Theoretical Computer Science 519 (2014) 140–154

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

General time consistent discounting

Tor Lattimore, Marcus Hutter

Research School of Computer Science, Australian National University, Australia

ARTICLE INFO

Keywords:
Rational agents
Sequential decision theory
General discounting
Time-consistency
Game theory

ABSTRACT

Modeling inter-temporal choice is a key problem in both computer science and economic theory. The discounted utility model of Samuelson is currently the most popular model for measuring the global utility of a time-series of local utilities. The model is limited by not allowing the discount function to change with the age of the agent. This is despite the fact that many agents, in particular humans, are best modelled with age-dependent discount functions. It is well known that discounting can lead to time-inconsistent behaviour where agents change their preferences over time. In this paper we generalise the discounted utility model to allow age-dependent discount functions. We then extend previous work in time-inconsistency to our new setting, including a complete characterisation of time-(in)consistent discount functions, the existence of sub-game perfect equilibrium policies where the discount function is time-inconsistent and a continuity result showing that "nearly" time-consistent discount rates lead to "nearly" time-consistent behaviour.

1. Introduction

A rational agent, by definition, should choose its actions to maximise its expected utility [17,18]. Discounting is used to construct a simple model of global utility as a weighted sum of local utilities (well-being experienced at each time-step). The weighting usually assigns greater value to earlier, rather than later, consumption.

The discounted utility (DU) model, first introduced by Saumuelson [21], provides a framework for making decisions about inter-temporal consumption. Samuelson assumed that the global utility of a sequence of local utilities could be expressed as follows

$$V_k = \sum_{t=k}^{\infty} d_{t-k} r_t \tag{1}$$

where $r_k, r_{k+1}, \ldots$ is an infinite sequence of expected local rewards (utilities), V_k is the global utility at time k and d_{t-k} is the weight assigned to consumption at time-step t when in time-step k.

This model has a number of consequences:

1. *Utility independence*: It assumes that global utility can be represented as a discounted sum of local utilities, which removes the possibility of preferring one utility structure over another. For example, there is no way to distinguish between a relatively flat well-being profile and a roller-coaster of ups and downs [4].
2. *Consumption independence*: In the simplistic models commonly used in economic theory, the instantaneous utility of consumption at time k is independent of previous consumption choices [13,22]. This means that if pizza is preferred to Chinese on one day then it will be preferred every day. It is not possible to account for getting sick of pizza.

E-mail addresses: tor.lattimore@anu.edu.au (T. Lattimore), marcus.hutter@anu.edu.au (M. Hutter).

0304-3975/$ – see front matter Crown Copyright © 2013 Published by Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.tcs.2013.09.022

3. *Age independence*: The DU model denies the possibility that the discount function may change over time.
4. *Time-inconsistency*: An agent choosing a plan to maximise its discounted utility for the future may continuously change its plan over time despite receiving no new information [24].

The first and third are limitations of the DU model while the last is a rational consequence of an agent acting to maximise its discounted utility (in some environments and with some discount functions). Despite these limitations, the simple DU model is widely used in both computer science and economics.

In this paper we address all points above while updating the work of [24] on time-inconsistency to our more general setting. The first two limitations are largely eliminated by using a more general model than typically considered by economists (see the first example in Section 2). The third limitation is removed by allowing the discount function to change over the lifetime of the agent. For each time-step k we assume the agent uses a (possibly) different discount function $\boldsymbol{d}^k$. Utility can then be written

$$V_k = \sum_{t=k}^{\infty} d_t^k r_t. \tag{2}$$

This allows an agent to become more (or less) farsighted over time, which is important for two reasons. First, because we frequently wish to model humans that operate in this way, i.e, children plan only a few months, or at most years ahead, whilst adults think also of retirement. Some studies have shown this experimentally by empirically estimating discount rates [9]. The second reason comes from computer science where we wish to construct agents that behave in certain ways. Allowing an increasing effective horizon may allow an agent to explore more effectively [11]. For example, it is well known that the Bayesian policy for learning unknown stochastic bandits with geometric discounting suffers from linear regret, while other algorithms enjoy logarithmic regret [7,16]. This occurs because a rational agent discounting geometrically has no incentive to explore more than a certain amount as the reward it receives from the extra knowledge occurs too far in the future. Under certain conditions, however, a Bayesian agent with a more farsighted discount function suffers sub-linear regret in the bandit setting, as well as more general environment classes [10].

It has been remarked that time-inconsistent behaviour can be a rational consequence of discounting. This has been used to explain inter-temporal preference reversals observed in humans. For example, many people express a preference for $50 in three years and three weeks over $20 in three years, but favour $20 now rather than $50 in three weeks [6]. This effect is a natural consequence of some discount functions (hyperbolic) but not others (exponential).

Strotz showed that if the same discount function is used in each time-step as in Eq. (1), then *only* exponential discounting is guaranteed to lead to time-consistent rational agents [24]. Strotz worked with continuous time and assumed the utility V at time k could be written as

$$V_k = \int_k^{\infty} d_{t-k} r(t)\, dt \tag{3}$$

where r is now a continuous function. Formally he showed that if d_{t-k} is not proportional to γ^{t-k} for some $\gamma \in (0,1)$, then there exists an environment where the policy that maximises utility changes over time.

We extend this work to a complete classification of time-consistent discount functions in the more general case where the discount function is permitted to change with age. However, rather than using deterministic choices of continuous consumption profiles as in [24], we use (stochastic) Markov decision processes to model our environments [3,23]. This has a number of implications:

1. Discrete time, rather than continuous.
2. Arbitrary utility and infinitely many consumption profiles choices.

The limitation of discrete time should not be seen as too problematic for two reasons. First, because it is possible to use discrete time to approximate continuous time. Second, because our main theorem regarding the characterisation of time-consistent discount rates in discrete environments is transferable to the continuous case with minimal effort.

Given that an agent may operate using a time-inconsistent discount function, it is reasonable to ask how they will behave, and also how they should behave. If the agent is unaware that its discount function is time-inconsistent then it will (if rational) take action to maximise its expected discounted utility at the present time (ignoring that it may not follow its own plan due to changing preferences). For time-inconsistent discount functions this can lead to extremely bad behaviour in some environments (see Section 3 for an example).

On the other hand, if the agent knows its discount function is time-inconsistent, then blindly acting as if it were not is irrational. In this case it may be optimal to take a course of action that restricts its own choices in the future to ensure its future self does not act poorly according to its current preferences. To illustrate the idea, a recovering alcoholic who knows they become myopic in the evenings may choose to pour their alcohol away in the morning and so remove temptation in the evening. This approach is known as pre-commitment, and is a common strategy employed by humans who know their preferences are changing [1].

The idea of pre-commitment is generalised using game theory where the players are the current and future selves of the agent whose preferences are changing. A number of authors have applied this idea in Strotz's setting to show the existence of game-theoretically optimal policies [8,19,20]. We extend their results to show the existence of equivalently optimal policies in our more general setting.

Previous work on generalised discount rates has been limited. Strotz, and others in the economics literature have considered discount rates of the form $d_t^k = d_{t-k}$ (usually monotonic decreasing). In the computer science literature there has been some analysis of discount rates of the form $d_t^k = d_t$ where $\sum_{t=1}^{\infty} d_t < \infty$ [2,10,12]. We eliminate all of these restrictions and allow arbitrary d_t^k, including no explicit requirements on summability.

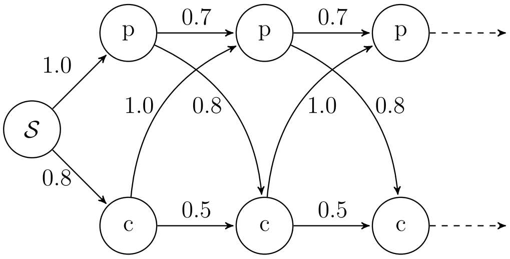
Our new contribution is a generalisation of Strotz's time-inconsistency results to arbitrary discount functions that change with age as in Eq. (2). This results in a large class of potentially interesting time-consistent discount functions (Theorem 13). We show that discount rates that are "nearly" time-consistent lead to policies that are only slightly differing in value (Theorem 15). Finally, we prove the existence of game-theoretically optimal policies for agents that know their discount rates are time-inconsistent (Theorem 19).

The paper is structured as follows. First the required notation is introduced (Section 2). Example discount functions and the consequences of time-inconsistent discount functions are then presented (Section 3). We next state and prove the main theorems, the complete classification of discount functions and the continuity result (Section 4). The game theoretic view of what an agent *should* do if it knows its discount function is changing is analysed (Section 5). Finally we offer some discussion and concluding remarks (Section 6).

