# Demystifying the Recency Heuristic in Temporal-Difference Learning

**Brett Daley**
brett.daley@ualberta.ca
Dept. of Computing Science
University of Alberta

**Marlos C. Machado**
machado@ualberta.ca
Dept. of Computing Science
University of Alberta
Canada CIFAR AI Chair

**Martha White**
whitem@ualberta.ca
Dept. of Computing Science
University of Alberta
Canada CIFAR AI Chair

## Abstract

The recency heuristic in reinforcement learning is the assumption that stimuli that occurred closer in time to an acquired reward should be more heavily reinforced. The recency heuristic is one of the key assumptions made by TD($\lambda$), which reinforces recent experiences according to an exponentially decaying weighting. In fact, all other widely used return estimators for TD learning, such as $n$-step returns, satisfy a weaker (i.e., non-monotonic) recency heuristic. Why is the recency heuristic effective for temporal credit assignment? What happens when credit is assigned in a way that violates this heuristic? In this paper, we analyze the specific mathematical implications of adopting the recency heuristic in TD learning. We prove that any return estimator satisfying this heuristic: 1) is guaranteed to converge to the correct value function, 2) has a relatively fast contraction rate, and 3) has a long window of effective credit assignment, yet bounded worst-case variance. We also give a counterexample where on-policy, tabular TD methods violating the recency heuristic diverge. Our results offer some of the first theoretical evidence that credit assignment based on the recency heuristic facilitates learning.

## 1 Introduction

The temporal credit-assignment problem in reinforcement learning (RL) is the challenge of determining which past actions taken by a decision-making agent contributed to a certain outcome (Minsky, 1961). Addressing the temporal credit-assignment problem effectively is paramount to efficient RL. Unfortunately, an optimal solution is likely infeasible for an agent acting in an arbitrary, unknown environment; perfect credit assignment would require precise knowledge of the environment's dynamics. Even then, the complexity of the problem grows enormously as the agent takes more actions over its lifetime. Instead, heuristics—simplifying rules or assumptions for credit assignment—can be adopted to make the problem more approachable. In the absence of any prior knowledge of the environment, a common and reasonable choice is the *recency heuristic*: "One assigns credit for current reinforcement to past actions according to how recently they were made" (Sutton, 1984, p. 94). The recency heuristic reflects the fact that there is likely to be a cause-and-effect relationship between actions and rewards that are close together in time.

In computational RL, the reinforcement signal is taken to be the temporal-difference (TD) error: the difference between the observed and expected reward earned by an action. TD($\lambda$) (Sutton, 1988) is the prime example of the recency heuristic; each TD error is applied to past actions in proportion to an exponentially decaying eligibility, achieving credit assignment that gracefully fades as the time between the action and TD error increases. This strategy, although simple, is highly effective and has been used by many recent algorithms (e.g., Schulman et al., 2015; Harb & Precup, 2016; Harutyunyan et al., 2016; Munos et al., 2016; van Seijen, 2016; Mahmood et al., 2017; Mousavi et al., 2017; Daley & Amato, 2019; Kozuno et al., 2021; Gupta et al., 2023; Tang et al., 2024).

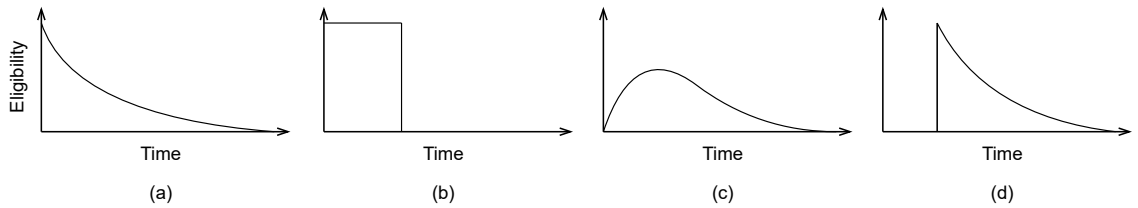


Figure 1: Illustrations of the eligibility curves for (a) $\lambda$-return, (b) $n$-step return, (c) inverted U-shape assignment inspired by Klopf (1972), and (d) time-delayed $\lambda$-return. The horizontal axis represents the elapsed time since the stimulus. Neither (c) nor (d) satisfy the recency heuristic.

However, the recency heuristic is, by definition, a simplifying assumption; one can imagine complex environments where non-recent credit assignment would theoretically be more beneficial. For example, if it were known that there is always some fixed delay between actions and their corresponding effects—especially when under partial observability (Kaelbling et al., 1998)—then this information could theoretically be exploited for faster learning. Klopf (1972), for instance, describes credit-assignment functions based on an inverted-U shape (see Figure 1c) that could achieve this exact effect. The shape of the credit-assignment curve encodes a prior belief over the likelihood of when a reward will arrive following an action, with the smooth distribution reflecting some uncertainty in the exact time of arrival. Klopf (1972) hypothesized that reactions in a firing neuron would leave it eligible to learn for a short duration. This later inspired the simplified spike-and-decay model of eligibility traces (Barto et al., 1983; Sutton, 1984) used by TD($\lambda$), which obeys the recency heuristic and has become a standard approach for credit assignment in computational RL.

Although there is potential for more efficient learning with non-recent credit assignment, it has not been tried in computational RL. Even alternatives to TD($\lambda$) that are not generally connoted with the recency heuristic, such as $n$-step TD methods (Cichosz, 1995), implement a crude form of recency heuristic: TD errors within some fixed time interval following an action are reinforced, while those outside are not. In fact, all other return estimators used for TD learning (which are constructed from $n$-step returns) satisfy some form of recency heuristic (see Section 5). We are not aware of any results that analyze what happens when TD updates do not follow the recency heuristic.

The goals of this paper are to understand the implications of forgoing the recency heuristic in TD learning, and to provide new insights into why assigning credit based on the recency heuristic has been so effective for RL. We test a model of non-recent credit assignment based on a short, time-delayed pulse inspired by Klopf's (1972) inverted-U function. Although this is one of the simplest and most benign forms of non-recency in TD learning, we show that it diverges under the favorable conditions of tabular, on-policy learning. We prove that the root cause of divergence is negative weights on some of the $n$-step returns in the return estimate, which appear whenever the recency heuristic is violated, and counteract learning by increasing the contraction modulus. In the off-policy setting, our analysis resolves the open problem by Daley et al. (2023) on the convergence of trajectory-aware eligibility traces. Finally, we show that satisfying the recency heuristic increases the effective credit-assignment window of a return estimate without increasing its bias and variance in the worst case, which partly explains the empirical success of methods like TD($\lambda$). Overall, our results demonstrate that the recency heuristic is not an overly simplistic assumption but is actually a crucial component in the mathematical basis of TD learning.

## 2 Background

We adopt the standard RL perspective of a decision-making agent learning in an unknown environment through trial and error (Sutton & Barto, 2018, Sec. 3.1). The agent-environment interface is modeled by a Markov decision process (MDP) formally described by the tuple $(\mathcal{S}, \mathcal{A}, p, r)$. The finite sets $\mathcal{S}$ and $\mathcal{A}$ contain the possible environment states and agent actions, respectively. At each time

step $t \geq 0$, the agent observes the current state of the environment, $S_t$, and takes an action, $A_t \in \mathcal{A}$, with probability $\pi(A_t|S_t)$, where $\pi$ is the agent's policy. Consequently, the environment state transitions to $S_{t+1} \in \mathcal{S}$ with probability $p(S_{t+1}|S_t, A_t)$, and the agent receives a reward, $R_t \stackrel{\text{def}}{=} r(S_t, A_t)$.

In prediction problems, the agent's objective is to learn the value function $v_\pi(s) \stackrel{\text{def}}{=} \mathbb{E}_\pi[G_t \mid S_t = s]$, where $G_t \stackrel{\text{def}}{=} \sum_{i=0}^{\infty} \gamma^i R_{t+i}$ is the observed discounted return. The constant $\gamma \in [0, 1]$ is called the discount factor and determines the agent's relative preference for delay rewards. In the rest of this section, we discuss various types of temporal-difference (TD) learning (Sutton, 1988), a common approach for prediction in reinforcement learning.

**TD($\lambda$) and the Recency Heuristic** TD methods estimate $v_\pi$ by iteratively reducing an error between predicted and observed returns, *bootstrapping* from the previous (biased) estimates in order to reduce variance. Let $v \colon \mathcal{S} \to \mathcal{A}$ be the agent's estimate of the value function, and define $V_t \stackrel{\text{def}}{=} v(S_t)$ for brevity. The TD error, defined as $\delta_t \stackrel{\text{def}}{=} R_t + \gamma V_{t+1} - V_t$, is the fundamental unit of reinforcement in TD methods. For instance, the simplest TD method, known as TD(0) or 1-step TD (Sutton, 1988), performs the update $v(S_t) \leftarrow V_t + \alpha_t \delta_t$, where $\alpha_t \in (0, 1]$ is the step size. TD(0) is a special case of TD($\lambda$) (Sutton, 1988), one of the earliest and most widely used TD methods. TD($\lambda$) is able to assign credit simultaneously to multiple states through the use of eligibility traces (Klopf, 1972; Barto et al., 1983; Sutton, 1984), a function $z \colon \mathcal{S} \to \mathbb{R}$ that tracks recent state visitations. On each time step, TD($\lambda$) performs the following updates:

$$z(s) \leftarrow \gamma\lambda\, z(s)\,,\ \forall\ s \in \mathcal{S}\,, \qquad z(S_t) \leftarrow z(S_t) + 1\,, \qquad v(s) \leftarrow v(s) + \alpha_t \delta_t\, z(s)\,,\ \forall\ s \in \mathcal{S}\,, \tag{1}$$

where $\lambda \in [0, 1]$ is the recency hyperparameter. Every eligibility trace is unconditionally decayed by a factor of $\gamma\lambda$, but only the trace for the current state is incremented. Then, every state is updated in proportion to its eligibility trace, using the current TD error. Eligibility traces are an efficient mechanism for assigning credit to recently visited states.

