Prospective and Retrospective Temporal Difference Learning

Peter Dayan
Gatsby Computational Neuroscience Unit, UCL
dayan@gatsby.ucl.ac.uk

Abstract

A striking recent finding is that monkeys behave maladaptively in a class of tasks in which they know that reward is going to be systematically delayed. This may be explained by a malign Pavlovian influence arising from states with low predicted values. However, by very carefully analyzing behavioural data from such tasks, La Camera & Richmond (*PLoS Computational Biology*, doi:10.1371/journal.pcbi.1000131) observed the additional important characteristic that subjects perform differently on states in the task that are equal distances from the *future* reward, depending on what has happened in the recent *past*. The authors pointed out that this violates the definition of state value in the standard reinforcement learning models that are ubiquitous as accounts of operant and classical conditioned behavior; they suggested and analyzed an alternative temporal difference model in which past and future are melded. Here, we show that, in fact, a standard temporal difference model can actually exhibit the same behavior, and that this avoids deleterious consequences for choice. At the heart of the model is the average reward per step, which acts as a baseline for measuring immediate rewards. Relatively subtle changes to this baseline occasioned by the past can markedly influence predictions and thus behavior.

Author Summary

When monkeys perform a sequence of identical tasks before getting a reward, they have been found to make many errors when they know the reward is far away. Oddly, these errors do not only depend on the number of trials to the future reward, they also depend, retrospectively, on the length of the sequence. A recent suggestion for modeling this result suggests an heuristic modification to an otherwise normative account. Here, we show an alternative way that the retrospective dependence could have arisen in a model based on estimating the average reward per step.

Introduction

Richmond and his colleagues (Liu and Richmond, 2000; Liu et al., 2000; Shidara and Richmond, 2002; Shidara et al., 1998; Sugase-Miyamoto and Richmond, 2005; Bowman et al., 1996; Liu et al., 2004; Ravel and Richmond, 2006; La Camera and Richmond, 2008) have investigated many behavioral and neural aspects of an appealingly simple problem for monkeys that they call the reward schedule task. The task is illustrated in figure 1A. Ostensively, it involves repeated performance of a simple vigilance task (releasing a touch-bar between 200-800ms after a visually presented dot turns from red to green). However, the monkey is not rewarded after each successful performance; rather reward comes at the end of schedules, *ie* after performing certain numbers of successful trials (up to four). The key finding is that if the monkey's position in the schedule is signalled uniquely (the 'Valid' condition) by a visual cue (the brightness of the elongated bar in

figure 1A), then it is found to perform poorly on trials it knows to be far from the reward, even though it has to repeat these trials until it acts correctly. If it executes the same sequence of tasks but without any reliable cue (the 'Random' condition), then it performs uniformly well.

This graded performance poses two important challenges to standard views of adaptive decision-making. The first is that performance should be systematically poor on some trials. This is surprising, since the monkey cannot advance along the schedule until it gets a task instance correct, and so bad performance merely delays the ultimate reward. One explanation for this is that it is an example of the maladaptive interference of Pavlovian predictions (in this case of the excess distance of the future reward) over instrumentally optimal behavior (Breland and Breland, 1961; Williams and Williams, 1969; Hershberger, 1986). Pavlovian predictions lead automatically to responses that are presumably programmed by evolution to be beneficial in natural environments. These result in animals approaching and engaging with stimuli predictive of more reward or of increasing reward, and withdrawing and disengagement from stimuli predictive of more punishment or of decreasing reward. Although the exact rules of this interference are not completely clear, there are many examples in the behavioral literature, and it has also been the subject of modeling (Dayan et al., 2006), relating it to various anomalies of neuroeconomic decision-making (Dayan and Seymour, 2008). Here, the trials that are farthest from the reward are comparatively worse than trials that are nearest, and so might inspire withdrawal or disengagement, thus harming performance.

The second challenge, which was described and analyzed by La Camera and Richmond (2008), is both more subtle and more pernicious. Consider the schedules involving three and four task instances. The second trial of the three-instance case (which is called `2/3`) is formally similar to the third trial of the four instance case (`3/4`), since in both cases, successful performance of two more tasks is necessary to get reward. All standard reinforcement learning methods (Sutton and Barto, 1998), and indeed more general approaches to optimal control (Puterman, 2005) are prospective, focusing on predicting and controlling *future* rewards. Apart, therefore, from any systematic biases arising from generalization along the dimension of the cue to the schedule state, they would therefore naively expect these tasks to be performed similarly irrespective of their different pasts along their different schedules, because they share a common future. Importantly, this is true of Pavlovian predictions just as much as quantities associated with operant actions. La Camera and Richmond (2008) analyzed the behavior of the monkeys, and showed that this principle (which they called the principle of invariance) is violated. Figure 1B shows the key data from one monkey in terms of error rates. Although there is substantial variability in the data, on average, the monkey performs much more accurately (and also more quickly) on the `3/4` trials than the `2/3` trials. The same is true for the other would-be matched pairs, except for the trials nearest the reward (`1/1, 2/2, 3/3, 4/4`).

Standard temporal difference (TD) learning schemes (Sutton, 1988) in reinforcement learning have been recently omnipresent as formal accounts of animal and human decision-making (e.g., (Daw and Doya, 2006; Johnson et al., 2007; Doya, 2008)). Part of the ubiquity arises from the fact that these schemes realize optimal or approximately optimal control control in a quite wide range of circumstances, and therefore provide a firm statistical, engineering and computational foundation for substantial bodies of psychological and neural data (Sutton and Barto, 1998; Bertsekas and Tsitsiklis, 1996). However, La Camera and Richmond (2008) suggested that this retrospective regard for the past rules them out, and advocated a heuristically-motivated alternative to standard TD learning, which we call retrospective-TD, in which values of states in the *past* as well as those in

the *future* are used to criticize *present* value predictions.

La Camera and Richmond (2008) showed that their new scheme does not greatly disturb some cases of decision-making. However, there is a legacy from earlier attempts in reinforcement learning to include information from the past as well as the future in defining learning rules (Baird, 1995) suggesting that this can be problematical in terms of leading to suboptimal choice. Indeed, we exhibit just such an example in the next section. Thus, retrospective-TD does have qualitatively different, and quantitatively worse behavior, than regular TD, and does not enjoy even the approximate general optimality guarantees of conventional TD.