2. Notation and problem setup

The general reinforcement learning (RL) setup involves an agent interacting sequentially with an environment where in each time-step t the agent chooses some action $a_t \in \mathcal{A}$, whereupon it receives a reward $r_t \in \mathcal{R} \subseteq \mathbb{R}$ and observation $o_t \in \mathcal{O}$. The environment can be formally defined as a probability distribution μ where $\mu(r_t o_t | a_1 r_1 o_1 a_2 r_2 o_2 \cdots a_{t-1} r_{t-1} o_{t-1} a_t)$ is the probability of receiving reward r_t and observation o_t having taken action a_t after history $h_{<t} := a_1 r_1 o_1 \cdots a_{t-1} r_{t-1} o_{t-1}$. For convenience, we assume for a given history $h_{<t}$ and action a_t, that r_t is fixed (not stochastic). We denote the set of all finite histories $\mathcal{H} := (\mathcal{A} \times \mathcal{R} \times \mathcal{O})^*$ and write $h_{1:t}$ and $h_{<t}$ for histories of length t and $t-1$ respectively. a_k, r_k, and o_k are the kth action/reward/observation tuple of history h and will be used without explicitly redefining them (there will always be only one history "in context").

A deterministic environment (where every value of $\mu(\cdot)$ is either 1 or 0) can be represented as a graph with edges for actions, rewards of each action attached to the corresponding edge, and observations in the nodes. For example, the deterministic environment below represents an environment where either pizza (p) or Chinese (c) must be chosen at each time-step (evening). An action leading to an upper node is `eat pizza` while the ones leading to a lower node are `eat chinese`. The rewards are for a consumer who prefers pizza to Chinese, but dislikes having the same food twice in a row. The starting node is marked as $\mathcal{S}$. This example, along with all those for the remainder of this paper, does not require observations. Note that this example demonstrates how the history-based models can overcome the consumption independence problem mentioned in the introduction.



The following assumption is required for clean results, but may be relaxed if an ϵ of slop is permitted in some results.

Assumption 1. We assume that $\mathcal{A}$ and $\mathcal{O}$ are finite and that $\mathcal{R} = [0, 1]$.

Definition 2 *(Policy).* A *policy* is a mapping $\pi : \mathcal{H} \to \mathcal{A}$ giving an action for each history. The set of all policies is denoted by Π.

Given policy π and history $h_{1:t}$ and $s \leqslant t$, then the probability of reaching history $h_{1:t}$ when starting from history $h_{<s}$ is $P(h_{1:t} | h_{<s}, \pi)$.

$$P(h_{1:t}|h_{<s},\pi) := \prod_{k=s}^{t} \mu\bigl(r_k o_k \big| h_{<k}\pi(h_{<k})\bigr). \tag{4}$$

If $s=1$ then we abbreviate and write $P(h_{1:t}|\pi) := P(h_{1:t}|h_{<1},\pi)$.

Definition 3 *(Expected rewards).* When applying policy π starting from history $h_{<t}$, the expected sequence of rewards $\boldsymbol{R}^{\pi}(h_{<t}) \in [0,1]^{\infty}$, is defined by

$$R^{\pi}(h_{<t})_k := \sum_{h_{t:k}} P(h_{1:k}|h_{<t},\pi) r_k.$$

If $k<t$ then $R^{\pi}(h_{<t})_k := 0$.

While the set of all possible $h_{t:k} \in (\mathcal{A}\times\mathcal{R}\times\mathcal{O})^k$ is uncountable due to the reward term, we sum only over the possible rewards which are determined by the action and previous history, and so this is actually a finite sum.

Definition 4 *(Discount vector).* A *discount vector* $\boldsymbol{d} \in [0,1]^{\infty}$ is a vector $[d_1, d_2, d_3, \ldots]$.

We do *not* insist that the discount vector be summable, $\sum_{t=k}^{\infty} d_t < \infty$.

Definition 5 *(Expected values).* The expected discounted reward (or utility or value) when using policy π starting in history $h_{<t}$ and discount vector $\boldsymbol{d}$ is

$$V_{\boldsymbol{d}}^{\pi}(h_{<t}) := \boldsymbol{R}^{\pi}(h_{<t})\cdot\boldsymbol{d} := \sum_{i=1}^{\infty} R^{\pi}(h_{<t})_i d_i = \sum_{i=t}^{\infty} R^{\pi}(h_{<t})_i d_i.$$

The sum can be taken to start from t since $R^{\pi}(h_{<t})_i = 0$ for $i<t$. As the scalar product is linear, a scaling of a discount vector has no affect on the ordering of the policies. Formally, if $V_{\boldsymbol{d}}^{\pi_1}(h_{<t}) \geqslant V_{\boldsymbol{d}}^{\pi_2}(h_{<t})$ then $V_{\alpha\boldsymbol{d}}^{\pi_1}(h_{<t}) \geqslant V_{\alpha\boldsymbol{d}}^{\pi_2}(h_{<t})$ for all $\alpha > 0$.

Definition 6 *(Optimal policy/value).* In general, an agent should choose a policy $\pi_{\boldsymbol{d}}^{*}$ to maximise $V_{\boldsymbol{d}}^{\pi}(h_{<t})$. This is defined as follows.

$$\pi_{\boldsymbol{d}}^{*} \in \Pi_{\boldsymbol{d}}^{*} := \arg\max_{\pi\in\Pi} V_{\boldsymbol{d}}^{\pi},$$
$$V_{\boldsymbol{d}}^{*}(h_{<t}) := V_{\boldsymbol{d}}^{\pi_{\boldsymbol{d}}^{*}}(h_{<t}),$$

where $\Pi_{\boldsymbol{d}}^{*}$ is the set of optimal policies maximising $V_{\boldsymbol{d}}^{\pi}$. Typically the choice of optimal policy is irrelevant, so if $\pi_{\boldsymbol{d}}^{*}$ is written without clarification, then it is chosen using some arbitrary rule.

An agent can use a different discount vector $\boldsymbol{d}^k$ at each time-step k. This motivates the following definition.

Definition 7 *(Discount matrix).* A *discount matrix* D is a $\infty\times\infty$ matrix with discount vector $\boldsymbol{d}^k$ for the kth column.

It is important that we distinguish between a discount matrix D (capital and italics), a discount vector $\boldsymbol{d}^k$ (bold and italics) used at time k, an arbitrary discount vector $\boldsymbol{d}$ (bold and italics), and a particular value in a discount vector d_t^k (just italics).

Definition 8 *(Sliding discount matrix).* A discount matrix D is *sliding* if $d_{k+t}^k = d_{t+1}^1$ for all $k, t \geqslant 1$.

Unfortunately, $\pi_{\boldsymbol{d}^k}^{*}$ need not exist without one further assumption.

Assumption 9. For all $k \geqslant 1$, $\lim_{t\to\infty}\sup_{\pi\in\Pi}\sum_{h_{<t}} P(h_{<t}|\pi) V_{\boldsymbol{d}^k}^{\pi}(h_{<t}) = 0$.

Assumption 9 appears somewhat arbitrary. We consider:

1. For summable $\boldsymbol{d}^k$ the assumption is true for all environments. With the exception of hyperbolic discounting, all commonly used discount vectors are summable.

2. For non-summable discount vectors $\boldsymbol{d}^k$ the assumption implies a restriction on the possible environments. In particular, they must return asymptotically lower rewards in expectation uniformly over all policies. This restriction is necessary to guarantee the existence of an optimal policy.

From now on, including in theorem statements, we only consider environments/discount vectors satisfying Assumptions 1 and 9. The following theorem then guarantees the existence of $\pi^*_{\boldsymbol{d}^k}$.

Theorem 10 *(Existence of optimal policy). $\pi^*_{\boldsymbol{d}^k}$ exists for any environment and discount vector $\boldsymbol{d}^k$ satisfying Assumptions 1 and 9.*

The proof of the existence theorem is in Appendix A.

Definition 11 *(Mixed policy).* The *mixed policy* is the policy where at each time-step t, the agent acts according to the possibly different policy $\pi^*_{\boldsymbol{d}^t}$.

$$\pi_D(h_{<t}) := \pi^*_{\boldsymbol{d}^t}(h_{<t}), \qquad \boldsymbol{R}_D(h_{<t}) := \boldsymbol{R}^{\pi_D}(h_{<t}).$$

We do not denote the mixed policy by π^*_D as it is arguably not optimal as discussed in Section 5. While non-unique optimal policies $\pi^*_{\boldsymbol{d}^k}$ at least result in equal discounted utilities, this is *not* the case for π_D. All theorems are proved with respect to any choice π_D.

The policies can be summarised as follows. $\pi^*_{\boldsymbol{d}}$ is the optimal policy with respect to discount vector $\boldsymbol{d}$. $\pi^*_{\boldsymbol{d}^k}$ is the optimal policy with respect to $\boldsymbol{d}^k$, which is the discount vector used at time-step k. π_D is the mixed policy constructed by following $\pi^*_{\boldsymbol{d}^k}$ at time-step k. Later we will introduce π^*_D, which is a game-theoretically optimal policy formalising the idea of pre-commitment discussed earlier.

Definition 12 *(Time-consistency).* A discount matrix D is *time consistent* if and only if for all environments and $k, j \in \mathbb{N}$, $\Pi^*_{\boldsymbol{d}^k} = \Pi^*_{\boldsymbol{d}^j}$.