The above updates are known as the *backward view* of TD($\lambda$). An alternative perspective is the *forward view*. Suppose we hold the value function and step size fixed, and track the cumulative update for a single state visitation. We would find that the state is updated according to

$$v(S_t) \leftarrow V_t + \alpha_t\big(G_t^\lambda - V_t\big)\,, \tag{2}$$

$$\text{where } G_t^\lambda \stackrel{\text{def}}{=} V_t + \sum_{i=0}^{\infty} (\gamma\lambda)^i \delta_{t+i}\,. \tag{3}$$

The forward and backward views are equivalent under the conditions described above (Sutton, 1988; Watkins, 1989). The quantity defined in Eq. (3) is known as the $\lambda$-return and represents the theoretical target of the TD($\lambda$) update. Although the forward view is acausal and not directly implementable as an online algorithm, it reveals the temporal relationship between a state and the degree to which future TD errors are reinforced. The exponential decay of Eq. (3) represents a form of *recency heuristic*, the assumption that the causality between events weakens as the time between them increases. Mathematically, the hyperparameter $\lambda$ controls the bias-variance trade-off by interpolating between high-bias 1-step TD ($\lambda = 0$) and high-variance Monte Carlo ($\lambda = 1$) methods (Kearns & Singh, 2000). As we show next, non-exponential implementations of the recency heuristic are also possible; however, they do not enjoy the same efficient implementation with eligibility traces.

**$n$-step Returns and Compound Returns** More generally, TD methods can be expressed as a forward-view update in terms of an arbitrary return estimate, $\hat{G}_t$:

$$v(S_t) \leftarrow V_t + \alpha_t\big(\hat{G}_t - V_t\big)\,. \tag{4}$$

This operation is known as a value backup, and we refer to the estimate $\hat{G}_t$ as its target. We already established in Eq. (2) that the $\lambda$-return, $G_t^\lambda$, is one possible target. Another common target is the $n$-step return (Watkins, 1989; Cichosz, 1995), defined as $G_t^{(n)} \stackrel{\text{def}}{=} \sum_{i=0}^{n-1} \gamma^i R_{t+i} + \gamma^n V_{t+n}$, where

$n \geq 1$ determines the length of the return. Just like the $\lambda$-return, the $n$-step return interpolates between high-bias TD ($n = 1$) and high-variance Monte Carlo ($n = \infty$) methods. Although not commonly used, the $n$-step return admits a forward-view cumulative error similar to Eq. (3):

$$G_t^{(n)} = V_t + \sum_{i=0}^{n-1} \gamma^i \delta_{t+i} \,, \tag{5}$$

This reveals that the $n$-step return also satisfies the recency heuristic, albeit a weaker notion than that of the $\lambda$-return (see Section 3). Nevertheless, it still fulfills the basic assumption that TD errors nearer in time to a given state should be reinforced, whereas those farther away should not. The $n$-step return is also useful as a fundamental building block for constructing other estimates. For instance, the $\lambda$-return from Eq. (3) is equivalent to a weighted average of $n$-step returns:

$$G_t^\lambda = (1-\lambda) \sum_{n=1}^{\infty} \lambda^{n-1} G_t^{(n)} \,. \tag{6}$$

More generally, we can consider arbitrary convex combinations of $n$-step returns, strictly generalizing both $\lambda$-returns and $n$-step returns. Let $(c_n)_{n=1}^\infty$ be a sequence of nonnegative weights such that $\sum_{n=1}^\infty c_n = 1$. We refer to the following estimate as a *convex* return:

$$G_t^{\mathbf{c}} \stackrel{\text{def}}{=} \sum_{n=1}^{\infty} c_n G_t^{(n)} \,. \tag{7}$$

When at least two weights are nonzero, a convex return becomes a weighted average of $n$-step returns known as a compound return (Watkins, 1989; Sutton & Barto, 2018; Daley et al., 2024). Examples of compound returns include $\lambda$-returns, $\gamma$-returns (Konidaris et al., 2011), and $\Omega$-returns (Thomas et al., 2015). In Section 5, we show that the definition of a convex return is inherently related to the recency heuristic. Prior to our work, convex returns were the most general form of return estimator for TD learning, but we generalize them further in Section 5.

**Value-Function Operators and Convergence Conditions** We have discussed a variety of TD methods based on forward-view return estimates, but we have not yet established what makes an estimate valid for learning. Convergence to $v_\pi$ is perhaps most easily seen from the perspective of value-function operators. An operator $\boldsymbol{H} \colon \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ transforms a value function. The most fundamental value-function operator is the Bellman operator (Bellman, 1957), defined as

$$\boldsymbol{T}_\pi \boldsymbol{v} \stackrel{\text{def}}{=} \boldsymbol{r} + \gamma \boldsymbol{P}_\pi \boldsymbol{v} \,, \quad \text{where } (\boldsymbol{P}_\pi \boldsymbol{v})(s) \stackrel{\text{def}}{=} \sum_{a \in \mathcal{A}} \pi(a|s) \sum_{s' \in \mathcal{S}} p(s'|s,a)\, \boldsymbol{v}(s') \,.$$

Note that $\boldsymbol{r}$ and $\boldsymbol{v}$ here are treated as vectors in $\mathbb{R}^{|\mathcal{S}|}$, and $\boldsymbol{P}_\pi$ is treated as a $|\mathcal{S}| \times |\mathcal{S}|$ stochastic matrix. Let $\boldsymbol{T}_\pi^n \boldsymbol{v} \stackrel{\text{def}}{=} \boldsymbol{T}_\pi \boldsymbol{T}_\pi^{n-1} \boldsymbol{v}$ and $\boldsymbol{T}_\pi^0 \boldsymbol{v} \stackrel{\text{def}}{=} \boldsymbol{v}$. The $n$-iterated Bellman operator, $\boldsymbol{T}_\pi^n$, corresponds to the $n$-step return. Hence, convex returns are associated with the operator $\boldsymbol{v} \mapsto \sum_{n=1}^\infty c_n \boldsymbol{T}_\pi^n \boldsymbol{v}$. More generally, every value backup like Eq. (4) is equivalent to the noisy application of some operator, $\boldsymbol{H}$, to an element of the value function. That is, a return estimate can be represented as $\hat{G}_t = (\boldsymbol{H}\boldsymbol{v})(S_t) + \omega_t$, where $\omega_t$ is zero-mean noise. TD updates can thus be expressed in the form

$$\boldsymbol{v}(s) \leftarrow \begin{cases} (1-\alpha_t)\, \boldsymbol{v}(s) + \alpha_t \big( (\boldsymbol{H}\boldsymbol{v})(s) + \omega_t \big) \,, & \text{if } s = S_t \,, \\ \boldsymbol{v}(s) \,, & \text{otherwise.} \end{cases} \tag{8}$$

To produce a TD method of the form of Eq. (4) that converges to $\boldsymbol{v}_\pi$ under general conditions (e.g., Bertsekas & Tsitsiklis, 1996, Proposition 4.4), it is required that $\boldsymbol{H}$ is a maximum-norm contraction mapping with $\boldsymbol{v}_\pi$ as its unique fixed point, and that the step sizes are annealed such that $\sum_{t=0}^\infty \alpha_t = \infty$ and $\sum_{t=0}^\infty \alpha_t^2 < \infty$ (Robbins & Monro, 1951). An operator $\boldsymbol{H}$ is a contraction mapping if and only if $\|\boldsymbol{H}\boldsymbol{v} - \boldsymbol{H}\boldsymbol{v}'\|_\infty \leq \beta \|\boldsymbol{v} - \boldsymbol{v}'\|_\infty$, where $\beta \in [0, 1)$ is the contraction modulus. All of the operators discussed so far satisfy these properties because they are convex combinations of $n$-step Bellman operators, each of which are contraction mappings around $\boldsymbol{v}_\pi$ with a modulus of $\gamma^n$.

## 3 Formalizing the Recency Heuristic

In this section, we precisely define the notion of the recency heuristic. We consider a general estimator for TD learning of the form

$$\hat{G}_t = V_t + \sum_{i=0}^{\infty} h_i \gamma^i \delta_{t+i}\,, \tag{9}$$

where $(h_i)_{i=0}^{\infty}$ is a sequence of real numbers. Although this may appear to be restrictive, we show in Section 5 that it can represent every valid return estimate (i.e., converges to $\boldsymbol{v}_\pi$) that comprises a linear combination of future rewards and state values, and thus is implementable as a TD method.

