Reinforcement Learning with Adaptive Temporal Discounting

Sahaj Singh Maini, Zoran Tiganj

Keywords: Adaptive Temporal Discounting, Weber-Fechner law, Log-compressed Timeline.

Summary

Conventional reinforcement learning (RL) methods often fix a single discount factor for future rewards, limiting their ability to handle diverse temporal requirements. We propose a framework that utilizes an interpretation of the value function as a Laplace transform. By training an agent across a spectrum of discount factors and applying an inverse transform, we recover a log-compressed representation of expected future reward. This representation enables post hoc adjustments to the discount function (e.g., exponential, hyperbolic, or finite horizon) without retraining. Furthermore, by precomputing a library of policies, the agent can dynamically select the policy that maximizes a newly specified discount objective at runtime, effectively constructing a hybrid policy to handle varying temporal objectives. The properties of this log-compressed timeline are consistent with human temporal perception as described by the Weber-Fechner law, theoretically enhancing efficiency in scale-free environments by maintaining uniform relative precision across timescales. We demonstrate this framework in a grid-world navigation task where the agent adapts to different time horizons.

Contribution(s)

1. We extend prior methods for computing log-compressed representations of expected future reward to a dynamic policy evaluation setting, showing that when a library of policies is available, this representation enables immediate re-evaluation of these policies under any desired discount function.
   Context: Prior work established methods for computing log-compressed representations of expected future reward (Momennejad & Howard, 2018; Tiganj et al., 2019; Tano et al., 2020; Masset et al., 2023). Our extension focuses on dynamical policy evaluation with arbitrary temporal discounting. We evaluate the approach under idealized conditions where true value functions are accessible, demonstrated through a grid-world navigation task.
2. Leveraging the Weber–Fechner link between log-time codes and human perception, we analytically show that log-compressed representations enable efficient decision-making in scale-free environments by maintaining uniform relative precision across timescales.
   Context: The Weber-Fechner law Fechner (1860/1912) is a widely referenced principle in psychophysics, stating that perceived magnitude is proportional to the logarithm of stimulus intensity, implying a logarithmic scale.

Reinforcement Learning with Adaptive Temporal Discounting

Sahaj Singh Maini, Zoran Tiganj
sahmaini@iu.edu, ztiganj@iu.edu

Department of Computer Science, Indiana University Bloomington

Abstract

Conventional reinforcement learning (RL) methods often fix a single discount factor for future rewards, limiting their ability to handle diverse temporal requirements. We propose a framework that utilizes an interpretation of the value function as a Laplace transform. By training an agent across a spectrum of discount factors and applying an inverse transform, we recover a log-compressed representation of expected future reward. This representation enables post hoc adjustments to the discount function (e.g., exponential, hyperbolic, or finite horizon) without retraining. Furthermore, by precomputing a library of policies, the agent can dynamically select the policy that maximizes a newly specified discount objective at runtime, effectively constructing a hybrid policy to handle varying temporal objectives. The properties of this log-compressed timeline are consistent with human temporal perception as described by the Weber-Fechner law, theoretically enhancing efficiency in scale-free environments by maintaining uniform relative precision across timescales. We demonstrate this framework in a grid-world navigation task where the agent adapts to different time horizons.

1 Introduction

In traditional RL setting, agents learn by maximizing a cumulative future reward, which is typically exponentially discounted over time using a fixed temporal discount factor (γ) (Sutton & Barto, 2018). Despite its widespread adoption, such discounting approach presents limitations, particularly in dynamic environments where the relevance of future rewards may vary over time. For example, if an agent encounters an emergency situation where immediate action is critical a high discount on future rewards is beneficial. Conversely, in investment scenarios, considering longer-term outcomes might be preferred in some cases, necessitating a lower discount rate while in other cases the focus might be on short-term investments where a higher discount rate would be preferred. Furthermore, while the exponential discounting has been widely used since exponentially discounted future reward can be efficiently computed using temporal difference (TD) learning thanks to the efficiency of Bellman equation, other types of temporal discounting are often preferred. Specifically, hyperbolic discounting has gained attention for its ability to more closely mimic human decision-making processes. Unlike exponential discounting, hyperbolic discounting reduces the value of rewards less steeply over time, which has been shown to reflect more accurately how people value immediate versus delayed rewards (Ainslie, 1975). This form of discounting suggests that people often display a preference for smaller-sooner rewards over larger-later rewards when the delays are short, but this preference can switch as the delays increase, a phenomenon often referred to as "temporal inconsistency" or "preference reversal" (Green & Myerson, 1994; Laibson, 1997). More generally, humans can adjust the discounting function to adapt to the temporal statistics of the environment (Redish, 2004; Frederick et al., 2002). These considerations highlight the need for more flexible approaches to temporal discounting in RL, where discounting can be adjusted dynamically after the agent has been trained.

1.1 Related work

Limitations of exponential discounting were addressed in a number of previous studies. Inspired by human decision-making, previous work mainly focused on hyperbolic discounting by integrating over a spectrum of γ values (Kurth-Nelson & Redish, 2009; Fedus et al., 2019; Masset et al., 2023; Tiganj et al., 2019; Tano et al., 2020). Tang et al. (2021) used Taylor expansion of discount rates to interpolate value functions of two distinct discount factors.

Going beyond integration across γ values, previous work proposed the use of inverse Laplace transform to convert a value function across the set of γ values into a function of future time (Momennejad & Howard, 2018; Tiganj et al., 2019; Tano et al., 2020; Masset et al., 2023). This was primarily applied to modeling of human decision making as existence of a mental timeline of the future has been supported by behavioral (Tiganj et al., 2022) and neural evidence (Cao et al., 2024). Proposed approach builds on this work and integrates the Laplace framework into a setting where agents need to select actions at every step. We propose that agents construct multiple timelines of the future and use policy selection at every step.

This approach is in general compatible with model-free and model-based setting. For example, Tano et al. (2020) used spectrum of γ values to compute a reward as a function of future time in a model-free setting with TD learning. Model-based RL learning of arbitrary discount functions, including hyperbolic was also done using function approximation of the optimal policy (Schultheis et al., 2022). Model based RL approaches used successor representation and statistical learning computed with a spectrum of γ values to estimate timeline of the future (Tano et al., 2020; Momennejad & Howard, 2018; Tiganj et al., 2019).

1.2 Summary of the proposed approach

Here we describe an RL method with adaptive temporal discounting, where the discounting function can be chosen after training and dynamically adjusted. This is achieved by computing expected reward as a function of future time τ^* for a policy π. Such function then enables temporal discounting with arbitrary functional forms, such as exponential and power-law as well as temporal windows (e.g., expected reward from time t_i to time t_j at a given state) by varying the weight given to different parts of the future timeline.

To compute expected reward as a function of future time for a given state under a policy π, we compute a set of values V for a spectrum exponentially discount rates γ. Using a linear transformation this set of values can then be converted from a function of γ into a function of future time, τ^* . From a set of candidate policies π, this approach can then select a policy that maximizes a reward under the desired temporal discounting.

In the limit where the value V is computed for all possible γ values (i.e., γ is a continuous variable) we show that this approach can select an optimal policy from a set of precomputed policies under a user-specified temporal discounting function. For practical applications, we consider a discrete set of geometrically spaced γ values. This gives rise to a log-compressed representation of the future time where the temporal resolution gradually decreases as a function of distance from the current state.