However, there are at least two routes by which information from the past of a schedule can affect present performance in regular, unaltered TD. One is that the Pavlovian misbehavior might include an immediate effect of the prediction error associated with the change from one state to the next, rather than only an effect via the learned values of states. That is, the TD prediction error depends on the difference between the values of two successive states. This difference (perhaps, when it is positive, acting via its dopaminergic report to the striatum (Satoh et al., 2003; Montague et al., 1996; Schultz et al., 1997)) could itself influence the monkeys' engagement with the trial. The prediction error will be 0 for deterministic transitions, but not for stochastic ones. In particular, in the first trials of the longer schedules (`1/3, 1/4`), there will be substantial negative prediction errors, given that one of the shorter schedules (`1/1, 1/2`) could have been picked instead. This could lead to particularly poor performance, as seen in figure 1.

The other route to an influence of the past on the present concerns the fact that in long schedules, subjects directly experience the recession of the past reward, and so on-line predictions of the average rate of reward will be decreasing as the monkey progresses. In the average reward framework, this quantity acts as a form of a comparator or variable baseline for ongoing prediction errors. Thus, as it decreases in long schedules, the effective learning signal will *increase*, leading to larger values, and thus better performance. Adaptive baselines are often thought of just in terms of variance reduction (Williams, 1992; Dayan and Sejnowski, 1993; Greensmith et al., 2004) or the balance between exploration and exploitation (Aston-Jones and Cohen, 2005). However, our results show that they can also provide a powerful way for past events to influence future-oriented predictions.

We first describe the reward schedule task more formally, along with La Camera and Richmond (2008)'s suggested solution. We then show a simple Markov chain in which the effect of the past in their algorithm leads to evidently suboptimal behavior. Next, we consider the two alternatives associated with conventional TD. One of these effects, namely the diminishing average rate of reward, appears to be more critical in accounting for the behavior. Finally we elucidate the consequences of this finding.

Results

The Reward Schedule Task and Retrospective-TD

Figure 1A-C shows the reward schedule task and the basic behavioral data in question. The overall data come from a large number of different individual experiments performed on 24 different monkeys, and a wide range of different stimuli reporting the state of the schedule. Following La Camera and Richmond (2008), we will work with aggregate results, which are representative.

The schedules in figures 1B;C are labelled as `trial/schedule`, so `1/1` is the sole trial in a schedule involving just one single vigilance task, `1/2` is the first trial of a two schedule task, and so forth. As mentioned, monkeys only progress along the schedule, from `1/1` to get the reward, or from `1/2` to `2/2` if they perform the vigilance task correctly, otherwise they just have to repeat the trial until they do get it right. Figure 1B shows the mean error rates for each trial in each schedule for one monkey. In fact, few monkeys are willing to perform a schedule of length four at all. Figure 1C shows the performance of the twelve monkeys on schedules of up to length three. These monkeys were selected because they all had significant differences between the error rates on the `1/2` and `2/3` trials. The substantial inter-subject variability is readily apparent.

The misbehavior of the subjects must somehow be occasioned (though not necessarily directly) by values $v(\mathtt{s})$ associated with the current state $\mathtt{s} = \{\tau/s\}$ (i.e., trial τ on schedule s). According to RL, these values are the long run expected utilities starting from each state. We consider two different definitions of long run utility. One is the discounted value, defined self-consistently as

$$v^{\gamma}(\mathtt{s}_t) = \langle r(\mathtt{s}_t) + \gamma v^{\gamma}(\mathtt{s}_{t+1}) \rangle \ . \tag{1}$$

Here, $r(\mathtt{s}_t)$ is the immediate reward associated with state $\mathtt{s}_t$ at time t. This reward is 0 except for being 1 on the state after the successful completion of all trials in a schedule, which we call `0/0`. Further, $\mathtt{s}_{t+1}$ is the next state. This is the same as $\mathtt{s}_t$ given unsuccessful completion of a trial, and $\tau + 1/s$ following successful completion or `0/0` at the end of a schedule. In the case that $\mathtt{s}_t$=`0/0`, we make $\mathtt{s}_{t+1}$ another unrewarded special state called `-/-`. Finally, we assume that $\mathtt{s}_{t+1}$ is equiprobably `1/1, 1/2, 1/3` or `1/4` following $\mathtt{s}_t$ =`-/-`. The angle brackets indicate that these values are averaged over the expected performance (i.e., including the chance of error), and $0 \leq \gamma < 1$ is the discount factor which downweights rewards in the distant future. The solid lines in figure 2A show the discounted values for all the schedule states in an analogous form to figure 1B for $\gamma = 0.4$, and for a simulated case in which there is no error.

A perhaps more natural way to define the long run utility in the task is in the form of ρ, which is the average reward per time step (Schwartz, 1993; Mahadevan, 1996; Puterman, 2005), and was used, for instance, by Niv et al. (2007) to study operant vigor. In this case, the state values $v^a(\mathtt{s}_t)$ are called differential values, and satisfy

$$v^a(\mathtt{s}_t) = \langle r(\mathtt{s}_t) + v^a(\mathtt{s}_{t+1}) \rangle - \rho \tag{2}$$

It is apparent that adding any constant to $v^a(\mathtt{s})$ for all $\mathtt{s}$ will leave this relationship fixed; thus it is conventional to set the differential value of one state to 0. We set $v^a($`-/-`$) = 0$. Figure 2B shows these average values in the same format as figure 2A (also without errors), assuming that the true

value of ρ is used. Later we consider the consequences of learning ρ from experience. Average case TD endows each timepoint with the same weight, and so the solid lines connecting the points within each single schedule are straight rather than curved.

Ignoring for the moment the errors the subjects make, figure 2A;B also shows the essential problem for RL that was identified by La Camera and Richmond (2008). Equations 1 and 2 are exclusively *forward*-looking. Thus the discounted or average values of states such as `2/3` and `3/4` are bound to be the same, and thus might be expected to lead to the same performance. The dashed lines in figures 2A;B connect these equivalent points in the schedules. That they are flat implies that performance based solely on $v^\gamma(\mathtt{s})$ or $v^a(\mathtt{s})$ would not vary by schedule. What is needed is some dependence on the *past* of the individual schedule.

The standard TD learning rule associated with the discounted definition in equation 1 is

$$\delta(\mathtt{s}_t) = r(\mathtt{s}_t) + \gamma v^\gamma(\mathtt{s}_{t+1}) - v^\gamma(\mathtt{s}_t)\ . \tag{3}$$

La Camera and Richmond (2008) suggested modifying it to give the retrospective-TD rule

$$\delta^r(\mathtt{s}_t) = r(\mathtt{s}_t) + \gamma v^\gamma(\mathtt{s}_{t+1}) + \sigma v^\gamma(\mathtt{s}_{t-1}) - v^\gamma(\mathtt{s}_t)\ . \tag{4}$$

where $\mathtt{s}_{t-1}$ is the state that *precedes* state $\mathtt{s}_t$. Figure 2C shows the consequence of doing this for the schedule case (for convenience, (again following La Camera and Richmond (2008), setting $\sigma = 0$ for `-/-` to reset the trace and not accumulating beyond each reward). By making $\sigma = 0.3 > 0$, the values of the states can depend on those of their predecessors, and so generate the sort of schedule dependence that is apparent in the data, with the decreasing dashed lines connecting prospectively, but not retrospectively, equivalent trials. Making the additional assumption that the value of a state is turned into the probability of an error at that state by a sigmoid relationship $P[\text{error}; \mathtt{s}] = 1/(1 + \exp(\beta v(\mathtt{s})))$, where β is an inverse temperature parameter, figure 2D shows the implied error rates with their retrospective dependence. Iterating this to find error rates and values that are mutually consistent (since the values depend on averages that include the effects of the errors) leads to a good match with the experimental data.