We can think of Eq. (9) as an abstract form of TD($\lambda$): one with an arbitrary stimulus-response model rather than the familiar exponential decay. At time $t$, the agent experiences an external stimulus modulated by the current environment state, $S_t$. Positive or negative reinforcement subsequently arrives in the form of the TD errors, $(\delta_t, \delta_{t+1}, \delta_{t+2}, \dots)$. Each weight, $h_i$, determines the agent's receptiveness, or eligibility, to learn from the TD error that occurs exactly $i$ steps after the initial stimulus. In this view, a return estimate, $\hat{G}_t$, is uniquely determined by the impulse response of a linear time-invariant system encoded by $(h_i)_{i=0}^{\infty}$. One possible interpretation of the recency heuristic, then, is a constraint on the impulse response such that it never increases after the initial stimulus. This gives us the following definition.

**Definition 3.1** (Weak Recency Heuristic)**.** *A return estimate satisfies the weak recency heuristic if and only if it has the form of Eq.* (9)*, and* $h_i \geq h_{i+1} \geq 0$ *holds for all* $i \geq 0$*.*

We show in Section 5 that this definition is highly related to the question of whether (and how fast) TD learning using this estimator converges in expectation, but it is slightly weaker than what is typically thought of as the recency heuristic. For instance, Sutton (1984, p. 94) is explicit that "Credit assigned should be a monotonically decreasing function of the time between action and reinforcement, approaching zero as this time approaches infinity." The credit-assignment function in Definition 3.1 is merely nonincreasing, and so we refer to it as the *weak* recency heuristic. Alternatively, we refer to the monotonically decreasing case as the *strong* recency heuristic, defined below.

**Definition 3.2** (Strong Recency Heuristic)**.** *A return estimate satisfies the strong recency heuristic if and only if it has the form of Eq.* (9)*, and* $h_i > h_{i+1} > 0$ *holds for all* $i \geq 0$*.*

Notice that Definition 3.2 implies Definition 3.1. We make the distinction between these more concrete with a few examples. The $\lambda$-return, used by TD($\lambda$), is the canonical example of the strong recency heuristic; its eligibility weights in Eq. (3) are strictly decreasing for any $\lambda \in (0, 1)$. In contrast, the $n$-step return remains equally receptive to the first $n$ TD errors, and then abruptly stops responding to the ones afterwards. However, these two updates are alike in that the weights never increase at any point: they both satisfy the weak recency heuristic. We could also imagine arbitrary weights in Eq. (9) that do not satisfy either definition of recency heuristic. For example, the inverted-U shape described by Klopf (1972) takes time to reach its peak value before falling back to zero, and thus violates Definitions 3.1 and 3.2. Similarly, we can take the standard spike-and-decay model of a $\lambda$-return and introduce a delay between the initial stimulus and the response. Both of these could exploit some known structure regarding the agent's environment, and may be more biologically plausible, but their mathematical implications are not yet known. These four examples are graphed in Figure 1. Notably, there are many more possibilities in Eq. (9), most of which have not yet been explored.

## 4 What Happens When the Recency Heuristic Is Violated?

We conduct an experiment to demonstrate that on-policy TD learning with a tabular value function can diverge when the recency heuristic is violated. This is surprising, since one view of the TD-error weights, $(h_i)_{i=0}^{\infty}$, is that they encode a belief over the time when rewards will arrive following a

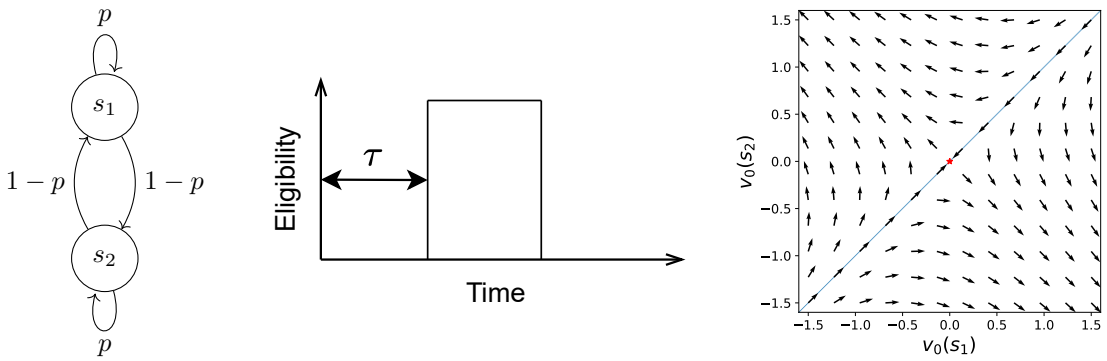


Figure 2: (Left) MRP for Counterexample 4.1; rewards are zero. (Center) Credit-assignment function for delayed TD(0). (Right) Expected update directions of Eq. (10) for $\tau = 1$, $\gamma = 0.9$, $p = 0.4$.

stimulus. Ideally, these weights could represent any shape for the credit-assignment function and the agent would still learn the correct value function, yet this does not appear to be the case.

We test perhaps the simplest possible example of non-recent credit assignment: an update based on a single, future TD error. More specifically, we generalize TD(0) by introducing a delay of $\tau \geq 0$:

$$v(S_t) \leftarrow V_t + \alpha_t \gamma^\tau \delta_{t+\tau} \,. \tag{10}$$

The impulse response for this method is generally given by $h_i = 1$ if $i = \tau$, and $h_i = 0$ otherwise. That is, the eligibility curve is a square pulse initiated exactly $\tau$ steps after the initial stimulus (see Figure 2, center). The operator corresponding to this update is $\boldsymbol{H} : \boldsymbol{v} \mapsto \boldsymbol{v} + (\gamma \boldsymbol{P}_\pi)^\tau (\boldsymbol{T}_\pi \boldsymbol{v} - \boldsymbol{v})$. The fixed point of this operator is $\boldsymbol{v}_\pi$ for any value of $\tau$ because $\boldsymbol{T}_\pi \boldsymbol{v}_\pi - \boldsymbol{v}_\pi = 0$.

Notice that this is a rather benign form of non-recent credit assignment; we are taking the simplest TD method and merely translating its impulse response along the time axis. More complex forms of non-recent credit assignment would consist of a superposition of multiple such updates, and so this example provides insight into other methods. Nevertheless, despite the simplicity of this method, we present a simple Markov reward process (MRP) that causes almost every value-function initialization to diverge away from $\boldsymbol{v}_\pi$.

**Counterexample 4.1.** *Consider a 2-state MRP with reward* $r(s, s') = 0$, $\forall\ s, s' \in \{s_1, s_2\}$. *Let* $p \in [0, 1]$ *be the self-transition probability (see Figure 2, left) and let* $\boldsymbol{v}_0$ *be the initial value function. If* $\tau = 1$, $\gamma = 0.9$, *and* $p = 0.4$, *then the TD update in Eq.* (10) *diverges whenever* $\boldsymbol{v}_0(s_1) \neq \boldsymbol{v}_0(s_2)$.

We give specific values of $\tau$, $\gamma$, and $p$ for the sake of the counterexample; however, it appears that divergence is inevitable for any $\tau > 0$ as $\gamma \to 1$ and $p \to 0$. The divergent behavior of the method is visualized in Figure 2 (right), where the arrows represent unit vectors pointing in the direction the expected update (i.e., $\boldsymbol{H}\boldsymbol{v} - \boldsymbol{v}$). Because the reward is zero for all transitions, $\boldsymbol{v}_\pi$ is the origin (red star) regardless of $\gamma$ and $p$. However, we see that every value-function initialization not on the blue line where $\boldsymbol{v}_0(s_1) = \boldsymbol{v}_0(s_2)$ progresses arbitrarily far away from the fixed point, $\boldsymbol{v}_\pi$.

Why does violating the recency heuristic in this easy problem cause divergence? The reason becomes more clear when we observe that $\gamma^\tau \delta_{t+\tau} = G_t^{(\tau+1)} - G_t^{(\tau)}$. Thus, an equivalent operator for Eq. (10) is $\boldsymbol{v} \mapsto \boldsymbol{v} + \boldsymbol{T}_\pi^{\tau+1} \boldsymbol{v} - \boldsymbol{T}_\pi^\tau \boldsymbol{v}$, whose worst-case contraction modulus is $1 + \gamma^{\tau+1} + \gamma^\tau$ by the triangle inequality—greater than 1. Although this does not automatically mean the operator will diverge, it does suggest that divergence is possible, and we see one instance of it here. It is important to note that this divergence is not due to sampling noise nor an uneven state distribution, as we are explicitly computing the expected result of the operator in both states. Furthermore, the phenomenon is not unique to this particular algorithm or problem, but generally arises whenever the weak recency heuristic is violated too much. We prove this formally in the next section.

## 5 Only Convex Returns Satisfy the Weak Recency Heuristic

Recall that convex returns are convex combinations of $n$-step returns: either compound returns or $n$-step returns themselves. In this section, we show this definition is logically equivalent to the weak recency heuristic; Definition 3.1 is satisfied if and only if a return estimate is convex (see Proposition 5.2).