Log-compressed time reflects the Weber-Fechner law which governs scale-invariance in human perception across a number of domains, including time: perceived magnitude is proportional to the logarithm of the true magnitude of the stimulus (Fechner, 1860/1912). In time perception the scale-invariance is manifested in behavioral tasks such as interval reproduction, where variability of reproduced interval by human participants is proportional to the duration of the interval. In other words, timing precision is proportional to the length of the interval itself, a phenomenon known as the scalar property (Gibbon, 1977; Wilkes, 2015). Such a logarithmically compressed representation of time is argued to be advantageous when the environment lacks a single dominant timescale (i.e., it is scale-free), so that events or correlations extend across multiple orders of magnitude in a self-similar manner (Howard & Shankar, 2018; Shankar & Howard, 2013). We show that the log-compressed

representation of the future time described here, captures the scalar property and enables efficient decision making under the assumption that variability in reward time (or spatial position) scales with its temporal (or spatial) distance.

2 Methods

We consider a standard RL framework modeled as a Markov Decision Process (MDP), defined by the tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R} \rangle$, where:

- $\mathcal{S}$ represents the state space, encompassing all possible states the agent may encounter,
- $\mathcal{A}$ denotes the action space, consisting of all actions available to the agent,
- $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the state transition probability function, specifying the probability $\mathcal{P}(s'|s, a)$ of transitioning to state $s' \in \mathcal{S}$ from state $s \in \mathcal{S}$ upon taking action $a \in \mathcal{A}$,
- $\mathcal{R} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ is the reward function, defining the immediate reward $\mathcal{R}(s, a, s')$ received when transitioning from state s to state s' via action a.

The agent's goal is to learn a policy $\pi(a|s)$, which maps states to a probability distribution over actions, maximizing the expected cumulative reward over time under a specified temporal discounting scheme. In the traditional RL formulation, this is captured by the state-value function with an exponential discount factor $\gamma \in [0, 1]$:

$$V_\gamma^\pi(s) = \mathbb{E}_\pi \left[\sum_{\tau^*=0}^{\infty} \gamma^{\tau^*} r_{t+\tau^*} \mid s_t = s \right], \quad \forall s \in \mathcal{S}, \tag{1}$$

where $r_{t+\tau^*} = \mathcal{R}(s_{t+\tau^*}, a_{t+\tau^*}, s_{t+\tau^*+1})$ is the reward at time step $t + \tau^*$, τ^* denotes future time steps relative to the current time t, and the expectation $\mathbb{E}_\pi$ is taken over trajectories induced by the policy π and the transition dynamics $\mathcal{P}$. The discount factor γ exponentially weights future rewards, prioritizing immediate rewards when $\gamma < 1$. However, this formulation fixes the temporal preference at training time, limiting adaptability to different discounting schemes or planning horizons post-training.

2.1 Computing Expected Rewards as a Function of Future Time

To enable this adaptability, we compute the state-value function $V_\gamma^\pi(s)$ across a discrete set of discount factors $\Gamma = \{\gamma_1, \gamma_2, \ldots, \gamma_m\}$, where each $\gamma_i \in [0, 1]$. For a specific discount factor γ_i, the value function is:

$$V_{\gamma_i}^\pi(s) = \mathbb{E}_\pi \left[\sum_{\tau^*=0}^{\infty} \gamma_i^{\tau^*} r_{t+\tau^*} \mid s_t = s \right]. \tag{2}$$

Define $g(\tau^*) = \mathbb{E}_\pi \left[r_{t+\tau^*} \mid s_t = s \right]$, which represents the expected reward at future time step τ^* under policy π, starting from state s at time t. This allows us to rewrite the value function as:

$$V_{\gamma_i}^\pi(s) = \sum_{\tau^*=0}^{\infty} \gamma_i^{\tau^*} g(\tau^*). \tag{3}$$

This expression resembles the discrete analog of the Laplace transform (or unilateral Z-transform) of the sequence $g(\tau^*)$, with γ_i acting as the transform variable. Equivalently, letting $\sigma_i = -\ln(\gamma_i)$ (so $\gamma_i = e^{-\sigma_i}$), we can express it as:

$$V_{\gamma_i}^\pi(s) = \sum_{\tau^*=0}^{\infty} e^{-\sigma_i \tau^*} g(\tau^*), \tag{4}$$

where $V^{\pi}_{\gamma_i}(s)$ can be interpreted as the Laplace transform of $g(\tau^*)$ evaluated at σ_i, i.e., $V^{\pi}_{\gamma_i}(s) = \mathcal{L}\{g(\tau^*)\}(\sigma_i)$. By computing $V^{\pi}_{\gamma_i}(s)$ for multiple $\gamma_i \in \Gamma$, we obtain a set of samples of this transform at points σ_i:

$$\mathbf{V}^{\pi}(s) = [V^{\pi}_{\gamma_1}(s), V^{\pi}_{\gamma_2}(s), \ldots, V^{\pi}_{\gamma_m}(s)]. \tag{5}$$

To recover the expected reward function $g(\tau^*)$ from these samples, we apply an inverse transform. With a finite set of γ_i, we approximate $g(\tau^*)$ for discrete $\tau^* = 0, 1, 2, \ldots$ using numerical methods, such as the Post inversion formula Post (1930) (see Sec. S1 for discussion on numerical stability and demonstrations of other approaches for computing the inverse that can provide better numerical stability than Post inversion formula). The Post method approximates $g(\tau^*)$ as:

$$g(\tau^*) \approx \frac{(-1)^k}{k!} \left(\frac{k}{\tau^*}\right)^{k+1} \frac{d^k}{d\sigma^k} V^{\pi}_{\gamma}(s) \bigg|_{\sigma = k/\tau^*}, \tag{6}$$

where k is a parameter chosen to balance accuracy and computational stability as discussed in the next section, and $\gamma = e^{-\sigma}$.

Once $g(\tau^*)$ is estimated, we can compute the value of the policy under any arbitrary temporal discounting function. Examples include finite horizon: for a horizon T, the cumulative expected reward is:

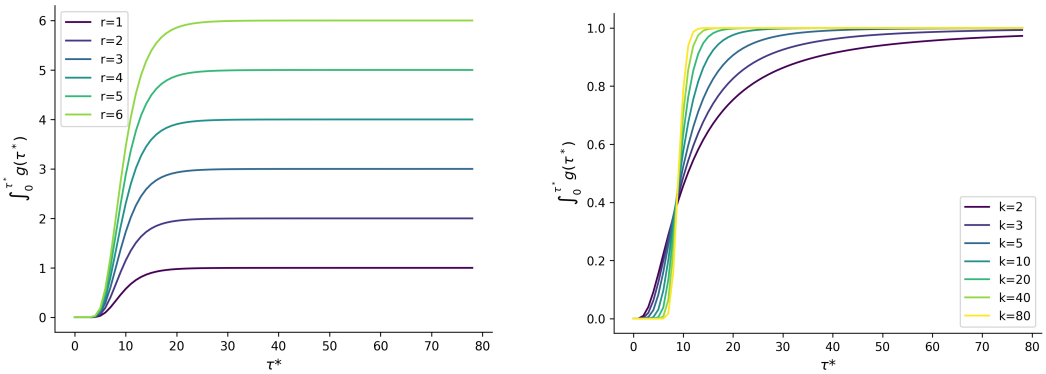
$$V^{\pi}_{0 \to T}(s) = \sum_{\tau^*=0}^{T} g(\tau^*), \tag{7}$$

and hyperbolic and power-law discounting: with a discount function $f(\tau^*) = \frac{1}{(1+K\tau^*)^{\beta}}$, the value is:

$$V^{\pi}_{\text{hyperbolic}}(s) = \sum_{\tau^*=0}^{\infty} \frac{1}{(1 + K\tau^*)^{\beta}} g(\tau^*), \tag{8}$$

where $K > 0$ controls the discounting rate.

