Published in final edited form as:
Neural Comput. 2010 June ; 22(6): 1511–1527. doi:10.1162/neco.2010.08-09-1080.

Hyperbolically Discounted Temporal Difference Learning

William H. Alexander and **Joshua W. Brown**
Dept. of Psychological and Brain Sciences, Indiana University, Bloomington IN

Abstract

Hyperbolic discounting of future outcomes is widely observed to underlie choice behavior in animals. Additionally, recent studies (Kobayashi & Schultz, 2008) have reported that hyperbolic discounting is observed even in neural systems underlying choice. However, the most prevalent models of temporal discounting, such as temporal difference learning, assume that future outcomes are discounted exponentially. Exponential discounting has been preferred largely because it is able to be expressed recursively, whereas hyperbolic discounting has heretofore been thought not to have a recursive definition. In this paper, we define a learning algorithm, hyperbolically discounted temporal difference (HDTD) learning which constitutes a recursive formulation of the hyperbolic model.

1 Introduction

A frequent decision faced by animals is whether to accept a small, immediate payoff for an action, or choose an action that will yield a better payoff in the future. Several factors may influence such decisions: the relative size of the possible rewards, the amount of delay between making a choice and receiving the more immediate reward, and the additional delay required to receive the greater reward.

Two possible explanations for temporal decision-making have been suggested. One hypothesis (Myerson & Green, 1995; Green & Myerson, 1996) is that delaying a reward introduces additional risks that an event may occur in the intervening time that will effectively prevent the animal from receiving the reward. A foraging animal, for instance, may find that a food item has been consumed by competitors or gone bad before the animal can retrieve the item. Alternatively, the appearance of a predator may preclude the animal from retrieving the food item. An animal should, therefore, select the option that maximizes the reward/risk ratio.

Another hypothesis (Kacelnik & Bateson, 1996) is that animals seek to maximize their average intake of food over time. In deciding between a small reward available immediately and a large reward that requires waiting (e.g., time for a food item to ripen) or travel (e.g., moving from a sparse patch of food to a richer one), the animal may be inclined to accept the lower-valued, immediate reward unless the delayed reward is large enough to justify the additional cost incurred in getting it. Under this hypothesis, any additional delay is acceptable to the animal provided the reward is large enough.

Both hypotheses, average reward and temporal discounting, have been formulated as models of real-time learning based on temporal difference (TD) learning. TD learning as originally formulated by Sutton & Barto (1990) discounts future rewards exponentially. Interpreted in

Address correspondence to: William H. Alexander, Dept. of Psychological & Brain Sciences, 1101 E Tenth St., Bloomington, IN 47405 USA, +1 812 856-3894 (office), +1 812 855-4691 (FAX), wialexan@indiana.edu.

terms of risk, this formulation of TD learning suggests that each unit of time added to the delay between a decision and the predicted outcome adds a fixed amount of risk that the predicted outcome won't occur. In contrast, an average reward variant of TD learning (Tsitsiklis & Van Roy, 1999; Tsitsiklis & Van Roy, 2002) attempts to maximize the rate of reward per time step. A key difference between these models is that average reward TD learning accounts for animal data showing preference reversals, whereas exponentially discounted TD learning does not (Green & Myerson, 1996).

A typical experiment in which animals exhibit preference reversals (e.g., Mazur, 1987) may involve an animal choosing between a large reward available at some fixed delay after a response, and a smaller reward available after a shorter, adjustable delay. When the animal selects the larger reward, the delay for the smaller reward is decreased, making it a more attractive option, and when the smaller reward is selected, its delay is increased. Eventually, the delay to the smaller reward will oscillate around a fixed point at which the animal selects the two options equally. At this point, if a fixed additional delay is added to the time required to receive either reward, the animal will tend to prefer the larger of the two. Conversely, if the time required is decreased by a fixed amount, the animal will prefer the smaller. This pattern is captured by average reward models, but not by exponentially discounted models of choice.

A wealth of data from humans, rats, pigeons, and monkeys suggests that animals discount future rewards hyperbolically. In terms of risk, this suggests that animals regard additional delays when a reward is proximal as incurring a greater risk that the reward will not occur than additional delays when a reward is temporally distant. Like average reward models, and unlike exponential discounting, in which each unit of time adds a fixed level of risk, models of hyperbolic discounting predict preference reversals as described above.

In this paper, we present a real-time model of hyperbolic discounting. Previous work has suggested that hyperbolic functions are not susceptible to computation by recursive methods (such as TD learning; Daw & Touretzky, 2000). However, by reinterpreting temporal discounting in terms of the level of risk per time step, we are able to define a variant of TD learning that discounts future rewards hyperbolically. Hyperbolically Discounted TD (HDTD) learning accounts for preference reversals, differential discounting based on reward size, as well as animal preference data which depend on sequences of reward delivery.