To illuminate the role of the weak recency heuristic, we first justify the general return estimator in Eq. (9). In particular, we show that estimates of this form correspond to the largest set of linear operators suitable for TD learning. This allows us to later analyze how the properties of these operators are affected by the choice of the weights, $(h_i)_{i=0}^{\infty}$, especially when these weights do not satisfy the recency heuristic.

To produce a TD method in the form of Eq. (4) that converges to $\boldsymbol{v}_\pi$ under general conditions, the return estimate $\hat{G}_t$ must correspond to a maximum-norm contraction mapping, $\boldsymbol{H}$, with its unique fixed point at $\boldsymbol{v}_\pi$ (recall Section 2). In addition to these requirements, we want a *sample-realizable* operator in order to create an implementable TD method: one that can be constructed from any rewards or state values following time $t$. To match existing TD methods, we assume that this operator is linear with respect to these quantities, giving us the following definition.

**Definition 5.1.** *A sample-realizable linear operator has the form* $\boldsymbol{H}\boldsymbol{v} = \sum_{i=0}^{\infty} a_i(\gamma \boldsymbol{P}_\pi)^i \boldsymbol{r} + b_i(\gamma \boldsymbol{P}_\pi)^i \boldsymbol{v}$, *where* $(a_i)_{i=0}^{\infty}$ *and* $(b_i)_{i=0}^{\infty}$ *are bounded sequences of real numbers.*

This definition covers all possible operators based on return estimates that can be constructed from a linear combination of sampled experiences: i.e., $\hat{G}_t = \sum_{i=0}^{\infty} a_i \gamma^i R_{t+i} + b_i \gamma^i V_{t+i}$. However, the vast majority of these operators will not meet our convergence criteria. In the following proposition, we reduce the space of operators by identifying only those whose fixed point is exactly $\boldsymbol{v}_\pi$.

**Proposition 5.1.** *For every sample-realizable operator* $\boldsymbol{H}$ *whose fixed point is* $\boldsymbol{v}_\pi$, *there exists a sequence of real numbers* $(h_i)_{i=0}^{\infty}$ *such that*

$$\boldsymbol{H}\boldsymbol{v} = \boldsymbol{v} + \sum_{i=0}^{\infty} h_i (\gamma \boldsymbol{P}_\pi)^i (\boldsymbol{T}_\pi \boldsymbol{v} - \boldsymbol{v}) \,. \tag{11}$$

*If we let* $c_n \stackrel{def}{=} h_{n-1} - h_n$ *for* $n \geq 1$, *then* $\boldsymbol{H}$ *also has the equivalent form*

$$\boldsymbol{H}\boldsymbol{v} = \left(1 - \sum_{n=1}^{\infty} c_n\right) \boldsymbol{v} + \sum_{n=1}^{\infty} c_n \boldsymbol{T}_\pi^n \boldsymbol{v} \,. \tag{12}$$

*Proof.* See Appendix A.1. □

Notice that Eq. (11) corresponds exactly to the sample estimate in Eq. (9) that we considered in Section 3 when defining the weak recency heuristic. We refer to these as *linear* returns. Hence, every linear return with $\boldsymbol{v}_\pi$ as its fixed point is expressible as a weighted sum of either TD errors or $n$-step returns, without loss of generality.

We now have a generic operator that is both sample realizable and has the correct fixed point, but it is not necessarily a contraction mapping without any conditions on its weights, $(h_i)_{i=0}^{\infty}$. Eq. (12) expresses the operator in terms of the $n$-step Bellman operators, facilitating the analysis of its contraction properties. Because $\boldsymbol{P}_\pi$ is a stochastic matrix, we have $\|\boldsymbol{P}_\pi\|_\infty = 1$, which also implies that $\|\boldsymbol{T}_\pi^n \boldsymbol{v} - \boldsymbol{T}_\pi^n \boldsymbol{v}'\|_\infty \leq \gamma^n \|\boldsymbol{v} - \boldsymbol{v}'\|_\infty$, for any $\boldsymbol{v}, \boldsymbol{v}' \in \mathbb{R}^{|\mathcal{S}|}$. Thus, by the triangle inequality,

$$\|\boldsymbol{H}\boldsymbol{v} - \boldsymbol{H}\boldsymbol{v}'\|_\infty \leq \left( \left| 1 - \sum_{n=1}^{\infty} c_n \right| + \sum_{n=1}^{\infty} |c_n| \gamma^n \right) \|\boldsymbol{v} - \boldsymbol{v}'\|_\infty \,, \tag{13}$$

and the contraction modulus is therefore $\beta = |1 - \sum_{n=1}^{\infty} c_n| + \sum_{n=1}^{\infty} |c_n| \gamma^n$. The operator is a contraction mapping if and only if $\beta < 1$.

Notice that Eq. (12) consists of two terms: the original value function scaled by $1 - \sum_{n=1}^{\infty} c_n$, and a linear combination of $n$-step returns. The first term can be eliminated without loss of generality by normalizing the sum of weights, i.e., by adding the constraint that $\sum_{n=1}^{\infty} c_n = 1$. This is because the first term changes only the magnitude of the update, which can be absorbed into the step size, $\alpha_t$, in Eq. (8). With this constraint in place, it follows that the weight of the first TD error is $h_0 = 1$ because of the telescoping series: $h_0 = \sum_{n=1}^{\infty} h_{n-1} - h_n = \sum_{n=1}^{\infty} c_n = 1$. The operator is now an affine combination of $n$-step Bellman operators, and so we refer to such return estimates as *affine* returns. Note that, since we have $h_0 = 0$ in Eq. (10) when $\tau > 0$, the divergent return estimate in Counterexample 4.1 is *not* an affine return, although it is linear. Affine returns look identical to convex returns from Eq. (7), but they are more general because they allow for negatively weighted $n$-step returns. We depict the hierarchical relationship between linear, affine, convex, compound, and $n$-step returns in Figure 3, and summarize their operators and corresponding sample estimates in Table 1.

This analysis provides a hint of why counterexamples like the one in Section 4 are possible; negative weights increase the contraction modulus due to the absolute value in Eq. (13). It turns out that such negative weights coincide exactly with the time steps on which the weak recency heuristic is violated, and therefore only convex returns satisfy the heuristic, as we show in the next proposition.

**Proposition 5.2.** *An affine return satisfies the weak recency heuristic if and only if it is a convex return (i.e., a compound return or an n-step return).*

*Proof.* See Appendix A.2. □

An immediate corollary of the above is that the weak recency heuristic is a sufficient condition for convergence, since both compound returns and $n$-step returns are already known to correspond to contraction mappings (Watkins, 1989, Sec. 7.2). This stems from the fact that a convex combination of $n$-step returns, each of which is contractive with modulus $\gamma^n$, must also be contractive: i.e., $\sum_{n=1}^{\infty} c_n \gamma^n \leq \gamma < 1$ for every choice of nonnegative weights that sum to one. In this view, the weak recency heuristic can be seen as a convergence test for TD learning, and explains some of its utility in computational RL: divergence is impossible under this heuristic.

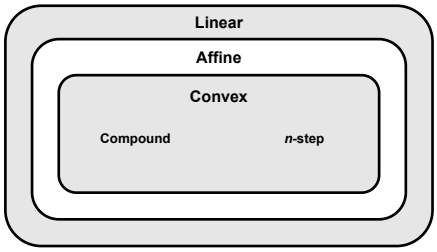


Figure 3: Hierarchical relationship between different return estimators. A return satisfies the weak recency heuristic if and only if it is a convex return: i.e., a compound or $n$-step return.

On the other hand, violating the weak recency heuristic increases the contraction modulus of the return estimator, with divergence possible if the violation becomes too extreme (e.g., Counterexample 4.1). This is because any time an $n$-step return has a negative weight, another $n$-step return must have a larger positive weight to counterbalance it and ensure the weights sum to 1 overall. This necessarily increases the contraction modulus in Eq. (13) due to the absolute value, underscoring yet another benefit of the weak recency heuristic. A convex return is not only guaranteed to converge regardless of its weights, but also has a faster contraction than a nonconvex (affine) return constructed from the same $n$-step returns.

## 6 Are Monotonically Decreasing Weights Necessary?

So far, we have focused on the weak recency heuristic: when the eligibility weights are nonincreasing. However, as we discussed in Section 3, the connotation of the recency heuristic is often that of strictly decreasing TD-error weights, i.e., the strong recency heuristic (Definition 3.2). This is why, for example, $\lambda$-returns are more strongly associated with a recency heuristic than $n$-step returns are. Does

| Name | Operator | Sample Estimate | Conditions |
|---|---|---|---|
| Linear | $\left(1-\sum_{n=1}^{\infty} c_n\right)\boldsymbol{v} + \sum_{n=1}^{\infty} c_n \boldsymbol{T}_\pi^n \boldsymbol{v}$ | $\left(1-\sum_{n=1}^{\infty} c_n\right) V_t + \sum_{n=1}^{\infty} c_n G_t^n$ | None |
| Affine | $\sum_{n=1}^{\infty} c_n \boldsymbol{T}_\pi^n \boldsymbol{v}$ | $\sum_{n=1}^{\infty} c_n G_t^n$ | $\sum_{n=1}^{\infty} c_n = 1$ and $\sum_{n=1}^{\infty} \lvert c_n \rvert \gamma^n < 1$ |
| Convex | $\sum_{n=1}^{\infty} c_n \boldsymbol{T}_\pi^n \boldsymbol{v}$ | $\sum_{n=1}^{\infty} c_n G_t^n$ | Affine and $c_n \geq 0, \forall\, n \geq 1$ |
| Compound | $\sum_{n=1}^{\infty} c_n \boldsymbol{T}_\pi^n \boldsymbol{v}$ | $\sum_{n=1}^{\infty} c_n G_t^n$ | Convex and $\exists\, c_i, c_j > 0$ |
| $n$-step | $\boldsymbol{T}_\pi^n v$ | $G_t^n$ | $n \geq 1$ |

Table 1: Summary of operators and sample estimates for the return estimators in Figure 3.

this distinction between weak and strong recency heuristics matter in practice? In this section, we conduct experiments indicating that the answer is yes, but in a surprising way; the smoothness of the weights do not appear to be significant, but the strong recency heuristic does imply that the return estimate consists of infinitely many $n$-step returns, which empirically improves credit assignment.

