ELSEVIER

Neurocomputing 32–33 (2000) 679–684

NEUROCOMPUTING

www.elsevier.com/locate/neucom

Behavioral considerations suggest an average reward TD model of the dopamine system[☆]

Nathaniel D. Daw[*], David S. Touretzky

Computer Science Department & Center for the Neural Basis of Cognition, Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA 15213, USA

Accepted 13 January 2000

Abstract

Recently there has been much interest in modeling the activity of primate midbrain dopamine neurons as signalling reward prediction error. But since the models are based on temporal-difference (TD) learning, they assume an exponential decline with time in the value of delayed reinforcers, an assumption long known to conflict with animal behavior. We show that a variant of TD learning that tracks variations in the average reward per timestep rather than cumulative discounted reward preserves the models' success at explaining neurophysiological data while significantly increasing their applicability to behavioral data. © 2000 Published by Elsevier Science B.V. All rights reserved.

Keywords: Dopamine; Exponential discounting; Temporal-difference learning

1. Introduction

Recently there has been much interest in reinforcement learning models of the dopamine system. These models [3,7,10] explain data on primate midbrain dopamine neurons (reviewed in [9]) in terms of temporal-difference (TD) learning [11]. As TD models have also been applied to animal behavior [12], the TD interpretation of the dopamine system has held out hope for a theory connecting neuronal responses to

[☆] This research was supported by NSF IRI-9720350. Nathaniel Daw is supported by an NSF Graduate Research Fellowship. The authors thank Rudolf Cardinal, Barry Everitt, Sham Kakade and Peter Dayan for helpful discussions.

[*] Corresponding author. Tel.: + 1-412-268-2582; fax: + 1-412-268-5576.

E-mail address: daw@cs.cmu.edu (N.D. Daw).

PII: S0925-2312(00)00232-0

high-level behavior: how animals learn to predict rewards and use their predictions to select optimal actions.

The existing models define optimality using an assumption common in reinforcement learning, that the value of a reward decreases exponentially in the length of its delay. But in experiments, animals' choices seem *not* to reflect such an assumption. And the predominant understanding of these experiments in psychology — that animals discount rewards hyperbolically in their delays — is incompatible with a reinforcement learning formulation.

We argue that an alternative explanation of these experiments, due to Kacelnik [5], is better suited to reinforcement learning, and hence to a combined model of dopamine and behavior. If animals' choices are governed by reward *rates*, rather than cumulative discounted rewards, then the learning task mirrors an area of active research in reinforcement learning: maximizing the average reward per time step. We show that a model based on such an algorithm [13] extends previous models' behavioral applicability.

2. TD models of dopamine

Recordings by Schultz and collaborators (reviewed in [9]) show that dopamine neurons in primate substantia nigra pars compacta (SNc) and ventral tegmental area (VTA) respond to rewards and reward-predicting stimuli. Several models [3,7,10] propose that this activity signals reward prediction error, as computed by TD learning [11]. TD systems learn to associate states of the world with a "value function" predicting future rewards, and can use these predictions to select actions maximizing the value function. The value function, $V_{\text{exp}}(t) = E[\sum_{\tau=t}^{\infty}\gamma^{\tau-t}r(\tau)]$ (where $r(\tau)$ is reward at time τ and $\gamma < 1$ is a discounting parameter), can be rewritten recursively as $V_{\text{exp}}(t) = E[r(t) + \gamma V_{\text{exp}}(t+1)]$, suggesting the error signal $\delta(t) = r(t) + \gamma V_{\text{exp}}(t+1) - V_{\text{exp}}(t)$. In the models, dopamine neuron firing signals $\delta(t)$, explaining several properties of the cells: response to unpredicted rewards, response transfer to earliest predictive stimuli, and baseline inhibition for missed rewards.

The assumption that rewards are discounted exponentially in their delays, as γ^{delay}, is behaviorally suspect but meets two algorithmic requirements. If $r(t)$ is bounded, $V_{\text{exp}}(t)$ converges. Eliminating discounting leads to the cumulative value function $V_{\text{cum}}(t) = E[\sum_{\tau=t}^{\infty}r(\tau)]$, which diverges. Also, $V_{\text{exp}}(t)$ is definable recursively; the online learning algorithm exploits the relationship between successive predictions and the immediately observable reward. Discounting delayed rewards hyperbolically, (as $1/(\theta + \text{delay})$ for some θ; popular in psychology) produces the value function $V_{\text{hyp}}(t) = E[\sum_{\tau=t}^{\infty}1/(\theta+\tau-t)r(\tau)]$, which both diverges and cannot be expressed recursively.

