Markov Decision Processes with Time-Varying Geometric Discounting

Jiarui Gan[1], Annika Hennes[2], Rupak Majumdar[3], Debmalya Mandal[3], Goran Radanovic[3]

[1] University of Oxford
[2] Heinrich-Heine-University Düsseldorf
[3] Max Planck Institute for Software Systems
jiarui.gan@cs.ox.ac.uk, annika.hennes@hhu.de, rupak@mpi-sws.org, dmandal@mpi-sws.org, gradanovic@mpi-sws.org

Abstract

Canonical models of Markov decision processes (MDPs) usually consider geometric discounting based on a constant discount factor. While this standard modeling approach has led to many elegant results, some recent studies indicate the necessity of modeling time-varying discounting in certain applications. This paper studies a model of infinite-horizon MDPs with time-varying discount factors. We take a game-theoretic perspective—whereby each time step is treated as an independent decision maker with their own (fixed) discount factor—and we study the *subgame perfect equilibrium* (SPE) of the resulting game as well as the related algorithmic problems. We present a constructive proof of the existence of an SPE and demonstrate the EXPTIME-hardness of computing an SPE. We also turn to the approximate notion of ϵ-SPE and show that an ϵ-SPE exists under milder assumptions. An algorithm is presented to compute an ϵ-SPE, of which an upper bound of the time complexity, as a function of the convergence property of the time-varying discount factor, is provided.

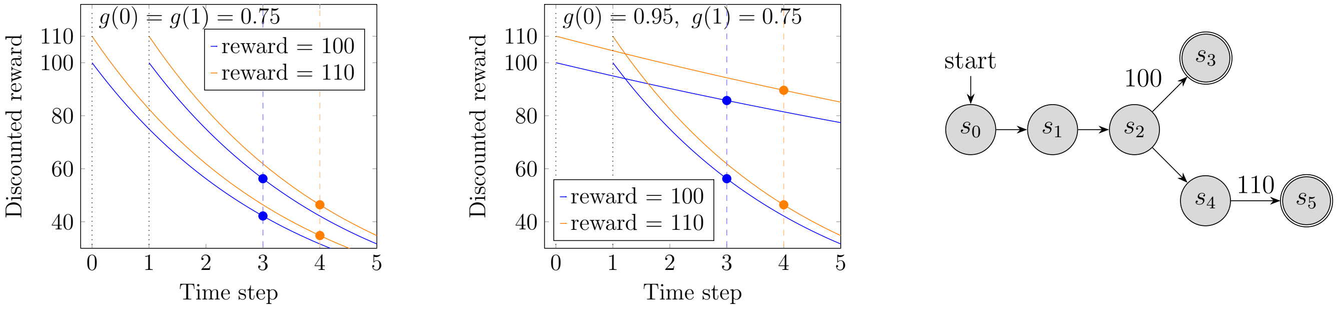
1 Introduction

Ever since Samuelson's foundational work introduced the discounted utility theory (Samuelson 1937), discounted utility models have played a central role in sequential decision making. Building on earlier work that recognized the influence of time on individuals' valuations of goods (e.g., see (Rae 1905; Jevons 1879; von Böhm-Bawerk 1922; Fisher 1930) and the discussion in (Loewe 2006)), Samuelson proposed a utility model in which a decision maker attempts to optimize the discounted sum of their utilities with a constant discount factor applied in every time step; this is known as geometric or exponential discounting.

Geometric discounting leads to many elegant and well-known results. In the context of Markov decision processes (MDPs) (Puterman 1994), it results in the decision maker's preferences over the policies being invariant over time. Moreover, it is key to the existence and polynomial-time computability of an optimal policy. These results have contributed greatly to the popularity and wide applicability of the MDPs. Nevertheless, in many applications, in particular those pertaining to human decision making under uncertainty, time-varying discount factors are essential for capturing long-run utilities. For example, it is shown in laboratory settings that human decision makers often exhibit time-inconsistent behavior: people prefer \$50 in three years plus three weeks to \$20 in three years, yet prefer \$20 now over \$50 in three weeks (Green, Fristoe, and Myerson 1994). Such behaviors are better explained through time-varying discount factors. Unfortunately, many of the aforementioned results break with time-varying discounting. As Strotz showed (Strotz 1955), geometric discounting with a constant discounting factor is the only discount function that satisfies dynamic- or time-consistency.

In this paper, we study a model of infinite-horizon MDPs with time-varying geometric discounting. Our model seizes the idea of geometric discounting, but generalizes the discount factor to a function of time. In each time step, the function produces a discount factor, and the agent's incentive is defined by the geometrically discounted sum of its future rewards with respect to this discount factor. Hence, the agent aims at optimizing a different objective in each time step. This changing incentive gives rise to a game-theoretic approach, proposed and studied in a series of works in the literature (Strotz 1955; Pollak 1968; Peleg and Yaari 1973; Lattimore and Hutter 2014; Jaśkiewicz and Nowak 2021; Lesmana and Pun 2021). Via this approach, the behavior of the sole agent in the process is interpreted as playing against its future selves in a sequential game. Analyzing the *subgame perfect equilibrium* (SPE) is therefore a naturally associated task, which we aim to address in this paper.

More concretely, Figure 1 presents an example that compares the behavior of this model to the standard model of geometric discounting, illustrating how time-varying geometric discounting can lead to time-inconsistency. In the example, the agent has to decide between getting a reward of 100 at time step 3 (option A) and getting a slightly increased reward of 110 one time step later (option B). This decision problem is captured by the MDP in Figure 1c. Under the standard geometric discounting, the preference order of these two outcomes does not change over time: as illustrated in Figure 1a, an agent that discounts its future with a constant factor 0.75 would always prefer option A, no matter at time step 0 or 1. This is not anymore the case with time-varying discounting. An agent who applies a discount factor



(a) Standard geometric discounting.

(b) Time-varying discounting.

(c) MDP (rewards are 0 if unspecified)

Figure 1: Time-consistent vs. time-inconsistent discounting. The blue (resp., orange) curves in (a) and (b) show how the agent values getting a reward of 100 (resp., 110) in future time steps. We examine the player's preferences at the beginning (s_0) and one time step later (s_1). The colored dots correspond to values of the two possible outcomes of the MDP in (c).

of 0.95 is more farsighted and prefers the higher but delayed reward (i.e. option B). But if the agent becomes more myopic one time step later on by applying a reduced discount factor of 0.75, the preference order would change, and the agent would no longer want to stick to its initial plan. This situation is illustrated in Figure 1b.