2 TD Learning

The goal of TD learning models is to learn the value of future rewards based on the current environmental state. The learned value of a state is the level of reward for that state, plus the discounted prediction of reward for subsequent states. The value at each state is updated proportionally to the discrepancy between the current value for that state and the combined value of the level of reward experienced at that state and future predictions. A common way to formalize this rule for updating is:

$$\delta_t = r_{t+1} - V_t + \gamma V_{t+1} \tag{2.1}$$

where r_{t+1} is the level of reward at time $t+1$, V_t is a reward prediction, and γ is a discounting factor. For $\gamma = 0$, the model only learns the value for the state at which it receives a reward. For $\gamma = 1$, the model learns the cumulative sum of future rewards.

For temporal difference models of simple conditioning experiments, a common tactic is to define a vector of states, *s*, such that each component of *s* represents a specific period of time following the onset of a CS. On each iteration of the model, the component of *s* corresponding with the current iteration *t* is set to 1, while other components are set to 0. The dynamics of this system are essentially a tapped delay line which tracks the amount of time since the presentation of a stimulus. On each iteration of such a model, the current value prediction is given as:

$$V_t = s_t \cdot w_t \tag{2.2}$$

where w_t is a weight representing the reward prediction at time *t*. The learning rule for calculating the temporal difference error associated with each state can be rewritten as:

$$\delta_t = r_{t+1} - V_t + V_{t+1} - (1-\gamma)V_{t+1} \tag{2.3}$$

While equivalent to exponentially-discounted TD learning as usually written, this formulation suggests an interpretation of TD learning in terms of risk. In the typical formulation of TD learning (2.1) γ is thought of as a discounting term, whereas in eq. 2.3, 1-γ is the hazard function of an exponential function. In the exponential case, the hazard function is constant and assumes that each unit of time involves the same level of risk as any other unit of time, while in hyperbolic discounting the hazard function varies with time; at times proximal to a reward, the hazard function is greater than at more distant times.

The intuition, then, is that a hyperbolically discounted variant of TD learning should include some means by which the hazard function is adjusted according to the temporal distance to a reward, so that the hazard function is greater at times nearest reward, when anticipated value is highest. This requires a way of estimating time remaining before an expected reward should occur. The time remaining until a reward is delivered can be approximated by the current value, V_t, which increases with temporal proximity to reward. This approach, while originally conceived of as an approximation, turns out to produce exactly hyperbolic discounting (see Appendix). The formulation of TD learning used here maintains estimates (via adjustable weights reflecting predictions of future reward) of both reward level and time until reward, which is approximated by the current discounted value. These predictions can be used to adjust the hazard function in a preliminary form of the HDTD learning rule:

$$\delta_t = r_{t+1} - V_t + V_{t+1} - \kappa V_t V_{t+1} \tag{2.4}$$

Here the term (1-γ) in (2.3) is replaced with κV_t to reflect the hyperbolically discounted form of TD (HDTD) learning, in which the discounting rate κ is modulated by current value V_t.

The non-recursive hyperbolic model of discounting is typically written as

$$V_t = \frac{R}{1+\kappa T} \tag{2.5}$$

where the parameter κ determines the level of discounting, and *T* is the delay to some reward, *R*. For a given value of κ, the HDTD model (2.4) learns the hyperbolically discounted value function given by the standard formalization of hyperbolic discounting (2.5), as shown in figure 1. In the Appendix, we supply a proof of this. Furthermore, the hazard function used for updating model weights in HDTD (κV_t) converges on the hazard function for the hyperbolic model, as shown in a proof in the Appendix.

An issue of generalizability arises, however, for reward magnitudes of varying sizes, as illustrated in figure 2a. In the preliminary formulation of the HDTD model (2.4), the discounting rate on each iteration is determined by a constant, κ, as well as the learned value function, V_t. As reward magnitude increases, so too does the value of V_t, which results in a *higher* discounting rate for higher magnitude rewards. The result is that the preliminary formulation of the HDTD model is incapable of showing preference reversals.

This issue can be resolved by scaling the discounting rate by the level of reward. (Myerson & Green, 1995) observed that rewards of unequal size are not discounted at the same rate. Specifically, larger rewards tend to be subject to less discounting than smaller rewards. This intuition can be implemented in the HDTD framework by dividing the hazard function from eq 2.4 by an estimate of the total magnitude per trial of a reward $\bar{r}$, where $\bar{r}$ is learned on successive trials by the delta rule $\bar{r} = \bar{r} + \alpha\ (R - \bar{r})$. Furthermore, it is not necessary to assume that the rate of discounting varies linearly with reward magnitude, so the denominator can be raised to a power σ. So the final formalization of the HDTD learning rule is:

$$\delta_t = r_{t+1} - V_t + V_{t+1} - \frac{\kappa V_t}{(bias + \bar{r})^{\sigma}} V_{t+1} \tag{2.6}$$

This formulation of the HDTD learning rule, unlike eq. 2.4, is capable of showing preference reversals (fig. 2b).

If the bias term is set to 0 and σ is set to 1, and we assume an *a priori* estimate of $\bar{r}$ where $\bar{r}$ is equal to the magnitude of the reward per trial, equation 2.6 results in the same effective rate of hyperbolic discounting regardless of reward size. That is, the equivalent non-recursive hyperbolic discounting model (2.5) is the same regardless of reward magnitude.

