wagner type learning rule (by setting $\gamma = 0$), the result is a roughly Gaussian response centered at the time of expected reward; see Figure 2a. This might be taken as suggestive of a mechanism for producing scalar timing behavior. Scalar timing, as described by Gibbon [12], refers to the standard deviation in animals' estimates of elapsed time scaling linearly with the length of the interval. However, we will not pursue this connection in this paper, as we are mainly interested here in the structure of responding at short durations (around 0.5 sec), a regime in which variability in responding is minimal, and for which the scalar property may not hold [12].

Blurring produced by broad variance, in combination with a value function that estimates future reward rather than expected reward at the current timestep, results in a response whose peak is shifted ahead of the reward; see Figure 2b. We have experimented with a multiscale version of spectral timing in which one set of feature detectors uses a fixed width $\sigma = 50$ ms, and a second set scales in proportion to the latency, $\sigma = \mu/3$. This allows the model to represent events precisely when there is low variance in the ISI, but still generalize over time, and acquire long latency events quickly via the broadly-tuned detectors that "look further ahead." Figure 2c shows the response of this multiscale model.

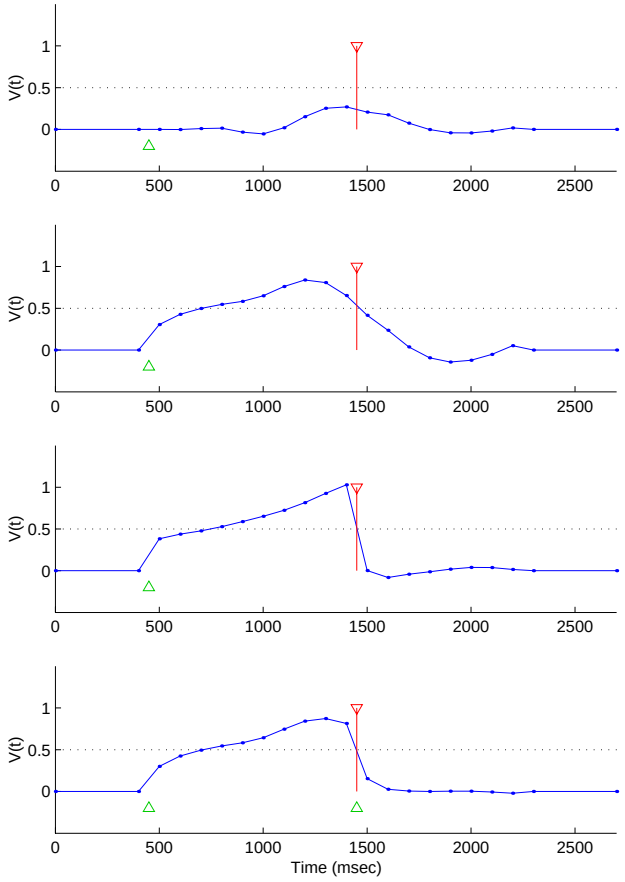

Figure 2: Models trained with 1 sec ISI. Triangles at bottom mark stimulus events; inverted triangles at top mark the US. (a) Rescorla-Wagner with scaled variance. $\sigma = \mu/3$, $\gamma = 0$. Remaining plots use TD with $\gamma = 0.9$. (b) TD with scaled variance. (c) TD with multiscale representation. (d) TD with scaled variance and explicit US representation.

Even with a multiscale representation, the model's prediction error will not go to zero if it is trained on a task where the ISI varies. Figure 3a shows responses of the model after training on ISIs of 900, 1000, and 1100 ms (200 trials each, interleaved.) The response remains elevated after the reward at 900 ms, and falls too soon when the reward comes at 1100 ms. However, by adding an explicit representation of the US as a working memory event, the model was able to produce a response properly tailored to each of the three ISIs, as shown in Figure 3b and also Figure 2d. The reasons for this are discussed in the next section.