3. Behavioral tests of discounting

A number of behavioral experiments have tested reward discounting. The experiments study how animals trade off reward magnitudes against delays by offering

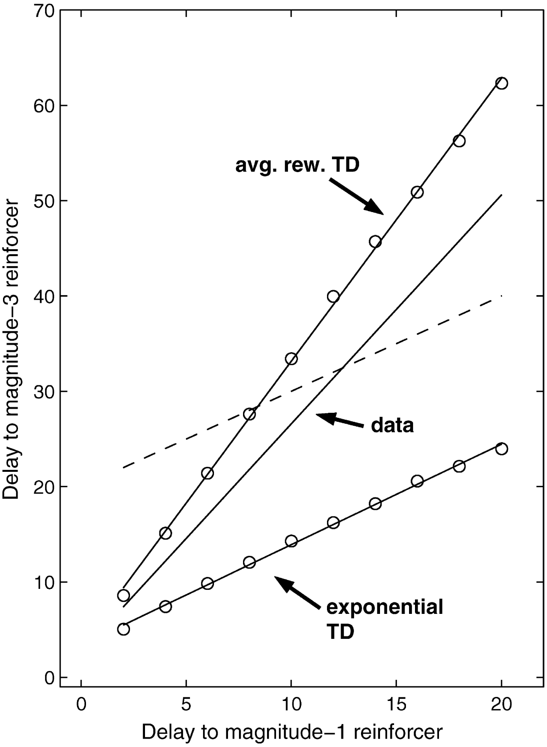


Fig. 1. Calculation of indifference points as a function of delay in the task of Mazur (1987). The line-marked "data" plots the average of Mazur's four subjects. Dashed line shows constant $D_2 - D_1$, giving a slope of one.

subjects repeated choices between a small reward of size R_1 after a short delay D_1, or a large reward R_2 after a long delay D_2. Under exponential discounting, both rewards' values decline multiplicatively by γ with each time step, so with fixed magnitudes, relative preference depends only on the *difference* in delays $D_2 - D_1$. Under hyperbolic discounting ($1/(\theta +$ delay)), delaying a reward one further timestep will reduce its value drastically after a short delay, but will have less proportional effect with longer initial delay. So unlike exponential discounting, this account predicts that any preference for the smaller of the two reinforcers should eventually reverse to a preference for the larger if D_1 and D_2 are both increased while maintaining $D_2 - D_1$. Such "preference reversals" are well demonstrated experimentally (reviewed in [1]).

Using pigeons, Mazur [6] studied "points of indifference": pairs of delays D_1 and D_2 for which fixed small and large rewards are equally likely to be chosen. With a reasonably small γ, exponential discounting predicts the function relating these delays should be $D_2 = \log_\gamma(R_1/R_2) + D_1$, a line with slope one. Hyperbolic discounting instead predicts $D_2 = (R_2/R_1)(\theta + D_1)$, a line with slope > 1 (since $R_2 > R_1$). Consistent with hyperbolic discounting, but not exponential, Mazur's measurements were well fit by lines with slopes ranging between two and three (Fig. 1, line-marked "data"). The existence of preference reversals can also be inferred from this data. If the indifference line separates the region (above) where the small reinforcer is preferred from the region (below) where the large reinforcer is preferred, then starting above the line and increasing both delays equally, which maintains their difference, traces a line of slope one (dashed line), which must cross the steeper-sloped separator to enter the region where the larger reinforcer is preferred. But it will never cross the slope-one separator of exponential discounting.