For environments in which reward estimates are initially unknown and subject to change, however, the bias term is necessary in order to avoid an undefined term (i.e., dividing by zero). An alternative approach would simply be to give the model an arbitrary initial estimate of $\bar{r}$ and allow it to adjust this estimate as described above; however, this may still result in an undefined term if $\bar{r}$ were to go to 0. For cases in which the bias term is non-zero, the equivalent non-recursive hyperbolic discounting model changes depending on the magnitude of $\bar{r}$. For relatively low magnitude rewards, the equivalent hyperbolic model has a discount factor κ lower than for high magnitude rewards. This is because the effective discount rate of the HDTD model is partially determined by the learned value function, V_t. When the reward magnitude per trial is small, the value function is similarly small, so that dividing by a constant bias term (plus $\bar{r}$) results in lower effective discounting, than when the reward magnitude and value function are large (although the discounting rate is lowered in both cases; it is simply lowered more for smaller magnitude rewards than larger).

This state of affairs, then, runs counter to our desire, which is that rewards with higher magnitude be discounted at a lower rate than low-magnitude rewards. Since the idealized situation of zero bias results in the same level of effective discounting for all reward magnitudes, and the inclusion of a bias term results in lower discounting for low-magnitude rewards relative to high-magnitude rewards, differential discounting based on reward size in the appropriate direction is due to the term σ. For the idealized case (bias = 0), a value of σ=1 would result in equivalent discounting rates for all levels of rewards, while values of σ>1 result in lower effective discounting as reward increases, and values of σ<1 result in higher effective discounting for larger rewards relative to smaller rewards. When a bias term is introduced, the precise value of σ which results in an equivalent discounting rate between two rewards of different magnitudes is shifted higher.

Figure 2b shows the hyperbolic value functions learned from equation (2.6) for r=1 (solid line) and r=2 (dashed line), and implies the presence of preference reversals. If a choice between the two rewards is made when the smaller reward is immediately available (vertical dashed line), the learned value of the immediate reward is greater. However, if the choice is made when the temporal distance to the smaller reward is greater (solid vertical line), the learned value for the greater reward is greater. Where the two value functions intersect is the point of indifference where each choice is equally likely to be made. This shows that HDTD is capable of preference reversals.

The parameter σ interacts in interesting ways with the level of reward predicted for a given trial. Of particular interest is that low values of σ (σ < 1, for example) yield an equivalent hyperbolic model (2.5) with a low discount factor for low levels of reward and a high discount factor for high levels of reward. Conversely, for high values of σ (e.g., σ = 2), the effective discount factor for low reward levels is higher than the effective discount factor for high reward levels.

Myerson & Green (1995) showed that, in humans, different rates of discounting based on reward size could be accounted for using two hyperbolic models with a single parameter each. In contrast, HDTD can reproduce the same hyperbolic curves with a single model containing two free parameters. Table 1 shows the best-fit hyperbolic models for a selection of individual subjects (from Green & Myerson, 1995), as well as the parameters κ and σ which produce the same two hyperbolic models using a single HDTD model (with the bias term equal to 1). These parameters can be determined analytically by solving the pair of equations

$$\frac{R_{high}\kappa}{(bias+R_{high})^{\sigma}}=\kappa_{high} \tag{2.7}$$

$$\frac{R_{low}\kappa}{(bias+R_{low})^{\sigma}}=\kappa_{low} \tag{2.8}$$

for κ and σ. This holds even when subjects appear to discount low rewards less heavily than high rewards (e.g., subject 7 in table 1). For intermediate levels of reward, the HDTD model predicts an effective discounting parameter falling between κ_{high} and κ_{low}. Whereas the standard hyperbolic model would require an additional model to be estimated for an intermediate reward condition, the HDTD model should be able to capture such data using the same estimates of κ and σ, suggesting that HDTD is more parsimonious. Further empirical tests of this are needed, however.

3 Average Reward vs. Hyperbolic Discounting

While we have shown that HDTD can exhibit preference reversals in accordance with animal data, this is not sufficient to differentiate HDTD from other models, such as average reward TD, which also exhibit preference reversals. To this end, we examine the behavior of HDTD and an implementation of average reward TD (Daw & Touretzky, 2000) in a context in which the order of reward delivery appears to influence preference. Brunner (1999) showed that rats tend to prefer reward sequences that “worsen” over time; given the choice between a reward sequence that delivers more food items at the beginning of the sequence than at the end (i.e., decreasing), and a reward sequence that delivers more food items at the end of the sequence than at the beginning (increasing), rats prefer the depleting sequence at short delays, and trend toward indifference between the two at long delays.