To test the question of whether the smoothness of the TD-error weights matters, we introduce the *sparse* $\lambda$-return, defined as

$$G_t^{\lambda,m} \stackrel{\text{def}}{=} \sum_{i=0}^{\infty} \gamma^i \lambda^{\lfloor \frac{i+m-1}{m} \rfloor} \delta_{t+i} = (1-\lambda) \sum_{k=1}^{\infty} \lambda^{k-1} G_t^{(m(k-1)+1)}\,, \tag{14}$$

where $m \geq 1$. The contraction modulus of this return is $\beta = \gamma(1-\lambda)\,/\,(1-\gamma^m\lambda)$. When $m = 1$, we simply recover the standard exponential decay of the $\lambda$-return from Eq. (6). However, for $m > 1$, the TD-error weights no longer satisfy the strong recency heuristic as they become more stepwise (see Figure 5). This implies that every $m-1$ out of $m$ $n$-step returns have zero weight. For example, setting $m = 2$ generates the TD-error weight sequence $(1,\ \lambda,\ \lambda,\ \lambda^2,\ \lambda^2,\ \ldots)$, which produces an exponential average of the odd $n$-step returns: $(G_t^{(1)}, G_t^{(3)}, G_t^{(5)}, G_t^{(7)}, G_t^{(9)}, \ldots)$. The reason we choose this form is because it isolates the effects of the two recency heuristics by keeping the type of weighted average consistent (i.e., exponential). If monotonicity is beneficial to learning, then we would expect to observe a performance degradation for sparse $\lambda$-returns ($m > 1$) compared to dense ($m = 1$).

Our experiment setup is a discounted variation ($\gamma = 0.99$) of the 19-state random walk from Sutton & Barto (2018, Sec. 12.1). In this environment, each episode starts with the agent in the center of a linear chain of 19 connected states (see Figure 4). The agent can move either left or right, and its behavior is fixed such that it randomly chooses either action with equal probability. Reaching either end of the chain terminates the episode and yields a reward: $-1$ for the left or $+1$ for the right.

We test three different degrees of sparsity for the $\lambda$-returns, adjusting $\lambda$ for each return to maintain the same contraction modulus in all cases: $(\lambda, m) \in \{(0.9, 1), (0.75, 3), (0.65, 5)\}$. The agents are trained for 10 episodes by applying offline value backups of the form Eq. (4) to every experience at the end of each episode. In Figure 6, we plot the root-mean-square (RMS) error, $\|\boldsymbol{v} - \boldsymbol{v}_\pi\|_2$,

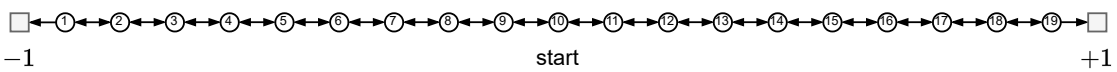


Figure 4: The 19-state random walk (Sutton & Barto, 2018, Sec. 12.1).

averaged over the 10 episodes versus the step size, $\alpha$, for each return. The final results are averaged over 400 trials with 95% confidence intervals indicated by shaded regions. Code is available online.[1]

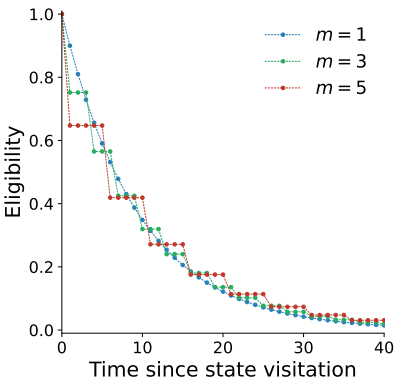


Figure 5: Impulse responses of $\lambda$-returns with varying degrees of sparsity.

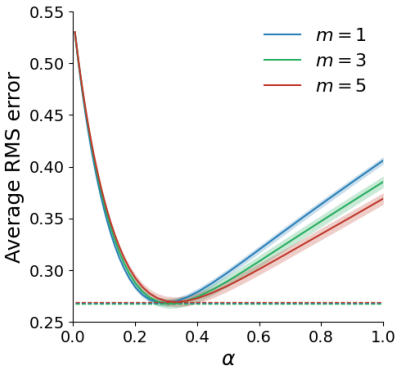


Figure 6: Random-walk performance of $\lambda$-returns with varying degrees of sparsity.

Because the three returns all have the same contraction modulus (i.e., expected convergence rate), their performance is nearly the same for small values of $\alpha$ which are able to average out the noise in the updates. Likewise, the returns share the same lowest error, as indicated by the dashed horizontal lines in Figure 6. However, as $\alpha$ gets larger, their performance begins to separate, achieving lower average error as the sparsity of the $\lambda$-return increases. Thus, even though the eligibility curves become more step-like as the sparsity is increased and they violate the strong recency heuristic, the overall performance of the return improves. This demonstrates that the monotonicity of the eligibility curves does not directly factor into the performance of the return estimators.

The main reason for the sparse $\lambda$-return's improvement appears to be that its eligibility initially decays faster than that of the dense $\lambda$-return, but then slower as time goes on (see Figure 5). This gives the eligibility curve a long-tailed characteristic which, in turn, propagates credit back in time more quickly. In fact, every return that satisfies the strong recency heuristic must have a similar characteristic, because Definition 3.2 implies that $c_n = h_{n-1} - h_n > 0$ for all $n \geq 1$, and thus Eq. (7) must correspond to a positively weighted average of infinitely many $n$-step returns. Although this property is not unique to the strong recency heuristic (e.g., the sparse $\lambda$-return has it but does not satisfy Definition 3.2), it does suggest a practical significance for this heuristic: it implies a longer horizon for credit assignment.

However, any benefit of a longer credit-assignment horizon is contingent on controlling the variance of the return. Fortunately, as we show in the following proposition, a long-tailed eligibility curve does not increase the worst-case variance when the contraction modulus is held constant.

**Proposition 6.1.** *Let* $\kappa_t \overset{def}{=} \max_{i,j \geq 0} \mathrm{Cov}[\delta_{t+i}, \delta_{t+j} \mid S_t]$. *The worst-case conditional variance of any convex return* $G_t^{\mathbf{c}}$ *with contraction modulus* $\beta$ *has the bound*

$$\mathrm{Var}[G_t^{\mathbf{c}} \mid S_t] \leq \left(\frac{1-\beta}{1-\gamma}\right)^2 \kappa_t \,. \tag{15}$$

*Proof.* See Appendix A.3. □

This bound is rather loose, but it is general. Eq. (15) implies that averages of $n$-step returns always have finite variance, even as the $n$-step returns become arbitrarily long. Furthermore, this upper bound depends only on the contraction modulus of the return itself and not the chosen weights for the average. Since the contraction modulus is proportional to the worst-case bias of the return by Eq. (13), we see that both the worst-case bias and worst-case variance of the $\lambda$-returns in our previous experiment remain the same regardless of sparsity. Thus, compound returns with a

[1] https://github.com/brett-daley/recency-heuristic

long-tailed eligibility curve are able to assign credit more quickly without negatively impacting the bias-variance trade-off[2] (at least, in a worst-case sense).

To test the effect of a longer credit-assignment horizon under a controlled contraction modulus, we repeat the previous random-walk experiment but with *truncated* $\lambda$-returns: $G^{\lambda}_{t:t+N} \stackrel{\text{def}}{=} V_t + \sum_{i=0}^{N-1} (\gamma\lambda)^i \delta_{t+i} = (1-\lambda)\sum_{n=1}^{N-1} \lambda^{n-1} G^{(n)}_t + \lambda^{N-1} G^{(N)}_t$, where $N \geq 1$ is the truncation length. The contraction modulus of this return is $\beta = \big((1-\gamma)(\gamma\lambda)^N + \gamma(1-\lambda)\big) \,/\, (1-\gamma\lambda)$. The eligibility curve for this return is a monotonically decreasing function, up until time $N$ when it abruptly falls to zero (see Figure 7). As $N \to \infty$, we recover the true $\lambda$-return, Eq. (6). We test three variants of this return: $(\lambda, N) \in \{(0.99, 10), (0.93, 20), (0.9, \infty)\}$. As before, all of these values are chosen to produce approximately the same contraction modulus. We plot the average RMS error in Figure 8, again averaged over 400 trials with 95% confidence intervals shaded. The performance is roughly identical when $\alpha$ is small, since the same contraction modulus guarantees the same expected performance. However, as $\alpha$ gets larger, the truncated returns perform poorly compared to the full $\lambda$-return. This suggests that the performance of the returns is strongly tied to longer $n$-step returns in the average, but only when the contraction moduli are equalized. This also supports our earlier hypothesis that the results observed with the sparse $\lambda$-returns in Figure 6 are due to their long-tail eligibility curves and not some other property such as monotonicity.