Since hyperbolic discounting is incompatible with TD, these results constrain TD models of the dopamine system's role in behavior. Kacelnik [5] argued that an undiscounted model, in which animals maximize *rate* of reward, could also explain the results. Since the rate is magnitude divided by time, on these experiments, rate maximization implies maximizing $R/(\text{ITI} + \text{delay})$, which resembles hyperbolic discounting with θ taken as the intertrial interval, ITI. Reinforcement learning algorithms for the discrete-time analog of this problem — maximizing undiscounted average reward per timestep — could be used to model both dopamine cell responses and behavior.

4. An average reward TD model of dopamine and behavior

A TD-like algorithm which tracks average rather than discounted reward was proposed by Tsitsiklis and Van Roy [13]. They redefine the TD value function as $V_{\text{avg}}(t) = E[\sum_{\tau=t}^{\infty} r(\tau) - r_{\text{avg}}(\tau)]$, where $r_{\text{avg}}(\tau)$ is a running estimate of the average reward per timestep. V_{avg} mirrors the cumulative value function $V_{\text{cum}}(t) = E[\sum_{\tau=t}^{\infty} r(\tau)]$, but rescales it by subtracting $r_{\text{avg}}(\tau)$ at each step to avoid divergence. By its difference from zero, V_{avg} estimates how much *better* or *worse* than average is the current reward expectation. The error signal is $\delta(t) = r(t) - r_{\text{avg}}(t) + V_{\text{avg}}(t+1) - V_{\text{avg}}(t)$, and $r_{\text{avg}}(t)$ must be learned separately, e.g. by $r_{\text{avg}}(t+1) = (1-v)r_{\text{avg}}(t) + vr(t)$.

We used this algorithm to model both the responses of dopamine cells and animal discounting behavior. We estimated the value function as $V_{\text{avg}}(t) = W(t) \cdot S(t)$, the dot product of trainable weights $W(t)$ and a state vector $S(t)$ that included both stimuli and recent history. $W(t)$ was updated by the delta rule, using the average reward TD error: $\Delta W(t) \propto \delta(t)S(t)$.

The model behaves almost identically to earlier TD models when tested on simulated tasks from Montague et al. [7] modeling dopamine responses. However, the average reward model also fits the behavioral data on discounting, which previous TD models did not. Fig. 1 compares the average reward and exponentially discounted TD algorithms on the task from [6]. For making choices, both models were augmented with a simple "actor" which used the TD predictions to learn choice preferences. Preference for the alternatives was represented by parameters $P_1(t)$ and $P_2(t)$. The first action was chosen with probability $Prob_1(t) = e^{P_1(t)}/(e^{P_1(t)} + e^{P_2(t)})$, and whichever action was chosen, its parameter was updated proportionately to the resultant TD error: $\Delta P(t) \propto \delta(t)$.

Both models were exposed to 350 series of 5000 trials each using different pairs of delays, and their preferences after training recorded. Indifference points were estimated by linearly interpolating between measurement pairs. The average reward algorithm's results were fit by a slope of 3 (Fig. 1, top line), similar to those measured by Mazur [6] (middle line), while the standard TD model (with $\gamma = 0.6$) produced a much lower slope (bottom line). These results also show that the average reward algorithm, but not the exponential one, could produce preference reversals (dashed line). In order to match the low y-intercepts of Mazur's data — controlled by the ITI under these models — we assumed an ITI shorter than that used by Mazur. Kacelnik [5] suggests this is necessary because animals may not mark time effectively between trials.

5. Discussion

We have reviewed data constraining models of the dopamine system's role in reward-guided behavior. The data support an average reward TD model of the dopamine system over previous exponentially discounted models. Our model preserves earlier models' success explaining dopamine responses, but extends the models' behavioral applicability, in part by connecting with existing psychological theories, for which rate is often a key variable (e.g. [4]).

The model requires the dopamine neurons to use an estimate of the average reward to compute the TD error. Simple learning rules could produce this estimate in an area such as central amygdala, which inhibits the VTA. Alternatively, average reward need not be computed separately; whatever cells (presumably hypothalamic) transmit primary rewards $r(t)$ to the dopamine neurons could exhibit fatigue, reducing their effective output to $r(t) - r_{\mathrm{avg}}(t)$.