We compared the fit between average reward TD and HDTD to the approximate rat choice preferences from Brunner (1999), experiment 1. A simple actor component, based on that described by Daw & Touretzky (2000), was added to each model to learn choice preferences. At each time step of a trial, preference weights for a reward sequence were updated by the temporal difference error term δ_t multiplied by a learning rate parameter (in this case 0.001). Each model experienced 2000 trials in each of 6 conditions: an increasing or decreasing reward schedule at delays of 0, 5, and 15 trial iterations. Each iteration of the model was interpreted as having a duration of 1 second. The reward schedules were chosen to approximate the schedules used by Brunner (fig. 3a). For increasing reward schedules, rewards occurred at 0, 10, 15, 17, and 18 seconds, plus the delay for that condition. Decreasing rewards occurred at 0, 1, 3, 8, and 18 seconds, plus the condition delay. Of interest is that both increasing and decreasing reward schedules have the same amount of reward over the same length of time; that is, the average reward for each is the same. The length of each trial was determined by the time of the last reward, plus an additional inter-trial interval that lasted between 1 and 20 seconds (randomly selected from a uniform distribution). Following training, the actor's learned choice preferences between increasing and decreasing reward schedules at each delay were computed by a softmax activation function

$$Prob.\ selecting\ w = \frac{e^{P_w \varphi}}{e^{P_w \varphi} + e^{P_b \varphi}} \tag{3.1}$$

where P_w is the learned preference weight for the decreasing reward schedule, P_b is the preference for the increasing reward schedule, and φ is a scaling factor. A low value of φ will cause the model to prefer all choices equally, while a high value of φ will cause the model to more highly prefer even slightly better options. Free parameters for the HDTD model were κ, σ, and φ, and the bias term was set to 1. Free parameters for the average reward model were the learning rate of the model, a parameter θ controlling the exponential online estimate of average reward (Daw & Touretzky, 2002), as well as φ.

Figure 3b shows the best fit of the average reward vs. HDTD models. As expected, the average reward TD model is indifferent to whether the reward schedule increases or decreases. The HDTD model not only captures the pattern of choice preferences better than does the average reward model, but it also fits the data better than does a previous variant of hyperbolic discounting, the parallel hyperbolic discount model (Brunner, 1999), which was found to asymptote well below the percent of choice preferences actually observed. A potential criticism is that there were only three data points in Brunner's experiment, while the HDTD model had three free parameters which were adjusted by the fitting routine. However, the average reward model also had three free parameters, and yielded a significantly worse fit than the HDTD model. It is not the case, therefore, that the HDTD model better accounts for the data by virtue of having more free parameters than the competing model.

4 Discussion

A key motivation for a hyperbolic discounting model of temporal difference learning is the ability of hyperbolic discounting, and not exponential discounting, to show preference reversals. Nonetheless, the general form of the HDTD equation (2.6) suggests that exponentially discounted TD learning could also, in principle, show preference reversals, provided that the exponential discounting factor is also scaled by the level of reward. In light of this, the mere fact of a model exhibiting such reversals is not sufficient reason to prefer one form of discounting to another. However, it has been observed that the pattern of preference reversals is better characterized by a hyperbolic function rather than an exponential for both group and individual data (Green & Myerson, 1996). Given this, there is a clear rationale for preferring a hyperbolic discounting model to exponential discounting.

Myerson & Green (1995) suggest two potential motivations for the hyperbolic model of temporal discounting. One motivation derives the hyperbolic form from the notion that an animal seeks to maximize the rate of reward, while the second motivation suggests that increases in the temporal distance to an outcome impose additional, increasing risk that the outcome will fail to occur. Both of these motivations result in the non-recursive model of hyperbolic discounting (2.5).

Average reward TD learning (Tsitsiklis & Van Roy, 1999; Tsitsiklis & Van Roy, 2002) extends the first motivation, rate maximization, to a TD learning framework, while the HDTD model does the same for the risk interpretation of discounting. Both models are able to exhibit preference reversals similar to those observed in human and animal behavior (Daw & Touretzky, 2000). While average reward TD learning is able to reproduce many predictions of hyperbolic discounting models of decision making, it is unable to account for animal data in which choice preferences are influenced by the pattern of reward delivery (Brunner, 1999). The HDTD model, however, is capable of reproducing such choice preferences. This suggests that the risk interpretation of temporal discounting, and not rate maximization, is correct.

Insofar as it is the goal of models of reinforcement learning to account for animal behavior and its possible neural corollaries, our proposed variant of TD learning is able to account for observed behavior not captured by exponentially discounted TD learning with a minimum of added complexity. Additionally, recent evidence has shown that not only does observed behavior correspond to hyperbolic discounting, but that the activity of midbrain dopamine neurons in response to a reward-predicting CS appears to decline hyperbolically (Kobayashi & Schultz, 2008) with increases in delay to a predicted reward. TD learning has provided a useful framework for understanding the activity of dopamine neurons, and HDTD extends this framework to include these recent findings.