La Camera and Richmond (2008) show directly that this retrospective-TD rule behaves just like regular TD in various circumstances. They also note that if retrospective-TD values are used to make choices between states (treating them like $\mathcal{Q}$-values (Watkins, 1989)), then there are some circumstances under which the behavior produced can be different from that implied by regular TD. However, in those problems (e.g., a reward schedule task in which an initial choice leads to one of two sequences of short or long delays to large or small rewards), regular TD would make the same choice as retrospective-TD, but for a different discount factor. La Camera and Richmond (2008) therefore conclude that "in simple choice tasks and in choice-schedule tasks, the context-sensitive model [retrospective-TD] predicts the same qualitative behavior as the standard TD model."

Unfortunately, this result is not generic. The consequence of corrupting the value of a state according to its past on paths, means that the state can look unreasonably attractive. Figure 3 shows an example of a Markov decision problem with exactly this characteristic. State 'c' is very valuable, and so, for a non-zero σ, dramatically boosts the value of its successor (state 'e'). This can make state 'e' appear better than state 'd', despite itself leading to a punishment rather than a reward.

The consequence is that subjects would make an error at their sole choice point, i.e., state 'b', between heading for state 'd' and state 'e'. The table in figure 3 shows this directly, comparing the values of discounted and retrospective-TD for the various states. No manipulation of the discounting for regular TD would have this effect. The amount by which $v(\text{'e'}) > v(\text{'d'})$, and thus the incorrect preference, scales with the reward at 'c'.

TD and the Past

The behavioral data from the subjects in the reward schedule tasks show that aspects of the past can play a role in determining choices, along with aspects of the future. However, before adopting retrospective-TD, along with its potentially suboptimal policies, it is important to consider how the past influences standard TD. We consider two factors. One is based on the observation that the immediate value of the temporal difference error signal itself might play a role in corrupting instrumental behavior. This explanation does not depend at all on the *values* of the equivalent schedule points being different – it is a 'performance' rather than a 'learning' effect. The second possibility does depend on learning, and is based on the nature and role of the reward rate ρ in temporal difference learning.

First, as we have discussed, there is not a normative reason for the poor performance of the monkeys on the distant trials of reward schedule tasks, and so one has to make a linking assumption as to the neural/psychological process underlying it. La Camera and Richmond (2008) suggested that it arises purely from the value. However, it is conceivable that the TD prediction error itself $\delta(\mathsf{s}_{t-1})$ in equation 3 might also be involved. That is, if the subjects go from a state s_{t-1} that is better, to one s_t that is worse, then, by creating a negative prediction error, this could induce poor performance, even if the actual value $v^\gamma(\mathsf{s}_t)$ or $v^a(\mathsf{s}_t)$ is by itself quite high. Note that the prediction error is 0 for deterministic transitions, but not necessarily for stochastic transitions such as that associated with the choice of schedule.

One (albeit far from conclusive) reason to think that this might have an influence is that trial-by-trial fluctuations in the activity of dopamine neurons, which putatively code for appetitive prediction errors (i.e., when states are getting *better*) are positively related to the vigor of responding (Satoh et al., 2003; Murschall and Hauber, 2006; Lex and Hauber, 2008). Vigorous responding is an obvious relative of strong engagement in a task, and thus, putatively low error rates. For instance, just such an effect has been observed in a human monetary incentive task, in which a cue indicating that an upcoming trial will actually be rewarded on successful completion leads to faster and more accurate choices, and also leads to a stronger BOLD signal in areas that are targets of dopaminergic processing (Wittmann et al., 2005) (although sufficiently excess vigor can, of course, itself lead to errors (Cools et al., 2005)) Something similar is true in a primate task involving mandatory, unrewarded, actions (Kawagoe et al., 2004).

Niv et al. (2007) provides a discussion of *operant* control over vigor and tonic levels of dopamine (Salamone and Correa, 2002), and notes the possibility that the tonic signal is realized partly (though not wholly (Goto and Grace, 2005)) through accumulated phasic signalling, having the same energising effect (Watanabe et al., 2001; Lauwereyns et al., 2002; Roitman et al., 2004). Whether pauses in dopamine firing (Bayer and Glimcher, 2005) or an aversive opponent system to dopamine (Solomon and Corbit, 1974; Grossberg, 1984; Daw et al., 2002) represent negative prediction errors

is not completely clear, but either could exert Pavlovian influences over instrumental vigor and choices via the nucleus accumbens (Reynolds and Berridge, 2002, 2008; Talmi et al., 2008; Balleine, 2005). Ravel and Richmond (2006) recorded the activity of dopamine neurons in the reward schedule task, showing a range of interesting responses, albeit ones that are hard to relate to the current analysis because of the possible existence of an opponent system.

An effect of this sort of the TD prediction error signal could play an important role in the reward schedule task. For instance, in the error free case, consider the prediction error associated with the transition from `-/-` to `1/3` versus the transition from `1/4` to `2/4` for the discounted case values in figure 2A. Since a shorter schedule could have been chosen, the transition to a relatively longer schedule (involving three trials) is a small negative step from $v^\gamma(\texttt{-/-}) = 0.067$ (an outcome that can be further enhanced when errors are also taken into account). This makes the TD prediction error affecting `1/3` ($\delta = -0.04$; for the numbers in the figure). By comparison, since it involves a deterministic transition, the TD prediction error affecting `2/4` is $\delta = 0$, since the effect of the unfortunate choice of a length four schedule has already been discounted through the low value $v^\gamma(\texttt{1/4}) = 0.026$. Thus performance of `1/3` could be comparatively impaired.

However, although this is a proof of principle for an effect of the past on regular TD signals, it seems unlikely to be the sole factor governing the behavioral findings in figure 1B. Even in simulations taking account of errors (not shown), the fact that `2/3` has a higher error rate than `3/4` turns out to be particularly difficult to accommodate, because the prior states (`1/3` and `2/4` respectively) are both already part of their schedules (rather than being external to the schedule as in `-/-`). Note, though, that few monkeys perform the schedules of length 4 at all, and indeed one of the three reported did not show a significant retrospective effect.