To summarize: stimulus events in our model are stored in a working memory buffer for three seconds. A separate set of feature detectors is supplied for each

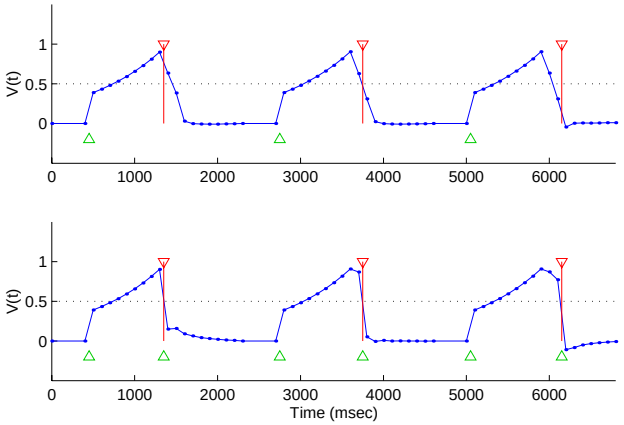

Figure 3: Models trained with variable ISI of 900, 1000, or 1100 ms. (a) Multiscale TD: $V(t)$ remains high after reward at 900 ms, falls too soon when reward comes at 1100 ms. (b) Adding an explicit US representation permits correctly-tailored responses.

event type. Each detector's response is a Gaussian function of the difference between its preferred latency and the age of the working memory event closest to that latency. Hence, if two instances of the same event type are present in the buffer, e.g., stimulus A followed two seconds later by a second occurrence of stimulus A, two sets of detectors will be active, one with near zero latencies, and one with latencies around 2 seconds. The outputs of the feature detectors are the input to the linear TD unit.

4 Responding

If a linear unit is trained to predict the CR, as in the Rescorla-Wagner model, then the output can be used directly to generate a conditioned response. However, in a TD model the unit is trained to predict $V(t)$, which rises smoothly as the time for reward approaches, then falls back to a baseline level afterwards. Thresholding the output in order to produce a well-timed response would introduce additional complexity to the model, since it's not clear how the threshold should be set. Furthermore, $V(t)$ may remain high for several time steps when a reward is imminent, yet an extended duration CR—or a train of successive CRs—might not be appropriate.

Some behavioral measures, such as freezing responses in fear conditioning, are sensitive to general future predictions and not particularly sharply timed. They could well be modeled as proportional to $V(t)$. But other behaviors are better timed. A paradigmatic example is the conditioned eyeblink experiment, in which rabbits are conditioned to a light or tone CS

preceding a puff of air directed at the eye or a shock to the eyelid; they time their CR to match the ISI. Such responses should perhaps be modeled as proportional to the predicted reinforcer $r(t)$, which can be estimated as $V(t) - V(t+1)$. Thus, we should expect reward whenever V drops by a substantial amount, indicating that the US is expected now. (To make an anticipatory response we would have to run the memory buffer forward a step to obtain $V(t+1)$, but this is not a problem.)

Moore et al. [13] observed that when the timing of the US is inherently uncertain, CR topography can vary between animals. They trained rabbits on an eyeblink task where the ISI between the light/tone CS and the shock US was either 300, 500, or 700 ms. Half the animals adopted a "failsafe" strategy in which the eye began closing at 200 ms, reached full closure by 300 ms, and remained closed past the 700 ms point. The other half followed a "conditional expectation" strategy in which the eye closed in increments as the probability of a shock increased. Another strategy, called "hedging," was observed in an experiment where there were only two possible times for the shock to occur, and the difference between them was 400 ms. In that case the animals blinked twice.

The step-like increases in eye closure seen in the conditional expectation response are difficult to explain with the basic TD model, since the correlation between CS onset and each of the three possible times for US arrival is identical. Moore et al. proposed a "marking" mechanism whereby separate cascades of event detectors would be initiated at the 300 and 500 ms intervals, since the US had been paired with a CS at those latencies. Hence, at 700 ms past CS onset there will be three cascades active, and their combined associative strengths will produce a larger output, resulting in greater eye closure.