This means that a time-consistent agent taking action $\pi^*_{\boldsymbol{d}^t}(h_{<t})$ at each time-step t will not change its plans. On the other hand, a time-inconsistent agent may at time 1 intend to take action a should it reach history $h_{<t}$ ($\pi^*_{\boldsymbol{d}^1}(h_{<t}) = a$). However upon reaching $h_{<t}$, it need not be true that $\pi^*_{\boldsymbol{d}^t}(h_{<t}) = a$.

3. Examples

In this section we review a number of common discount matrices and give an example where a time-inconsistent discount matrix causes very bad behaviour.

Constant horizon. Constant horizon discounting is where the agent only cares about the future up to H time-steps away, defined by $d^k_t = [\![t - k < H]\!]$.[1] Shortly we will see that the constant horizon discount matrix can lead to very bad behaviour in some environments.

Fixed lifetime. Fixed lifetime discounting is where an agent knows it will not care about any rewards past time-step m, defined by $d^k_t = [\![t < m]\!]$. Unlike the constant horizon method, a fixed lifetime discount matrix is time-consistent. Unfortunately, the definition requires knowledge of the lifetime of the agent ahead of time and also makes asymptotic analysis meaningless.

Hyperbolic. $d^k_t = 1/(1 + \kappa(t - k))$. The parameter κ determines how farsighted the agent is with smaller values leading to more farsighted agents. Hyperbolic discounting is often used in economics with some experimental studies explaining human time-inconsistent behaviour by suggesting that we discount hyperbolically [25]. The hyperbolic discount matrix is not summable, so may be replaced by the following (similar to [11]), which has similar properties for β close to 1.

$$d^k_t = 1/\bigl(1 + \kappa(t - k)\bigr)^\beta \text{with } \beta > 1.$$

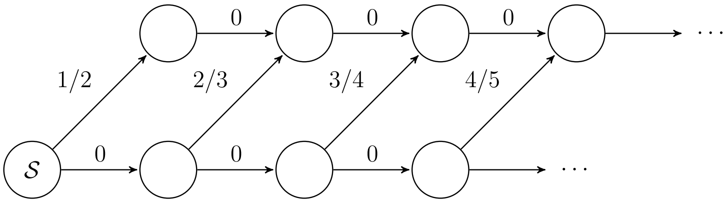
Geometric. $d^k_t = \gamma^t$ with $\gamma \in (0, 1)$. Geometric discounting is the most commonly used discount matrix. Philosophically it can be justified by assuming an agent will die (and not care about the future after death) with probability $1 - \gamma$ at each time-step. Another justification for geometric discount is its analytic simplicity – it is summable and leads to time-consistent policies. It also models fixed interest rates.

[1] $[\![expr]\!] = 1$ if *expr* is true and 0 otherwise.

Power. $d^k_t = t^{-\beta}$ with $\beta > 1$. Power discounting is an example of a discount matrix with increasing effective horizon. At time-step t an agent using power discounting will consider $O(t)$ time-steps into the future. Power discounting is attractive since it is time-consistent without having to pre-commit to a fixed (effective) horizon. It also leads to so-called self-optimising Bayesian policies [10]. Note that power discounting is beyond the scope of the original DU model of Samuelson, which is a justification for using the more general model introduced in this paper [21].

No discounting. $d^k_t = 1$, for all k, t. Legg and Hutter point out that discounting future rewards via an explicit discount matrix is unnecessary since the environment can capture both temporal preferences for early (or late) consumption, as well as the risk associated with delaying consumption [14,15]. Of course, this discount matrix is not summable, but can be made to work by insisting that all environments satisfy Assumption 9. This approach is elegant in the sense that it eliminates the need for a discount matrix, essentially admitting far more complex preferences regarding inter-temporal rewards than a discount matrix allows. On the other hand, a discount matrix gives the "controller" an explicit way to adjust the myopia of the agent.

Time-inconsistency. To illustrate the potential consequences of time-inconsistent discount matrices we consider the policies of several agents acting in the following deterministic environment.



Let agent A use a constant horizon discount matrix with $H = 2$ and agent B a geometric discount matrix with some discount rate γ.

In the first time-step agent A prefers to move right with the intention of moving up in the second time-step for a reward of 2/3. At the second time-step, however, the agent will change its plan by moving right again. This continues indefinitely, so agent A will always delay moving up and receives zero reward forever.

Agent B acts very differently. Let π_t be the policy in which the agent moves right until time-step t, then up and right indefinitely. $V^{\pi_t}_{\boldsymbol{d}^k}(h_{<1}) = \gamma^t \frac{(t+1)}{(t+2)}$. This value does not depend on k and so the agent will move right until $t = \arg\max\{\gamma^t \frac{(t+1)}{t+2}\} < \infty$ when it will move up and receive a reward.

The actions of agent A are an example of the worst possible behaviour arising from time-inconsistent discounting. Nevertheless, agents with a constant horizon discount matrix are used in all kinds of problems. In particular, agents in zero sum games where fixed depth mini-max searches are common. In practise, serious time-inconsistent behaviour for game-playing agents seems rare, presumably because most strategic games don't have a reward structure similar to the example above.

4. Theorems

The main theorem of this paper is a complete characterisation of time consistent discount matrices.

Theorem 13 *(Characterisation). Let D be a discount matrix, then the following are equivalent.*

1. *D is time-consistent (Definition 12).*
2. *For each k there exists an $\alpha_k > 0$ such that $d^k_t = \alpha_k d^1_t$ for all $t \geqslant k \in \mathbb{N}$.*

Recall that a discount matrix is sliding if $d^k_t = d^1_{t-k+1}$. Theorem 13 can be used to show that if a sliding discount matrix is used as in [24] then the only time-consistent discount matrix is geometric. Let D be a time-consistent sliding discount matrix. By Theorem 13 and the definition of sliding, $\alpha_2 d^1_{t+1} = d^2_{t+1} = d^1_t$. Therefore $\alpha_2 d^1_2 = d^1_1$ and $\alpha_2^2 d^1_3 = \alpha_2 d^1_2 = d^1_1$ and similarly, $\alpha_2^{t-1} d^1_t = d^1_1 \Rightarrow d^1_t \propto \gamma^t$ with $\gamma = 1/\alpha_2$, which is geometric discounting. This is the analogue to the results of [24] converted to our setting.

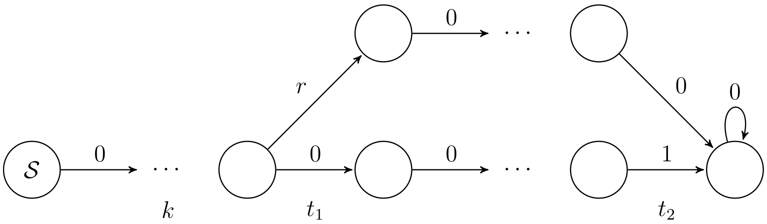
The theorem can also be used to construct time-consistent discount rates. Let $\boldsymbol{d}^1$ be a discount vector, then the discount matrix defined by $d^k_t := d^1_t$ for all $t \geqslant k$ will always be time-consistent, for example, the *fixed lifetime* discount matrix with $d^k_t = 1$ if $t \leqslant H$ for some horizon H. Indeed, all time-consistent discount rates can be constructed in this way (up to irrelevant scaling).

Proof of Theorem 13. $2 \Rightarrow 1$: This direction follows easily from linearity of the scalar product.

$$\begin{aligned}\Pi^*_{\boldsymbol{d}^k}(h_{<t}) &\equiv \arg\max_\pi V^\pi_{\boldsymbol{d}^k}(h_{<t}) \equiv \arg\max_\pi \boldsymbol{R}^\pi(h_{<t}) \cdot \boldsymbol{d}^k \\ &\overset{(a)}{=} \arg\max_\pi \boldsymbol{R}^\pi(h_{<t}) \cdot \alpha_k \boldsymbol{d}^1 = \arg\max_\pi \alpha_k \boldsymbol{R}^\pi(h_{<t}) \cdot \boldsymbol{d}^1 \\ &= \arg\max_\pi \boldsymbol{R}^\pi(h_{<t}) \cdot \boldsymbol{d}^1 \equiv \Pi^*_{\boldsymbol{d}^1}(h_{<t}),\end{aligned}$$

as required. Equality (a) follows from the assumption that $d^k_t = \alpha_k d^1_t$ for all $t \geqslant k$ and because $\boldsymbol{R}^\pi(h_{<t})_i = 0$ for all $i < t$.

$1 \Rightarrow 2$: We use the deterministic environment below where the agent has a choice between earning reward r at time t_1 or reward 1 at time t_2. In this environment there are only two policies, π_1 and π_2, where $\boldsymbol{R}^{\pi_1}(h_{<k}) = r\boldsymbol{e}_{t_1}$ and $\boldsymbol{R}^{\pi_2}(h_{<k}) = \boldsymbol{e}_{t_2}$ with $\boldsymbol{e}_i$ the infinite vector with all components zero except the ith, which is 1.