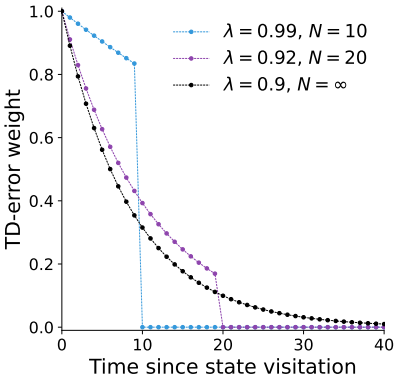


Figure 7: Eligibility curves of $\lambda$-returns with varying degrees of truncation.

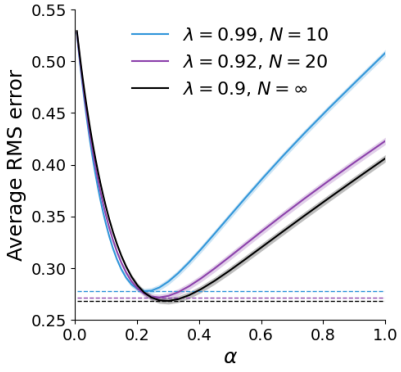


Figure 8: Random-walk performance of $\lambda$-returns with varying degrees of truncation.

To summarize, satisfying the strong recency heuristic creates a compound return consisting of infinitely many $n$-step returns—a long-tailed eligibility curve. This improves the effective window of credit assignment without exacerbating variance (in a conservative sense), as long as the contraction modulus is held constant. However, this property is not unique to the strong recency heuristic; for instance, sparse $\lambda$-returns violate this heuristic, but are still averages of infinitely many $n$-step returns, and outperform dense $\lambda$-returns in Figure 6. These insights help explain why smooth averages like the $\lambda$-return are often effective in practice, even if not strictly necessary for good performance.

## 7 Off-Policy Learning and Other Extensions

The weak recency heuristic is closely tied to an open problem on the convergence of off-policy eligibility traces (Daley et al., 2023, Sec. 5.3). Off-policy learning occurs whenever the agent's policy for action selection, $b$, differs from the policy for return estimation, $\pi$. Let $\rho_{t+i} \stackrel{\text{def}}{=} \pi(A_{t+i}|S_{t+i}) \,/\, b(A_{t+i}|S_{t+i})$ be the importance-sampling ratio. Daley et al. (2023) proved that satisfying $h_i \rho_{t+i+1} \geq h_{i+1} \geq 0$, $\forall\, i \geq 0$, is sufficient for the off-policy update analogous to Eq. (9) to converge to $\boldsymbol{v}_\pi$, where the TD-error weights can generally be *trajectory aware* (i.e., dependent on past state-action pairs). The open problem is to determine whether this condition is necessary as well.

[2]In fact, it is likely such long-tailed returns have a *positive* impact on the bias-variance trade-off by reducing variance, under an additional assumption that the TD-error variances are roughly uniform (see Daley et al., 2024, Sec. 6).

Notably, the condition is exactly the off-policy generalization of the weak recency heuristic (Definition 3.1), since $\rho_{t+i} = 1$ when $\pi = b$. Based on the analysis in Section 5, we know that it is sometimes possible to violate this heuristic and still converge, and so *the condition is sufficient but not necessary*. We provide more details in Appendix B, where we also extend our theory to state-dependent eligibilities (e.g., Yu, 2012; White & White, 2016) and function approximation (Tsitsiklis & Van Roy, 1997). These results show that the possibility of divergence like in Counterexample 4.1 is a general phenomenon of TD learning when not utilizing a recency heuristic.

## 8 Conclusion

Although non-recent credit assignment should theoretically be possible and useful in certain learning environments, it does not seem readily compatible with our current formulation of TD learning. In particular, violating the recency heuristic manifests as negative weights on some of the $n$-step components of the return target. These negative weights appear to counteract learning by increasing the contraction modulus, without offering a clear benefit to learning, and potentially culminating in divergence as demonstrated by Counterexample 4.1. The fact that divergence is possible in such a favorable setting—an on-policy, tabular MRP with fully observable states—points to the severity of this issue. Indeed, as we discussed in Section 7, this issue persists in more challenging settings including off-policy learning and function approximation. Successfully implementing new forms of credit assignment that do not strictly follow the recency heuristic will likely require rethinking how we formulate the reinforcement signal in computational RL. Our theory will provide a good starting point for algorithmic development in this direction.

Another major finding is that the recency heuristic is not merely a simple protocol for addressing the temporal credit-assignment problem, but also has intrinsic importance for learning value functions. The existence of diverging counterexamples illuminates the critical role of nonincreasing weights on the TD errors—the weak recency heuristic. The logical equivalence between this heuristic and the return estimate's ability to be expressed as a convex combination of $n$-step returns unifies two fundamental yet seemingly disparate ideas in RL. More specifically, convex returns were the most general return estimates for TD learning identified before our work, and so it is surprising to find they coincide exactly with another foundational concept in RL: the recency heuristic. This appears to be a novel, unifying perspective between the forward and backward views of TD learning with arbitrary return estimates. In the off-policy setting, the weak recency heuristic is equivalent to the convergence condition for eligibility traces discovered by Daley et al. (2023), providing more evidence for its importance in learning value functions.

Finally, our results help to further explain the strong empirical performance and continued popularity of TD($\lambda$), along with its many variants, for nearly four decades. Our experiments suggest that the smoothness of TD($\lambda$)'s exponential decay is not directly responsible for this success; rather, all compound returns (including $\lambda$-returns) that average an infinite number of $n$-step returns are able to distribute credit over a longer period without exacerbating the maximum bias or variance. These results confirm the intuition that "the fading strategy [of TD($\lambda$)] is often the best [versus $n$-step TD methods]" (Sutton & Barto, 2018, p. 304), though non-exponential fading strategies are also viable.

### Acknowledgments

We thank Rich Sutton for helpful discussions and insights. This research is supported in part by the Alberta Machine Intelligence Institute (Amii), the Natural Sciences and Engineering Research Council of Canada (NSERC), and the Canada CIFAR AI Chair Program.

## References


Andrew G. Barto, Richard S. Sutton, and Charles W. Anderson. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics*, 13(5):834–846, 1983.

Richard Bellman. *Dynamic Programming*. Princeton University Press, 1957.

Dimitri P. Bertsekas and John N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996.

Pawel Cichosz. Truncating temporal differences: On the efficient implementation of TD($\lambda$) for reinforcement learning. *Journal of Artificial Intelligence Research*, 2:287–318, 1995.

Brett Daley and Christopher Amato. Reconciling $\lambda$-returns with experience replay. In *Neural Information Processing Systems (NeurIPS)*, 2019.

Brett Daley, Martha White, Christopher Amato, and Marlos C. Machado. Trajectory-aware eligibility traces for off-policy reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2023.

Brett Daley, Martha White, and Marlos C. Machado. Averaging $n$-step returns reduce variance in reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2024.

Dhawal Gupta, Scott M Jordan, Shreyas Chaudhari, Bo Liu, Philip S. Thomas, and Bruno Castro da Silva. From past to future: Rethinking eligibility traces. *arXiv*, 2312.12972, 2023.

Jean Harb and Doina Precup. Investigating recurrence and eligibility traces in deep Q-networks. In *NeurIPS Deep Reinforcement Learning Workshop*, 2016.

Anna Harutyunyan, Marc G. Bellemare, Tom Stepleton, and Rémi Munos. Q($\lambda$) with off-policy corrections. In *International Conference on Algorithmic Learning Theory (ALT)*, 2016.

Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1-2):99–134, 1998.

Michael J. Kearns and Satinder Singh. Bias-variance error bounds for temporal difference updates. In *Conference on Learning Theory (COLT)*, 2000.

A. Harry Klopf. Brain function and adaptive systems: A heterostatic theory. Technical report, Air Force Cambridge Research Laboratories, 1972.

George Konidaris, Scott Niekum, and Philip S. Thomas. TD$_\gamma$: Re-evaluating complex backups in temporal difference learning. In *Neural Information Processing Systems (NeurIPS)*, 2011.

Tadashi Kozuno, Yunhao Tang, Mark Rowland, Rémi Munos, Steven Kapturowski, Will Dabney, Michal Valko, and David Abel. Revisiting Peng's Q($\lambda$) for modern reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2021.

A. Rupam Mahmood, Huizhen Yu, and Richard S. Sutton. Multi-step off-policy learning without importance sampling ratios. *arXiv*, 1702.03006, 2017.