Time-inconsistent behavior may result from different forms of discounting. It may arise from a consistent way of planning but an inconsistent treatment of future time steps. In hyperbolic discounting, for example, the agent assigns a fixed sequence of varying discount factors to future time steps relative to the current step. In other words, the agent plans with consistency (using the same sequence over time) but treats the future inconsistently (discounting different time steps differently). In contrast, in our model, the agent discounts the future using a constant factor, but this factor might change over time. Human beings, for example, may start with very low discount factors when they are young, but increasingly think more about the future as they grow older. Conceivably, a young person, probably through observing this in elder people, knows that their way of discounting will change when they go into middle age. But still, they cannot do anything about their urge for immediate rewards. Likewise, the agent in our model knows how its rate changes in the future and tries to find a compromise between all the different preferences. This motivates the use of the SPE as our solution concept.

Contributions

Besides introducing a model of MDP with time-varying discounting, we make the following technical contributions.

- We present a constructive proof for the existence of an SPE. Our proof differs from another non-constructive approach in the literature (Lattimore and Hutter 2014) which uses the compactness of the underlying space to argue about convergence.
- From our constructive proof, an algorithm for computing an SPE can be readily extracted. Meanwhile, we demonstrate that the problem of computing an SPE is EXPTIME-hard even in restricted settings.
- In order to circumvent some of the assumptions needed to construct an exact SPE, we turn to the relaxed notion of ϵ-SPE. We show that an ϵ-SPE exists under strictly milder assumptions and present an algorithm to compute an ϵ-SPE. Using a continuity argument of the value functions, we also derive an upper bound on the time complexity of the algorithm, as a function of the convergence property of the time-varying discount factors.

Related Work

A large body of experimental evidence suggests that human behavior is not characterized by geometric discounting with a constant discount factor. Empirical findings that do not support the hypothesis that discounting is consistent over time have been reported (Thaler 1981; Benzion, Rapoport, and Yagil 1989; Redelmeier and Heller 1993; Green, Fristoe, and Myerson 1994; Kirby and Herrnstein 1995; Millar and Navarick 1984), implying dynamic inconsistency of human preferences. Prior work has also proposed and studied different forms of discounting, such as hyperbolic (Herrnstein 1961; Ainslie 1992; Loewenstein and Prelec 1992) or quasi-hyperbolic discounting (Phelps and Pollak 1968; Laibson 1997), which are considered to be more aligned with human behavior (Ainslie 1975; Green, Fry, and Myerson 1994; Kirby 1997). Interpretations of discounting functions as uncertainty over hazard rates were also proposed (Sozou 1998; Dasgupta and Maskin 2005).

Focusing on sequential decision making under uncertainty, this paper closely relates to the line of work that studies non-geometric discount factors and dynamic inconsistency in Markov decision processes and stochastic games (Shapley 1953; Alj and Haurie 1983; Puterman 1994; Nowak 2010; Jaśkiewicz and Nowak 2021; Lesmana and Pun 2021). Some recent works studied dynamic inconsistency using a game-theoretic framework akin to the ones from (Strotz 1955; Pollak 1968; Peleg and Yaari 1973), where their focus is on the existence of an equilibrium in randomized stationary Markov perfect strategies (Jaśkiewicz and Nowak 2021), or an SPE in a finite horizon setting (Lesmana and Pun 2021). Arguably, the closest work to ours is the work of Lattimore and Hutter (Lattimore and Hutter

2014), which considers "age-dependent" (time-varying) geometric discounting functions. The characterization results therein prove the existence of an SPE in the resulting game. Our results are complementary: we provide a constructive proof of the existence result and additionally study the computational complexity of the problem setting.

MDP-based settings similar to ours have also been considered in reinforcement learning (Sutton 1995; Sutton et al. 2011; White 2017; Pitis 2019; Fedus et al. 2019). Increasing the efficacy of learning by using multiple discount factors has been explored previously (Burda et al. 2018; Romoff et al. 2019). It is also worth mentioning settings that use a weighted reward criterion (e.g., (Filar and Vrieze 2012)), where the objective can be expressed as the weighted sum of two value functions with different discount factors.

2 The Model

We consider an infinite horizon MDP $\mathcal{M} = (S, A, R, P, s_{\text{start}}, \gamma)$, where S is a finite state space of the environment, with $|S| = n$, and $A = \bigcup_{s \in S} A_s$ is a union of finite action spaces, with each A_s being the set of actions available in state s. Moreover, $R\colon S \times A \to \mathbb{R}$ is a reward function, such that when action a is taken in state s, a reward $R(s,a)$ will be generated, and the state of the environment transitions according to the transition function $P\colon S \times A \to S$, with probability $P(s,a,s')$ to another state $s' \in S$. Finally, $s_{\text{start}} \in S$ is a starting state, and $\gamma \in [0,1)$ is a discount factor that is applied for defining the *cumulative reward* of a policy π, i.e., the discounted sum of rewards obtained over an infinite horizon:

$$\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \cdot R(s_t, a_t) \,\middle|\, s_0 = s_{\text{start}}, \pi\right], \tag{1}$$

where the expectation is over the trajectory $(s_t, a_t)_{t=0}^{\infty}$ generated by starting from s_{start} and following π subsequently.

Two types of policies will be of interest in this paper: *static policies* and *dynamic policies*. A static policy $\pi : S \to A$ assigns a (deterministic) action to each state, so that using policy π, an agent performs action $\pi(s)$ whenever the environment is in state s, irrespective of the time step. In contrast, a dynamic policy is time-dependent and is defined as a sequence $\boldsymbol{\pi} = (\pi_t)_{t=0}^{\infty}$ of static policies. At each time step t, the static policy π_t is employed to determine the action to take. Hence, dynamic policies are a generalization of static policies.[1]

One may also consider policies that depend on the history (i.e., the trajectory of states and actions $s_0, a_0, s_1, a_1, \ldots, s_{t-1}, a_{t-1}$ generated so far), but as it shall be clear this is unnecessary for the problems studied in this paper: the underlying process is Markovian and the agent always observes the state of the environment. Moreover, it is well-known that when the discount factor γ is a constant, to maximize the cumulative reward defined in (1) it suffices to consider static policies. Though this does not hold when the optimization horizon is finite (Shirmohammadi et al. 2019) or, as we will study in this paper, when γ varies with time.

Constant Discount Factor and Optimality

When a constant discount factor is applied, the optimality of a static policy with respect to (1) can be characterized using the value function V_γ^π defined as:

$$V_\gamma^\pi(s) := Q_\gamma^\pi(s, \pi(s)) \tag{2}$$

for all $s \in S$, where

$$\begin{aligned} Q_\gamma^\pi(s,a) &:= R(s,a) + \gamma \cdot \mathbb{E}_{s' \sim P(s,a,\cdot)} V_\gamma^\pi(s') \\ &= R(s,a) + \gamma \cdot \sum_{s' \in S} P(s,a,s') \cdot V_\gamma^\pi(s') \end{aligned} \tag{3}$$

is called the Q-function. The value of $V_\gamma^\pi(s)$ corresponds to the expected sum of rewards when starting in state s and following policy π. A static policy π^* is optimal if for every state s and every action $a \in A$, it holds that

$$V_\gamma^{\pi^*}(s) = \max_{a \in A_s} Q_\gamma^{\pi^*}(s,a). \tag{4}$$

We denote by Π the set of all static policies, and by Π_γ^* the set of all optimal policies with respect to a constant γ. It is well known that $\Pi_\gamma^* \neq \emptyset$ for all $\gamma \in [0,1)$, and one can compute a policy in Π_γ^* in polynomial time (Rincón-Zapatero and Rodríguez-Palmero 2003).

It will also be useful to introduce the notion of equivalent policies. Two policies are deemed equivalent if their value functions are identical for all states and discount factors.

Definition 2.1 (Equivalent policies). *Two static policies $\pi_1, \pi_2 \in \Pi$ are* equivalent *if for all $s \in S$ and all $\gamma \in [0,1)$ it holds that $V_\gamma^{\pi_1}(s) = V_\gamma^{\pi_2}(s)$. We write $\pi_1 \sim \pi_2$ if π_1 and π_2 are equivalent.*

Time-Varying Discounting—Game-Theoretic View

We generalize the above definition to *MDPs with time-varying discounting* (hereafter, MDPs for simplicity) by replacing the constant factor γ by a *discount function* $g : \mathbb{N} \to [0,1)$, such that $g(t)$ is the discount factor the agent applies at time step t. We will only consider discount functions that converge to a value in $[0,1]$ when $t \to \infty$ in this paper.

Time-varying discounting changes the agent's incentive over time and as a result the agent behaves as if they are different agents. Hence, we apply a game-theoretic view and view the MDP as a sequential game played by countably many players. Every player is associated with a time step $t \in \mathbb{N}$ and decides on a static policy π_t to use at that particular time step. Moreover, player t represents the agent's incentive at time step t and cares about the subsequent cumulative reward with respect to the (constant) discount factor $g(t)$, i.e.,

$$u_t(\boldsymbol{\pi}|s) := \mathbb{E}\left[\sum_{t'=t}^{\infty} g(t)^{t'-t} \cdot R(s_{t'}, a_{t'}) \,\middle|\, s_t = s, \boldsymbol{\pi}\right] \tag{5}$$

[1] More generally, a static policy can also choose randomized actions, i.e., $\pi : S \to \Delta(A)$. Nevertheless, it is without loss of generality to consider only deterministic policies with respect to all the results in our paper. Hence, unless otherwise clarified, all static policies considered are deterministic ones, whereas we do allow the agent to use randomized static policies.

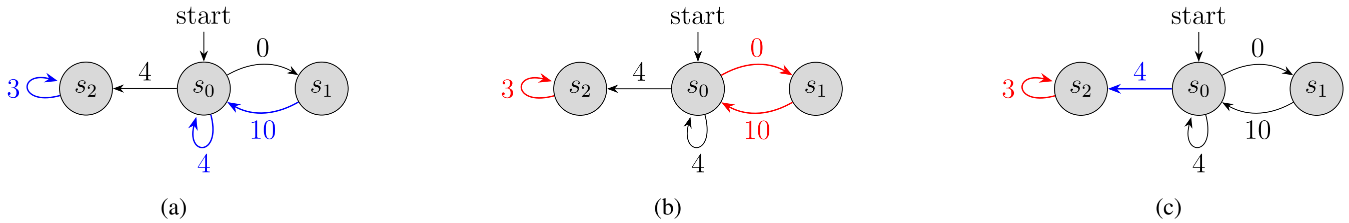


Figure 2: Optimal static policies for geometric discounting with different values of γ and an SPE. (a) and (b) illustrate the optimal policies for $\gamma_1 = 0.1$ (blue) and $\gamma_2 = 0.8$ (red), respectively. (c) illustrates an SPE under time-varying discounting, with $g(0) = \gamma_1$ and $g(i) = \gamma_2$ for all $i \geq 1$. The action chosen by player 0 is depicted in blue, the actions chosen by all subsequent players are in red.

when the environment is in state s before player t is to take an action, and the other players $t+1, t+2, \dots$ subsequently act according to $\pi_{t+1}, \pi_{t+2}, \dots$ given by the dynamic policy $\boldsymbol{\pi} = (\pi_t)_{t=0}^{\infty}$. In other words, each player has the same geometric-discounting-style vision as that defined in (1). The function u_t can be viewed as the utility function of player t conditioned on s_t, and $\boldsymbol{\pi}$ as the players' strategy profile. The discount factor stays constant for this particular player, but it might be different for different players. We will analyze the *subgame-perfect equilibrium* (SPE) of the resulting game, which is the standard solution concept for sequential games (Osborne et al. 2004).

Definition 2.2 (SPE). *A dynamic policy $\boldsymbol{\pi} = (\pi_t)_{t=0}^{\infty}$ is an SPE if for all $t \in \mathbb{N}$ and $s \in S$ it holds that: $u_t(\boldsymbol{\pi}|s) \geq u_t(\boldsymbol{\pi}'|s)$ if $\pi'_i = \pi_i$ for all $i \in \mathbb{N} \setminus \{t\}$.*

In other words, in an SPE, from any time step t onward, the players' policies form a Nash equilibrium of the subsequent subgame, no matter what s_t is. Note that the above definition takes the same form as a Nash equilibrium of the players' policies because every player t only plays at time step t throughout the game.

We can use value functions to characterize an SPE: a dynamic policy $\boldsymbol{\pi}^*$ is an SPE if it holds that

$$V^{\boldsymbol{\pi}^*}_{g(t),t}(s) = \max_{a \in A_s} Q^{\boldsymbol{\pi}^*}_{g(t),t}(s,a), \tag{6}$$

for all $t = 0, 1, \dots$ and $s \in S$, where for any $\boldsymbol{\pi}$ we define

$$V^{\boldsymbol{\pi}}_{\gamma,t}(s) \coloneqq Q^{\boldsymbol{\pi}}_{\gamma,t}(s, \pi_t(s)), \text{ and} \tag{7}$$

$$Q^{\boldsymbol{\pi}}_{\gamma,t}(s,a) \coloneqq R(s,a) + \gamma \sum_{s' \in S} P(s,a,s') V^{\boldsymbol{\pi}}_{\gamma,t+1}(s'). \tag{8}$$

Namely, each player t has a value function $V^{\boldsymbol{\pi}}_{g(t),i}$ and a Q-function $Q^{\boldsymbol{\pi}}_{g(t),i}$ for each time step i, defined with respect to their own discount factor $g(t)$. We can make two observations below: the first observation follows by definition (i.e., (5)), and the second holds as the dynamic policy essentially degenerates to a static one in the stated situation.

Observation 2.3. *$u_t(\boldsymbol{\pi}|s) \equiv V^{\boldsymbol{\pi}}_{g(t),t}(s)$ for all $s \in S$.*

Observation 2.4. *Let $\boldsymbol{\pi} = (\pi_t)_{t=0}^{\infty}$ be a dynamic policy and $\pi \in \Pi$ be a static policy. If $\pi_t = \pi$ for all $t \geq T$, then $V^{\boldsymbol{\pi}}_{\gamma,t}(s) = V^{\pi}_{\gamma}(s)$ for all $t \geq T$ and all $s \in S$.*

We analyze the problem of computing an SPE. Since a solution to this problem is a dynamic policy over an infinite horizon, it is not immediately clear whether a solution admits any concise representation. We therefore consider only the first step ($t = 0$) and the following decision problem: *For a given action $a \in A$, is there an SPE $\boldsymbol{\pi}$ such that $\pi_0(s_0) = a$?* (More formally, see the definition below.) We refer to this problem as SPE-START.

Definition 2.5 (SPE-START). *An instance of* SPE-START *is given by a tuple $(\mathcal{M}, a^\dagger)$, consisting of an MDP $\mathcal{M} = (S, A, R, P, s_{start}, g)$ (with a time-varying discount function g) and an action $a^\dagger \in A$. It is a yes-instance if there exists an SPE $\boldsymbol{\pi}$ such that $\pi_0(s_{start}) = a^\dagger$; and a no-instance, otherwise.*

It is straightforward that when g is a constant function, an SPE corresponds to an optimal policy for the MDP. Yet, it appears that SPE-START is computationally more demanding than computing an optimal policy in a constant-discounting MDP: as we will show in the paper, SPE-START is EXPTIME-hard, whereas the latter is well-known to be solvable in polynomial time.

An Illustrating Example

We conclude our description of the model with an illustrating example. Consider the MDP given in Figure 2, and two different discount factors, $\gamma_1 = 0.1$ and $\gamma_2 = 0.8$. Let s_0 be the starting state. As we are solely considering deterministic state transitions in this example, we will identify actions starting in a certain state with the state it changes to, e.g. $\pi(s_0) = s_1$ denotes the action that causes a transition from s_0 to s_1.

As shown in Figure 2a and Figure 2b, the optimal static policy $\pi^*_{\gamma_1}$ of an agent who applies a constant discount factor γ_1 in the classical setting, is given by $\pi^*_{\gamma_1}(s_0) = s_0$, $\pi^*_{\gamma_1}(s_1) = s_0$ and $\pi^*_{\gamma_1}(s_2) = s_2$. For γ_2, the optimal static policy differs from only in state s_0, namely $\pi^*_{\gamma_2}(s_0) = s_1$.

We want to compute an SPE for the setting where the first player discounts with γ_1 and all subsequent players discount with γ_2. Since the discount factor is constant from time step 1 onward, we can fix the policies of players $1, 2, \dots$ to $\pi^*_{\gamma_2}$ to derive an SPE. Player 0 knows all future players' policies and can influence future rewards merely by choosing the current action. Starting in s_0, the action given by policy $\pi^*_{\gamma_1}$ would be the one that transitions to itself. As player

0 is fairly short-sighted, knowing that player 1 will choose an action leading to reward 0 in the subsequent time step, it prefers taking an immediate reward of 4 and transitioning into state s_2 to stay there forever and keep receiving rewards of 3. Hence, an SPE is given by $\boldsymbol{\pi}_{\text{SPE}} = (\pi', \pi^*_{\gamma_2}, \pi^*_{\gamma_2}, \ldots)$, where $\pi'(s_0) = s_2$; see Figure 2c.

3 Existence of an SPE

We investigate the existence of an SPE. Indeed, the following result has already answered this question in affirmative.

Theorem 3.1 (Lattimore and Hutter 2014)**.** *An (exact) SPE always exists.*

The result applies even without our assumption on the convergence of the discount function but it is obtained via a non-constructive approach: by reasoning about a sequence of policies that are optimal for the truncated versions of the problem, each with a longer horizon than their predecessors. Hence, the proof does not yield any algorithm or procedure for obtaining an SPE. We next provide a constructive proof. The proof also forms the basis for deriving a complexity upper bound of SPE-START as we will demonstrate later.

We will in particular show that there exists an SPE that is *eventually constant*, i.e., there exists a time step after which all the subsequent players use the same static policy. A mild assumption is needed for our proof: g converges to a value outside of a set of *degenerate points* defined below.

Definition 3.2 (Degenerate point)**.** *A discount factor γ is called a* degenerate point *if Π^*_γ contains more than one non-equivalent policy (see Definition 2.1), i.e., $|\Pi^*_\gamma/\sim| > 1$, where $\Pi^*_\gamma/\sim$ is the set of equivalence classes on Π^*_γ under the equivalence relation $\sim$ defined in Definition 2.1, i.e., an element $\{\pi' \in \Pi : \pi' \sim \pi\} \in \Pi^*_\gamma/\sim$ contains all policies equivalent to π.*

In what follows, let

$$\Gamma := \left\{\gamma \in [0,1) : |\Pi^*_\gamma/\sim| > 1\right\}$$

be the set of degenerate points in $[0,1)$. The assumption above ensures that the players will eventually adopt the same behavior after some time step T, as if the subsequent process is a constant-discounting MDP. From that point on, the dynamic policy can be represented as a static one and we can use backward induction to derive policies for players in the previous time steps. More formally, the main result of this section is formulated as follows.

Theorem 3.3. *Suppose that $\gamma^* := \lim_{t\to\infty} g(t)$ exists and $\gamma^* \in [0,1] \setminus \Gamma$. Then there exists an SPE $\boldsymbol{\pi}$ that is eventually constant, i.e., there exists a number $T \in \mathbb{N}$ such that $\pi_t = \pi_T$ for all $t \geq T$.*

Proof of Theorem 3.3

The key to the proof is to argue that $\Pi^*_{g(t)}$ is eventually constant after a certain time step T. With this property, we can pick an arbitrary $\tilde{\pi} \in \Pi^*_{g(T)}$ and assign it to all the players $t \geq T$. This forms an SPE for the subgame starting at T, and according to Observation 2.4, we can use $V^{\pi}_{g(t),T}$ as a basis and use backward induction to construct $\pi_{T-1}, \pi_{T-2}, \ldots, \pi_0$ as the optimal policies of players $T-1, T-2, \ldots, 0$ with respect to $V^{\pi}_{g(T-1),T}, V^{\pi}_{g(T-2),T-1}, \ldots, V^{\pi}_{g(0),1}$, respectively. The approach is summarized in Algorithm 1.

Algorithm 1: Constructing an SPE $\boldsymbol{\pi} = (\pi_t)^{\infty}_{t=0}$, given that $\pi_t = \tilde{\pi}$ for all $t \geq T$

```
Input  : a static policy $\tilde{\pi} \in \Pi$, and a time step $T \in \mathbb{N}$
Output: an SPE $\boldsymbol{\pi} = (\pi_t)^{\infty}_{t=0}$
for $t = T-1, T-2, \ldots, 0$ do
  Compute $V^{\tilde{\pi}}_{g(t)}$ defined according to (2) and (3);
  $V_{t,T}(s) \leftarrow V^{\tilde{\pi}}_{g(t)}(s)$ for all $s \in S$;
  // so $V_{t,T} = V^{\boldsymbol{\pi}}_{g(t),T}$ (Observation 2.4)
  for $i = T-1, T-2, \ldots, t$ do
    for each $s \in S, a \in A_s$ do
      $Q_{t,i}(s,a) \leftarrow$
        $R(s,a) + \gamma \sum_{s' \in S} P(s,a,s') \cdot V_{t,i+1}(s')$;
      $V_{t,i}(s) \leftarrow Q_{t,i}(s, \pi_{t+1}(s))$;     //so $V_{t,i}{=}V^{\boldsymbol{\pi}}_{g(t),i}$
        in (7)
  for each $s \in S$ do
    $\pi_t(s) \leftarrow$ arbitrary action in
      $\arg\max_{a \in A_s} Q_{t,t}(s, \pi_{t+1}(s))$;
```

Now to show that $\Pi^*_{g(t)}$ is eventually constant, we argue that the set Γ of degenerate points is finite (Lemma 3.4). Since $\gamma^* \notin \Gamma$, there must be a neighbourhood of γ^* in $\mathbb{R}$ which does not intersect Γ. After a certain time step, the tail of g will be contained inside this neighbourhood, so Lemma 3.6 then implies that $\Pi^*_{g(t)}$ becomes constant after a finite number of time steps.

Lemma 3.4. *Γ is a finite set.*

Proof. Define

$$h^s_{\pi_1,\pi_2}(\gamma) := V^{\pi_1}_\gamma(s) - V^{\pi_2}_\gamma(s) \tag{9}$$

By definition, for any $\gamma \in \Gamma$, there exist $\pi_1, \pi_2 \in \Pi^*_\gamma$ such that $\pi_1 \not\sim \pi_2$, which means that $h^s_{\pi_1,\pi_2}(\gamma) = 0$ for all $s \in S$. Hence, $|\Gamma|$ is bounded from above by the number of γs such that $h^s_{\pi_1,\pi_2}(\gamma) = 0$ for some $s \in S$ and some $\pi_1, \pi_2 \in \Pi$ that are not equivalent. By Lemma 3.5, $h^s_{\pi_1,\pi_2}(\gamma) = \Psi(\gamma)/\Phi(\gamma)$, where both $\Psi(\gamma)$ and $\Phi(\gamma)$ are polynomial functions of γ with finite degrees. Hence, the number of zeros of $h^s_{\pi_1,\pi_2}(\gamma)$ is finite. □

Lemma 3.5. *Let $\pi_1, \pi_2 \in \Pi$ be two policies and $\gamma \in [0,1)$. The function $h^s_{\pi_1,\pi_2}(\gamma)$ can be written as*

$$h^s_{\pi_1,\pi_2}(\gamma) = \Psi(\gamma)/\Phi(\gamma), \tag{10}$$

where Ψ and Φ are polynomials of γ with finite degrees.[2]

Lemma 3.6. *For any interval $I \subseteq [0,1)$ such that $I \cap \Gamma = \emptyset$, we have $\Pi^*_\gamma = \Pi^*_{\gamma'}$ for any $\gamma, \gamma' \in I$.*

[2] Omitted proofs can be found in the full version of this paper.

Proof. Without loss of generality, suppose that $\gamma < \gamma'$ and, for the sake of contradiction, $\Pi^*_\gamma \neq \Pi^*_{\gamma'}$. The fact that $\gamma, \gamma' \notin \Gamma$ directly implies that $\Pi^*_\gamma \cap \Pi^*_{\gamma'} = \emptyset$ as both sets contain only equivalent policies. Pick arbitrary $\pi \in \Pi^*_\gamma$ and $\pi' \in \Pi^*_{\gamma'}$. We have $h_{\pi,\pi'}(\gamma) > 0$ and $h_{\pi,\pi'}(\gamma') < 0$. As $h_{\pi,\pi'}$ is a continuous function, there exists a $\tilde{\gamma} \in (\gamma, \gamma') \subseteq I$ such that $h_{\pi,\pi'}(\gamma') = 0$, which implies $|\Pi^*_{\tilde{\gamma}} / \sim| > 1$ and hence $\tilde{\gamma} \in \Gamma$. This contradicts the assumption that $I \cap \Gamma = \emptyset$. □

4 Complexity of SPE-Start

Consider using Algorithm 1 to construct an SPE. It requires specifying T in the input which we have not yet described how to obtain. Indeed, this replies on the specific format of g. In addition to the computational cost of obtaining T, Algorithm 1 includes $O(T^2 \cdot |S|)$ iterations, so the overall time complexity also depends on the magnitude of T. The latter cost prevents the algorithm from being efficient if T is exponential in the size of the problem, so the question is whether there are better algorithms that solve SPE-Start without going through all the iterations. It turns out that this is in general not possible: as we will show next, even for discount functions that admit efficient computation of T, computing an SPE can be EXPTIME-hard.

EXPTIME-Hardness

We show that SPE-Start is EXPTIME-hard even when $g : \mathbb{N} \to [0, 1)$ is a *down-step* function defined as follows.

$$g(t) \coloneqq \begin{cases} \gamma & \text{if } t \leq T \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

where $\gamma \in (0, 1)$ and $T \in \mathbb{N}$ is encoded in binary. Arguably this is one of the simplest forms of time-varying discounting, and g can be encoded as (γ, T).

Note that (11) does not define a finite horizon MDP. Instead, it defines a game where eventually all the players from time step T onward exhibit a discount factor of 0. We will show that SPE-Start is EXPTIME-hard even when the discount function is restricted to this simple form. The proof uses a reduction from the following problem, termed ValIt (value iteration), which is known to be EXPTIME-complete (Shirmohammadi et al. 2019).

Definition 4.1 (ValIt)**.** *An instance of* ValIt*, given by* $(\mathcal{M}, a^\dagger, T)$*, consists of an MDP* $\mathcal{M} = (S, A, R, P, s_{\mathit{start}}, \gamma)$ *with constant discount factor* $\gamma > 0$*, an action* $a^\dagger \in A$*, and finite time horizon* $T \in \mathbb{N}$ *encoded in binary. It is a yes-instance if there exists a dynamic policy* $\boldsymbol{\pi}$ *such that* $\pi_0(s_{\mathit{start}}) = a^\dagger$ *and* $\pi_t(s) \in \arg\max_{a \in A_s} Q_t(s, a)$ *for all* $t = 0, \ldots, T-1$ *and* $s \in S$*, where*

$$Q_t(s, a) \coloneqq R(s, a) + \gamma \sum_{s' \in S} P(s, a, s') \cdot V_{t+1}(s'), \tag{12}$$

$$V_t(s) \coloneqq \max_{a \in A_s} Q_t(s, a), \tag{13}$$

and $Q_T(s, a) \equiv 0$*. Otherwise, it is a no-instance.*

The V_t functions in the above definition are akin to the value functions defined in (4) but with a time-dependency. Using ValIt, we prove the following result.

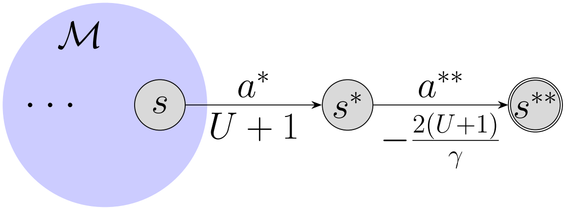


Figure 3: Reduction from ValIt to SPE-Start. The blue disk represents the original MDP $\mathcal{M}$ in the ValIt instance, and the outer nodes indicate how to extend $\mathcal{M}$ to an MDP for the SPE-Start instance, where the discount rate is fixed to the original discount rate γ in the first T time steps and set to 0 afterwards. Labels above edges are action names and labels below are rewards.

Theorem 4.2. SPE-Start *is EXPTIME-hard even when the discount function is a down-step function.*

Proof sketch. We reduce ValIt to SPE-Start. The main idea of the reduction is to construct an SPE-Start instance where all players $t > T$ will stick to the same static policies regardless of policies chosen by the preceding players. Figure 3 illustrates the MDP in the SPE-Start instance. A chain consisting of two states s^* and s^{**} is appended to every state s in the ValIt instance. The high reward at a^* ensures that a^* is the *dominant* action for player T, who has $g(T) = 0$ and only cares about the immediate reward; whereas the high penalty at a^{**} ensures that a^* is a *dominated* action for all players $t = 0, \ldots, T$, who have $g(t) = \gamma$. Hence, for every player $t \leq T$ in the SPE-Start instance, the process is equivalent to an MDP with time horizon $T + 1$. The procedure to derive $\pi_T, \pi_{T-1}, \ldots, \pi_0$ in an SPE using backward induction is the same as computing the value functions of the ValIt instance. Every SPE is then associated to an optimal policy of ValIt. □

We remark that the binary encoding of T plays a crucial role in the EXPTIME-hardness of SPE-Start. Indeed, if T is encoded in unary or is a constant, the hardness will disappear. In general, an efficient algorithm for computing an SPE is possible but requires the assumption of g converging fast enough to an interval between two consecutive numbers in Γ. To ease part of the intricacies introduced by the requirement, a practical approach which we will present next is by considering the approximate notion of the SPE, the ϵ-SPE.

5 Approximate SPEs

The ϵ-SPE, defined below, assumes that the players are reluctant to deviate as long as the potential improvement is smaller than some $\epsilon > 0$.

Definition 5.1 (ϵ-SPE)**.** *A dynamic policy* $\boldsymbol{\pi} = (\pi_t)_{t=0}^\infty$ *forms an* ϵ*-SPE if for all* $t \in \mathbb{N}$ *it holds that for all* $s \in S$*:* $u_t(\boldsymbol{\pi}|s) \geq u_t(\boldsymbol{\pi}'|s) - \epsilon$ *for all* $\boldsymbol{\pi}' = (\pi'_t)_{t=0}^\infty$ *such that* $\pi'_i = \pi_i$ *for all* $i \in \mathbb{N} \setminus \{t\}$*.*

The notion allows us to relax the assumption that g converges to a point outside of Γ and allows us to derive an upper bound of the computational complexity, too. Indeed, the existence of an ϵ-SPE, in particular an *eventually constant* one, does not require this assumption. Instead, we use the following continuity argument.

Lemma 5.2. *Suppose that all rewards are bounded by M. Then for any discount factors $\gamma, \tilde{\gamma} \in [0, 1)$ and static policy $\pi \in \Pi$, we have the following bound for all $s \in S$:*

$$|V_{\gamma}^{\pi}(s) - V_{\tilde{\gamma}}^{\pi}(s)| \leq \frac{2M \cdot |S| \cdot |\gamma - \tilde{\gamma}|}{(1 - \max\{\gamma, \tilde{\gamma}\})^3 \cdot (1 - \min\{\gamma, \tilde{\gamma}\})}.$$

Theorem 5.3. *Suppose that $\gamma^* := \lim_{t\to\infty} g(t)$ exists and $\gamma^* \in [0, 1]$. For any $\epsilon > 0$, there exists an ϵ-SPE $\boldsymbol{\pi}$ that is* eventually constant, *i.e., there exists a $T \in \mathbb{N}$ such that $\pi_t = \pi_T$ for all $t \geq T$.*

Computing an ϵ-SPE

The ϵ slackness introduced by ϵ-SPE appears to suggest that it suffices to consider a finite time horizon: a player can cut off the time horizon up to a certain (finite) future time step, beyond which the sum of the discounted rewards is sufficiently small to be ignored. This is nevertheless not the case. Even if the horizon is cut off, there are still infinitely many players in the game and each player's payoff is influenced by the subsequent players before the cutting off point. Hence, cutting off the time horizon does not reduce our consideration to a finite number of time steps.

To compute an ϵ-SPE, we use the continuity argument in Lemma 5.2. If we can pin down a time step t after which the tail of g is contained in a sufficiently small interval, we can use $g(t)$ to compute an SPE for the subgame as if g is constant after t. This approximates an actual SPE provided that the tail of $g(t)$ is sufficiently small. Hence, the time complexity depends on the rate at which g converges. In accordance with our existence proof, let d be a number such that

$$\min_{\gamma\in\Gamma} \left| \lim_{t\to\infty} g(t) - \gamma \right| \geq d.$$

To derive a general result, we also assume that there is an oracle $\mathcal{A}$ that, for any given $\delta > 0$, computes a time step T such that $|g(t) - g(T)| \leq \delta$ for all $t \geq T$. More specifically, we introduce the following notion called (α, β)-convergence for the discount function.

Definition 5.4 ((α, β)-convergence)**.** *Let $\alpha : \mathbb{R}^2 \to \mathbb{R}$ and $\beta : \mathbb{R}^2 \to \mathbb{R}$. A class $\mathcal{G}$ of discount functions is* (α, β)-convergent *if there is an oracle $\mathcal{A}$ such that: for any $g \in \mathcal{G}$ and any $\delta > 0$ with bit-size d, $\mathcal{A}$ computes an integer T in time $\alpha(|g|, d)$ such that $|g(t) - g(T)| \leq \delta$ for all $t \geq T$, and $T \leq \beta(|g|, d)$, where $|g|$ denotes the bit-size of the representation of g.*

For example, the class of down-step functions, defined in (11) and encoded as (γ, T) (in binary), is (α, β)-convergent with $\alpha(x, y) = x$ and $\beta(x, y) = O(2^x)$. Our next lemma provides a lower bound on the distance between any two points in the set Γ.

Theorem 5.5. *Suppose that g is (α, β)-convergent and $\lim_{t\to\infty} g(t) < 1 - c$ for a known constant c. Then an ϵ-SPE can be computed in time $\alpha(|g|, d) + poly(|A|, |S|) \cdot (\beta(|g|, d))^2$, where $d = \log(M|S|/\epsilon) + o(1)$.*

Algorithm 2: Computing an ϵ-SPE

```
Input  : $\epsilon > 0$
Output: an $\epsilon$-SPE $\boldsymbol{\pi} = (\pi_t)_{t=0}^{\infty}$
$D \leftarrow c^4 \cdot \min\{\epsilon/4M|S|,\ c\}$;
$T \leftarrow \mathcal{A}(D)$;
$\tilde{\pi} \leftarrow$ an arbitrary policy in $\Pi^*_{g(T)}$;
$\pi_t \leftarrow \tilde{\pi}$, for all $t = T, T+1, \ldots$;
Run Algorithm 1 on input $\tilde{\pi}$ to construct $\pi_0, \ldots, \pi_{T-1}$;
```

Proof. We use Algorithm 2 to compute an ϵ-SPE. To see that it correctly computes an ϵ-SPE, it suffices to argue that $(\pi_T, \pi_{T+1}, \ldots)$ form an ϵ-SPE for the subgame after T.

Indeed, for any player $t \geq T$, we have $|g(t) - g(T)| \leq D \leq \frac{c^4 \cdot \epsilon}{2M|S|}$. Hence, according to Lemma 5.2, we have $|V^{\pi}_{g(t)}(s) - V^{\pi}_{g(T)}(s)| \leq \epsilon/2$ for any static policy π and $s \in S$. Let $\pi \in \Pi^*_{g(t)}$. We have

$$V^{\pi}_{g(t)}(s) - V^{\tilde{\pi}}_{g(t)}(s) \leq V^{\pi}_{g(T)}(s) - V^{\tilde{\pi}}_{g(T)}(s) + \\ |V^{\pi}_{g(t)}(s) - V^{\pi}_{g(T)}(s)| + |V^{\tilde{\pi}}_{g(t)}(s) - V^{\tilde{\pi}}_{g(T)}(s)| \leq \epsilon,$$

where $\tilde{\pi} \in \Pi^*_{g(T)}$ as in Algorithm 2. Moreover, since the optimal static policy is at least as good as any dynamic policy for player t. This means that for any strategy profile $\boldsymbol{\pi}'$ resulting from a deviation of player t,

$$u_t(\boldsymbol{\pi}'|s) - u_t(\pi_T, \pi_{T+1}, \ldots|s) \leq V^{\pi}_{g(t)}(s) - V^{\tilde{\pi}}_{g(t)}(s) = \epsilon.$$

Hence, Algorithm 2 generates an ϵ-SPE.

To see the time complexity of the algorithm, note that it takes time $\alpha(|g|, \log D)$ to run $\mathcal{A}$. In addition to that, the time it takes to run Algorithm 1 is bounded by $(\beta(|g|, d))^2 \cdot poly(|A|, |S|)$. □

We remark that Theorem 5.5 only requires the mild assumption of a known constant gap between 1 and the limit point of g. If c is unknown or the gap cannot be bounded by a constant, an ϵ-SPE can be computed via a more sophisticated algorithm with a higher time complexity. We provide this algorithm in the full version of this paper for theoretical interest.

Via Theorem 5.5, an exponential upper bound of the complexity of computing an ϵ-SPE that is can be derived when g is the down-step function defined in (11) (for which $\beta(|g|, d) = 2^{O(|g|)}$). This does not require any assumption on the convergence of g with respect to Γ. Better bounds can be derived if g converges faster, e.g., $\beta(|g|, d) = d$ or even $2^{O(d)}$, or when g is not a variable of the model.

6 Conclusion

We study a model of infinite-horizon MDPs with time-varying discounting. Our model seizes the idea of geometric discounting, but with time-varying discount factors, and it allows for a game-theoretic interpretation. We study the SPE of the underlying game. Results on the existence and computation of an exact or an ϵ-SPE are presented. Future work can be done to consider other types of discount functions, such as the ones described in (Lattimore and Hutter 2014).

Acknowledgements

The authors would like to thank the anonymous reviewers for their insightful comments. This project was supported by DFG project 389792660 TRR 248–CPEC. Part of the work was done when Jiarui Gan was a postdoc at MPI-SWS, when he was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 945719). Part of the work was done when Annika Hennes was an intern at MPI-SWS. Further, she was partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 456558332. Goran Radanovic's research was, in part, funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 467367360.