Our simulations suggest an alternative way to model conditional expectation: by simply including US events in working memory. The units encoding US events develop negative weights that cancel the excitation provided by CS event units, so that values of $V(t)$ after the US has arrived are zero. This reflects the fact that the task involves only one US per trial. If the US comes at 300 ms, there should be no expectaton of a US at 500 or 700 ms, and our model does not respond at those times. Now the memory unit coding for a CS 700 ms ago has a stronger correlation with the US than earlier units, because those trials where the US doesn't arrive have been accounted for. Hence the value of $V(t)$ will be larger if the trial reaches 700 ms with no reward. Figure 4 shows the result.

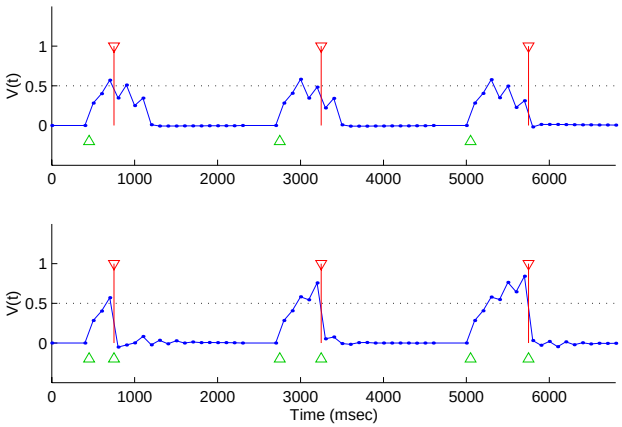

Figure 4: Responses in a model of the eyeblink experiment of Moore et al. [13] in which the is US delivered 300, 500, or 700 ms after CS onset. $\gamma = 0.7$ in both plots. (a) TD with multiscale representation produces a descending response pattern. (b) TD with an explicit US representation shows the correct ascending pattern. The leftmost curve has only one peak, and the middle curve only two, because the memory of the US at 300 or 500 ms (respectively) suppresses further responding.

5 Configural Learning

A number of studies have looked at mechanisms for recognizing configurations of cues, which could account for animals' ability to solve the negative patterning problem. See Pearce [14] for a review. The most general solution for a TD model would be to replace the single linear unit with a multi-layer backpropagation network, where the input layer was the memory buffer, the single output unit computed $V(t)$, and one or more hidden layers computed arbitrarily complex functions of the current memory state. The problem with this approach is that even small multilayer networks require considerable training time compared to the number of trials needed to demonstrate basic conditioning effects in animals. Unless we assume that the animal can learn offline, by mentally rehearsing recent trials while waiting for the next trial to begin, it will not be possible to train a network quickly enough using backpropagation to match the animal behavioral data.

An alternative to a general multilayer network is to adopt a specialized rule for creating units that recognize cue configurations. Pearce [14] gave one such rule, for configurations that consisted of a set of simultaneously presented stimuli. Since we are attempting to combine configural learning with a real-time con-

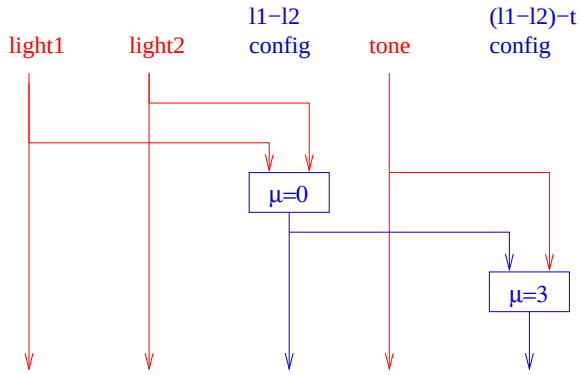

Figure 5: Configural units encoding a compound stimulus and a sequence of stimuli.

ditioning model, we will take timing into account explicitly.

We employ a very simple configural unit that looks for a conjunction of two events in working memory. Event 1 must be a stimulus that has just arrived (zero latency); event 2 can be either another sensory stimulus, or the firing of a lower-numbered configural unit. In this way, chains of configural units can be composed to recognize compound stimuli and/or sequences of stimuli.