Our model suggests an untested prediction about the response of dopamine neurons to unpredicted rewards. For rewards delivered on a Poisson schedule, $\delta(t)$ reduces to $r(t) - r_{\mathrm{avg}}(t)$. If the dopamine response is sensitive to the *magnitude* of $\delta(t)$ (itself an untested prediction), then the responses to rewards, and baseline firing between rewards, should *decrease* as the rate of reward delivery *increases*. This holds even for exponentially discounted TD, so it could provide a generic test of the TD approach.

One might also attempt to bring previous TD models into line with the behavioral results by increasing γ to such a level that the value function approaches V_{cum}, since choices maximizing cumulative reward will also maximize average reward. But the level of γ (at least 0.99) required to reproduce the behavioral measurements is impractical: as γ approaches one and the value function approaches divergence, the algorithm converges increasingly slowly.

Future work in models of dopamine and behavior includes incorporating attentional effects, important both to dopamine responses and conditioning behavior [2]. Also, the history, state and action space representations of TD models are too poor to represent the hierarchical structure of real instrumental conditioning [8]. The changes we suggest are orthogonal to these issues, and could combine with other proposals for addressing them.

References

[1] C.M. Bradshaw, E. Szabadi, Choice between delayed reinforcers in a discrete-trials schedule, Q. J. Exp. Psychol. B 44 (1992) 1–16.

[2] P. Dayan, T. Long, Statistical models of conditioning, NIPS 10 (1997) 117–124.

[3] J.C. Houk, J.L. Adams, A.G. Barto, A model of how the basal ganglia generate and use neural signals that predict reinforcement, in: J.C. Houk, J.L. Davis, D.G. Beiser (Eds.), Models of Information Processing in the Basal Ganglia, MIT Press, Cambridge, 1995, pp. 249–270.

[4] C.R. Gallistel, J. Gibbon, Time, rate and conditioning (2000), forthcoming.

[5] A. Kacelnik, Normative and descriptive models of decision making: time discounting and risk sensitivity, in: G.R. Bock, G. Cardew (Eds.), Characterizing Human Psychological Adaptations, Wiley, Chichester, 1997, pp. 51–70.

[6] J.E. Mazur, An adjusting procedure for studying delayed reinforcement, in: M.L. Commons, J.E. Mazur, J.A. Nevin, H. Rachlin (Eds.), Quantitative Analyses of Behavior, Vol. V, Erlbaum, Hillsdale, 1987, pp. 55–73.

[7] P.R. Montague, P. Dayan, T.J. Sejnowski, A framework for mesencephalic dopamine systems based on predictive Hebbian learning, J. Neurosci. 16 (1996) 1936–1947.

[8] L.M. Saksida, S.M. Raymond, D.S. Touretzky, Shaping robot behavior using principles from instrumental conditioning, Robot. Autonom. System 22 (1998) 231–249.

[9] W. Schultz, Predictive reward signal of dopamine neurons, J. Neurophys. 80 (1998) 1–27.

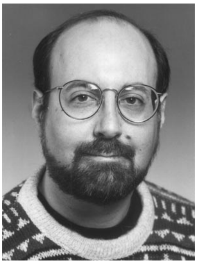
[10] W. Schultz, P. Dayan, P.R. Montague, A neural substrate of prediction and reward, Science 275 (1997) 1593–1599.

[11] R.S. Sutton, Learning to predict by the method of temporal differences, Mach. Learning 3 (1988) 9–44.

[12] R.S. Sutton, A.G. Barto, A temporal-difference model of classical conditioning, in: Proceedings of the Ninth Annual Conference of the Cognitive Science Society, Erlbaum, Hillsdale, 1987, pp. 355–378.

[13] J.N. Tsitsiklis, B. Van Roy, Average cost temporal-difference learning, Automatica 35 (1999) 319–349.

Nathaniel D. Daw is a Ph.D. student in the Computer Science Department and Center for the Neural Basis of Cognition at Carnegie Mellon University. He received his undergraduate degree in Philosophy of Science from Columbia University in 1996. His research interests include the algorithmic and representational foundations of learning in the brain.



David S. Touretzky is a Senior Research Scientist in the Computer Science Department and the Center for the Neural Basis of Cognition at Carnegie Mellon University. He received his Ph.D. in Computer Science from Carnegie Mellon. Dr. Touretzky's research interests include representations of space in the rodent brain and computational models of animal learning.