Fig.1a shows how value increases as a function of τ^*: without temporal discounting the value reaches the true magnitude of the reward. Fig. 2 illustrates the impact of finite horizon on the value for different distances from the goal.



(a) (b)

Figure 1: Value function as a result of integrating expected reward at future time steps for different values of reward (with $k = 10$) (a) and parameter k (b). The reward is located at a distance of 10 steps from the agent.

Overall, this approach eliminates the need to retrain the policy for different temporal preferences, offering a flexible framework for post-hoc adaptation.

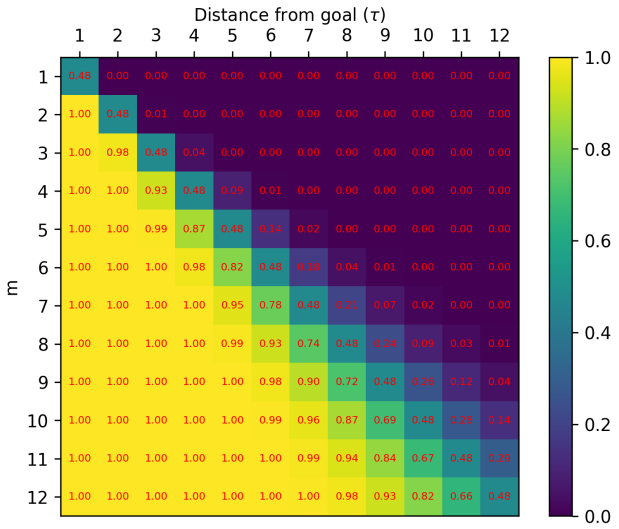


Figure 2: Value of $\int_{\tau^*=0}^{\tau^*=m} g(\tau^*)$ across varying deadlines (m) and distance from rewarding goal state (τ) for a terminal reward of magnitude 1 for $k = 30$.

2.2 Emergence of Log-Compression in the Estimation of Future Time

In this section, we investigate how the discrete approximation employed to estimate the expected reward function $g(\tau^*)$ naturally yields a log-compressed representation of future time. This property arises from the mathematical structure of the inverse transform approximation and carries important implications for an agent's temporal perception and decision-making.

To illustrate the emergence of log-compression we take advantage of linearity of the Laplace and inverse Laplace transform and first derive its impulse response and later generalize that to arbitrary reward functions. To compute the impulse response, we consider an environment where a single reward of magnitude 1 is positioned at a temporal distance τ from the current state, indicating the number of steps to reach the reward. The expected reward at a future time step τ^* is:

$$g(\tau, \tau^*) = \frac{1}{\tau} \frac{k^{k+1}}{k!} \left(\frac{\tau}{\tau^*}\right)^{k+1} e^{-k\left(\frac{\tau}{\tau^*}\right)}. \tag{9}$$

The result has a form of gamma distribution (Weisstein, 2004), and has been widely used in learning temporal relationships (Hopfield & Tank, 1990; Grossberg & Schmajuk, 1989; Shankar & Howard, 2012; Jacques et al., 2021).

As shown by Shankar & Howard (2012) taking the partial derivative of the impulse response of $g(\tau, \tau^*)$ with respect to τ^* and setting it to zero, we find that it is a unimodal function of time t with a maximum at $\tau^* \frac{k}{k+1}$ (in the limit when $k \to \infty$, the peak is exactly at τ^*, see Appendix S3 for derivation). Expected reward as a function of future time for reward magnitude of 1 and distance of 25 steps is shown in Fig. 3 for different values of k. We used 50 log-spaced τ^* values ranging from 1 to 50.

To demonstrate that the representation is indeed log-compressed, we express the width of the unimodal impulse response of $g(\tau, \tau^*)$ through the coefficient of variation $CV = \frac{1}{\sqrt{k+1}}$, which is a ratio of standard deviation and mean (see Appendix S4 for derivation). Critically, CV does not depend on τ and τ^*, implying that the width of the unimodal basis functions increases linearly with their peak time indicating log-compression.

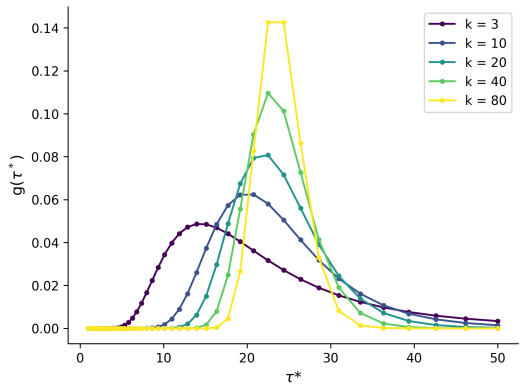


Figure 3: Expected reward at future time steps for different values of k. The reward of magnitude 1 is presented at distance 25. Increasing k results in more narrow peaks that approach closer to the value of the reward (since $\tau^*_{\text{peak}} = \tau k/(k+1)$).

If discrete time steps $\tau_1^*, \tau_2^*, \ldots, \tau_m^*$ are log-spaced, i.e., $\tau_i^* = e^{ci}$ (where c sets the spacing), then $\log(\tau_i^*) = ci$, yielding constant intervals on a logarithmic scale. Peaks at $\tau^*_{\text{peak}} = \frac{k}{k+1}\tau$ are also log-spaced, as $\log(\tau^*_{\text{peak}}) = \log\left(\frac{k}{k+1}\right) + \log(\tau)$. With constant CV, peak widths scale with τ^*_{peak}, maintaining consistent relative width on a log scale, thus compressing distant time steps relative to near ones.

For temporal discounting with a finite horizon we provide lemmas and proofs for analytical solutions for three different integral bounds: 0 to m (Fig. 4a), m to n (Fig. 4b) and m to ∞ (Fig. 4c). Proofs for the lemmas are in Supplemental Information (Sec. S2).