Here are some examples of configurations that can be learned. A pair of co-occurring stimuli (say, simultaneous presentation of two lights) would be encoded by a configural unit whose event 1 was *light-1* and event 2 was *light-2*, with an inter-event interval of zero. When these events both appear in working memory at about the same time, the configural unit fires. The configural unit's activation is recorded as another event in working memory, with the same temporal tag as the unit's event 1 stimulus.

If we then want to learn a configuration consisting of the two lights followed 3 sec later by a tone, another configural unit is constructed whose event 1 is the tone, and whose event 2 is the configural unit for the pair of lights. The inter-event interval for this configural unit would be 3 sec. See Figure 5.

The range of acceptable intervals between events 1 and 2 for a particular configural unit is described by a mean μ and variance σ^2. We initialize μ to the time difference between the actual events that led to the creation of the unit. We initialize σ to $\mu/3$, based on scalar timing effects observed in animals [12]. Using the fact that $\sigma^2 = E\left[(x-\mu)^2\right] = E[x^2] - \mu^2$, we can adapt these parameters online.

If a constant ISI is used then σ^2 will go to zero, so it is necessary to impose a lower bound to preserve some generalization ability. We maintain $\sigma \geq \mu/3$, and to prevent fluctuations associated with a small sample size, we put off adaptation until $n > 5$.

Due to the adaptation process, the order of presentation of training trials can affect the way the model categorizes the input. Consider two types of trials where stimulus A is followed by stimulus B after a delay. If we initially train on A-B sequences with a 1 sec inter-stimulus interval, the configural unit for that sequence will have $\sigma = 0.5$ secs. Subsequent training on A-B sequences with a 0.5 sec ISI will activate the same configural unit. On the other hand, if the model is initially trained with the 0.5 sec ISI, then $\sigma = 0.25$ secs, and the configural unit will not become active for A-B sequences with 1 sec delays. Instead, a new configural unit will be created, and the model will treat the two sequences as distinct event types.

6 Classical Conditioning on the AIBO

The Sony AIBO is an autonomous robot with color vision and a variety of other sensors, and sixteen degrees of freedom of motion. For our initial experiments with the AIBO we used touch switches on the bottoms of the robot's feet, a touch switch on its back, and one under its chin as conditioned stimuli. A touch switch on the head sensed the reward stimulus (i.e., a "pat on the head.") For the conditioned response we used either a head movement, or flashing of the LEDs that serve as "eyes" in the robot's face. (These colored LEDs, with accompanying sound effects, are used by AIBO applications to signal emotional states of the robot.)

The robot processes events at roughly 30 Hz, which is more than adequate temporal resolution. Initially there was no response to the conditioned stimuli, but by pairing them with a subsequent reward, the robot learned to make a CR at the time a reward should be expected. Using versions of the learning architecture described above, we successfully demonstrated learned inhibition, negative patterning, and second order conditioning on the AIBO.

7 Summary and Conclusions

Our inability to manually deliver precisely timed stimuli forced us to deal with the issues of spectral timing and temporal generalization. Another problem that arises in real-world implementations is the richness of stimuli encountered, which threatens a combinatorial explosion of configural units.

We are exploring heuristics to limit the number of configural units built, or to prune units that do not contribute significantly to reduction of prediction error. One possible heuristic would be to concentrate

the creation of configural units on trials where the prediction error is significant. Another is to limit the amount of redundancy in the population. In Figure 5, in addition to the units shown, a naive algorithm will also create configural units for *light1-tone* and *light2-tone*. Redundancy supports generalization by allowing the model to respond to partial patterns, so it should not be eliminated entirely. But some limit, perhaps empirically determined, seems necesary

Attention is an important issue in the creation of configural units. Ideally we would like a long memory buffer (much longer than 3 secs) to learn long duration sequences, or tasks involving long ISIs. But then, in tasks with short inter-trial intervals, the buffer doesn't empty between trials, so the memory state is complex and there is an explosion in the number of configural unit candidates. Heuristics will be required to cope with this complexity, e.g., by focusing on more recent events in the buffer unless older ones show important correlations with error. Touretzky and Saksida [15] describe heuristics for judging the saliency of working memory elements when constructing conjunctive representations.

Acknowledgments

This work was funded by National Science Foundation grants IIS-9978403 and DGE-9987588, and by a grant from the Sony Corporation. We thank Aaron Courville for useful discussions and for help with the AIBO robot, and Scott Lenser of the CMU Robosoccer group for additional robot help.

References

[1] A. G. Barto and R. S. Sutton, "Time-derivative models of Pavlovian conditioning," in *Learning and Computational Neuroscience: Foundations of Adaptive Networks* (M. Gabriel and J. Moore, eds.), pp. 497–537, Cambdridge, MA: MIT Press, 1990.

[2] R. A. Rescorla and A. R. Wagner, "A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement," in *Classical Conditioning II: Theory and Research* (A. H. Black and W. F. Prokasy, eds.), New York: Appleton-Century-Crofts, 1972.

[3] B. Widrow and M. A. Lehr, "Perceptrons, Adalines, and backpropagation," in *The Handbook of Brain Theory and Neural Networks* (M. A. Arbib, ed.), pp. 719–724, The MIT Press, 1995.

[4] R. S. Sutton, "Learning to preduct by the method of temporal differences," *Machine Learning*, vol. 3, pp. 9–44, 1988.

[5] J. C. Houk, J. L. Adams, and A. G. Barto, "A mode of how the basal ganglia generate and use neural signals that predict reinforcement," in *Models of Information Processing in the Basal Ganglia* (J. C. Houk, J. L. Davis, and D. G. Beiser, eds.), ch. 13, pp. 249–270, MIT Press, 1995.

[6] P. R. Montague, P. Dayan, and T. J. Sejnowski, "A framework for mesencephalic dopamine systems based on predicive Hebbian learning," *Journal of Neuroscience*, vol. 16, no. 5, pp. 1936–1947, 1996.

[7] W. Schultz, P. Dayan, and P. R. Montague, "A neural substrate of prediction and reward," *Science*, vol. 275, pp. 1593–1599, 1997.

[8] N. D. Daw and D. S. Touretzky, "Behavioral results suggest an average reward TD model of dopamine neurons," *Neurocomputing*, vol. 32, pp. 679–684, 2000.

[9] J. N. Tsitsiklis and B. V. Roy, "Average cost temporal-difference learning," *Automatica*, vol. 35, pp. 319–349, 1999.

[10] A. Kacelnik, "Normative and descriptive models of decision making: time, discounting, and risk sensitivity," in *Characterizing Human Psychological Adaptations* (G. R. BBlock and G. Cardew, eds.), pp. 51–70, Wiley, 1997.

[11] S. Grossberg and N. A. Schmajuk, "Neural dynamics of adaptive timing and temporal discrimination during associative learning," *Neural Networks*, vol. 2, pp. 79–102, 1989.

[12] J. Gibbon, "Scalar expectancy theory and Weber's law in animal timing," *Psychological Review*, vol. 107, no. 2, pp. 289–344, 1977.

[13] J. W. Moore, J.-S. Choi, and D. H. Brunzell, "Predictive timing under temporal uncertainty: The TD model of the conditioned reponse," in *Timing of Behavior: Neural, Computational, and Psychological Perspectives* (D. A. Rosenbaum and C. E. Collyer, eds.), pp. 3–34, MIT Press, 1998.

[14] J. M. Pearce, "Similarity and discrimination: a selective review and a connectionist model," *Psychological Review*, vol. 101, no. 4, pp. 587–607, 1994.

[15] D. S. Touretzky and L. M. Saksida, "Operant conditioning in Skinnerbots," *Adaptive Behavior*, vol. 5, no. 3/4, pp. 219–247, 1997.