Another way that the past may influence the present arises through the ρ term in the average case TD rule of equation 2. In figure 2B, the exact value of ρ was used given knowledge of the (error-free) policy. However, in practice, this has to be learned, along with the average TD values. One common way to implement this is to define two learning processes (Tsitsiklis and Van Roy, 2002). One is a standard delta or Rescorla-Wagner rule producing a running estimate of ρ_t

$$\rho_{t+1} = \rho_t + \eta^\rho(r(\mathtt{s}_t) - \rho_t) \tag{5}$$

where η^ρ is the learning rate associated with the running average. The other learning process implements TD learning, but with the TD prediction error being

$$\delta(\mathtt{s}_t) = r(\mathtt{s}_t) - \rho_t + v^a(\mathtt{s}_{t+1}) - v^a(\mathtt{s}_t) \tag{6}$$

In this formulation, ρ_t acts as a form of comparator or baseline for the reward $r(\mathtt{s}_t)$, and so small values of ρ_t lead to larger differential values, and vice-versa. One of the differences between prospectively equivalent states such as `2/3` and `3/4` is that, retrospectively, the latter is farther from the reward in the reward schedule task (an outcome that is exacerbated by the poor performance of the monkey at the previous points in the schedules). Thus, following equation 5, the value of ρ_t will be lower by this point, and so, over learning, the values will come to satisfy $v^a(\texttt{3/4}) > v^a(\texttt{2/3})$. This will make performance at `3/4` better than at `2/3`. The same applies for the other context-dependent differences. This did not arise in figure 2B, since the values there were based on the true, infinite-horizon, value of ρ that was not estimated.

Figure 4A shows this effect in the case that there is no error in performance, in terms of the values of ρ at each point of each trial (for $\eta^\rho = 0.1$). Figure 4B shows that the *decreasing* values of ρ as the schedules progress duly turn into *increasing* values of $v^a(\mathtt{s})$ as predicted from this analysis, with the dashed lines joining equivalent trials no longer being flat.

In considering the effect of the values on behavior, the average reward case requires one further assumption, since the absolute level of the values is undetermined. This led us to set one state (`-/-`) to have the value 0; however it also means that there is a potentially arbitrary mapping to the error rate. This particularly affects $v^a(\mathtt{1/4})$, which becomes negative. Given that the minimum observed error rate is 40%, and that there is substantial variability among the subjects, we simply used the crude $P[\text{error}|\mathtt{s}] = \min\{1/(1+\exp(\beta v^a(\mathtt{s})), 0.4\}$ rather than attempting to match the averages in detail. Using this, figure 4C shows the final, self-consistent, error rates, together with the empirical values from figure 1B (squares) for comparison. The qualitative match is readily apparent – given the variability between different animals, we did not attempt to fit the data precisely. Indeed, the qualitative results are not strongly dependent on the parameters – although the smaller the learning rate η^ρ governing ρ, the smaller this effect.

Finally, one might wonder about the behavior of this rule in the instrumental choice task of figure 3. In this particular problem, the effect of making η^ρ very high is actually to depress the value of state 'e', from the times that it is preceded by state 'c' with its large reward. Thus the choice at state 'b' will actually tend to be more correct. In other decision problems, choice can be skewed. However, if η^ρ is decreased to make ρ a longer-run average, then these problems should not affect the optimal solution. To put it another way, the key difference between this explanation and that of retrospective-TD is that the underlying definition of the values is exactly as it should be under properly prospective TD models, the rule for learning ρ concerns the modality of achieving the optimal values. By comparison, retrospective-TD alters the representational *goal* of the values, away from a normatively reasonable basis. Similarly, this form of average TD can cope with the equivalence of schedule states in the random cue condition of the reward schedule task.

Discussion

In this paper, we have considered the interesting challenge to TD that comes from the retrospectivity apparent in the reward schedule task. Although it is possible to address the challenge through the heuristic modification to TD (retrospective-TD) suggested by La Camera and Richmond (2008), this is not normatively based, and can lead to suboptimal behavior, weakening one of the main original strengths of TD.

We therefore showed two ways that retrospective factors can influence the otherwise prospective TD rule. One involves detailed consideration of the provenance of the malign Pavlovian influence over instrumental behavior, and the suggestion that the TD prediction error itself might affect the monkey's engagement with the task, and hence error rate. The other concerns the effect that the long-run average reward has on learning TD values in an average-case RL setting, acting as a sort of baseline. As this long-run average changes through learning, it acts to inject retrospective information about the time since the reward, and hence the length of the current schedule, into the values. It thereby has the same effect as retrospective-TD in this task. However, since it is nothing

more than an instantiation of regular TD, it is benign with respect to policy choices.

This magnitude and nature of this effect is controlled by the learning rate η^ρ for the long run average reward. That subjects might adjust this learning rate in the light of the task is suggested by various experiments in monkey and human decision-making (Lau and Glimcher, 2005; Corrado et al., 2005; Behrens et al., 2007) and may offer a route for testing our account. Along with the well-known effect that in choice tasks there is a sampling bias that leads to variance aversion (Niv et al., 2002; March, 1996; Hertwig et al., 2004; Weber et al., 2004), this influence of past events on future predictions is a reminder of the complexities of on-line learning.

Although average-case TD is a little less common than discounted TD for application to computational problems and neural data, it is actually more reasonable for ongoing tasks that do not have either an obvious end point or a natural timescale. The idea that monkeys are trying to maximize this average rate underlies substantial work on temporal decision-making in monkeys and humans (Gold and Shadlen, 2007; Ratcliff and Smith, 2004), and there is evidence that a population of anterior cingulate neurons in the macaque represents a form of online estimate of the quantity in a complex reinforcement learning task (Seo and Lee, 2007). Further, average-case and discounted RL are closely related for large values of the discount factor γ. Indeed the pair of rules 5 and 6 have been shown to arise from a particular form of regular TD, in which an appropriate representation of the context exactly substitutes for the effect of ρ (Tsitsiklis and Van Roy, 2002).