Marvin L. Minsky. Steps toward artificial intelligence. In *Proceedings of the Institute of Radio Engineers (IRE)*, 1961.

Seyed Sajad Mousavi, Michael Schukat, Enda Howley, and Patrick Mannion. Applying Q($\lambda$)-Learning in deep reinforcement learning to play Atari games. In *AAMAS Adaptive Learning Agents Workshop*, pp. 1–6, 2017.

Rémi Munos, Tom Stepleton, Anna Harutyunyan, and Marc G. Bellemare. Safe and efficient off-policy reinforcement learning. In *Neural Information Processing Systems (NeurIPS)*, 2016.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. In *International Conference on Learning Representations (ICLR)*, 2015.

Richard S. Sutton. *Temporal Credit Assignment in Reinforcement Learning*. PhD thesis, University of Massachusetts Amherst, 1984.

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

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2nd edition, 2018.

Yunhao Tang, Mark Rowland, Rémi Munos, Bernardo Ávila Pires, and Will Dabney. Off-policy distributional Q($\lambda$): Distributional RL without importance sampling. *arXiv*, 2402.05766, 2024.

Philip S. Thomas, Scott Niekum, Georgios Theocharous, and George Konidaris. Policy evaluation using the $\Omega$-return. *Neural Information Processing Systems (NeurIPS)*, 2015.

John N. Tsitsiklis and Benjamin Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690, 1997.

Harm van Seijen. Effective multi-step temporal-difference learning for non-linear function approximation. *arXiv*, 1608.05151, 2016.

Christopher J. C. H. Watkins. *Learning from Delayed Rewards*. PhD thesis, University of Cambridge, 1989.

Martha White and Adam White. A greedy approach to adapting the trace parameter for temporal difference learning. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pp. 557–565, 2016.

Huizhen Yu. Least squares temporal difference methods: An analysis under general conditions. *SIAM Journal on Control and Optimization*, 50(6):3310–3343, 2012.

# A Proofs

This section contains the omitted proofs of all propositions in the paper.

## A.1 Proof of Proposition 5.1

**Proposition 5.1.** *For every sample-realizable operator $\boldsymbol{H}$ whose fixed point is $\boldsymbol{v}_\pi$, there exists a sequence of real numbers $(h_i)_{i=0}^{\infty}$ such that*

$$\boldsymbol{H}\boldsymbol{v} = \boldsymbol{v} + \sum_{i=0}^{\infty} h_i(\gamma \boldsymbol{P}_\pi)^i(\boldsymbol{T}_\pi \boldsymbol{v} - \boldsymbol{v})\,. \tag{11}$$

*If we let $c_n \stackrel{def}{=} h_{n-1} - h_n$ for $n \geq 1$, then $\boldsymbol{H}$ also has the equivalent form*

$$\boldsymbol{H}\boldsymbol{v} = \left(1 - \sum_{n=1}^{\infty} c_n\right)\boldsymbol{v} + \sum_{n=1}^{\infty} c_n \boldsymbol{T}_\pi^n \boldsymbol{v}\,. \tag{12}$$

*Proof.* It is given that $\boldsymbol{H}$ is sample realizable. Without loss of generality, we consider an alternative parameterization of Definition 5.1 that spans the same space of linear operators. There exist sequences of real numbers $(x_i)_{i=0}^{\infty}$ and $(y_i)_{i=0}^{\infty}$ such that

$$\begin{aligned}
\boldsymbol{H}\boldsymbol{v} &= \boldsymbol{v} + \sum_{i=0}^{\infty}(\gamma \boldsymbol{P}_\pi)^i\Big[x_i(\boldsymbol{r} + \gamma \boldsymbol{P}_\pi \boldsymbol{v}) - y_i \boldsymbol{v}\Big] \\
&= \boldsymbol{v} + \sum_{i=0}^{\infty}(\gamma \boldsymbol{P}_\pi)^i\Big[x_i(\boldsymbol{r} + \gamma \boldsymbol{P}_\pi \boldsymbol{v} - \boldsymbol{v}) + (x_i - y_i)\boldsymbol{v}\Big] \\
&= \boldsymbol{v} + \sum_{i=0}^{\infty}(\gamma \boldsymbol{P}_\pi)^i\Big[x_i(\boldsymbol{T}_\pi \boldsymbol{v} - \boldsymbol{v}) + (x_i - y_i)\boldsymbol{v}\Big] \\
&= \boldsymbol{v} + \sum_{i=0}^{\infty} x_i(\gamma \boldsymbol{P}_\pi)^i(\boldsymbol{T}_\pi \boldsymbol{v} - \boldsymbol{v}) + \sum_{i=0}^{\infty}(x_i - y_i)(\gamma \boldsymbol{P}_\pi)^i \boldsymbol{v}\,.
\end{aligned}$$

Because $\boldsymbol{T}_\pi \boldsymbol{v}_\pi = \boldsymbol{v}_\pi$, it follows that $\boldsymbol{H}\boldsymbol{v}_\pi = \boldsymbol{v}_\pi + \sum_{i=0}^{\infty}(x_i - y_i)(\gamma \boldsymbol{P}_\pi)^i \boldsymbol{v}_\pi$. To ensure that $\boldsymbol{v}_\pi$ is the fixed point of $\boldsymbol{H}$ (i.e., that $\boldsymbol{H}\boldsymbol{v}_\pi = \boldsymbol{v}_\pi$), we must make the remaining sum zero. However, this happens only when $x_i = y_i$, $\forall\ i \geq 0$. Thus, we substitute $h_i = x_i$ and $h_i = y_i$ to get Eq. (11).

To derive Eq. (12), we apply the fact that $h_i = \sum_{n=i+1}^{\infty} c_n$ due to the telescoping series. We complete the proof by rewriting Eq. (11) as

$$\begin{aligned}
\boldsymbol{H}\boldsymbol{v} &= \boldsymbol{v} + \sum_{i=0}^{\infty}\left(\sum_{n=i+1}^{\infty} c_n\right)(\gamma \boldsymbol{P}_\pi)^i(\boldsymbol{T}_\pi \boldsymbol{v} - \boldsymbol{v}) \\
&= \boldsymbol{v} + \sum_{n=1}^{\infty} c_n \sum_{i=0}^{n-1}(\gamma \boldsymbol{P}_\pi)^i(\boldsymbol{T}_\pi \boldsymbol{v} - \boldsymbol{v}) \\
&= \boldsymbol{v} + \sum_{n=1}^{\infty} c_n(\boldsymbol{T}_\pi^n \boldsymbol{v} - \boldsymbol{v}) \\
&= \left(1 - \sum_{n=1}^{\infty} c_n\right)\boldsymbol{v} + \sum_{n=1}^{\infty} c_n \boldsymbol{T}_\pi^n \boldsymbol{v}\,.
\end{aligned}$$

The second equality interchanged the sums using the rule $\sum_{i=0}^{\infty}\sum_{n=i+1}^{\infty} = \sum_{n=1}^{\infty}\sum_{i=0}^{n-1}$. The third equality followed from the $n$-step Bellman operator expansion: $\boldsymbol{T}_\pi^n \boldsymbol{v} = \boldsymbol{v} + \sum_{i=0}^{n-1}(\gamma \boldsymbol{P}_\pi)^i(\boldsymbol{T}_\pi \boldsymbol{v} - \boldsymbol{v})$. □

### A.2 Proof of Proposition 5.2

**Proposition 5.2.** *An affine return satisfies the weak recency heuristic if and only if it is a convex return (i.e., a compound return or an $n$-step return).*

*Proof.* Recall that $c_n = h_{n-1} - h_n$. Therefore, the affine operator from Eq. (12) is equal to

$$\boldsymbol{H}\boldsymbol{v} = \sum_{n=1}^{\infty} (h_{n-1} - h_n)\, \boldsymbol{T}_\pi^n \boldsymbol{v}\,. \tag{16}$$

If the weak recency heuristic (Definition 3.1) holds, then we have $h_{n-1} \geq h_n \implies h_{n-1} - h_n \geq 0$, for all $n \geq 1$. Thus, Eq. (16) is a convex combination of $n$-step returns, because we have $\sum_{n=1}^{\infty} h_{n-1} - h_n = \sum_{n=1}^{\infty} c_n = 1$ for an affine return.