Lemma 1: The value of a state at distance τ from the reward given the deadline m can be computed as:

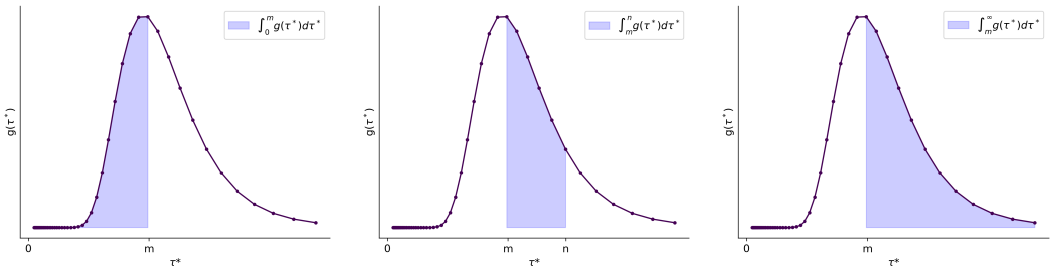
$$\int_0^m g(\tau, \tau^*) d\tau^* = e^{-k\frac{(\tau)}{m}} \sum_{i=0}^{k-1} \frac{(k\frac{(\tau)}{m})^i}{i!}. \tag{10}$$

Lemma 2: The state-value perceived by the agent between steps $\tau^* = m$ and $\tau^* = n$ in the future where $m < n$ when the agent is at distance τ from the reward can be computed as follows:

$$S(n, c, x) = e^{cx} \sum_{i=0}^{n} (-1)^{n-i} \frac{n!}{i! c^{n-i+1}} x^i$$
$$\int_m^n g(\tau, \tau^*) d\tau^* = (-1)^{k+1} (\frac{k^{k+1}}{k!})[S(k-1, k, \frac{-\tau}{n}) - S(k-1, k, \frac{-\tau}{m})]. \tag{11}$$

Lemma 3: The value perceived by the agent between step m and ∞ in the future when the agent is at a distance τ from the reward can be computed as follows:

$$\int_m^\infty g(\tau, \tau^*) d\tau^* = 1 - e^{-k\frac{(\tau)}{m}} \sum_{i=0}^{k-1} \frac{(k\frac{(\tau)}{m})^i}{i!}. \tag{12}$$



(a) (b) (c)

Figure 4: Illustration of expected reward as a function of the future time integrated over different bounds: 0 to m (a), m to n (b) and m to ∞ (c). In this example, the reward of magnitude 1 is located 25 steps from the current state (marked as 0). The total area under each curve corresponds to the magnitude of the reward. Analytical derivation for each integral is provided in the Supplemental Information (Sec. S2).

2.3 Policy Selection Under Arbitrary Discounting Functions

In this subsection, we extend our framework to select the optimal policy from a set of precomputed policies under any user-specified temporal discounting function. This method leverages the expected reward function $g(\tau^*)$ introduced earlier, enabling adaptability to diverse temporal preferences without requiring retraining or additional environment interactions.

Given a set of precomputed policies $\Pi = \{\pi_1, \pi_2, \ldots, \pi_p\}$, we assume each policy π_k has been evaluated to compute state-value functions $V_{\gamma_i}^{\pi_k}(s)$ for a discrete set of discount factors Γ. As described above, we approximate the expected reward function $g_k(\tau^*) = \mathbb{E}_{\pi_k}[r_{t+\tau^*} \mid s_t = s]$ for each policy π_k using the inverse transform applied to the sampled values $V_{\gamma_i}^{\pi_k}(s)$.

For a user-specified temporal discounting function $f : \mathbb{N} \to \mathbb{R}^+$, which assigns weights to rewards based on their temporal distance τ^*, we compute the value of each policy π_k as:

$$V_f^{\pi_k}(s) = \sum_{\tau^*=0}^{\infty} f(\tau^*) g_k(\tau^*). \tag{13}$$

This value represents the expected cumulative reward under policy π_k adjusted by the discounting function f. The optimal policy from the set Π is then selected as:

$$\pi^* = \arg\max_{\pi_k \in \Pi} V_f^{\pi_k}(s). \tag{14}$$

This approach allows the agent to adapt to arbitrary discounting schemes post-training. Under idealized conditions—where the set of discount factors Γ is continuous and $g_k(\tau^*)$ is perfectly recovered—the selected policy π^* is optimal within Π (not globally optimal) for the given f (see Appendix S5 for proof of optimality).

In practice, Γ is finite, and $g_k(\tau^*)$ is approximated as $\hat{g}_k(\tau^*)$. Thus, we compute $\hat{V}_f^{\pi_k}(s) = \sum_{\tau^*=0}^{\infty} f(\tau^*)\hat{g}_k(\tau^*)$, and the selected policy is optimal with respect to these approximations.

2.4 Dynamic Policy Selection

In this subsection, we present an extension to our framework that enables *dynamic policy selection* at each time step. This approach allows the agent to adapt its strategy based on the current state and a user-specified temporal discounting function f.

At each time step t, the agent, in state s_t, computes the value of each policy π_k from that state under the discounting function f. This value is given by:

$$V_f^{\pi_k}(s_t) = \sum_{\tau^*=0}^{\infty} f(\tau^*) g_k(\tau^* \mid s_t), \tag{15}$$

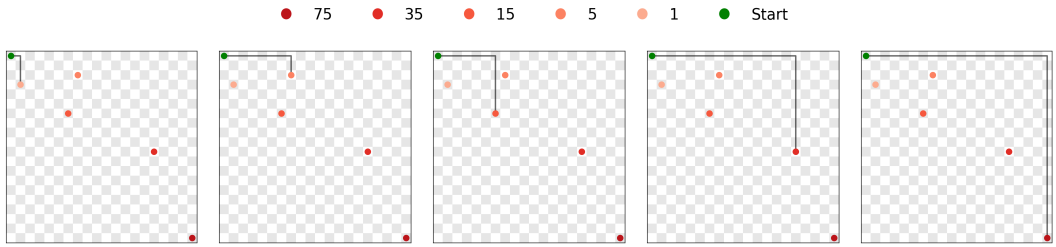
where $g_k(\tau^* \mid s_t)$ represents the expected reward at time $t + \tau^*$ under policy π_k, starting from s_t. The agent selects the policy π_{k^*} that maximizes this value:

$$\pi_{k^*} = \arg\max_{\pi_k \in \Pi} V_f^{\pi_k}(s_t). \tag{16}$$

The action at s_t is then $a_t = \pi_{k^*}(s_t)$. Upon transitioning to the next state s_{t+1}, the agent repeats this process, recomputing $V_f^{\pi_k}(s_{t+1})$ for all $\pi_k \in \Pi$ and selecting the optimal policy for s_{t+1}, which may differ from the previous choice.

This dynamic process constructs a *hybrid policy* π_{hybrid}, where the action at each state s_t is determined by the policy π_{k^*} that maximizes $V_f^{\pi_k}(s_t)$ at that time. The hybrid policy π_{hybrid} is not necessarily an element of Π, as it may combine actions from multiple policies depending on the state, where $\pi_{k^*} = \arg\max_{\pi_k \in \Pi} V_f^{\pi_k}(s_t)$. By greedily selecting the action from the policy that maximizes the value at each state, π_{hybrid} locally optimizes the expected cumulative reward under f. In environments where different policies perform better in different regions of the state space, this approach can outperform any single policy in Π.

We illustrate this approach through a grid-world example (Fig. 5). The agent starts from a corner of the grid. Five rewards are placed at different locations in the environment such that rewards with a larger magnitude are placed further from the agent's start location. We first compute five different policies and corresponding values such that each of the policies leads the agent to a different reward. When the agent starts navigating, it is given a number of steps to complete the task (finite horizon). The agent follows Alg. 1, which at each step uses the precomputed reward timelines, $g_k(\tau^*|s_t)$ to evaluate the value of each policy under the current horizon. The agent always selects a policy that leads to the largest reward and takes the corresponding action. In our example, the agent always reached the largest possible reward given the available number of steps (note that this is not guaranteed, as we investigate in the following subsections). We emphasize that agents based on traditional exponential discounting with a single value of γ would always converge to the same reward since those agents do not have the capability to take into account the available horizon.