Of course, even with its alternative way of addressing the retrospectivity, our account retains exactly La Camera and Richmond (2008)'s explanation for the essential maladaptivity of the high error rates themselves. The execution of apparently instrumentally irrelevant actions also arose in Niv et al. (2007)'s study of operant vigor. Animals do perform actions that are clearly incorrect with respect to the experimenter's definition of the task, such as checking food cups when they have not heard the food drop, visiting a water spout when hungry rather than thirsty, or grooming. Niv et al. (2007) suggested that this arises from a stochastic policy, including small values for these choices. In fact, the sigmoid policy adopted by La Camera and Richmond (2008) is equivalent to this, assuming that the Q-value of performing incorrectly is 0, and that of performing correctly is $v(\mathtt{s})$. The latter is unusual, since the Q-value should be more like the value of the next state (plus any reward along the way); however, as La Camera and Richmond (2008) point out, this would lead to a policy in the random cue case that does not match the subjects' behavior. Unfortunately, our understanding of the details of Pavlovian influences over operant actions is not sufficiently advanced to distinguish such instrumental and classical routes to error. Further, if anything, the essential effect in Niv et al. (2007)'s operant model of the fact that ρ decreases over long schedules would be to decrease, rather than increase, operant vigor, again reinforcing the Pavlovian/operant difference.

The involvement of dopamine in both the reward schedule task (Liu et al., 2004; Ravel and Richmond, 2006) and operant vigor (Niv et al., 2007; Satoh et al., 2003; Salamone and Correa, 2002; Salamone et al., 2007; Lex and Hauber, 2008; Murschall and Hauber, 2006) is suggestive. However, Niv et al. (2007)'s account involves tonic rather than phasic dopamine, and this would presumably not distinguish different trials in a schedule. However, although there is clearly some separation between tonic and phasic dopamine signals (Goto and Grace, 2005), the latter may contribute to the former, perhaps underlying the correlation between phasic activity and response vigor noted by Satoh et al. (2003). Unfortunately, the whole collection of threads associated with tonic and pha-

sic dopamine, Pavlovian influences over instrumental behavior such as Pavlovian-instrumental transfer, and the nucleus accumbens has yet to be satisfactorily tied.

One tonic effect that might play a relevant role arises from the design of presenting blocks of valid and invalid cues. The overall reward rate is substantially higher in the latter blocks than the former because of the excess errors induced by the valid cues. This could account for a remaining issue for the average case TD rule concerning the absolute error rate for the invalid cues. With invalid cues, there is effectively just one single value for all the states in the schedules, since the randomization renders them effectively indistinguishable. The average TD rule thus correctly leads to almost equal error rates for all states in the schedules (since the retrospective effect of the changing estimates of the average reward rate, ρ_t, depends mostly on learning). However, in the error free case, one can show that the differential value of the apparently single invalid state is approximately the average reward rate ρ. By comparison, the differential values of the last states of each schedule are approximately $1 - 2\rho$. These will only be equal (thus leading to equal error rates, as very approximately observed in the data) for $\rho = 1/3$, which is an overestimate. However, if the increased overall reward rate in the invalid blocks leads to a greater overall engagement with the task, via tonic dopamine signalling, then this might lead the error rates to match. To put the effect another way, the valid cue blocks may induce a form of learned helplessness (Seligman and Maier, 1967; Maier et al., 2006), with the subjects being unable to eliminate the delay to the reward arising from the intervening tasks.

There are at least two experimental approaches which could readily be used to compare retrospective-TD with the suggestions here. First, it would be important to test the role of performance versus learning. One way to do this would be to 'break' the schedules very occasionally and at random, changing which state comes before which one. This should induce ongoing prediction errors without substantially affecting the values. If the error rate at a state depends on the precise previous transition leading there, then this would vote in favor of performance considerations over learning. The current data actually includes trials of this sort (one could compare the first entry to a state versus the repeat, following an error); however, this is a rather special transition involving the same state twice, and furthermore has been observed an overwhelming number of times. Thus its results might not be conclusive. Successive trials in the invalid blocks could also be revealing.

Second, it would be interesting to try tasks of the sort suggested in figure 3. If the monkey's behavior at state 'b' indeed reveals a preference for state 'e' over state 'd', because of the *prior* reward in state 'c', then this would be a strong vote for retrospective-TD, particularly because of its non-normativity. Of course, this particular example involves relatively extreme differences in the immediate rewards at different states (a requirement imposed by the relatively small value of σ). It would be important to create a range of tasks with similar properties.

Conclusions

We have studied the apparently anomalous, retrospectively sensitive, performance of monkeys in a task involving substantial, signalled, delays to rewards. We considered how information about the past infects standard temporal difference learning methods of reinforcement learning. One of the effects of this is to manipulate the future-oriented predictions in a manner that obviates the

requirement for a heuristic, non-normative, learning rule. Finally, we suggested some directions for experimental test of these hypotheses.

Acknowledgements

I am very grateful to Nathaniel Daw, Giancarlo La Camera, Alex Lerchner and Barry Richmond for very helpful discussions, and to them and two anonymous reviewers for their very careful reading of and suggestions on a previous version of this paper. I am also most grateful to Tim Behrens for his support.

References

Aston-Jones, G. and Cohen, J. D. (2005). Adaptive gain and the role of the locus coeruleus-norepinephrine system in optimal performance. *J Comp Neurol*, 493(1):99–110.

Baird, L. (1995). Residual Algorithms: Reinforcement Learning with Function Approximation. In Prieditis, A. and Russell, S., editors, *International Conference on Machine Learning*, pages 30–37, San Francisco, CA. Morgan Kaufmann.

Balleine, B. W. (2005). Neural bases of food-seeking: affect, arousal and reward in corticostriatolimbic circuits. *Physiol Behav*, 86(5):717–730.

Bayer, H. M. and Glimcher, P. W. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron*, 47(1):129–141.

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., and Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nat Neurosci*, 10(9):1214–1221.

Bertsekas, D. P. and Tsitsiklis, J. N. (1996). *Neuro-Dynamic Programming (Optimization and Neural Computation Series, 3)*. Athena Scientific.

Bowman, E. M., Aigner, T. G., and Richmond, B. J. (1996). Neural signals in the monkey ventral striatum related to motivation for juice and cocaine rewards. *J Neurophysiol*, 75(3):1061–1073.

Breland, K. and Breland, M. (1961). The misbehavior of organisms. *American Psychologist*, 16(9):681–84.

Cools, R., Blackwell, A., Clark, L., Menzies, L., Cox, S., and Robbins, T. W. (2005). Tryptophan depletion disrupts the motivational guidance of goal-directed behavior as a function of trait impulsivity. *Neuropsychopharmacology*, 30(7):1362–1373.

Corrado, G. S., Sugrue, L. P., Seung, H. S., and Newsome, W. T. (2005). Linear-nonlinear-poisson models of primate choice dynamics. *J Exp Anal Behav*, 84(3):581–617.

Daw, N. D. and Doya, K. (2006). The computational neurobiology of learning and reward. *Curr Opin Neurobiol*, 16(2):199–204.

Daw, N. D., Kakade, S., and Dayan, P. (2002). Opponent interactions between serotonin and dopamine. *Neural Netw*, 15(4-6):603–616.