References

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

Ainslie, G. 1992. *Picoeconomics: The strategic interaction of successive motivational states within the person*. Cambridge University Press.

Alj, A.; and Haurie, A. 1983. Dynamic equilibria in multi-generation stochastic games. *IEEE Transactions on Automatic Control*, 28(2): 193–203.

Benzion, U.; Rapoport, A.; and Yagil, J. 1989. Discount rates inferred from decisions: An experimental study. *Management science*, 35(3): 270–284.

Burda, Y.; Edwards, H.; Storkey, A.; and Klimov, O. 2018. Exploration by random network distillation. In *International Conference on Learning Representations*.

Dasgupta, P.; and Maskin, E. 2005. Uncertainty and hyperbolic discounting. *American Economic Review*, 95(4): 1290–1299.

Fedus, W.; Gelada, C.; Bengio, Y.; Bellemare, M. G.; and Larochelle, H. 2019. Hyperbolic discounting and learning over multiple horizons. *arXiv preprint arXiv:1902.06865*.

Filar, J.; and Vrieze, K. 2012. *Competitive Markov decision processes*. Springer Science & Business Media.

Fisher, I. 1930. *Theory of interest: as determined by impatience to spend income and opportunity to invest it*. Augustusm Kelly Publishers, Clifton.

Green, L.; Fristoe, N.; and Myerson, J. 1994. Temporal discounting and preference reversals in choice between delayed outcomes. *Psychonomic Bulletin & Review*, 1(3): 383–389.

Green, L.; Fry, A. F.; and Myerson, J. 1994. Discounting of delayed rewards: A life-span comparison. *Psychological science*, 5(1): 33–36.

Herrnstein, R. J. 1961. Relative and absolute strength of response as a function of frequency of reinforcement. *Journal of the experimental analysis of behavior*, 4(3): 267.

Jaśkiewicz, A.; and Nowak, A. S. 2021. Markov decision processes with quasi-hyperbolic discounting. *Finance and Stochastics*, 25(2): 189–229.

Jevons, W. S. 1879. *The theory of political economy*. Macmillan and Company.

Kirby, K. N. 1997. Bidding on the future: Evidence against normative discounting of delayed rewards. *Journal of Experimental Psychology: General*, 126(1): 54.

Kirby, K. N.; and Herrnstein, R. J. 1995. Preference reversals due to myopic discounting of delayed reward. *Psychological science*, 6(2): 83–89.

Laibson, D. 1997. Golden eggs and hyperbolic discounting. *The Quarterly Journal of Economics*, 112(2): 443–478.

Lattimore, T.; and Hutter, M. 2014. General time consistent discounting. *Theoretical Computer Science*, 519: 140–154.

Lesmana, N. S.; and Pun, C. S. 2021. A Subgame Perfect Equilibrium Reinforcement Learning Approach to Time-inconsistent Problems. *Available at SSRN 3951936*.

Loewe, G. 2006. The development of a theory of rational intertemporal choice. *Papers: revista de sociologia*, 195–221.

Loewenstein, G.; and Prelec, D. 1992. Anomalies in intertemporal choice: Evidence and an interpretation. *The Quarterly Journal of Economics*, 107(2): 573–597.

Millar, A.; and Navarick, D. J. 1984. Self-control and choice in humans: Effects of video game playing as a positive reinforcer. *Learning and Motivation*, 15(2): 203–218.

Nowak, A. 2010. On a noncooperative stochastic game played by internally cooperating generations. *Journal of optimization theory and applications*, 144(1): 88–106.

Osborne, M. J.; et al. 2004. *An introduction to game theory*, volume 3. Oxford University Press New York.

Peleg, B.; and Yaari, M. E. 1973. On the existence of a consistent course of action when tastes are changing. *The Review of Economic Studies*, 40(3): 391–401.

Phelps, E. S.; and Pollak, R. A. 1968. On second-best national saving and game-equilibrium growth. *The Review of Economic Studies*, 35(2): 185–199.

Pitis, S. 2019. Rethinking the discount factor in reinforcement learning: A decision theoretic approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 7949–7956.

Pollak, R. A. 1968. Consistent planning. *The Review of Economic Studies*, 35(2): 201–208.

Puterman, M. L. 1994. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc.

Rae, J. 1905. *The sociological theory of capital: being a complete reprint of the new principles of political economy, 1834*. Macmillan.

Redelmeier, D. A.; and Heller, D. N. 1993. Time preference in medical decision making and cost-effectiveness analysis. *Medical Decision Making*, 13(3): 212–217.

Rincón-Zapatero, J. P.; and Rodríguez-Palmero, C. 2003. Existence and uniqueness of solutions to the Bellman equation in the unbounded case. *Econometrica*, 71(5): 1519–1555.

Romoff, J.; Henderson, P.; Touati, A.; Brunskill, E.; Pineau, J.; and Ollivier, Y. 2019. Separating value functions across time-scales. In *International Conference on Machine Learning*, 5468–5477. PMLR.

Samuelson, P. A. 1937. A note on measurement of utility. *The Review of Economic Studies*, 4(2): 155–161.

Shapley, L. S. 1953. Stochastic games. *Proceedings of the National Academy of Sciences*, 39(10): 1095–1100.

Shirmohammadi, M.; Balaji, N.; Kiefer, S.; Novotny, P.; and Pérez, G. A. 2019. On the Complexity of Value Iteration. In *46th International Colloquium on Automata, Languages, and Programming, ICALP 2019, July 9-12, 2019, Patras, Greece*.

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

Strotz, R. H. 1955. Myopia and inconsistency in dynamic utility maximization. *The review of economic studies*, 23(3): 165–180.

Sutton, R. S. 1995. TD models: Modeling the world at a mixture of time scales. In *Machine Learning Proceedings 1995*, 531–539. Elsevier.

Sutton, R. S.; Modayil, J.; Delp, M.; Degris, T.; Pilarski, P. M.; White, A.; and Precup, D. 2011. Horde: A scalable real-time architecture for learning knowledge from unsupervised sensorimotor interaction. In *The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, 761–768.

Thaler, R. 1981. Some empirical evidence on dynamic inconsistency. *Economics letters*, 8(3): 201–207.

von Böhm-Bawerk, E. 1922. *Capital and interest: A critical history of economical theory*. Brentano.

White, M. 2017. Unifying task specification in reinforcement learning. In *International Conference on Machine Learning*, 3742–3750. PMLR.