(a) T=4 (b) T=9 (c) T=12 (d) T=25 (e) T=38

Figure 5: Paths chosen by the model for a varying number of available steps, T. The colors represent reward locations and correspond to rewards of different magnitudes. Higher rewards are located further. Given a longer horizon, the agent chooses rewards located further to obtain a larger reward.

2.5 Efficient Performance in Scale-Free Environments

In environments lacking a single dominant timescale—often called *scale-free*—reward timing may follow a power-law pattern in its temporal correlations, for instance $\mathbb{E}[\, r_t \, r_{t+\tau^*} \,] \propto (\tau^*)^{-\alpha}$ with

Algorithm 1 Dynamic Policy Selection

Input: Set of policies $\Pi = \{\pi_1, \pi_2, \ldots, \pi_p\}$, temporal discounting function f, initial state s_0, maximum time steps T, horizon for each time step $h_0, h_1, ...h_{T-1}$, approximated expected reward functions $g_k(\tau^* \mid s)$ for each $\pi_k \in \Pi$ and state s
Set $s \leftarrow s_0$
Set $t \leftarrow 0$
while $t < T$ and s is not terminal **do**
    for each $\pi_k \in \Pi$ **do**
        Compute $V_f^{\pi_k}(s) \leftarrow \sum_{\tau^*=0}^{h_t} f(\tau^*) g_k(\tau^* \mid s)$
    end for
    Select $\pi_{k^*} = \arg\max_{\pi_k \in \Pi} V_f^{\pi_k}(s)$
    Take action $a = \pi_{k^*}(s)$
    Observe next state s' and reward r
    Set $s \leftarrow s'$
    Set $t \leftarrow t + 1$
end while

$\alpha > 0$. This implies that reward timing variances scale with the square of their mean, $\mathrm{Var}(\tau) \propto \tau^2$—a relationship observed in interval-timing tasks in humans and other animals.

To illustrate this, suppose the agent can pursue two rewards: one at $\tau = 10$ steps (with variance ± 1 step), and another at $\tau = 100$ steps (with variance ± 10 steps). In our log-compressed timeline, each reward is represented by a unimodal function centered near its mean arrival time, whose "width" is proportional to that mean. Thus, the reward at 10 steps exhibits a narrow peak, while the reward at 100 steps has a broader peak. When integrating the expected reward up to a finite horizon $T = 50$, that is

$$V_{0\to T}^{\pi}(s) \;=\; \sum_{\tau^*=0}^{T} g(\tau^*), \tag{17}$$

the entire distribution for the near reward lies within the 50-step horizon. In contrast, only a portion of the broader distribution at 100 steps overlaps the first 50 steps. However, if that distant reward is large enough, the probability of it arriving earlier than 100 steps can contribute a higher total expected value than the smaller near reward. By preserving the full variance structure around each expected time, the agent's scale-adaptive representation captures this partial overlap automatically.

The *efficiency* of such a log-compressed representation can be framed in terms of minimizing the maximum *relative* error under a scale-invariant prior $p(\tau) \propto 1/\tau$. Specifically, for the expected squared relative error,

$$\mathrm{Error} \;=\; \mathbb{E}\left[\left(\frac{\hat{g}(\tau^*) - g(\tau^*)}{g(\tau^*)}\right)^2\right], \tag{18}$$

this approach ensures that

$$\frac{\sqrt{\mathrm{Var}(\hat{g}(\tau^*))}}{g(\tau^*)} \;=\; \mathrm{CV} \;=\; \frac{1}{\sqrt{k+1}}, \tag{19}$$

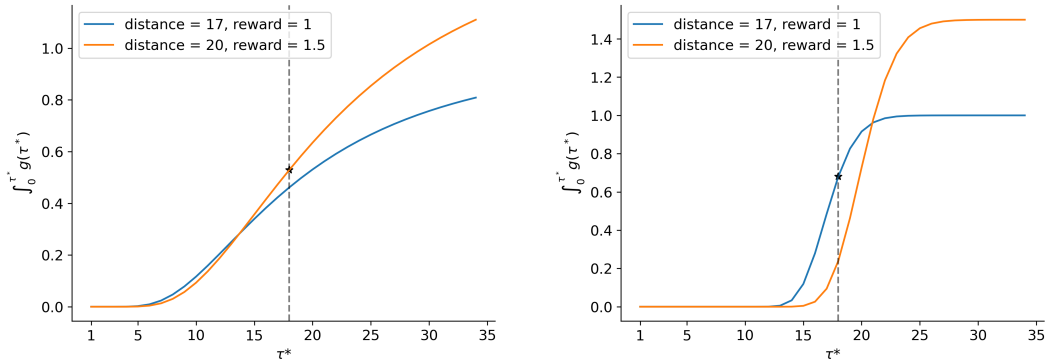
remains constant across τ^*.

By contrast, a representation with uniform precision across a linear timeline (e.g., using basis functions of a fixed width) would face a difficult trade-off. To accurately represent near-future rewards, it would require narrow basis functions, leading to an extremely inefficient tiling of the distant future with a large number of redundant units. Conversely, using wide basis functions to cover the distant future would blur near-term events, resulting in poor temporal resolution for immediate decisions. The log-compressed timeline avoids this dilemma by naturally adapting its resolution to the timescale. The log compression allocates resources proportionally to the relevant timescale, ensuring uniform relative precision for near and distant rewards—particularly advantageous in scale-free

settings and under limited time horizons. We stress that this efficiency argument is theoretical; empirical validation in environments with demonstrable scale-free reward structures is a necessary next step for future work.

2.6 Controlling the Precision of Reward Estimation

The parameter k in the log-compressed representation governs the precision of reward estimation, with the coefficient of variation given by CV. Larger values of k reduce the CV (Fig. 3), enhancing accuracy but increasing computational demands and impact stability of the inverse transform (since k controls the order of the numerical derivative in the Post inversion, Eq. 6). The change in expected future reward as a function of future time for different values of k is shown in Fig. 1b. The choice of k can impact the choice that the agent makes. This is illustrated in Fig 6a where with $k = 3$ agent prefers a closer but smaller reward and in Fig 6b where the only change is that $k = 80$ but the agent now prefers more distant but larger reward. This flexibility in k reflects a tunable trade-off between precision and computational cost, allowing the representation to adapt to different environments or resource constraints. We emphasize that this does not undermine the representation's core efficiency principle, which ensures uniform relative precision across τ^* for any chosen k. This variability in k also offers a parallel to human timing behavior, where the scalar property implies a constant CV that differs across individuals.



(a) $k = 3$ (b) $k = 80$

Figure 6: Comparison of value function for different values of k in the presence of rewards of different magnitudes at different distances.