Dayan, P., Niv, Y., Seymour, B., and Daw, N. D. (2006). The misbehavior of value and the discipline of the will. *Neural Netw*, 19(8):1153–1160.

Dayan, P. and Sejnowski, T. (1993). The Variance of Covariance Rules for Associative Matrix Memories and Reinforcement Learning. *Neural Computation*, 5(2):205–209.

Dayan, P. and Seymour, B. (2008). Values and actions in aversion. In Glimcher, P., Camerer, C., Poldrack, R., and Fehr, E., editors, *Neuroeconomics: Decision making and the brain*, pages 175–191. Academic Press, New York, NY.

Doya, K. (2008). Modulators of decision making. *Nat Neurosci*, 11(4):410–416.

Gold, J. I. and Shadlen, M. N. (2007). The neural basis of decision making. *Annu Rev Neurosci*, 30:535–574.

Goto, Y. and Grace, A. A. (2005). Dopaminergic modulation of limbic and cortical drive of nucleus accumbens in goal-directed behavior. *Nat Neurosci*, 8(6):805–812.

Greensmith, E., Bartlett, P., and Baxter, J. (2004). Variance Reduction Techniques for Gradient Estimates in Reinforcement Learning. *The Journal of Machine Learning Research*, 5:1471–1530.

Grossberg, S. (1984). Some normal and abnormal behavioral syndromes due to transmitter gating of opponent processes. *Biol Psychiatry*, 19(7):1075–1118.

Hershberger, W. (1986). An approach through the looking-glass. *Animal Learning & Behavior*, 14:443–451.

Hertwig, R., Barron, G., Weber, E. U., and Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, 15(8):534–539.

Johnson, A., van der Meer, M. A. A., and Redish, A. D. (2007). Integrating hippocampus and striatum in decision-making. *Curr Opin Neurobiol*, 17(6):692–697.

Kawagoe, R., Takikawa, Y., and Hikosaka, O. (2004). Reward-predicting activity of dopamine and caudate neurons–a possible mechanism of motivational control of saccadic eye movement. *J Neurophysiol*, 91(2):1013–1024.

La Camera, G. and Richmond, B. J. (2008). Modeling the violation of reward maximization and invariance in reinforcement schedules. *PLoS Comput Biol*, 4(8):e1000131.

Lau, B. and Glimcher, P. W. (2005). Dynamic response-by-response models of matching behavior in rhesus monkeys. *Journal of the Experimental Analysis of Behavior*, 84(3):555–579.

Lauwereyns, J., Watanabe, K., Coe, B., and Hikosaka, O. (2002). A neural correlate of response bias in monkey caudate nucleus. *Nature*, 418(6896):413–417.

Lex, A. and Hauber, W. (2008). Dopamine d1 and d2 receptors in the nucleus accumbens core and shell mediate pavlovian-instrumental transfer. *Learn Mem*, 15(7):483–491.

Liu, Z., Murray, E. A., and Richmond, B. J. (2000). Learning motivational significance of visual cues for reward schedules requires rhinal cortex. *Nat Neurosci*, 3(12):1307–1315.

Liu, Z. and Richmond, B. J. (2000). Response differences in monkey te and perirhinal cortex: stimulus association related to reward schedules. *J Neurophysiol*, 83(3):1677–1692.

Liu, Z., Richmond, B. J., Murray, E. A., Saunders, R. C., Steenrod, S., Stubblefield, B. K., Montague, D. M., and Ginns, E. I. (2004). Dna targeting of rhinal cortex d2 receptor protein reversibly blocks learning of cues that predict reward. *Proc Natl Acad Sci U S A*, 101(33):12336–12341.

Mahadevan, S. (1996). Average reward reinforcement learning: Foundations, algorithms and empirical results. *Machine Learning*, 22:1–38.

Maier, S. F., Amat, J., Baratta, M. V., Paul, E., and Watkins, L. R. (2006). Behavioral control, the medial prefrontal cortex, and resilience. *Dialogues Clin Neurosci*, 8(4):397–406.

March, J. G. (1996). Learning to be risk averse. *Psychological Review*, 103(2):309–319.

Montague, P. R., Dayan, P., and Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive hebbian learning. *J Neurosci*, 16(5):1936–1947.

Murschall, A. and Hauber, W. (2006). Inactivation of the ventral tegmental area abolished the general excitatory influence of Pavlovian cues on instrumental performance. *Learning & Memory*, 13:123–126.

Niv, Y., Daw, N. D., Joel, D., and Dayan, P. (2007). Tonic dopamine: opportunity costs and the control of response vigor. *Psychopharmacology (Berl)*, 191(3):507–520.

Niv, Y., Joel, D., Meilijson, I., and Ruppin, E. (2002). Evolution of reinforcement learning in uncertain environments: A simple explanation for complex foraging behaviors. *Adaptive Behavior*, 10(1):5–24.

Puterman, M. L. (2005). *Markov Decision Processes: Discrete Stochastic Dynamic Programming (Wiley Series in Probability and Statistics)*. Wiley-Interscience.

Ratcliff, R. and Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychol Rev*, 111(2):333–367.

Ravel, S. and Richmond, B. J. (2006). Dopamine neuronal responses in monkeys performing visually cued reward schedules. *Eur J Neurosci*, 24(1):277–290.

Reynolds, S. M. and Berridge, K. C. (2002). Positive and negative motivation in nucleus accumbens shell: bivalent rostrocaudal gradients for gaba-elicited eating, taste "liking"/"disliking" reactions, place preference/avoidance, and fear. *J Neurosci*, 22(16):7308–7320.

Reynolds, S. M. and Berridge, K. C. (2008). Emotional environments retune the valence of appetitive versus fearful functions in nucleus accumbens. *Nat Neurosci*, 11(4):423–425.

Roitman, M. F., Stuber, G. D., Phillips, P. E. M., Wightman, R. M., and Carelli, R. M. (2004). Dopamine operates as a subsecond modulator of food seeking. *Journal of Neuroscience*, 24(6):1265–1271.

Salamone, J. D. and Correa, M. (2002). Motivational views of reinforcement: implications for understanding the behavioral functions of nucleus accumbens dopamine. *Behav Brain Res*, 137(1-2):3–25.

Salamone, J. D., Correa, M., Farrar, A., and Mingote, S. M. (2007). Effort-related functions of nucleus accumbens dopamine and associated forebrain circuits. *Psychopharmacology (Berl)*, 191(3):461–482.

Satoh, T., Nakai, S., Sato, T., and Kimura, M. (2003). Correlated coding of motivation and outcome of decision by dopamine neurons. *J Neurosci*, 23(30):9913–9923.

Schultz, W., Dayan, P., and Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, 275(5306):1593–1599.