Several brain areas have been identified which seem to show anticipatory activity related to the prediction of an imminent reward. These areas include ventral striatum (Schultz, Apicella et al., 1992), anterior cingulate cortex (Amador, Schlag-Rey et al., 2000), orbitofrontal cortex(Schultz, Tremblay et al., 2000), and putamen (Schultz, Apicella et al., 1993). In the context of TD learning, this anticipatory activity appears to correspond with the learned value function (Suri & Schultz, 2001, e.g., figure 1). An interesting property of the hyperbolic discount function, however, is that its hazard function is simply a multiple of the function itself (Sozou, 1998). This suggests that the activity of areas of the brain which have previously been identified as encoding value predictions may actually signal a measure of risk as a function of time. The hyperbolic model, however, also suggests a means by which areas coding value can be distinguished from those whose activity simply reflects a hyperbolic hazard function. For different levels of reward, a value-predicting area should show differential activity, while a hazard function neuron will have the same pattern of activity for different levels of reward. This follows from the hyperbolic hazard function $\frac{\kappa}{1+\kappa T}$ which is the same regardless of reward size. It is not certain, however, that the brain does in fact maintain such hazard representations, and more research is needed to answer this question.

Additional parameters in the HDTD model may also have interpretations in terms of neuromodulatory systems, such as serotonin, whose role in reinforcement learning and decision-making is an ongoing research concern (Schweighofer, Bertin et al., 2008). In the HDTD model, a new parameter, σ, is introduced which modulates the balance of discounting between low and high rewards. Previous work has suggested that the serotonin is involved in reinforcement discounting; low levels of serotonin are associated with impulsive behavior, suggestive of high discounting for high value, delayed rewards. The HDTD model makes a novel prediction in this regard. If σ is related to the serotonergic system, it suggests that not only should high rewards be discounted more for low levels of serotonin, but also that low value rewards should be discounted less.

Appendix A

In the main text, we present the HDTD model in a descriptive manner and suggest that it is equivalent to the non-recursive formulation of the hyperbolic model of discounting. Here, we show the formal equivalence between the HDTD model and the hyperbolic model of discounting, and justify our interpretation of the model in terms of risk. We proceed in three steps. First, in theorem 1, we show that the hyperbolic discounting model has an exact recursive definition. Second, using the recursive formulation of hyperbolic discounting, we derive the HDTD learning rule presented in the main text. Finally, in theorem 2, we show that the quantity we describe as a hazard function $\frac{\kappa V_t}{R}$ in the main text is equivalent to the hyperbolic hazard function in the simple case of Δt=1.

Recursive definition of the hyperbolic model

Consider the hyperbolic discounting model:

$$V_t=\frac{R}{1+\kappa T} \tag{A.1}$$

Of note, the value V_t of R after hyperbolic discounting by time is decreased by scaling with the denominator on the right hand side, which is one plus a constant multiplied by temporal distance.

The hyperbolic discounting model is defined recursively for any $\{T, t\} \in \mathbb{Q}^+ \cup \{0\}$ (where $\mathbb{Q}^+$ is the set of rational, positive numbers), as

$$\begin{gathered} V_t=R \text{ if } T=0 \\ V_t=\frac{V_{t+\Delta t}}{1+\frac{\Delta t\kappa V_{t+\Delta t}}{R}} \text{otherwise} \end{gathered} \tag{A.2}$$

The origin of equation (A.2) can be seen in the functional similarity with equation (A.1), in which the discounted reward V_t at time t is smaller (i.e. reward is more distant in the future). This smaller value V_t is obtained by starting with the value $V_{t+\Delta t}$ and decreasing it by scaling with the denominator on the right hand side, which is one plus a constant $\frac{\Delta t\kappa}{R}$, multiplied by temporal distance. Here, the recursion is effected by representing temporal distance by $V_{t+\Delta t}$ instead of T as in equation (A.1)).

Let $T = -t+C$, where C is a constant, which implies that ΔT = $-\Delta t$, constrained by $T \geq 0$. This change of variables implies, from equation (A.1), that:

$$V_{t-\Delta t}=\frac{R}{1+\kappa(T+\Delta T)} \tag{A.3}$$

Theorem 1. For all rational, positive numbers T, the hyperbolic function $V_t=\frac{R}{1+\kappa T}$ from equation (A.1) is a solution to the recursive equation (A.2).

Proof. The proof is by induction over T for rational, positive numbers and 0. We proceed first by demonstrating that the base case T=0 is true:

$$\textit{Base Case}{:}V_0=\frac{R}{1+\kappa 0}.$$

By definition, $V_0 f$=R

$$\begin{gathered}R=\frac{R}{1+\kappa 0}\\ R=\frac{R}{1}\\ R=R\end{gathered}$$

Hence equation (A.1) is a solution to (A.2) in the special case of *T*=0. In order to demonstrate by induction that the recursive hyperbolic model is equivalent to the non-recursive hyperbolic model for all T, we assume that the *inductive hypothesis* $V_t=\frac{R}{1+\kappa T}$ is true, and show that the relationship holds for $V_{t-\Delta t}$ in equation (A.2).

$$\textit{Inductive hypothesis}{:}\text{assume } V_t=\frac{R}{1+\kappa T} \tag{A.4}$$

Then by extension of (A.4),

$$V_{t-\Delta t}=\frac{R}{1+\kappa(T+\Delta t)} \tag{A.5}$$

It is required to show that (A.4) and (A.5) together provide a solution to (A.2).