To complete the proof, we also show the contrapositive. Consider an affine return that is not a convex combination of $n$-step returns. Consequently, it must have at least one negatively weighted $n$-step return: there exists some $k \geq 1$ such that $c_k < 0$. However, this implies that $h_{k-1} - h_k < 0$, and therefore $h_{k-1} < h_k$, so the weak recency heuristic is violated. We conclude that an affine return satisfies the weak recency heuristic if and only if it is a convex return. □

### A.3 Proof of Proposition 6.1

**Proposition 6.1.** *Let $\kappa_t \stackrel{def}{=} \max_{i,j\geq 0} \mathrm{Cov}[\delta_{t+i}, \delta_{t+j} \mid S_t]$. The worst-case conditional variance of any convex return $G_t^{\mathbf{c}}$ with contraction modulus $\beta$ has the bound*

$$\mathrm{Var}[G_t^{\mathbf{c}} \mid S_t] \leq \left(\frac{1-\beta}{1-\gamma}\right)^2 \kappa_t\,. \tag{15}$$

*Proof.* First, note that $\mathrm{Var}[\hat{G}_t \mid S_t] = \mathrm{Var}[\hat{G}_t - V_t \mid S_t]$ for any return estimate, $\hat{G}_t$, since $V_t$ is deterministic given state $S_t$. This allows us to derive an upper bound on the covariance between two $n$-step returns with lengths $n_1$ and $n_2$ using Eq. (5):

$$\begin{aligned}
\mathrm{Cov}[G_t^{(n_1)}, G_t^{(n_2)} \mid S_t] &= \mathrm{Cov}\left[\sum_{i=0}^{n_1-1} \gamma^i \delta_{t+i}, \sum_{j=0}^{n_2-1} \gamma^j \delta_{t+j} \,\middle|\, S_t\right] \\
&= \sum_{i=0}^{n_1-1}\sum_{j=0}^{n_2-1} \gamma^{i+j} \mathrm{Cov}[\delta_{t+i}, \delta_{t+j} \mid S_t] \\
&\leq \sum_{i=0}^{n_1-1}\sum_{j=0}^{n_2-1} \gamma^{i+j} \kappa_t \\
&= \Gamma(n_1)\,\Gamma(n_2)\,\kappa_t\,,
\end{aligned}$$

where $\Gamma(n) \stackrel{\text{def}}{=} (1-\gamma^n)\,/\,(1-\gamma)$ is the $n$-th partial sum of the geometric series. Because $\sum_{n=1}^{\infty} c_n = 1$ and $\beta = \sum_{n=1}^{\infty} c_n \gamma^n$ for a convex return, we also have

$$\sum_{n=1}^{\infty} c_n\, \Gamma(n) = \sum_{n=1}^{\infty} c_n \left(\frac{1-\gamma^n}{1-\gamma}\right) = \frac{1 - \sum_{n=1}^{\infty} c_n \gamma^n}{1-\gamma} = \frac{1-\beta}{1-\gamma}\,.$$

Therefore, we derive the following upper bound on the variance of a convex return:

$$\begin{aligned}
\mathrm{Var}[G_t^{\boldsymbol{c}} \mid S_t] &= \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} \mathrm{Cov}[c_i G_t^{(i)}, c_j G_t^{(j)} \mid S_t] \\
&= \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} c_i c_j \mathrm{Cov}[G_t^{(i)}, G_t^{(j)} \mid S_t] \\
&\leq \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} c_i c_j \, \Gamma(i)\, \Gamma(j)\, \kappa_t \\
&= \left(\frac{1-\beta}{1-\gamma}\right)^2 \kappa_t \,,
\end{aligned}$$

which completes the proof. □

# B Extensions

This section contains extensions of our theory to off-policy learning, state- or trajectory-dependent eligibility traces, and function approximation.

## B.1 Function Approximation

Our results easily generalize to the case where the value function is approximated by a linear parametric function: $V_t = \boldsymbol{x}_t^\top \boldsymbol{w}_t$, where $\boldsymbol{w}_t \in \mathbb{R}^d$ is the value-function weights, and $\boldsymbol{x}_t \in \mathbb{R}^d$ is a feature vector corresponding to state $S_t$. Because $\frac{\partial}{\partial \boldsymbol{w}} V_t\big|_{\boldsymbol{w}=\boldsymbol{w}_t} = \boldsymbol{x}_t$, the semi-gradient TD update becomes

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t + \alpha_t \left(\hat{G}_t - V_t\right) \boldsymbol{x}_t \,.$$

Let $\boldsymbol{X} \in \mathbb{R}^{|\mathcal{S}| \times d}$ be the matrix whose rows correspond to the feature vectors for every state in $\mathcal{S}$. Because $\boldsymbol{H}\boldsymbol{v}$ generally cannot be represented exactly by the function approximator, the estimate $\hat{G}_t$ corresponds to a composite linear operator $\boldsymbol{\Pi}\boldsymbol{H}$, where $\boldsymbol{\Pi}$ is a projection operator onto the set $\{\boldsymbol{X}\boldsymbol{w} \mid \boldsymbol{w} \in \mathbb{R}^d\}$ under the state weighting induced by the MDP's stationary distribution (Tsitsiklis & Van Roy, 1997). Furthermore, $\boldsymbol{\Pi}$ is nonexpansive, linear, and independent of $\boldsymbol{w}_t$ (Tsitsiklis & Van Roy, 1997, proof of Lemma 6); hence, if $\boldsymbol{H}$ is a contraction mapping, then so is $\boldsymbol{\Pi}\boldsymbol{H}$ with the same maximum contraction modulus. This implies that violating the weak recency heuristic too much can still increase the contraction modulus and cause divergence, just like in Counterexample 4.1.

In the case of *nonlinear* function approximation, the existence of counterexamples is certain, as even TD(0) diverges for at least one function (Tsitsiklis & Van Roy, 1997, Fig. 1).

## B.2 State-Dependent Eligibility Traces

The general return estimate considered by our work, Eq. (9), determines the eligibility weights solely based on the elapsed time since the initial state. Additionally, we can have weights that depend on the actual states experienced on each time step (e.g., Yu, 2012; White & White, 2016). A return estimate in this case has the form

$$\hat{G}_t = V_t + \sum_{i=0}^{\infty} h_i(S_{t+i})\, \gamma^i \delta_{t+i} \,, \tag{17}$$

where $h_i \colon \mathcal{S} \to \mathbb{R}$ is now a weighting function over the state space. This estimate satisfies the weak recency heuristic if

$$h_i(s) \geq h_{i+1}(s') \geq 0 \,, \quad \forall\, i \geq 0 \,, \quad \forall\, s, s' \in \mathcal{S} \,.$$

The operator corresponding to Eq. (17) is $(\boldsymbol{H}\boldsymbol{v})(s) = \mathbb{E}_\pi[\hat{G}_t \mid S_t = s]$, i.e., a convex combination of the estimates in Eq. (17). Therefore, it too satisfies the weak recency heuristic, except that the weight at each time step is an average of random variables and cannot be explicitly written without additional information about the MDP. Other than this minor difference, we see that the results for state-based eligibility curves are analogous to the strictly time-based eligibility curves discussed in our paper.

### B.3 Off-Policy Learning and Trajectory-Aware Eligibility Traces

A further generalization of the state-dependent eligibility traces discussed in the previous section is trajectory-aware eligibility traces (Daley et al., 2023). These have been studied in the context of off-policy learning with action values, where the agent estimates the action-value function $q_\pi(s,a) \stackrel{\text{def}}{=} \mathbb{E}[G_t \mid (S_t, A_t) = (s,a)]$. Additionally, it is assumed that the agent samples actions from a behavior policy, $b$, that differs from the target policy, $\pi$. The off-policy bias resulting from the mismatch between behavior and target distributions must be corrected to converge to $\boldsymbol{q}_\pi$.

Let $\mathcal{F}_{t:t+i} \stackrel{\text{def}}{=} (S_{t+j}, A_{t+j})_{j=0}^{i}$ be the partial history of the MDP from time $t$ to $t+i$. Additionally, let $\delta_t^\pi \stackrel{\text{def}}{=} R_t + \gamma \bar{V}_{t+1} - q(S_t, A_t)$ denote the mean TD error using action values, where $\bar{V}_t \stackrel{\text{def}}{=} \sum_{a' \in \mathcal{A}} \pi(a'|S_t)\, q(S_t, a')$. A trajectory-aware return estimate has the form

$$\hat{G}_t = V_t + \sum_{i=0}^{\infty} h_i(\mathcal{F}_{t:t+i})\, \gamma^i \delta_{t+i}\,,$$

where $h_i \colon (\mathcal{S} \times \mathcal{A})^i \to \mathbb{R}$ is a weighting function over partial histories. The corresponding operator is $(\boldsymbol{H}\boldsymbol{q})(s,a) = \mathbb{E}_\mu[\hat{G}_t \mid (S_t, A_t) = (s,a)]$. For the operator to converge to $\boldsymbol{q}_\pi$, it is sufficient to satisfy the following condition (Daley et al., 2023, Theorem 5.2):

$$h_i(\mathcal{F}_{t:t+i})\, \rho_{t+i+1} \geq h_{i+1}(\mathcal{F}_{t:t+i+1}) \geq 0\,, \;\; \forall\, i \geq 0\,, \;\; \forall\, t \geq 0\,, \tag{18}$$

where $\rho_{t+i} \stackrel{\text{def}}{=} \pi(A_{t+i}|S_{t+i}) \,/\, b(A_{t+i}|S_{t+i})$ is the importance-sampling ratio. An open problem is whether this condition is necessary in addition to being sufficient (Daley et al., 2023, Sec. 5.3). Rather interestingly, this condition is the off-policy analog of the weak recency heuristic, since $\mathbb{E}_\mu[\rho_{t+i+1} \mid (S_t, A_t)] = 1$ and therefore the inequality equates to Definition 3.1 in expectation. Based on our analysis in Section 5, the heuristic can be slightly violated without increasing the contraction modulus above 1, still allowing the operator to sometimes converge to $\boldsymbol{q}_\pi$. We thus settle the open problem in the negative: the condition in Eq. (18) is *sufficient but not necessary* for the operator to converge to its fixed point.