3 Discussion

This paper introduces a framework for reinforcement learning (RL) that separates temporal discounting from the training phase, allowing adjustments to the discount function after training without the need for retraining. By employing a log-compressed representation of expected future rewards, derived from value functions computed across a range of discount factors, the approach enables agents to adapt flexibly to various temporal preferences—such as exponential, hyperbolic, or finite-horizon discounting. Additionally, it enables dynamic policy selection by utilizing a precomputed library of policies to construct a hybrid strategy that adjusts to varying temporal objectives. These contributions enhance the adaptability of RL systems, which could prove useful in applications where temporal preferences vary, such as robotics or financial decision-making. Furthermore, the log-compressed representation aligns conceptually with human perception of time, as described by the Weber-Fechner law, suggesting a potential avenue for modeling human-like decision processes in artificial agents.

Despite its potential, the framework has several limitations that warrant consideration:

- Approximation errors: The reliance on an approximated inverse Laplace transform, based on a finite set of discount factors, introduces possible inaccuracies in reconstructing the desired discount function. The quality of this approximation depends heavily on the number and distribution of discount factors chosen, which may not generalize across all environments.
- Computational overhead: Reconstructing the reward timeline and selecting policies dynamically at each step can demand significant computational resources, particularly in large or complex state spaces. The feasibility of scaling this approach to high-dimensional tasks, such as continuous control, has yet to be demonstrated.
- Policy library dependence: The success of dynamic policy selection hinges on the diversity and quality of the precomputed policy library. If the library lacks policies suited to a specific discount objective, performance may fall short compared to a policy trained specifically for that purpose.
- Efficiency and comparison to human perception: Claims of efficiency in scale-free environments and alignment with human perception are theoretical; additional experiments, such as comparisons to human behavior or dynamic deadline shifts, could strengthen these, as could scalability tests beyond grid-worlds.

Inspired by human perception and decision making, this work offers a step toward greater flexibility in RL, enabling agents to adapt to varying temporal preferences and objectives.

Acknowledgments

We are grateful to Dr. Roni Khardon for insightful discussions.

References

George Ainslie. Specious reward: a behavioral theory of impulsiveness and impulse control. *Psychological Bulletin*, 82(4):463, 1975.

Rui Cao, Ian M Bright, and Marc W Howard. Ramping cells in the rodent medial prefrontal cortex encode time to past and future events via real laplace transform. *Proceedings of the National Academy of Sciences*, 121(38):e2404169121, 2024.

Gustav Fechner. *Elements of Psychophysics. Vol. I.* Houghton Mifflin, 1860/1912.

William Fedus, Carles Gelada, Yoshua Bengio, Marc G Bellemare, and Hugo Larochelle. Hyperbolic discounting and learning over multiple horizons. *arXiv preprint arXiv:1902.06865*, 2019.

Shane Frederick, George Loewenstein, and Ted O'donoghue. Time discounting and time preference: A critical review. *Journal of economic literature*, 40(2):351–401, 2002.

John Gibbon. Scalar expectancy theory and weber's law in animal timing. *Psychological review*, 84 (3):279, 1977.

Leonard Green and Joel Myerson. Exponential versus hyperbolic discounting of delayed outcomes: Risk and waiting time. *American Zoologist*, 34(4):496–505, 1994.

Stephen Grossberg and Nestor A Schmajuk. Neural dynamics of adaptive timing and temporal discrimination during associative learning. *Neural networks*, 2(2):79–102, 1989.

John J Hopfield and David W Tank. Neural computation by time concentration, June 26 1990. US Patent 4,937,872.

Marc W Howard and Karthik H Shankar. Neural scaling laws for an uncertain world. *Psychological review*, 125(1):47, 2018.

Brandon Jacques, Zoran Tiganj, Marc W Howard, and Per B Sederberg. Deepsith: Efficient learning via decomposition of what and when across time scales. *Advances in neural information processing system*, 2021.

Zeb Kurth-Nelson and A David Redish. Temporal-difference reinforcement learning with distributed representations. *PLoS One*, 4(10):e7362, 2009.

David Laibson. Golden eggs and hyperbolic discounting. *The Quarterly Journal of Economics*, 112 (2):443–477, 1997.

Paul Masset, Pablo Tano, HyungGoo R Kim, Athar N Malik, Alexandre Pouget, and Naoshige Uchida. Multi-timescale reinforcement learning in the brain. *bioRxiv*, 2023.

András Mészáros and Miklós Telek. Concentrated matrix exponential distributions with real eigenvalues. *Probability in the Engineering and Informational Sciences*, 36(4):1171–1187, 2022.

Ida Momennejad and Marc W Howard. Predicting the future with multi-scale successor representations. *BioRxiv*, pp. 449470, 2018.

Emil L Post. Generalized differentiation. *Transactions of the American Mathematical Society*, 32 (4):723–781, 1930.

A David Redish. Addiction as a computational process gone awry. *Science*, 306(5703):1944–1947, 2004.

Matthias Schultheis, Constantin A. Rothkopf, and Heinz Koeppl. Reinforcement learning with non-exponential discounting. In *Advances in Neural Information Processing Systems*, 2022.

Karthik H Shankar and Marc W Howard. A scale-invariant internal representation of time. *Neural Computation*, 24(1):134–193, 2012.

Karthik H Shankar and Marc W Howard. Optimally fuzzy temporal memory. *The Journal of Machine Learning Research*, 14(1):3785–3812, 2013.

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

Yunhao Tang, Mark Rowland, Rémi Munos, and Michal Valko. Taylor expansion of discount factors. In *International Conference on Machine Learning*, pp. 10130–10140. PMLR, 2021.

Pablo Tano, Peter Dayan, and Alexandre Pouget. A local temporal difference code for distributional reinforcement learning. *Advances in neural information processing systems*, 33:13662–13673, 2020.

Zoran Tiganj, Samuel J Gershman, Per B Sederberg, and Marc W Howard. Estimating scale-invariant future in continuous time. *Neural Computation*, 31(4):681–709, 2019.

Zoran Tiganj, Inder Singh, Zahra G Esfahani, and Marc W Howard. Scanning a compressed ordered representation of the future. *Journal of Experimental Psychology: General*, pp. In Press, 2022.

Eric W Weisstein. Gamma distribution. *https://mathworld. wolfram. com/*, 2004.

Jason T Wilkes. *Reverse first principles: Weber's law and optimality in different senses*. PhD thesis, UNIVERSITY OF CALIFORNIA, SANTA BARBARA, 2015.

Supplementary Materials

The following content was not necessarily subject to peer review.

S1 Numerical Approximations of the Inverse Laplace Transform

In this section, we describe a numerical approximation for computing the inverse Laplace transform using the CME-R (Complex Matrix Exponentials with Real eigenvalues) method (introduced by Mészáros & Telek (2022)). This method provides better numerical stability when compared to Post's method (Post, 1930) (Fig S1).

Let $\beta, \eta \in \mathbb{R}^M$, where $\tau_i^* \in \mathbb{R}$, $\sigma_{ij} = \frac{\beta_j}{\tau_i^*}$ and $\gamma_{ij} = e^{-\sigma_{ij}}$. The approximation takes the form:

$$g(\tau_i^*) \approx \sum_{k=1}^{M} \frac{\eta_k}{\tau_i^*} V^{\pi}_{\gamma_{i,k}}. \tag{20}$$

Here, M represents the maximum number of function evaluations and serves a similar role to the parameter k in the Post's method. Both η and β are real-valued tensors. It is worth noting that Equation 20 requires higher numerical precision as M increases to obtain accurate inverse Laplace transform results. The detailed methodology for computing the parameters η and β is thoroughly documented in the work of Mészáros & Telek (2022).