From (A.2),

$$V_{t-\Delta t}=\frac{V_t}{1+\frac{\Delta t\kappa V_t}{R}}$$

By application of the inductive hypothesis, we substitute $\frac{R}{1+\kappa T}$ for V_t

$$V_{t-\Delta t}=\frac{\frac{R}{1+\kappa T}}{1+\frac{\Delta t\kappa\frac{R}{1+\kappa T}}{R}}$$

and show that $V_{t-\Delta t}=\frac{R}{1+\kappa(T+\Delta t)}$:

$$\begin{array}{c}
\frac{\frac{R}{1+\kappa T}}{1+\frac{\Delta t\kappa\frac{R}{1+\kappa T}}{R}}=\frac{R}{1+\kappa(T+\Delta t)}\\
\frac{\frac{R}{1+\kappa T}}{1+\frac{\Delta t\kappa}{1+\kappa T}}=\frac{R}{1+\kappa(T+\Delta t)}\\
\frac{\frac{R}{1+\kappa T}}{\frac{1+\kappa T}{1+\kappa T}+\frac{\Delta t\kappa}{1+\kappa T}}=\frac{R}{1+\kappa(T+\Delta t)}\\
\frac{\frac{R}{1+\kappa T}}{\frac{1+\kappa T+\Delta tk}{1+\kappa T}}=\frac{R}{1+\kappa(T+\Delta t)}\\
\frac{R}{1+\kappa T+\Delta tk}=\frac{R}{1+\kappa(T+\Delta t)}\\
\frac{R}{1+\kappa(T+\Delta t)}=\frac{R}{1+\kappa(T+\Delta t)}
\end{array}$$

hence by induction ∀ $\{T, t\} \in \mathbb{Q}^{+} \cup \{0\}$, $V_t=\frac{R}{1+\kappa T}$ is a solution to the recursive equation (A.2).

QED.

Derivation of the HDTD model

Theorem 1 says that the hyperbolic model has an exact, recursive definition. We can now use this recursive definition to obtain the HDTD model in the form of a Bellman equation. First, note that the recursive model in equation (A.2) can be written equivalently as

$$\begin{array}{c}
V_t=R \text{ if } T=0\\
V_t=V_{t+\Delta t}-\frac{\Delta t\kappa V_t V_{t+\Delta t}}{R}\text{otherwise}
\end{array}$$

This will be important when we confirm that the hyperbolic hazard function is the same as the HDTD hazard function in theorem 2 (below).

At convergence, predictions learned by the HDTD model, $\hat{V}_t$, should satisfy the definition above. If, however, a prediction is off, the prediction is updated in proportion to the amount it deviates from the ideal estimate – essentially a temporal difference error:

$$\begin{array}{c}
\delta_t=R-\widehat{V}_t \text{ if } T=0\\
\delta_t=\widehat{V}_{t+\Delta t}-\widehat{V}_t-\frac{\Delta t\kappa\widehat{V}_t\widehat{V}_{t+\Delta t}}{R}\text{otherwise}
\end{array}$$

Note that $\hat{V}_{t+\Delta t}$ itself is also a prediction learned by the model. These can be combined into a single learning rule:

$$\delta_t=r_t+\widehat{V}_{t+\Delta t}-\widehat{V}_t-\frac{\Delta t\kappa\widehat{V}_t\widehat{V}_{t+\Delta t}}{R}$$

where $r_t = R$ if T=0, and 0 otherwise. The prediction at time T, then, is updated according to

$$\widehat{V}_t=\widehat{V}_t+\alpha\delta_t$$

where α is the learning rate parameter.

Hyperbolic hazard function

In the main text, we refer to the quantity $\frac{\kappa V_t}{R}$ as the HDTD hazard function, in the simple case of $\Delta t = 1$. We now show that, at convergence, this quantity works out to the hazard function of the hyperbolic model.

Theorem 2. The hyperbolic hazard function is identical to the hazard function of the HDTD equation (2.4) at convergence. In a general sense, this follows from Theorem 1, in that if the functions are identical, then their hazard functions must be identical. In mathematical terms, $\forall\ R, \kappa$, the HDTD hazard function $\frac{\kappa V_t}{R}$ from equation (2.4) is identical to the hyperbolic hazard function $\frac{k}{1+\kappa T}$.

Preliminaries. An alternate way of writing the hyperbolic discounting function is as the value of an immediate reward multiplied by the hyperbolic survivor function, $\frac{1}{1+\kappa T}$ (Sozou, 1998). The hazard function, defined as the negative derivative of the survivor function divided by the survivor function, gives us the hyperbolic hazard function, $\frac{\kappa}{1+\kappa T}$, which is itself a hyperbola.

Proof. From Theorem 1, we defined in equation (A.1) that

$$V_t = \frac{R}{1+\kappa T}$$

Substituting into the HDTD hazard function and setting it equal to the hyperbolic hazard function (defined above) we get

$$\frac{\kappa \frac{R}{1+\kappa T}}{R} = \frac{k}{1+\kappa T}$$
$$\frac{\kappa \frac{1}{1+\kappa T}}{1} = \frac{k}{1+\kappa T}$$
$$\frac{k}{1+\kappa T} = \frac{k}{1+\kappa T}$$

Q.E.D

References

Amador N, Schlag-Rey M, et al. Reward-predicting and reward-detecting neuronal activity in the primate supplementary eye field. J Neurophysiol 2000;84(4):2166–2170. [PubMed: 11024104]

Brunner D. Preference for sequences of rewards: further tests of a parallel discounting model. Behavioural Processes 1999;45(1–3):87–99.