Schwartz, A. (1993). Thinking locally to act globally: A novel approach to reinforcement learning. In *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society*, pages 906–911, Hillsdale, NJ. Lawrence Erlbaum Associates.

Seligman, M. E. and Maier, S. F. (1967). Failure to escape traumatic shock. *J Exp Psychol*, 74(1):1–9.

Seo, H. and Lee, D. (2007). Temporal filtering of reward signals in the dorsal anterior cingulate cortex during a mixed-strategy game. *J Neurosci*, 27(31):8366–8377.

Shidara, M., Aigner, T. G., and Richmond, B. J. (1998). Neuronal signals in the monkey ventral striatum related to progress through a predictable series of trials. *J Neurosci*, 18(7):2613–2625.

Shidara, M. and Richmond, B. J. (2002). Anterior cingulate: single neuronal signals related to degree of reward expectancy. *Science*, 296(5573):1709–1711.

Solomon, R. L. and Corbit, J. D. (1974). An opponent-process theory of motivation. i. temporal dynamics of affect. *Psychol Rev*, 81(2):119–145.

Sugase-Miyamoto, Y. and Richmond, B. J. (2005). Neuronal signals in the monkey basolateral amygdala during reward schedules. *J Neurosci*, 25(48):11071–11083.

Sutton, R. (1988). Learning to predict by the methods of temporal differences. *Machine Learning*, 3(1):9–44.

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement Learning: An Introduction (Adaptive Computation and Machine Learning)*. The MIT Press.

Talmi, D., Seymour, B., Dayan, P., and Dolan, R. J. (2008). Human pavlovian-instrumental transfer. *J Neurosci*, 28(2):360–368.

Tsitsiklis, J. and Van Roy, B. (2002). On average versus discounted reward temporal-difference learning. *Machine Learning*, 49(2):179–191.

Watanabe, M., Cromwell, H. C., Tremblay, L., Hollerman, J. R., Hikosaka, K., and Schultz, W. (2001). Behavioral reactions reflecting differential reward expectations in monkeys. *Experimental Brain Research*, 140(4):511–518.

Watkins, C. J. C. H. (1989). *Learning from delayed rewards*. PhD thesis, King's College, Cambridge University.

Weber, E. U., Shafir, S., and Blais, A.-R. (2004). Predicting risk sensitivity in humans and lower animals: risk as variance or coefficient of variation. *Psychol Rev*, 111(2):430–445.

Williams, D. R. and Williams, H. (1969). Auto-maintenance in the pigeon: sustained pecking despite contingent non-reinforcement. *J Exp Anal Behav*, 12(4):511–520.

Williams, R. (1992). Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning. *Reinforcement Learning*, 8:229–256.

Wittmann, B. C., Schott, B. H., Guderian, S., Frey, J. U., Heinze, H.-J., and Dzel, E. (2005). Reward-related fmri activation of dopaminergic midbrain is associated with enhanced hippocampus-dependent long-term memory formation. *Neuron*, 45(3):459–467.

Figure legends

Figure 1: The reward schedule task. A) The left figure shows the sequence of events in the basic vigilance task, in which a monkey is prepared by a red spot to react to a green spot in order to get reward (the drop). Any error implies the task has to be repeated. The right figure shows a schedule of length two, in which the vigilance task has to be completed twice to get the reward. Any error in the first task (called `1/2`) requires it to be repeated. The display includes a bar which, in this case, is black for the trial next to the reward, and gray for `1/2`. In the Random condition, the same schedules apply, but the colors of the bars are meaningless, and variable on a trial-by-trial basis. B) Error rate of one monkey (called monkey B in La Camera and Richmond (2008)) in the four schedules (valid cues: black dots; random cues: asterisks). Each schedule is shown in order, so the sequence along the x-axis is `1/1, 1/2, 2/2, 1/3, 2/3, ...`. The context dependence is that trials that are equal numbers of tasks from the reward (such as `1/2` and `2/3`) have different error rates. The asterisks show that this depends on the monkey having information as to where in the schedule it is. C) Data from 12 monkeys who performed schedules of length 3 with significant context dependence showing the substantial variability. Figure segments taken from La Camera and Richmond (2008)

Figure 2: TD and retrospective-TD. A) The value of all the states in the schedules assuming discounted TD with $\gamma = 0.4$ (here, the values could accumulate across trials; note the y-axis is flipped). Equivalent states in the schedules are joined by dashed lines, and have the same values. B) The same quantities for average case TD, setting $v^a(\texttt{-/-}) = 0$ and using the analytical value of ρ. C) The values under retrospective-TD with $\sigma = 0.3$ (for consistency with the simplest definition in La Camera and Richmond (2008), the reward is actually $1/\gamma$, and there is no accumulation across the reward state `0/0`). Here, the context-dependence is apparent in that the dashed lines are decreasing. D) The retrospective-TD values are turned into choices using $P[\text{error}; \texttt{s}] = 1/(1 + \exp(\beta v(\texttt{s}))$ with $\beta = 3.2$.

Figure 3: Retrospective-TD and choice. The figure shows a simple Markov decision problem with a single choice (at node 'b'). Nodes are labelled by their states; reward values are shown on their exteriors. The table shows the values for retrospective-TD (using $\gamma = 0.4; \sigma = 0.3$, setting $\sigma = 0$ at 'a') and standard discounted TD (with $\gamma = 0.4$). The optimal choice at 'b' is to go to 'd', but this appears to reverse under retrospective-TD, since the large reward at 'c' skews the value of 'e'. The values assume that the choice at 'b' depends deterministically on the sign of $v(\text{'d'}) - v(\text{'e'})$.

Figure 4: The retrospective effects of ρ. A) Using the learning rule of equation 5, the value of ρ *decreases* as the number of steps from the reward *increases*. B) The effect of this is to increase the values of the later trials in each schedule, breaking their prospective equivalence.(A) and (B) are both in a case without errors, and $\eta^\rho = 0.1$. C) If the values actually determine errors through a truncated sigmoid (with the maximum error rate of 40%), then (including the effect of the errors on $v^a(\texttt{s})$, with the learning rate for the values also being 0.1), the self-consistent error rates qualitatively resemble the data from figure 1B (replotted as squares). Here $\beta = 5$.