First we show that $d^1_t > 0 \Leftrightarrow d^k_t > 0$ for all $t \geqslant k$. Suppose without loss of generality that $d^1_t > 0$ and $d^k_t = 0$ for some $t \geqslant k$. Then let $r = 0$ in the environment above and $t_2 = t$. Then $V^{\pi_1}_{\boldsymbol{d}^k}(h_{<k}) = V^{\pi_2}_{\boldsymbol{d}^k}(h_{<k}) = 0$ implies that π_1 and π_2 are both optimal with respect to $\boldsymbol{d}^k$. On the other hand, $V^{\pi_1}_{\boldsymbol{d}^1}(h_{<k}) = 0$ while $V^{\pi_2}_{\boldsymbol{d}^1}(h_{<k}) = d^1_t > 0$, which implies that π_1 is not optimal with respect to $\boldsymbol{d}^k$, which contradicts the assumption that D is time-consistent.

We now prove the main result by contradiction. Suppose there exists a k such

$$\forall \alpha > 0,\ \exists t \geqslant k\colon\quad d^k_t \neq \alpha d^1_t.$$

Let t_1 be such that $d^k_{t_1}, d^1_{t_1} > 0$ and $\alpha := d^k_{t_1}/d^1_{t_1}$. Now let $t_2 > t_1$ be such that $\beta := d^k_{t_2}/d^1_{t_2} \neq \alpha$, which exists by assumption. Therefore

$$\frac{d^k_{t_2}}{d^k_{t_1}} = \frac{\beta}{\alpha}\frac{d^1_{t_2}}{d^1_{t_1}} \neq \frac{d^1_{t_2}}{d^1_{t_1}}.$$

Now let $r := (d^k_{t_2}/d^k_{t_1} + d^1_{t_2}/d^1_{t_1})/2$. Then since $d^1_{t_2}/d^1_{t_1} \neq d^k_{t_2}/d^k_{t_1}$, either

$$d^1_{t_2}/d^1_{t_1} < r < d^k_{t_2}/d^k_{t_1} \quad \text{or} \quad d^k_{t_2}/d^k_{t_1} < r < d^1_{t_2}/d^1_{t_1}.$$

In the first case,

$$\begin{aligned}&V^{\pi_1}_{\boldsymbol{d}^1}(h_{<k}) \equiv r d^1_{t_1} > d^1_{t_2} \equiv V^{\pi_2}_{\boldsymbol{d}^1}(h_{<k}) \quad \text{and} \\ &V^{\pi_1}_{\boldsymbol{d}^k}(h_{<k}) \equiv r d^k_{t_1} < d^k_{t_2} \equiv V^{\pi_2}_{\boldsymbol{d}^k}(h_{<k}),\end{aligned}$$

which is a contradiction of time-consistency since the agent discounting with respect to $\boldsymbol{d}^1$ prefers π_2 while the agent discounting with respect to $\boldsymbol{d}^k$ prefers π_1. Case 2 is identical except the preferences are reversed. Therefore for all k there exists an $\alpha_k > 0$ such that $d^k_t = \alpha_k d^1_t$ for all $t \geqslant k$ as required. □

In Section 3 we saw an example where time-inconsistency led to very bad behaviour. The discount matrix causing this was very time-inconsistent. Is it possible that an agent using a "nearly" time-consistent discount matrix can exhibit similar bad behaviour? For example, could rounding errors when using a geometric discount matrix seriously affect the agent's behaviour? The following theorem shows that this is not possible. First we require a measure of the cost of time-inconsistent behaviour. The regret experienced by the agent at time one from following policy π_D rather than $\pi^*_{\boldsymbol{d}^1}$ is $V^*_{\boldsymbol{d}^1}(h_{<1}) - V^{\pi_D}_{\boldsymbol{d}^1}(h_{<1})$. The regret measures the difference in expected discounted rewards from following π_D rather than the preferred $\pi^*_{\boldsymbol{d}^1}$ where rewards are discounted with respect to the discount vector of the agent at time one. We also need a distance measure on the space of discount vectors.

Definition 14 *(Distance measure).* Let $\boldsymbol{d}^k$ and $\boldsymbol{d}^j$ be discount vectors, then define a distance measure Δ by

$$\Delta\big(\boldsymbol{d}^k, \boldsymbol{d}^j\big) := \sum_{i=\max\{k,j\}}^{\infty} \big|d_i^k - d_i^j\big|.$$

Note that this is almost the taxicab metric, but the sum is restricted to $i \geqslant \max\{k, j\}$.

Theorem 15 *(Continuity). Suppose $\epsilon \geqslant 0$ and $\Delta_{k,j} := \Delta(\boldsymbol{d}^k, \boldsymbol{d}^j)$ then*

$$V^*_{\boldsymbol{d}^1}(h_{<1}) - V^{\pi_D}_{\boldsymbol{d}^1}(h_{<1}) \leqslant \epsilon + \sum_{k=1}^{t-1} \Delta_{k,k+1}$$

with $t = \min\{t\colon \sup_{\pi\in\Pi} \sum_{h_{<t}} P(h_{<t}|\pi) V^{\pi}_{\boldsymbol{d}^1}(h_{<t}) \leqslant \epsilon\}$, which for $\epsilon > 0$ is guaranteed to exist by Assumption 9.

Theorem 15 implies that the regret of the agent at time one in its future time-inconsistent actions is bounded by the sum of the differences between the discount vectors used at different times. If these differences are small then the regret is also small. For example, it implies that small perturbations (such as rounding errors) in a time-consistent discount matrix lead to minimal bad behaviour.

The proof can be found in Appendix A. It relies on proving the result for finite horizon environments and showing that this extends to the infinite case by using a horizon t after which the actions of the agent are no longer important.

Example 16. Let D be a time-consistent summable discount matrix with $d_t^k = d_t^1$. For computation reasons it may be necessary to use a modified discount matrix $\tilde{D}$ defined by

$$\tilde{d}_t^k = \begin{cases} d_t^k & \text{if } t - k < N, \\ 0 & \text{otherwise.} \end{cases}$$

This may occur if values can only be computed by a depth N lookahead. Now $\Delta_{k,k+1} = d^1_{k+N}$, so if Theorem 15 is applied to $\tilde{D}$ with $\epsilon = 0$, then $t = N$ (since $V^*_{\tilde{\boldsymbol{d}}^1}(h_{<N}) = 0$) and the theorem shows that

$$V^*_{\tilde{\boldsymbol{d}}^1}(h_{<1}) - V^{\pi_{\tilde{D}}}_{\tilde{\boldsymbol{d}}^1}(h_{<1}) \leqslant \sum_{k=1}^{N-1} \Delta_{k,k+1} = \sum_{i=N}^{2N-2} d_i^1.$$

However D is summable, so $\lim_{N\to\infty} \sum_{i=N}^{2N-2} d_i^1 = 0$. Therefore we can always choose an N (sufficiently large) to guarantee a small regret.

This result is already well known for a single time-step [3], but the example shows that errors do not compound in a very undesirable way over time.

The bound in Theorem 15 is tight in the following sense.

Theorem 17. *For any $t \in \mathbb{N}$ and any sufficiently small $\alpha > 0$ there exists an environment and discount matrix such that*

$$(t-2)(1-\alpha) \leqslant V^*_{\tilde{\boldsymbol{d}}^1}(h_{<1}) - V^{\pi_{\tilde{D}}}_{\tilde{\boldsymbol{d}}^1}(h_{<1}) \leqslant t - 2 + 2\alpha,$$

where t is as in Theorem 15 with $\epsilon = 0$ and $\sum_{k=1}^{t-1} \Delta_{k,k+1} = t - 2 + 2\alpha$.

Theorem 17 shows that there exists a discount matrix and environment where the regret due to time-inconsistency is nearly equal to the bound given by Theorem 15.