Daw ND, Touretzky DS. Behavioral considerations suggest an average reward TD model of the dopamine system. Neurocomputing: An International Journal 2000;32–33:679–684.

Daw ND, Touretzky DS. Long-term reward prediction in TD models of the dopamine system. Neural Comput 2002;14(11):2567–2583. [PubMed: 12433290]

Green L, Myerson J. Exponential Versus Hyperbolic Discounting of Delayed Outcomes: Risk and Waiting Time. Amer. Zool 1996;36(4):496–505.

Kacelnik A, Bateson M. Risky Theories--The Effects of Variance on Foraging Decisions. Amer. Zool 1996;36(4):402–434.

Kobayashi S, Schultz W. Influence of reward delays on responses of dopamine neurons. J Neurosci 2008;28(31):7837–7846. [PubMed: 18667616]

Mazur JE. An adjusting procedure for studying delayed reinforcement. Commons, Michael L 1987;5:55–73.

Myerson J, Green L. Discounting of delayed rewards: Models of individual choice. J Exp Anal Behav 1995;64(3):263–276. [PubMed: 16812772]

Schultz W, Apicella P, et al. Reward-related activity in the monkey striatum and substantia nigra. Prog Brain Res 1993;99:227–235. [PubMed: 8108550]

Schultz W, Apicella P, et al. Neuronal activity in monkey ventral striatum related to the expectation of reward. J Neurosci 1992;12(12):4595–4610. [PubMed: 1464759]

Schultz W, Tremblay L, et al. Reward processing in primate orbitofrontal cortex and basal ganglia. Cereb Cortex 2000;10(3):272–284. [PubMed: 10731222]

Schweighofer N, Bertin M, et al. Low-serotonin levels increase delayed reward discounting in humans. J Neurosci 2008;28(17):4528–4532. [PubMed: 18434531]

Sozou PD. On hyperbolic discounting and uncertain hazard rates. Proceedings of the Royal Society of London. Series B: Biological Sciences 1998;265(1409):2015–2020.

Suri RE, Schultz W. Temporal difference model reproduces anticipatory neural activity. Neural Comput 2001;13(4):841–862. [PubMed: 11255572]

Sutton RS, Barto AG. Time-derivative models of Pavlovian reinforcement. Gabriel, Michael. 1990

Tsitsiklis JN, Van Roy B. Average cost temporal-difference learning. Automatica 1999;35(11):1799–1808.

Tsitsiklis JN, Van Roy B. On average versus discounted reward temporal-difference learning. Machine Learning 2002;49(2–3):179–191.

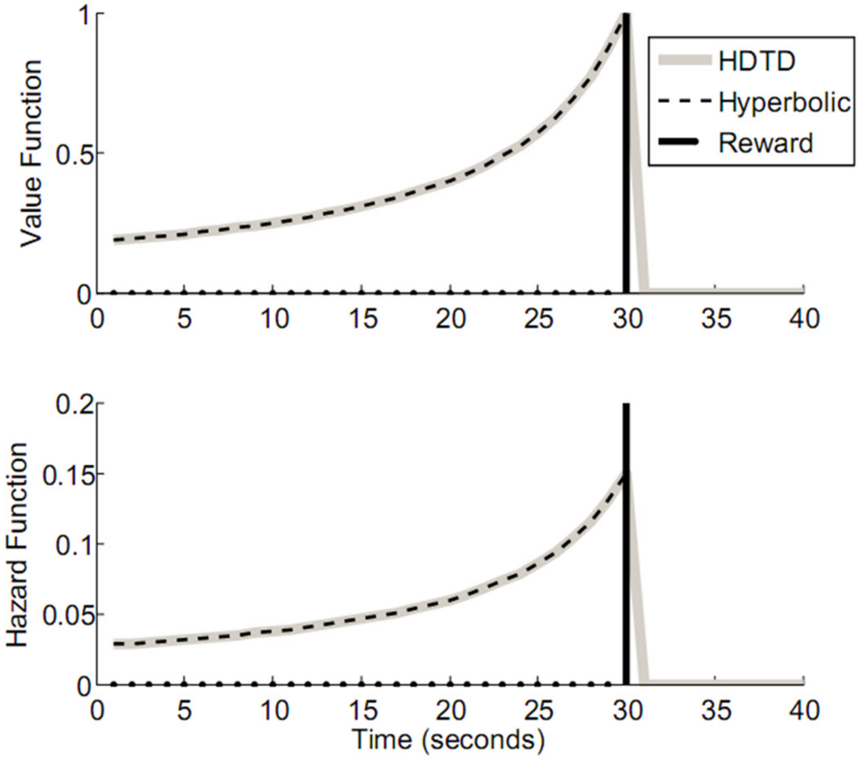

Figure 1.