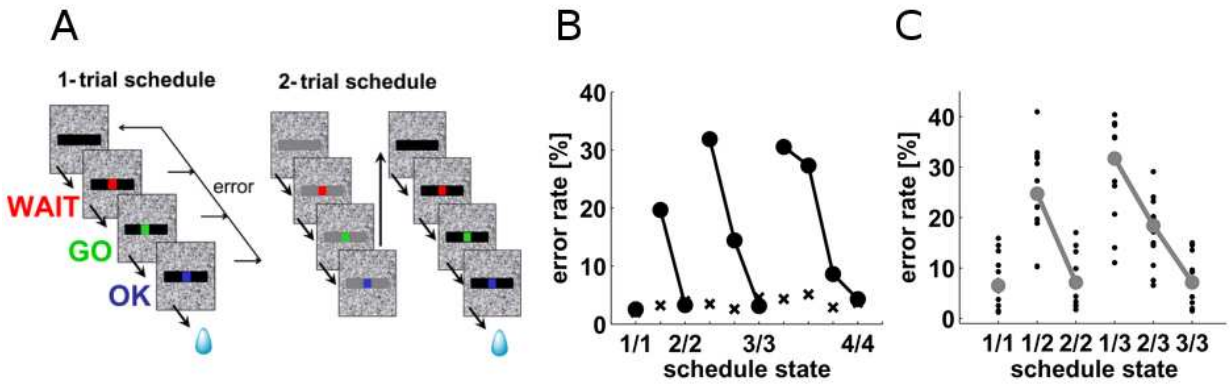

A
1-trial schedule
2-trial schedule
WAIT
GO
OK
error
B
error rate [%]
40
30
20
10
0
1/1
2/2
3/3
4/4
schedule state
C
error rate [%]
40
30
20
10
0
1/1 1/2 2/2 1/3 2/3 3/3
schedule state

Figure 1:

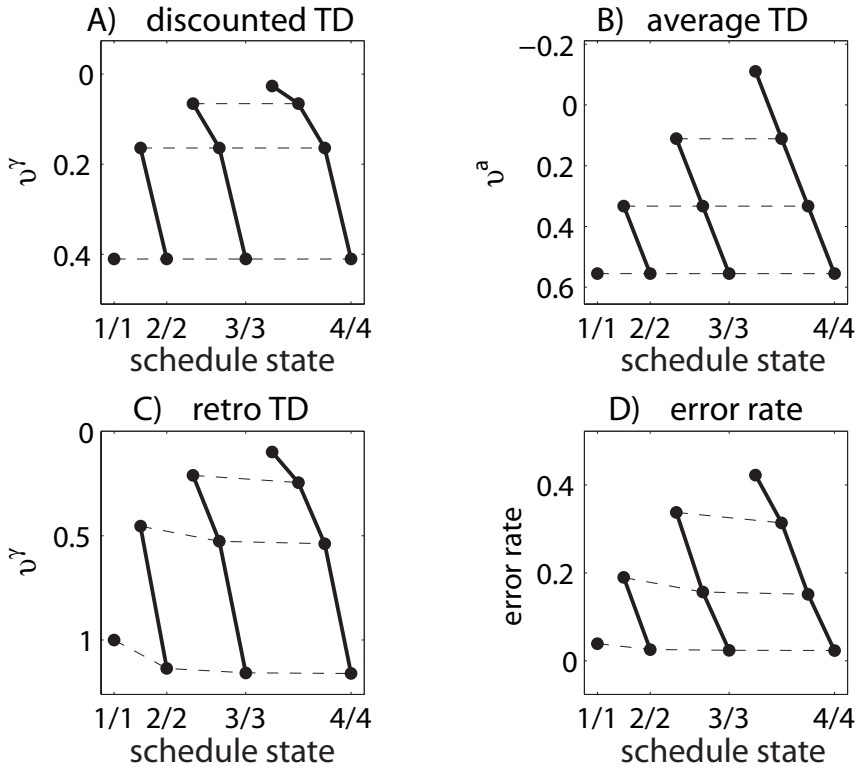
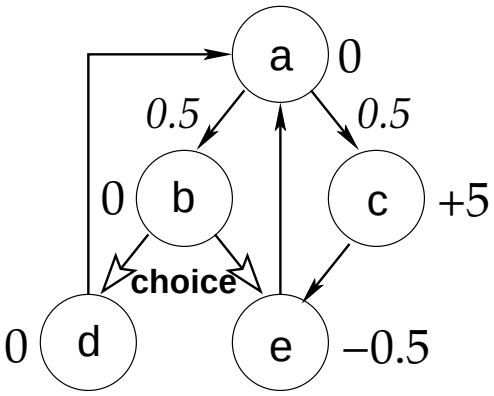

A) discounted TD
B) average TD
C) retro TD
D) error rate
schedule state
1/1 2/2 3/3 4/4
error rate

Figure 2:



state	retro	discount
a	1.3	1.0
b	0.80	0.16
c	5.8	5.0
d	**0.77**	**0.41**
e	**1.0**	**−0.09**

Figure 3:

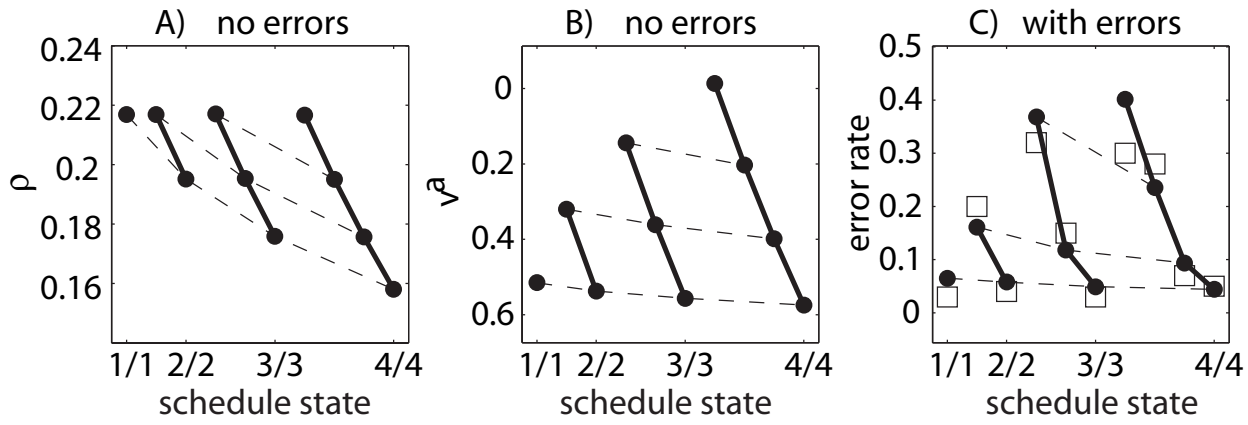

A) no errors
0.24
0.22
0.2
0.18
0.16
ρ
1/1 2/2 3/3 4/4
schedule state
B) no errors
0
0.2
0.4
0.6
v^a
1/1 2/2 3/3 4/4
schedule state
C) with errors
0.5
0.4
0.3
0.2
0.1
0
error rate
1/1 2/2 3/3 4/4
schedule state

Figure 4: