Operant behavior suggests attentional gating of dopamine system inputs [1]

Nathaniel D. Daw & David S. Touretzky

Computer Science Department & Center for the Neural Basis of Cognition, Carnegie Mellon University, Pittsburgh PA 15213

Abstract

Neurophysiological recording experiments in the dopamine system by Schultz and colleagues (6) suggest that neurons there are involved in learning to predict rewards and assess behaviors using the temporal-difference algorithm. One aspect of this theory which is undeveloped and experimentally underconstrained is its assumption of an exhaustive input representing all stimuli and their history over time. We use the algorithm to model operant choice between concurrent variable interval schedules—a key animal conditioning experiment—and show that animals' subtly suboptimal performance resembles the behavior of the algorithm with a more limited input representation. This limitation may reflect the operation of an attentional mechanism gating the inputs to the dopamine system.

Key words: dopamine; operant conditioning; temporal-difference learning

1 Introduction

Neuronal activity in the dopamine system is well characterized by a temporal-difference (TD) learning model (6). In this model, dopamine activity reports an error signal for the prediction of reward, useful for selecting actions which maximize expected reward. Since TD is an optimization algorithm, the model suggests a neural basis for optimality-based accounts of animal behavior (2).

A feature of the TD model which is underconstrained by the neurophysiological data is the nature of the input representation or "state space" from which the algorithm constructs a reward prediction. The model builds its prediction from an exhaustive list of information about all stimuli in the environment

[1] This research was supported by NSF IRI-9720350 and IIS-9978403. Nathaniel Daw is supported by an NSF Graduate Research Fellowship. The authors thank Sham Kakade and Peter Dayan for helpful discussions.

and their prior history. In a realistic situation, such a vast state space would cause the algorithm to perform poorly.

We search for behavioral constraints on the dopamine model's state space by investigating TD models of one of the key results thought to suggest that animals can make optimal choices: rate matching on free operant choice between concurrent variable interval schedules. In our models, a TD learner using the complete state space outperforms animals; the animals' poorer choices resemble those of a TD learner using a reduced state space that lacks information about the intervals between some events.

This inadequacy in the prediction system's input representation may reflect the action in an operant conditioning context of attentional mechanisms previously postulated (3) to explain classical conditioning experiments. In these theories, stimuli compete for attention on the basis of their reliability as reward predictors; such competition may occur between stimulus representations in dopamine system input structures such as the nucleus accumbens (3).

2 TD model of dopamine

Response properties of dopamine neurons in primate substantia nigra pars compacta and ventral tegmental area resemble the error signal $\delta(t)$ by which TD learns to estimate a value function mapping the state of the world at time t to a prediction $V(t)$ of reward expected in the future (6). In a version of the model with strong connections to behavior (2), the value function is defined as cumulative, average-adjusted expected reward: $V(t) = E\left[\sum_{\tau=t}^{\infty} r(t) - \rho\right]$, where $r(t)$ is the reward at time t and ρ the average reward per timestep. The error by which an estimate of $V(t)$ is incrementally improved is $\delta(t) = V(t+1) - V(t) + r(t) - \rho$. The model uses a variation on policy iteration to guide action selection: a behavioral policy is repeatedly improved by favoring choices which lead to the states that predict the most reward.

The model estimates $V(t)$ linearly as $\mathbf{w}(t) \cdot \mathbf{x}(t)$, the product of a trainable weight vector $\mathbf{w}$ and a stimulus vector $\mathbf{x}$. But there are few constraints on how the state of the world should map to a stimulus vector $\mathbf{x}(t)$; the model assumes a set of binary vectors $\mathbf{x}_{ij}(t)$ encoding the stimuli present at t along with all previous history: $\mathbf{x}_{ij}(t) = 1$ if the stimulus i was present at time $t-j$.

3 VI and matching

Some of the most robust quantitative data in instrumental conditioning come from free operant choice between concurrent variable interval (VI) schedules (4). Under a VI schedule, an animal's response on a lever is reinforced (e.g. with food) if the interval since the *last* reinforced response exceeds some threshhold, chosen randomly from a Poisson distribution. In the *concurrent* VI task, independent VI schedules with different payoff rates run simultaneously on two levers, allowing study of the distribution of responses between

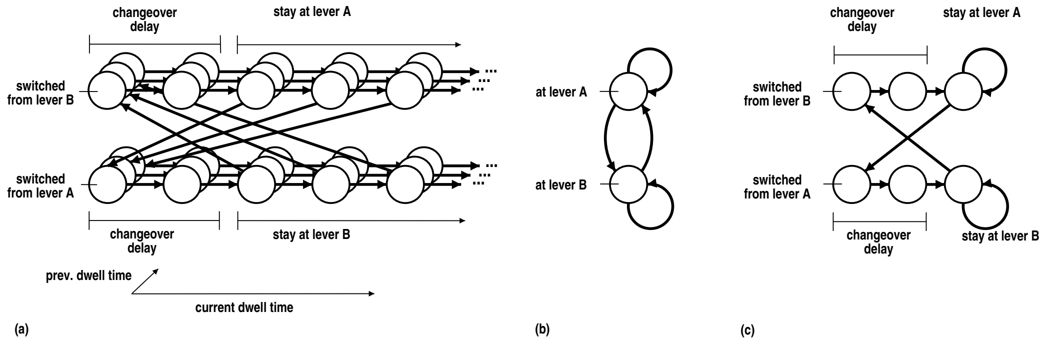

Fig. 1. State spaces for models of concurrent VI. (a) Complete state space, with states tracking the conjunction of current and previous dwell times. Some transitions omitted. (b) Minimal state space representing only the current side. (c) State space used in our TD model: states track dwell time during the COD but ignore it thereafter. Transitions for lever switches during the COD omitted.

levers. The optimal policy is not to devote all responses to the richer lever, but to divert some attention to the poorer lever as well. The reason is that, the longer the poorer lever is ignored, the more likely it is that its interval has expired, causing a reward to be waiting there for collection. Fast cycling between alternatives is discouraged by a *changeover delay* (COD): a short unrewarded period enforced whenever the subject switches levers.

The classic result (4) is that animals closely match the proportion of responses R_n on each lever to the proportion of rewards r_n received there: $r_1/(r_1 + r_2) = R_1/(R_1 + R_2)$. Equivalently, since the responses on a lever are roughly proportional to the time T_n spent there, animals allocate their responses so that the reward rates on each lever match each other: $r_1/T_1 = r_2/T_2$.

Theories of matching have focused on the fact that under some allocation strategies, matching is close to optimal in terms of the overall reward rate $(r_1 + r_2)/(T_1 + T_2)$ received (1). In the melioration model (5), matching results from a gradient ascent optimization scheme under which the subject tries to improve the overall reward rate by shifting responses toward the alternative that is paying off better. This model is conceptually similar to the TD approach, suggesting that TD dopamine models might extend to matching behavior.

4 TD models of the VI task

Modeling the concurrent VI task requires a state space with enough information to predict the chance of reward from any situation. Rather than modeling individual lever presses, we assume that animals leverpress frequently during a visit to a lever and model only their decisions to stay or switch levers. Under this approximation, the only state needed to predict the chance of reward during a long bout on a lever is which lever the animal is visiting, since we assume the subject responds quickly enough to collect rewards at roughly the constant, Poisson rate at which they are made available.

More state is required to predict the expected payoff after switching sides.

When a subject decides to switch to a new lever, there is a chance that an uncollected reward has become available there during the time away. This chance depends on how much time has passed since the *last* response on the new lever (that is, how long the animal dwelt at the old lever). But any pending reward won't be delivered until the COD expires. The probability of reward after switching sides is thus a function of dwell time at the *new* lever (which determines how much of the COD has passed) and dwell time at the *old* lever (which determines the likelihood that a reward is waiting at the new lever). Hence, the state space required to fully predict the reward available is a conjoint representation of the current and previous dwell times (figure 1a). This state space can be expressed in the manner of the TD dopamine model by taking the "stimuli" $\mathbf{x}_{i0}$ to be the events of switching from left to right and right to left, so that the $\mathbf{x}_{ij}$s correspond to different dwell times. Extra stimulus elements would also be needed to represent the conjunction of $\mathbf{x}_{1j}$s with $\mathbf{x}_{2k}$s.

Using such a state space, we solved using policy iteration for the optimal action selection strategy that a TD learner would find. Candidate policies are labelings of the arrows in the state space diagram with the probability at each state of staying or switching sides. The optimal policy is deterministic: respond on a single lever until the chance that uncollected reward is available on the other lever rises sufficiently, then switch. Figure 2a shows histograms of the durations of visits to each lever as the policy is executed. These dwell time distributions are sharply peaked, reflecting deterministic switching.

However, animals' switching behavior appears to be mostly Poisson (3). After an initial period of not much switching, probably corresponding to the COD, the chance of leaving a lever is constant as dwell time increases. This pattern produces dwell-time distributions (resembling those modeled in figure 2b) that are not so peaked, but instead descend linearly on a log plot. Strategies of this kind do not pay off as well as the optimal deterministic strategy, so animals do not perform as well as the TD model using the full state space.

The TD model can be made to exhibit Poisson switching consistent with animal behavior if it is constrained to work in a less informative state space. After all, animals' switching, being mostly Poisson, is mostly *not* sensitive to accumulated dwell time, so the model's state space need not represent it. Psychological models of concurrent VI behavior usually assume a two-state state space (figure 1b), representing only which lever is currently being visited. In this state space, a behavioral policy is fully specified by assigning a Poisson leaving rate to each side.

In fact, this state space is a bit *too* impoverished to account for animals' behavior, because it neglects the COD. Empirically, animals rarely switch after short dwell times, presumably because they learn to wait out the COD. Only for longer dwell times is switching Poisson. Therefore, the state space suggested by the behavior (figure 1c) tracks dwell time for the duration of the COD, then reverts to a Poisson mode in which further dwell time is ignored.

We used policy iteration to find the strategies that a TD learner operating

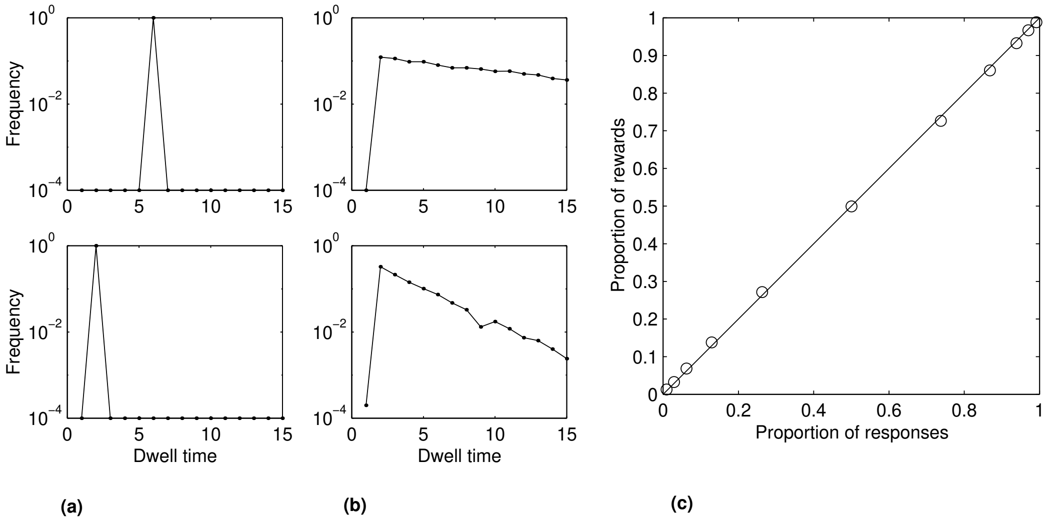

Fig. 2. Characteristics of concurrent VI policies learned by TD. (a) Using the full state space, the model produces sharply peaked dwell time distributions when a lever delivering rewards with an 8 second average inter-reward interval (top) is pitted against one with a 16 sec. avg. IRI (bottom). (b) In the reduced state space, the model produces dwell time distributions in accord with empirical measurements. Few dwell times are shorter than the COD, and at longer dwell times, switching is Poisson (distribution descends linearly on a log plot). (c) The proportion of responses on a lever with an 8 sec. IRI matches the proportion of rewards received from it when pitted successively against levers with IRIs ranging from 8-256 secs.

in this reduced state space would choose, that is to discover the best of the suboptimal policies available. (We used the additional assumption, common in models of this task, that the leaving rates from the two sides must sum to a constant.) The policies' dwell-time distributions (figure 2b) have characteristics similar to the empirical distributions (3): the system learns to avoid switching during the COD, and thereafter reverts to Poisson switching. Moreover, for a variety of pairs of payoff rates, the learned policy displays rate matching consistent with animal behavior (figure 2c).

5 Discussion

We have shown how rate matching, as well as switching policy, on the concurrent VI task could result from a TD algorithm constrained to work with only partial information about past events, rather than the unrealistically comprehensive history representation used in several dopamine system models. The state space in which animals seem to be learning is oddly inconsistent, though: subjects track dwell time during the COD, but afterwards seem to ignore it, even though continuing to mark time would earn them more rewards.

Why should this information be ignored? One possibility is that an event's representation in the state space is determined by its independent predictive value. In this task, the most predictive information by far is the time since switching to the current lever. This is because it can forecast the occurrence of a large spike in the probability of reward that occurs at the moment the COD expires, when any reward which became available during the subject's time away from the lever is delivered. How much chance there is that reward will occur here depends on how long the subject spent away from the current lever. But the dopamine model treats the same event happening at different times in the past as separate elements in the stimulus vector. The particular knowledge that, say, one last left the current lever 28 seconds ago is only occasionally predictively useful — when that interval happens to coincide with the expiry of a COD on the return to the lever — and then only in conjunction with knowledge about *when* the COD will expire. By comparison, the knowledge that one arrived at the current lever five seconds ago, if the COD is five seconds, is consistently predictive alone. In consequence, the animal may ignore the weakly informative knowledge about past dwell times in favor of more consistently useful information about the progression of the COD, producing the appearance of a state space like that shown in figure 1c.

Such competition between stimuli has been proposed by Dayan and collaborators (3). They argue that attentional effects in classical conditioning follow from the principle that a stimulus' contribution to prediction should be weighted by its predictive *reliability*. The selective inattention to events in operant conditioning that we model here may result from a similar principle.

Applied to dopamine models, the behavior considered in this paper suggests that the sensory and memory representations in dopamine system input structures such as nucleus accumbens and prefrontal cortex are not as exhaustive as assumed in previous models, but rather show attentional modulation. In particular, these results provide further behavioral support for the suggestion (3) that stimuli compete for representation in the nucleus accumbens.

References

[1] W.M. Baum, Optimization and the matching law as accounts of instrumental behavior. *J. Exp. Anal. Behav.* 36 (1981) 387-403.

[2] N.D. Daw and D.S. Touretzky, Behavioral considerations suggest an average reward TD model of the dopamine system, *Neurocomputing* 32 (2000) 679-684.

[3] P. Dayan, S. Kakade and P.R. Montague, Learning and selective attention, *Nature Neurosci.*, forthcoming, (2000).

[3] J. Gibbon, Dynamics of time matching, *Psychonomic Blltn. & Rev.* 2 (1995) 208-215.

[4] R.J. Herrnstein, Relative and absolute strength of response as a function of frequency of reinforcement, *J. Exp. Anal. Behav.* 4 (1961) 267-272.

[5] R.J. Herrnstein and W. Vaughan, Melioration and behavioral allocation, in: J.E.R. Staddon, ed., *Limits to action.* (Academic Press, New York, 1980) 143-176.

[6] W. Schultz, P. Dayan and P.R. Montague, A neural substrate of prediction and reward, *Science.* 275 (1997) 1593-1598.