Learned value and hazard functions for the HDTD model compared with same from the non-recursive hyperbolic discounting model (κ = 0.15). For a reward given at t=30 (vertical line), both the hyperbolic discounting model and HDTD have the same value function. The HDTD model learns the appropriate value function over the course of multiple (1000) trials. Similarly, the HDTD hazard function corresponds exactly with the hyperbolic discounting hazard function.

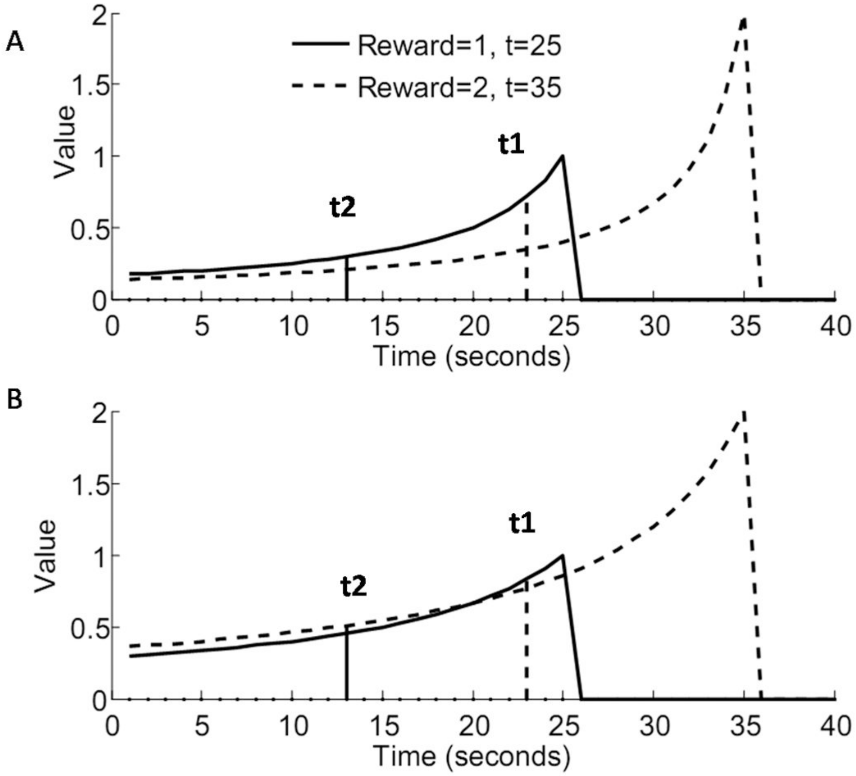

Figure 2.

Behavior of the HDTD model (A) when the discounting factor is not scaled by estimated reward per trial (eq. 2.4, $\kappa = 0.2$), and (B) when the discounting factor is scaled by the estimated reward per trial(eq. 2.6, $\kappa = 0.2$, $\sigma = 1$). The HDTD model reverses preferences (B) depending on the temporal proximity of two unequal rewards. When a small reward is immediately available (t1), the value function for that reward (solid line) is higher than for a larger delayed reward (dashed line). However, when the distance to both rewards is increased (t2), the preferences reverse; the value function for the larger reward is higher than for the smaller.

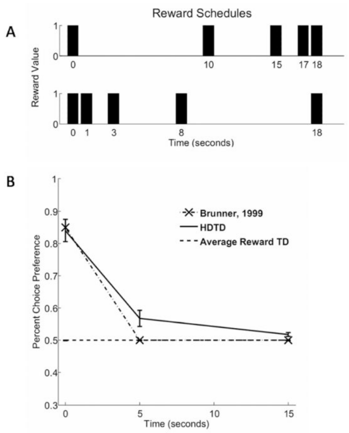

Figure 3.

The HDTD model and average reward TD learning were fit to data from Brunner, 1999. A) Rewards were delivered according to two schedules, increasing (top) and decreasing (bottom). The average reward for both schedules is the same. B) The average reward TD model is indifferent to reward schedule, while the HDTD model strongly prefers the decreasing reward schedule at short delays, in accordance with Brunner, 1999. The best-fit parameters for the HDTD model are κ = 0.544, σ = 0.741, and φ = 54.85. Parameters found for the average reward TD model were θ = 0.0010, φ = 0.9841, and α (learning parameter) = 0.0986. The fit of the HDTD model yielded a mean-square error of 0.0050, while the fit of the average reward model yielded a MSE of 0.1226. Data were approximated from Brunner, 1999, figure 1.

Table 1

	Hyperbolic Models		Equivalent HDTD Model	
Subject	κ_{low} (reward=1,000)	κ_{high} (reward=10,000)	κ	σ
1	0.065	0.008	35.1117	1.9106
2	0.025	0.007	1.1454	1.5534
7	3.941	8.580	0.3828	0.66238
9	0.008	0.009	0.005638	0.94922

A selection of subjects from Myerson & Green (1995). Subjects' data were fit by two hyperbolic models for a low and high potential reward condition. A single HDTD model can be found for each subject that fits both the low and high reward hyperbolic models (see text).