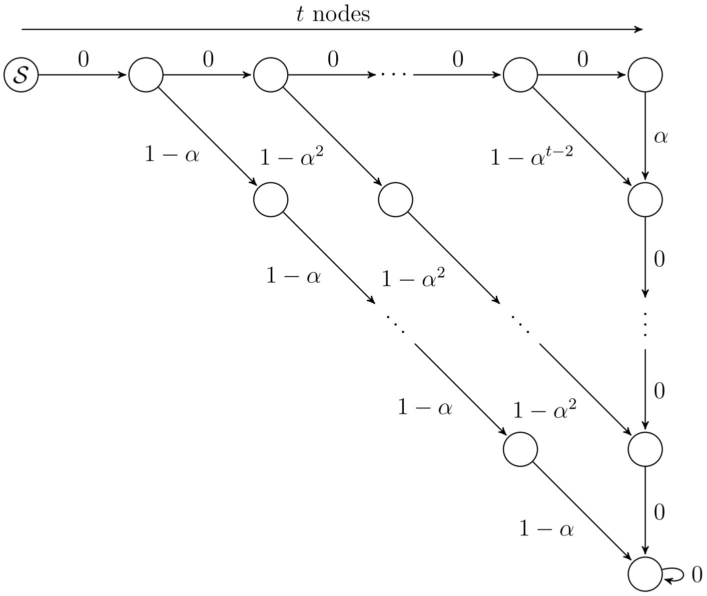
Proof of Theorem 17. Define D by

$$d_i^k := \begin{cases} 1 & \text{if } k < i < t, \\ \alpha & \text{if } i = t \text{ and } k = t-1, \\ 0 & \text{otherwise.} \end{cases} \qquad \underbrace{\begin{array}{l} \boldsymbol{d}^1 = 0,\ \ 1,\ \ 1,\ \ 0,\ \ 0 \\ \boldsymbol{d}^2 = 0,\ \ 0,\ \ 1,\ \ 0,\ \ 0 \\ \boldsymbol{d}^3 = 0,\ \ 0,\ \ 0,\ \ \alpha,\ \ 0 \\ \boldsymbol{d}^4 = 0,\ \ 0,\ \ 0,\ \ 0,\ \ 0 \end{array}}_{\text{Example if } t=4}$$

Observe that

$$\Delta\left(\boldsymbol{d}^k, \boldsymbol{d}^{k+1}\right) = \begin{cases} 1 & \text{if } k < t-2, \\ 1+\alpha & \text{if } k = t-2, \\ \alpha & \text{if } k = t-1, \\ 0 & \text{otherwise.} \end{cases}$$

Therefore $\sum_{k=1}^{t-1} \Delta(\boldsymbol{d}^k, \boldsymbol{d}^{k+1}) = t - 2 + 2\alpha$. Now consider the environment below.



For sufficiently small α, the agent at time-step one will plan to move right and then down leading to $\boldsymbol{R}_{\boldsymbol{d}^1}^{\pi_{\boldsymbol{d}}^*}(h_{<1}) = [0, 1-\alpha, 1-\alpha, \ldots]$ and $V_{\boldsymbol{d}^1}^*(h_{<1}) = (t-2)(1-\alpha)$.

To compute $\boldsymbol{R}_D$, note that $d_k^k = 0$ for all k. Therefore the agent in time-step k doesn't care about the next instantaneous reward, so prefers to move right with the intention of moving down in the next time-step when the rewards are slightly better. This leads to $\boldsymbol{R}_D(h_{<t}) = \alpha \mathbf{e}_t$ and so $V_{\boldsymbol{d}^1}^{\pi_D}(h_{<1}) = 0$. Therefore

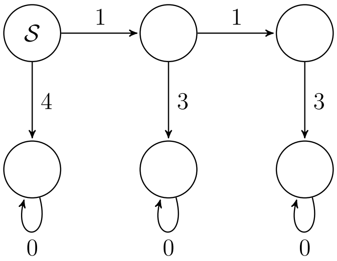
$$V_{\boldsymbol{d}^1}^*(h_{<1}) - V_{\boldsymbol{d}^1}^{\pi_D}(h_{<1}) = (t-2)(1-\alpha).$$

Furthermore, $V_{\boldsymbol{d}^1}^*(h_{<1}) - V_{\boldsymbol{d}^1}^{\pi_D}(h_{<1}) \leqslant t - 2 + 2\alpha$ by Theorem 15 with $\epsilon = 0$, which completes the proof. □

5. Game theoretic approach

What should an agent do if it knows it is time inconsistent? One option is to treat its future selves as “opponents” in an extensive game. The game has one player per time-step who chooses the action for that time-step only. At the end of the game the agent will have received a reward sequence $\boldsymbol{r} \in \mathcal{R}^\infty$. The utility given to the kth player is then $\boldsymbol{r} \cdot \boldsymbol{d}^k$. So each player in this game wishes to maximise the discounted reward with respect to a different discounting vector.

For example, let $\boldsymbol{d}^1 = [2, 1, 2, 0, 0, \ldots]$ and $\boldsymbol{d}^2 = [*, 3, 1, 0, 0, \ldots]$ and consider the environment below.



Initially, the agent has two choices. It can either move down to guarantee a reward sequence of $\boldsymbol{r} = [4, 0, 0, \ldots]$ which has utility of $\boldsymbol{d}^1 \cdot [4, 0, 0, \ldots] = 8$ or it can move right in which case it will receive a reward sequence of either $\boldsymbol{r}' = [1, 3, 0, 0, \ldots]$

with utility 5 or $\boldsymbol{r}'' = [1, 1, 3, 0, 0, \ldots]$ with utility 9. Which of these two reward sequences it receives is determined by the action taken in the second time-step. However this action is chosen to maximise utility with respect to discount sequence $\boldsymbol{d}^2$ and $\boldsymbol{d}^2 \cdot \boldsymbol{r}' > \boldsymbol{d}^1 \cdot \boldsymbol{r}''$. This means that if at time 1 the agent chooses to move right, the final reward sequence will be $[1, 3, 0, 0, \ldots]$ and the final utility with respect to $\boldsymbol{d}^1$ will be 5. Therefore the rational thing to do in time-step 1 is to move down immediately for a utility of 8.

The technique above is known as backwards induction. A variant of Kuhn's theorem proves that backwards induction finds sub-game perfect equilibria in finite extensive games [18]. For arbitrary (infinite) extensive games a sub-game perfect equilibrium need not exist, but we prove a theorem for our particular class of infinite games.

A sub-game perfect equilibrium policy is one the players could agree to play, and subsequently have no incentive to renege on their agreement during play. It isn't always philosophically clear that a sub-game perfect equilibrium policy *should* be played. For a deeper discussion, including a number of good examples, see [18].

Definition 18 *(Sub-game perfect equilibria).* A policy π^*_D is a sub-game perfect equilibrium policy if and only if for each t, $V^{\pi^*_D}_{\boldsymbol{d}^t}(h_{<t}) \geqslant V^{\tilde\pi}_{\boldsymbol{d}^t}(h_{<t})$, for all $h_{<t}$, where $\tilde\pi$ is any policy satisfying $\tilde\pi(h_{<i}) = \pi^*_D(h_{<i}) \forall h_{<i}$ where $i \neq t$.

Theorem 19 *(Existence of sub-game perfect equilibrium policy). For all environments and discount matrices D satisfying Assumptions 1 and 9 there exists at least one sub-game perfect equilibrium policy π^*_D.*

Many results in the literature of game theory almost prove this theorem. Our setting is more difficult than most because we have countably many players (one for each time-step) and exogenous uncertainty. Fortunately, it is made easier by the very particular conditions on the preferences of players for rewards that occur late in the game (Assumption 9). The closest related work appears to be that of Drew Fudenberg in [5], but our proof is very different. The proof idea is to consider a sequence of environments identical to the original environment but with an increasing bounded horizon after which reward is zero. By Kuhn's Theorem [18] a sub-game perfect equilibrium policy must exist in each of these finite games. However the space of policies is compact (Lemma 24) and so this sequence of sub-game perfect equilibrium policies contains a convergent subsequence converging to some policy π. Then it is not hard to show that π is a sub-game prefect equilibrium policy in the original environment.

Proof of Theorem 19. For each $t \in \mathbb{N}$ choose π_t to be a sub-game perfect equilibrium policy in the modified environment obtained by setting $r_i = 0$ if $i > t$. That is, the environment that always gives zero reward after time t. Note that π_t exists by backwards induction where the policy is chosen arbitrarily for histories longer than t. Since Π is compact, the sequence $\pi_1, \pi_2, \ldots$ has a convergent subsequence $\pi_{t_1}, \pi_{t_2}, \ldots$ converging to some π and satisfying:

1. $\pi_{t_i}(h_{<k}) = \pi(h_{<k})$, for all $h_{<k}$ where $k \leqslant i$.
2. π_{t_i} is a sub-game perfect equilibrium policy in the modified environment with reward $r_k = 0$ if $k > t_i$.