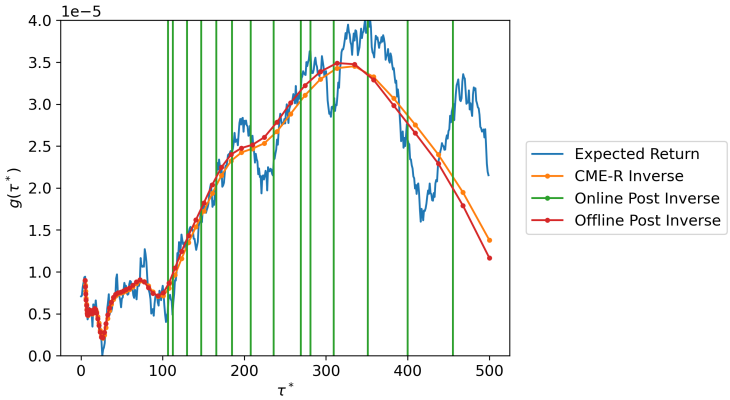


Figure S1: Comparing expected return between CME-R inverse, Offline Post Inverse and Online Post Inverse. While this figure represents a general property of the inverse Laplace transform for a time-series input, in our case, the input (blue) is a time-varying reward function. Despite high temporal variability of the input (reward), CME-R (orange) is able to construct a faithful log-compressed estimate with temporal resolution gradually decaying from 0 (the current state) towards the future. Offline Post inverse (red) is an analytically computed offline solution that convolves the input function with the set of impulse responses defined in Eq. 9. (It serves as a ground truth since it does not suffer from numerical issues.) A high match between CME-R and the Offline Post inverse indicates the good fidelity of the CME-R representation. On the other hand, Online Post inverse (computed online using Eq. 6) suffers from numerical instabilities when the input signal has high temporal variability (green).

S2 Lemmas and Proofs

We define $\tau^*, \tau \in [0, \infty]$. τ is the distance from the reward and τ^* refers to the number of steps in the future from the agent's current position and $\sigma = \frac{k}{\tau^*}$. If we assume there is a single terminal

reward with magnitude 1 in an environment, then the value of state s at distance τ from the reward is given by $\gamma^\tau = (e^{-\sigma})^\tau = e^{-\sigma\tau}$. Then we can compute the expected reward at time step τ_i^* in the future as follows:

$$\begin{aligned} g(\tau, \tau^*) &= \mathcal{L}^{-1}\{V^\pi_{\gamma=e^{-\sigma\tau}}(s)\} \\ &= \mathcal{L}^{-1}\{e^{-\sigma\tau}\} \\ &= \frac{(-1)^k}{k!}\sigma^{k+1}\frac{d^k}{d\tau^k}e^{-\sigma\tau} \\ &= \frac{(-1)^k}{k!}\sigma^{k+1}(-\tau)^k e^{-\sigma\tau} \\ &= \frac{1}{\tau}\frac{k^{k+1}}{k!}(\frac{\tau}{\tau^*})^{k+1}e^{-k(\frac{\tau}{\tau^*})} \end{aligned}$$

Lemma 1: The value of a state at distance τ from the reward given the deadline m can be computed as:

$$\int_0^m g(\tau, \tau^*)d\tau^* = e^{-k\frac{(\tau)}{m}}\sum_{i=0}^{k-1}\frac{(k\frac{(\tau)}{m})^i}{i!}. \tag{21}$$

Proof:

$$\begin{aligned} \int_0^m g(\tau, \tau^*)d\tau^* &= \int_0^m \frac{1}{\tau}\frac{k^{k+1}}{k!}(\frac{\tau}{\tau^*})^{k+1}e^{-k(\frac{\tau}{\tau^*})}d\tau^* \\ &= \frac{1}{\tau}\frac{k^{k+1}}{k!}\int_0^m (\frac{\tau}{\tau^*})^{k+1}e^{-k(\frac{\tau}{\tau^*})}d\tau^* \end{aligned}$$

Substituting $u = \frac{\tau}{\tau^*}$,

$$\begin{aligned} &= -\frac{k^{k+1}}{k!}\int_\infty^{\frac{\tau}{m}} u^{k-1}e^{-ku}du \\ &= \frac{k^{k+1}}{k!}\int_{\frac{\tau}{m}}^\infty u^{k-1}e^{-ku}du \\ &= \frac{k^{k+1}}{k!}[\int_0^\infty u^{k-1}e^{-ku}du - \int_0^{\frac{(\tau)}{m}} u^{k-1}e^{-ku}du] \end{aligned}$$

Using $\int_0^\infty y^m e^{-ky}dy = \frac{\Gamma(m+1)}{k^{m+1}}$,

$$\begin{aligned} &= \frac{k^{k+1}}{k!}[\frac{\Gamma(k)}{k^k} - \int_0^{\frac{(\tau)}{m}} u^{k-1}e^{-ku}du] \\ &= 1 - \frac{k^{k+1}}{k!}\int_0^{\frac{(\tau)}{m}} u^{k-1}e^{-ku}du \end{aligned}$$

Using $\int_0^b x^n e^{-ax} dx = \frac{n!}{a^{n+1}} \left[1 - e^{-ab} \sum_{i=0}^{i=n} \frac{(ab)^i}{i!}\right]$,

$$= 1 - \frac{k^{k+1}}{k!} \frac{(k-1)!}{k^k} \left[1 - e^{-k\frac{(\tau)}{m}} \sum_{i=0}^{i=k-1} \frac{(k\frac{(\tau)}{m})^i}{i!}\right]$$
$$= 1 - 1 \left[1 - e^{-k\frac{(\tau)}{m}} \sum_{i=0}^{i=k-1} \frac{(k\frac{(\tau)}{m})^i}{i!}\right]$$
$$= e^{-k\frac{(\tau)}{m}} \sum_{i=0}^{i=k-1} \frac{(k\frac{(\tau)}{m})^i}{i!}$$

Lemma 2: The state-value perceived by the agent between steps $\tau^* = m$ and $\tau^* = n$ in the future where $m < n$ when the agent is at distance τ from the reward can be computed as follows:

$$S(n, c, x) = e^{cx} \sum_{i=0}^{n} (-1)^{n-i} \frac{n!}{i! c^{n-i+1}} x^i$$
$$\int_m^n g(\tau, \tau^*) d\tau^* = (-1)^{k+1} (\frac{k^{k+1}}{k!})[S(k-1, k, \frac{-\tau}{n}) - S(k-1, k, \frac{-\tau}{m})]. \tag{22}$$

Proof:

$$\int_m^n g(\tau, \tau^*) d\tau^* = \int_m^n \frac{1}{\tau} \frac{k^{k+1}}{k!} (\frac{\tau}{\tau^*})^{k+1} e^{-k(\frac{\tau}{\tau^*})} d\tau^*$$
$$= (-1)^{k+1} \frac{1}{\tau} \frac{k^{k+1}}{k!} \int_m^n (\frac{-\tau}{\tau^*})^{k+1} e^{-k(\frac{\tau}{\tau^*})} d\tau^*$$