We write $\tilde V^{\pi_{t_i}}$ for the value function of π_{t_i} in the t_i-modified environment. It is now shown that π is a sub-game perfect equilibrium policy. Fix a $t \in \mathbb{N}$ and let $\tilde\pi$ be an arbitrary policy with $\tilde\pi(h_{<k}) = \pi(h_{<k})$ for all $h_{<k}$ where $k \neq t$. Now define policies $\tilde\pi_{t_i}$ by

$$\tilde\pi_{t_i}(h_{<k}) = \begin{cases} \tilde\pi(h_{<k}) & \text{if } k \leqslant i, \\ \pi_{t_i}(h_{<k}) & \text{otherwise.} \end{cases}$$

By point 1 above, $\tilde\pi_{t_i}(h_{<k}) = \pi_{t_i}(h_{<k})$ for all $h_{<k}$ where $k \neq t$. Now for all $i > t$ we have

$$\begin{aligned} V^{\pi}_{\boldsymbol{d}^t}(h_{<t}) &\geqslant V^{\pi_{t_i}}_{\boldsymbol{d}^t}(h_{<t}) - \left|V^{\pi}_{\boldsymbol{d}^t}(h_{<t}) - V^{\pi_{t_i}}_{\boldsymbol{d}^t}(h_{<t})\right| && (5)\\ &\geqslant \tilde V^{\pi_{t_i}}_{\boldsymbol{d}^t}(h_{<t}) - \left|V^{\pi}_{\boldsymbol{d}^t}(h_{<t}) - V^{\pi_{t_i}}_{\boldsymbol{d}^t}(h_{<t})\right| && (6)\\ &\geqslant \tilde V^{\tilde\pi_{t_i}}_{\boldsymbol{d}^t}(h_{<t}) - \left|V^{\pi}_{\boldsymbol{d}^t}(h_{<t}) - V^{\pi_{t_i}}_{\boldsymbol{d}^t}(h_{<t})\right| && (7)\\ &\geqslant V^{\tilde\pi}_{\boldsymbol{d}^t}(h_{<t}) - \left|V^{\pi}_{\boldsymbol{d}^t}(h_{<t}) - V^{\pi_{t_i}}_{\boldsymbol{d}^t}(h_{<t})\right| \\ &\quad - \left|V^{\tilde\pi_{t_i}}_{\boldsymbol{d}^t}(h_{<t}) - \tilde V^{\tilde\pi_{t_i}}_{\boldsymbol{d}^t}(h_{<t})\right| - \left|V^{\tilde\pi_{t_i}}_{\boldsymbol{d}^t}(h_{<t}) - V^{\tilde\pi}_{\boldsymbol{d}^t}(h_{<t})\right| && (8)\end{aligned}$$

where (5) follows from arithmetic. (6) since $V \geqslant \tilde V$. (7) since π_{t_i} is a sub-game perfect equilibrium policy. (8) by arithmetic. We now show that the absolute value terms in (8) converge to zero. Since $V^{\pi}(\cdot)$ is continuous (Lemma 25) in π and $\lim_{i\to\infty} \pi_{t_i} = \pi$ and $\lim_{i\to\infty} \tilde\pi_{t_i} = \tilde\pi$, we obtain $\lim_{i\to\infty}[|V^{\pi}_{\boldsymbol{d}^t}(h_{<t}) - V^{\pi_{t_i}}_{\boldsymbol{d}^t}(h_{<t})| + |V^{\tilde\pi_{t_i}}_{\boldsymbol{d}^t}(h_{<t}) - V^{\tilde\pi}_{\boldsymbol{d}^t}(h_{<t})|] = 0$. Furthermore, by Assumption 9 and the fact that $\tilde\pi_{t_i}$ converges to $\tilde\pi$, $\lim_{i\to\infty} |V^{\tilde\pi_{t_i}}_{\boldsymbol{d}^t}(h_{<t}) - \tilde V^{\tilde\pi_{t_i}}_{\boldsymbol{d}^t}(h_{<t})| = 0$. Therefore taking the limit as i goes to infinity in (8) shows that $V^{\pi}_{\boldsymbol{d}^t}(h_{<t}) \geqslant V^{\tilde\pi}_{\boldsymbol{d}^t}(h_{<t})$ as required. □

In general, π_D^* need not be unique, and different sub-game perfect equilibrium policies can lead to different utilities. This is a normal, but unfortunate, problem with the sub-game perfect equilibrium solution concept. The policy is unique if for all players the value of any two arbitrary policies is different. Also, if $\forall k(V_{\boldsymbol{d}^k}^{\pi_1} = V_{\boldsymbol{d}^k}^{\pi_2} \Rightarrow \forall j V_{\boldsymbol{d}^j}^{\pi_1} = V_{\boldsymbol{d}^j}^{\pi_2})$ is true then the non-unique sub-game equilibrium policies have the same values for all agents. Unfortunately, neither of these conditions is necessarily satisfied in our setup. The problem of how players might choose a sub-game perfect equilibrium policy appears surprisingly understudied. We feel it provides another reason to avoid the situation altogether by using time-consistent discount matrices. The following example illustrates the problem of non-unique sub-game perfect equilibrium policies.

Example 20. Consider the example in Section 3 with an agent using a constant horizon discount matrix with $H = 2$. There are exactly two sub-game perfect equilibrium policies, π_1 and π_2 defined by,

$$\pi_1(h_{<t}) = \begin{cases} up & \text{if } t \text{ is odd,} \\ right & \text{otherwise,} \end{cases} \qquad \pi_2(h_{<t}) = \begin{cases} up & \text{if } t \text{ is even,} \\ right & \text{otherwise.} \end{cases}$$

Note that the reward sequences (and values) generated by π_1 and π_2 are different with $\boldsymbol{R}^{\pi_1}(h_{<1}) = [1/2, 0, 0, \ldots]$ and $\boldsymbol{R}^{\pi_2}(h_{<1}) = [0, 2/3, 0, 0, \ldots]$. If the players choose to play a sub-game perfect equilibrium policy then the first player can choose between π_1 and π_2 since they have the first move. In that case it would be best to follow π_2 by moving right as it has a greater return for the agent at time 0 than π_1.

For time-consistent discount matrices we have the following proposition.

Proposition 21. *If D is time-consistent, then $V_{\boldsymbol{d}^k}^* = V_{\boldsymbol{d}^k}^{\pi_D} = V_{\boldsymbol{d}^k}^{\pi_D^*}$ for all k and choices of $\pi_{\boldsymbol{d}^k}^*$ and π_D and π_D^*.*

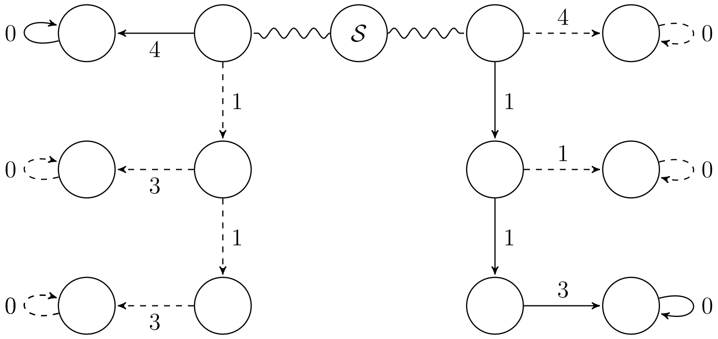
The proof that $V_{\boldsymbol{d}^k}^* = V_{\boldsymbol{d}^k}^{\pi_D^*}$ is trivial by noting that if D is time-consistent, then all players in the game-theory setting have the same preferences. Therefore any diversion from the optimal is non-credible. That $V_{\boldsymbol{d}^k}^* = V_{\boldsymbol{d}^k}^{\pi_D}$ follows from a straightforward generalisation of Theorem 15 to discount vectors $\boldsymbol{d}^k$ rather than just $\boldsymbol{d}^1$.

Is it possible that backwards induction is simply expected discounted reward maximisation in another form? The following theorem shows this is not the case and that sub-game perfect equilibrium policies are a rich and interesting class worthy of further study in these (and more general) settings.

Theorem 22. *$\exists D$ such that $\pi_D^* \neq \pi_{\boldsymbol{d}}^*$, for all $\boldsymbol{d}$.*

The result is proven using a simple counter-example. The idea is to construct a stochastic environment where the first action leads the agent to one of two sub-environments, each with probability half. These environments are identical to the example at the start of this section, but one of them has the reward 1 (rather than 3) for the history *right*, *down*. It is then easily shown that π_D^* is not the result of an expectimax expression because it behaves differently in each sub-environment, while any expectimax search (irrespective of discounting) will behave the same in each.

Proof. Observe the stochastic environment below where in the first time-step the agent moves either right or left with probability $\frac{1}{2}$ irrespective of action. In either case it receives 0 reward. Otherwise the environment is deterministic as in other diagrams.



Now let $\boldsymbol{d}^1 = [1, 0, 0, 0, \ldots]$, $\boldsymbol{d}^2 = [*, 2, 1, 2, 0, \ldots]$ and $\boldsymbol{d}^3 = [*, *, 3, 1, 0, \ldots]$. Using this we can compute the equilibrium policy by backwards induction. The agent has no (meaningful) choice of actions at time-step 1. Actions that form the sub-game perfect equilibrium policy are marked as solid arrows. The sub-game perfect equilibrium policy is unique for all histories except the first (where both actions are part of a sub-game perfect equilibrium policy). We omit the computation, but notice that the left sub-game is the same as the example at the start of this section, so has the same analysis.

Now suppose there exists a $\boldsymbol{d}$ such that $\pi^*_{\boldsymbol{d}}$ is equal to the sub-game perfect equilibrium policy. From the left part of the environment we see that $\boldsymbol{d} \cdot ([0,4,0,0] - [0,1,1,3]) > 0$ while from the right part of the environment we see that $\boldsymbol{d} \cdot ([0,4,0,0] - [0,1,1,3]) < 0$, which is a contradiction. Therefore no such $\boldsymbol{d}$ exists. □

6. Discussion

Summary. We have generalised Samuelson's DU model to allow the discount vector used by an agent to vary with its age. In combination with a flexible model choice this extension allowed us to address all four points made in the introduction of this paper while substantially increasing the number of time-consistent discount matrices.

Theorem 13 gives a characterisation of time-(in)consistent discount matrices and shows that all time-consistent discount matrices follow the simple form of $d^k_t = d^1_t$. Theorem 15 shows that using a discount matrix that is nearly time-consistent produces mixed policies with low regret. This is useful for a few reasons, including showing that small perturbations, such as rounding errors, in a discount matrix cannot cause major time-inconsistency problems. It also shows that "cutting off" time-consistent discount matrices after some fixed depth – which makes the agent potentially time-inconsistent – doesn't affect the policies too much, provided the depth is large enough. When a discount matrix is very time-inconsistent then taking a game theoretic approach may dramatically decrease the regret in the change of policy over time.

Some comments on the policies $\pi^*_{\boldsymbol{d}^k}$ (policy maximising expected $\boldsymbol{d}^k$-discounted reward), π_D (mixed policy using $\pi^*_{\boldsymbol{d}^k}$ at each time-step k) and π^*_D (sub-game perfect equilibrium policy).

1. A time-consistent agent should play policy $\pi^*_{\boldsymbol{d}^k} = \pi_D$ for any k. In this case, every optimal policy $\pi^*_{\boldsymbol{d}^k}$ is also a sub-game perfect equilibrium policy π^*_D.
2. π_D will be played by an agent that believes it is time-consistent, but may not be. This can lead to very bad behaviour as shown in Section 3.
3. An agent may play π^*_D if it knows it is time-inconsistent, and also knows exactly how (i.e., it knows $\boldsymbol{d}^k$ for all k at every time-step). This policy is arguably rational, but comes with its own problems, especially non-uniqueness as discussed.

Assumptions. We made a number of assumptions about which we make some brief comments.

1. Assumption 1, which states that $\mathcal{A}$ and $\mathcal{O}$ are finite, guarantees the existence of an optimal policy. Removing the assumption would force us to use ϵ-optimal policies, which shouldn't be a problem for the theorems to go through with an additive ϵ slop term in some cases.
2. Assumption 9 only affects non-summable discount vectors. Without it, even ϵ-optimal policies need not exist and all the machinery will break down.
3. The use of discrete time greatly reduced the complexity of the analysis. Given a sufficiently general model, the set of environments with continuous time should include all "discrete" environments with transitions only occurring at times $t = 1, 2, 3, \ldots$ and flat reward/observation signals between transitions. For this reason the proof of Theorem 13 should go through essentially unmodified. The same may not be true for Theorems 15 and 19. The former may be fixable with substantial effort (and perhaps should be true intuitively). The latter has been partially addressed, with a positive result in [8,19,20,24].

Open questions.

1. Given a discount matrix D, for which environment is the regret, $V^*_{\boldsymbol{d}^1}(h_{<1}) - V^{\pi_D}_{\boldsymbol{d}^1}(h_{<1})$, maximised?
2. Improve the solution concept of sub-game perfect equilibrium policies in Section 5. What is the rational choice of sub-game perfect equilibrium?
3. Extend all results to the continuous case in all generality.

Acknowledgements

We thank the reviewers for valuable feedback on earlier drafts.

Appendix A. Technical proofs

Before the proof of Theorem 10 we require a definition and two lemmas.

Definition 23. Let Π be the set of all policies and define a metric Δ on Π by $T(\pi_1, \pi_2) := \min_{t \in \mathbb{N}}\{t\colon\ \exists h_{<t} \text{ s.t } \pi_1(h_{<t}) \neq \pi_2(h_{<t})\}$ or ∞ if $\pi_1 = \pi_2$ and $\Delta(\pi_1, \pi_2) := \exp(-T(\pi_1, \pi_2))$.

T is the time-step at which π_1 and π_2 first differ. Now augment Π with the topology induced by the metric D.

Lemma 24. *Π is compact.*

Proof. We proceed by showing Π is totally bounded and complete. Let $\epsilon = \exp(-t)$ and define an equivalence relation by $\pi \sim \pi'$ if and only if $T(\pi_1, \pi_2) \geqslant t$. If $\pi \sim \pi'$ then $\Delta(\pi, \pi') \leqslant \epsilon$. Note that $\Pi/\sim$ is the set of cylinders each consisting of policies that are identical up to time-step t. This is set is finite because the sets of observations and actions are finite and although the set of rewards is infinite, the reward r_k depends deterministically on $h_{<k}$. Now choose a representative from each class in $\Pi/\sim$ to create a finite set $\bar{\Pi}$. Now $\bigcup_{\pi \in \bar{\Pi}} B_\epsilon(\pi) = \Pi$, where $B_\epsilon(\pi)$ is the ball of radius ϵ around π. Therefore Π is totally bounded.

Next we show Π is complete. Let $\pi_1, \pi_2, \ldots$ be a Cauchy sequence with $\Delta(\pi_i, \pi_{i+j}) < \exp(-i)$ for all $j > 0$. Therefore $\pi_i(h_{<k}) = \pi_{i+j}(h_{<k}) \forall h_{<k}$ with $k \leqslant i$, by the definition of Δ. Now define π by $\pi(h_{<t}) := \pi_t(h_{<t})$ and note that $\pi_i(h_{<j}) = \pi(h_{<j}) \forall j \leqslant i$ since $\pi_i(h_{<k}) = \pi_k(h_{<k}) \equiv \pi(h_{<k})$ for $k \leqslant i$. Therefore $\lim_{i\to\infty} \pi_i = \pi$ and so Π is complete. Finally, Π is compact by the Heine–Borel theorem. □

Lemma 25. *Let $\boldsymbol{d}$ be a fixed discount vector and $h_{<k}$ a history sequence of length $k - 1$. When viewed as a function from Π to $\mathbb{R}$, $V^\pi_{\boldsymbol{d}}(h_{<k})$ is continuous. (Under Assumption 9 and using the metric of Definition 23.)*

Proof. Let π be an arbitrary policy and π' satisfy $\Delta(\pi, \pi') < \exp(-t)$ for some $t \in \mathbb{N}$. By the definition of Δ, π and π' are identical until time-step t.

$$\begin{aligned} V^\pi_{\boldsymbol{d}}(h_{<k}) - V^{\pi'}_{\boldsymbol{d}}(h_{<k}) &= \boldsymbol{d} \cdot \left[\boldsymbol{R}^\pi(h_{<k}) - \boldsymbol{R}^{\pi'}(h_{<k})\right] \\ &= \sum_{i=t}^\infty d_i \left[R^\pi(h_{<k})_i - R^{\pi'}(h_{<k})_i\right] \\ &= \sum_{i=t}^\infty d_i \left[\sum_{h_{k:i}} \left(P(h_{1:i}|h_{<k}, \pi) - P\left(h_{1:i}\middle|h_{<k}, \pi'\right)\right) r_i\right]. \end{aligned}$$

Now

$$\begin{aligned} \sum_{i=t}^\infty d_i \sum_{h_{k:i}} P(h_{1:i}|h_{<k}, \pi) r_i &= \sum_{h_{k:t-1}} P(h_{<t}|h_{<k}, \pi) \sum_{i=t}^\infty d_i \sum_{h_{t:i}} P(h_{1:i}|h_{<t}, \pi) r_i \\ &= \sum_{h_{k:t-1}} P(h_{<t}|h_{<k}, \pi) V^\pi_{\boldsymbol{d}}(h_{<t}) \\ &\leqslant \sigma_t := \sup_{\pi \in \Pi} \sum_{h_{k:t-1}} P(h_{<t}|h_{<k}, \pi) V^\pi_{\boldsymbol{d}}(h_{<k}). \end{aligned}$$

Similarly

$$\sum_{i=t}^\infty d_i \sum_{h_{k:i}} P\left(h_{1:i}|h_{<k}, \pi'\right) r_i \leqslant \sigma_t.$$

Therefore

$$-\sigma_t \leqslant V^\pi_{\boldsymbol{d}}(h_{<k}) - V^{\pi'}_{\boldsymbol{d}}(h_{<k}) \leqslant \sigma_t.$$

The proof is completed by noting that $\lim_{t\to\infty} \sigma_t = 0$ by Assumption 9. □

Proof of Theorem 10. Let Π be the space of all policies with the metric of Definition 23. By Lemmas 24 and 25 Π is compact and V is continuous. Therefore $\arg\max_\pi V^\pi_{\boldsymbol{d}^k}(h_{<1})$ exists by the extreme value theorem. □

Proof of Theorem 15. Let t be the ϵ-effective horizon as defined in the statement of Theorem 15 and π_k be the policy obtained by following π_D until time-step k and then following $\pi^*_{\boldsymbol{d}^k}$. We use the shorthand $\boldsymbol{R}^\pi := \boldsymbol{R}^\pi(h_{<1}) \in [0, 1]^\infty$. Then

$$\begin{aligned} & V^*_{\boldsymbol{d}^1}(h_{<1}) - V^{\pi_D}_{\boldsymbol{d}^1}(h_{<1}) \\ &\quad \equiv \left(\boldsymbol{R}^{\pi_1} - \boldsymbol{R}^{\pi_D}\right) \cdot \boldsymbol{d}^1 & \text{(A.1)} \\ &\quad = \left(\boldsymbol{R}^{\pi_1} - \boldsymbol{R}^{\pi_D}\right) \cdot \left(\boldsymbol{d}^1 - \boldsymbol{d}^2\right) + \left(\boldsymbol{R}^{\pi_1} - \boldsymbol{R}^{\pi_D}\right) \cdot \boldsymbol{d}^2 & \text{(A.2)} \\ &\quad \leqslant \left(\boldsymbol{R}^{\pi_1} - \boldsymbol{R}^{\pi_D}\right) \cdot \left(\boldsymbol{d}^1 - \boldsymbol{d}^2\right) + \left(\boldsymbol{R}^{\pi_2} - \boldsymbol{R}^{\pi_D}\right) \cdot \boldsymbol{d}^2 & \text{(A.3)} \end{aligned}$$

$$\leqslant \left(\boldsymbol{R}^{\pi_t} - \boldsymbol{R}^{\pi_D}\right) \cdot \boldsymbol{d}^t + \sum_{\tau=1}^{t-1} \left(\boldsymbol{R}^{\pi_\tau} - \boldsymbol{R}^{\pi_D}\right) \cdot \left(\boldsymbol{d}^\tau - \boldsymbol{d}^{\tau+1}\right) \tag{A.4}$$

$$\leqslant \epsilon + \sum_{\tau=1}^{t-1} \left(\boldsymbol{R}^{\pi_\tau} - \boldsymbol{R}^{\pi_D}\right) \cdot \left(\boldsymbol{d}^\tau - \boldsymbol{d}^{\tau+1}\right) \tag{A.5}$$

$$\leqslant \epsilon + \sum_{\tau=1}^{t-1} \Delta_{\tau,\tau+1}, \tag{A.6}$$

where Eq. (A.1) is the definition of the value function. Eq. (A.2) is algebra. Eq. (A.3) follows because π_1 and π_2 are identical up to time-step 1 at which point, if discounting with respect to $\boldsymbol{d}^2$, it is better to follow $\pi^*_{\boldsymbol{d}^2}$ than $\pi^*_{\boldsymbol{d}^1}$, which implies that $\boldsymbol{R}^{\pi_2} \cdot \boldsymbol{d}^2 \geqslant \boldsymbol{R}^{\pi_1} \cdot \boldsymbol{d}^2$. Eq. (A.4) by iterating the argument in Eq. (A.3). Eq. (A.5) by noting that π_D and π_t are identical until time-step t and then using the definition of the effective horizon. Eq. (A.6) follows by the definition of $\Delta_{\tau,\tau+1}$ and because $\boldsymbol{R}^\pi \in [0,1]^\infty$ for any π. □

Appendix B. Table of notation

Symbol	**Description**
D	Discount matrix (d^k_t)
$\boldsymbol{d}^k$	Discount vector k
d^k_t	The tth component of discount vector $\boldsymbol{d}^k$ (at time k reward r_t is discounted by d^k_t)
k, t	Indices. k usually referring to a discount vector used at fixed time k; t usually a time index.
i	Summing index
ϵ, δ	Small real numbers greater than zero
π, π', π_i	Policies
Π	The space of all policies
$\mathcal{A}$, $\mathcal{O}$, $\mathcal{R}$	Action, reward and observation spaces
$h_{1:t}$, $h_{<t}$	History sequences of length t and $t-1$
$P(h_{1:t} \mid h_{<k}, \pi)$	The probability observing history $h_{1:t}$ given history $h_{<k}$ while following policy π
$\mathbb{N}$, $\mathbb{R}$	The natural and real numbers respectively
$B_\epsilon(\cdot)$	A ball of radius ϵ
$\boldsymbol{R}^\pi(h_{<t})$	The expected reward sequence when following π from history $h_{<t}$
$\pi^*_{\boldsymbol{d}^k}$	The optimal policy when using discount vector $\boldsymbol{d}^k$
π_D	The mixed policy using discount matrix D
π^*_D	The sub-game perfect equilibrium policy using discount matrix D
$\Pi^*_{\boldsymbol{d}^k}$	The set of (equal valued) optimal policies with respect to discount vector $\boldsymbol{d}^k$
$V^*_{\boldsymbol{d}^k}(h_{<t})$	The value of the optimal policy $\pi^*_{\boldsymbol{d}^k}$
γ	Discount rate for geometric discounting
α_k	A real valued scaling factor on a discount vector
κ	Discount rate for hyperbolic discounting
H	Horizon for constant depth discounting
m	Lifespan for fixed lifetime discounting
$\Delta(\pi_1, \pi_2)$	The distance between policies π_1 and π_2 using the metric of Definition 23
$\Delta(\boldsymbol{d}^k, \boldsymbol{d}^j)$	The distance measure between discount vectors $\boldsymbol{d}^k$ and $\boldsymbol{d}^j$ as defined by Definition 14

References

[1] D. Ariely, K. Wertenbroch, Procrastination, deadlines, and performance: Self-Control by precommitment, Psychol. Sci. 13 (3) (May 2002) 219–224.

[2] D. Berry, B. Fristedt, Bandit Problems: Sequential Allocation of Experiments, Chapman and Hall, 1985.

[3] D. Bertsekas, J. Tsitsiklis, Neuro-Dynamic Programming, 1st edition, Athena Scientific, 1996.

[4] S. Frederick, G. Oewenstein, T. O'Donoghue, Time discounting and time preference: A critical review, J. Econ. Lit. 40 (2) (2002).

[5] D. Fudenberg, Subgame-perfect equilibria of finite and infinite-horizon games, J. Econ. Theory 31 (2) (1983).

[6] L. Green, N. Fristoe, J. Myerson, Temporal discounting and preference reversals in choice between delayed outcomes, Psychon. Bull. Rev. 1 (3) (1994) 383–389.

[7] J. Gittins, Bandit processes and dynamic allocation indices, J. R. Stat. Soc. B 41 (2) (1979) 148–177.

[8] S. Goldman, Consistent plans, Rev. Econ. Stud. 47 (3) (1980) 533–537.

[9] G. Harrison, M. Lau, M. Williams, Estimating individual discount rates in Denmark: A field experiment, Am. Econ. Rev. 92 (5) (2002) 1606–1617.

[10] M. Hutter, Self-optimizing and Pareto-optimal policies in general environments based on Bayes-mixtures, in: Proc. 15th Annual Conf. on Computational Learning Theory (COLT'02), Sydney, in: LNAI, vol. 2375, Springer, Berlin, 2002, pp. 364–379.

[11] M. Hutter, Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability, Springer, Berlin, 2005.

[12] M. Hutter, General discounting versus average reward, in: Proc. 17th International Conf. on Algorithmic Learning Theory (ALT'06), Barcelona, in: LNAI, vol. 4264, Springer, Berlin, 2006, pp. 244–258.

[13] T. Koopmans, Stationary ordinal utility and impatience, Econometrica 28 (2) (1960) 287–309.

[14] S. Legg, Machine super intelligence, PhD thesis, IDSIA, University of Lugano, 2008.

[15] S. Legg, M. Hutter, Universal intelligence: A definition of machine intelligence, Minds Mach. 17 (4) (2007) 391–444.
[16] T. Lai, H. Robbins, Asymptotically efficient adaptive allocation rules, Adv. Appl. Math. 6 (1) (1985) 4–22.
[17] P. Norvig, S. Russell, Artificial Intelligence: A Modern Approach, 3rd edition, Prentice Hall Series in Artificial Intelligence, Prentice Hall, 2010.
[18] M. Osborne, A. Rubinstein, A Course in Game Theory, The MIT Press, 1994.
[19] R. Pollak, Consistent planning, Rev. Econ. Stud. 35 (2) (1968) 201–208.
[20] B. Peleg, M. Yaari, On the existence of a consistent course of action when tastes are changing, Rev. Econ. Stud. 40 (3) (1973) 391–401.
[21] P. Samuelson, A note on measurement of utility, Rev. Econ. Stud. 4 (2) (1937) 155–161.
[22] P. Samuelson, Wages and interest: A modern dissection of Marxian economic models, Am. Econ. Rev. 47 (6) (1957) 884–912.
[23] R. Sutton, A. Barto, Reinforcement Learning: An Introduction, MIT Press, 1998.
[24] R. Strotz, Myopia and inconsistency in dynamic utility maximization, Rev. Econ. Stud. 23 (3) (1955) 165–180.
[25] R. Thaler, Some empirical evidence on dynamic inconsistency, Econ. Lett. 8 (3) (1981) 201–207.