Substituting $u = \frac{-\tau}{\tau^*}$,

$$= (-1)^{k+1} \frac{1}{\tau} \frac{k^{k+1}}{k!} \int_{\frac{-\tau}{m}}^{\frac{-\tau}{n}} u^{k+1} e^{ku} \frac{\tau}{u^2} du$$
$$= (-1)^{k+1} \frac{1}{\tau} \frac{k^{k+1}}{k!} \int_{\frac{-\tau}{m}}^{\frac{-\tau}{n}} u^{k-1} e^{ku} du$$

Using $S(n, c, x) = \int x^n e^{cx} dx = e^{cx} \sum_{i=0}^{n} (-1)^{n-i} \frac{n!}{i! c^{n-i+1}} x^i$,

$$= (-1)^{k+1} (\frac{k^{k+1}}{k!})[S(k-1, k, \frac{-\tau}{n}) - S(k-1, k, \frac{-\tau}{m})$$

Lemma 3: The value perceived by the agent between step m and ∞ in the future when the agent is at a distance τ from the reward can be computed as follows:

$$\int_m^\infty g(\tau, \tau^*) d\tau^* = 1 - e^{-k\frac{(\tau)}{m}} \sum_{i=0}^{k-1} \frac{(k\frac{(\tau)}{m})^i}{i!}. \tag{23}$$

Proof:

$$\int_m^\infty g(\tau, \tau^*) d\tau^* = \int_0^\infty g(\tau^*) d\tau^* - \int_0^m g(\tau^*) d\tau^*$$
$$= (\frac{1}{\tau} \frac{k^{k+1}}{k!} \int_0^\infty (\frac{\tau}{\tau^*})^{k+1} e^{-k(\frac{\tau}{\tau^*})} d\tau^*) - (\int_0^m g(\tau^*) d\tau^*)$$

Substituting $u = \frac{\tau}{\tau^*}$,

$$= (\frac{-k^{k+1}}{k!}\int_{\infty}^{0} u^{k-1}e^{-ku}du) - (\int_0^m g(\tau^*)d\tau^*)$$
$$= (\frac{k^{k+1}}{k!}\int_{0}^{\infty} u^{k-1}e^{-ku}du) - (\int_0^m g(\tau^*)d\tau^*)$$

Using $\int_0^\infty y^m e^{-ky}dy = \frac{\Gamma(m+1)}{k^{m+1}}$,

$$= 1 - \int_0^m g(\tau^*)d\tau^*$$

Using Lemma 1,

$$= 1 - e^{-k\frac{(\tau)}{m}} \sum_{i=0}^{k-1} \frac{(k\frac{(\tau)}{m})^i}{i!}$$

S3 Derivation of the Peak Time of the Impulse Response

$$\frac{\partial \mathcal{T}(\tau, \tau^*)}{\partial \tau^*} = 0.$$

Define $u = \frac{\tau}{\tau^*}$, so the function becomes $\mathcal{T}(\tau, \tau^*) = \frac{1}{\tau}\frac{k^{k+1}}{k!}u^{k+1}e^{-ku}$. Since τ is fixed, we differentiate with respect to u, where $\tau^* = \frac{\tau}{u}$ and $\frac{d\tau^*}{du} = -\frac{\tau}{u^2}$. Using the chain rule:

$$\frac{\partial \mathcal{T}}{\partial \tau^*} = \frac{\partial \mathcal{T}}{\partial u} \cdot \frac{du}{d\tau^*},$$

with $\frac{du}{d\tau^*} = -\frac{\tau}{(\tau^*)^2}$. Compute the derivative:

$$\frac{\partial \mathcal{T}}{\partial u} = \frac{1}{\tau}\frac{k^{k+1}}{k!}\left[(k+1)u^k e^{-ku} - ku^{k+1}e^{-ku}\right] = \frac{1}{\tau}\frac{k^{k+1}}{k!}e^{-ku}u^k\left[(k+1) - ku\right].$$

Setting this to zero:

$$(k+1) - ku = 0 \quad \Rightarrow \quad u = \frac{k+1}{k}.$$

Since $u = \frac{\tau}{\tau^*}$, we solve for τ^*:

$$\tau^* = \frac{k}{k+1}\tau.$$

Thus, the peak occurs at $\tau^*_{\text{peak}} = \frac{k}{k+1}\tau$, showing a linear relationship with τ, where the factor $\frac{k}{k+1}$ approaches 1 as k grows, enhancing accuracy.

S4 Derivation of CV of the Impulse Response

The mean of the impulse response of $g(\tau, \tau^*)$ is:

$$\begin{aligned}
\mu &= \int_0^\infty t\tilde{f}(s;t)dt \\
&= \int_0^\infty t\frac{1}{t}\frac{k^{k+1}}{k!}\left(\frac{t}{\tau^*}\right)^{k+1} e^{-k\frac{t}{\tau^*}}dt \\
&= \frac{k^{k+1}}{k!}\int_0^\infty \left(\frac{t}{\tau^*}\right)^{k+1} e^{-k\frac{t}{\tau^*}}dt \\
&= \tau\frac{k+1}{k}.
\end{aligned}$$

The standard deviation of the impulse response of $g(\tau, \tau^*)$ is:

$$\begin{aligned}
\sigma &= \sqrt{\int_0^\infty (t-\mu)^2\tilde{f}(s;t)dt} \\
&= \sqrt{\int_0^\infty (t-\mu)^2\frac{1}{t}\frac{k^{k+1}}{k!}\left(\frac{t}{\tau^*}\right)^{k+1} e^{-k\frac{t}{\tau^*}}dt} \\
&= \tau\frac{\sqrt{k+1}}{k}.
\end{aligned}$$

Finally, the coefficient of variation is then:

$$CV = \frac{\sigma}{\mu} = \frac{1}{\sqrt{k+1}}.$$

S5 Proof of Optimality for Policy Selection

Consider a finite set of policies Π and a discounting function $f(\tau^*) \geq 0$ with $\sum_{\tau^*=0}^{\infty} f(\tau^*) < \infty$. For each policy $\pi_k \in \Pi$, the true value under f is:

$$V_f^{\pi_k}(s) = \sum_{\tau^*=0}^{\infty} f(\tau^*)g_k(\tau^*),$$

where $g_k(\tau^*) = \mathbb{E}_{\pi_k}[r_{t+\tau^*} \mid s_t = s]$ is the exact expected reward at time τ^*, fully determined by π_k, $\mathcal{P}$, and $\mathcal{R}$. In the limit of continuous Γ, the inverse Laplace transform recovers $g_k(\tau^*)$ exactly from $V_\gamma^{\pi_k}(s) = \sum_{\tau^*=0}^{\infty} \gamma^{\tau^*} g_k(\tau^*)$, as $V_\gamma^{\pi_k}(s)$ is the Laplace transform of $g_k(\tau^*)$ evaluated at $\sigma = -\ln(\gamma)$. Thus, $V_f^{\pi_k}(s)$ precisely captures the expected discounted reward under discount function f.

Since Π is finite, there exists a policy $\pi^* \in \Pi$ such that:

$$V_f^{\pi^*}(s) = \max_{\pi_k \in \Pi} V_f^{\pi_k}(s).$$

The selection $\pi^* = \arg\max_{\pi_k \in \Pi} V_f^{\pi_k}(s)$ identifies this policy, proving that π^* is optimal within Π for the given f under these idealized conditions.