Time Consistent Discounting

Tor Lattimore[1] and Marcus Hutter[1,2]

[1] Research School of Computer Science
Australian National University
[2] ETH Zürich
{tor.lattimore,marcus.hutter}@anu.edu.au

Abstract. A possibly immortal agent tries to maximise its summed discounted rewards over time, where discounting is used to avoid infinite utilities and encourage the agent to value current rewards more than future ones. Some commonly used discount functions lead to time-inconsistent behavior where the agent changes its plan over time. These inconsistencies can lead to very poor behavior. We generalise the usual discounted utility model to one where the discount function changes with the age of the agent. We then give a simple characterisation of time-(in)consistent discount functions and show the existence of a rational policy for an agent that knows its discount function is time-inconsistent.

Keywords: Rational agents, sequential decision theory, general discounting, time-consistency, game theory.

1 Introduction

The goal of an agent is to maximise its expected utility; but how do we measure utility? One method is to assign an instantaneous reward to particular events, such as having a good meal, or a pleasant walk. It would be natural to measure the utility of a plan (policy) by simply summing the expected instantaneous rewards, but for immortal agents this may lead to infinite utility and also assumes rewards are equally valuable irrespective of the time at which they are received.

One solution, the discounted utility (DU) model introduced by Samuelson in [12], is to take a weighted sum of the rewards with earlier rewards usually valued more than later ones.

There have been a number of criticisms of the DU model, which we will not discuss. For an excellent summary, see [1]. Despite the criticisms, the DU model is widely used in both economics and computer science.

A discount function is time-inconsistent if plans chosen to maximise expected discounted utility change over time. For example, many people express a preference for \$110 in 31 days over \$100 in 30 days, but reverse that preference 30 days later when given a choice between \$110 tomorrow or \$100 today [4]. This behavior can be caused by a rational agent with a time-inconsistent discount function.

Unfortunately, time-inconsistent discount functions can lead to extremely bad behavior and so it becomes important to ask what discount functions are time-inconsistent.

J. Kivinen et al. (Eds.): ALT 2011, LNAI 6925, pp. 383–397, 2011.

Previous work has focussed on a continuous model where agents can take actions at any time in a continuous time-space. We consider a discrete model where agents act in finite time-steps. In general this is not a limitation since any continuous environment can be approximated arbitrarily well by a discrete one. The discrete setting has the advantage of easier analysis, which allows us to consider a very general setup where environments are arbitrary finite or infinite Markov decision processes.

Traditionally, the DU model has assumed a sliding discount function. Formally, a sequence of instantaneous utilities (rewards) $R = (r_k, r_{k+1}, r_{k+2}, \cdots)$ starting at time k, is given utility equal to $\sum_{t=k}^{\infty} d_{t-k} r_t$ where $\boldsymbol{d} \in [0,1]^{\infty}$. We generalise this model as in [6] by allowing the discount function to depend on the age of the agent. The new utility is given by $\sum_{t=k}^{\infty} d_t^k r_t$. This generalisation is consistent with how some agents tend to behave; for example, humans becoming temporally less myopic as they grow older.

Strotz [13] showed that the only time-consistent sliding discount function is geometric discounting. We extend this result to a full characterisation of time-consistent discount functions where the discount function is permitted to change over time. We also show that discounting functions that are "nearly" time-consistent give rise to low regret in the anticipated future changes of the policy over time.

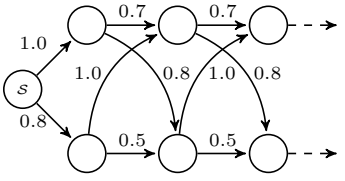
Another important question is what policy should be adopted by an agent that knows it is time-inconsistent. For example, if it knows it will become temporarily myopic in the near future then it may benefit from paying a price to pre-commit to following a particular policy. A number of authors have examined this question in special continuous cases, including [3, 10, 11, 13]. We modify their results to our general, but discrete, setting using game theory.

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 analyzed (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 that 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 $h_{1:t}$ to be a history of length t, $h_{<t}$ to

be a history of length $t-1$. 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 on the right represents an environment where either pizza or pasta 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 pasta`. The rewards are for a consumer who prefers pizza to pasta, 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.

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 1 (Policy). *A policy is a mapping $\pi : \mathcal{H} \to \mathcal{A}$ giving an action for each history.*

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

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

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

Definition 2 (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*

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

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

Note while the set of all possible $h_{t:k} \in (\mathcal{A} \times \mathcal{R} \times \mathcal{O})^{k-t+1}$ 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 3 (Discount Vector). *A discount vector $\boldsymbol{d}^k \in [0, 1]^\infty$ is a vector $\left[d_1^k, d_2^k, d_3^k, \cdots\right]$ satisfying $d_t^k > 0$ for at least one $t \geq k$.*

The apparently superfluous superscript k will be useful later when we allow the discount vector to change with time. We do *not* insist that the discount vector be summable, $\sum_{t=k}^{\infty} d_t^k < \infty$.

Definition 4 (Expected Values). *The expected discounted reward (or utility or value) when using policy π starting in history $h_{<t}$ and discount vector $\boldsymbol{d}^k$ is*

$$V^{\pi}_{\boldsymbol{d}^k}(h_{<t}) := \boldsymbol{R}^{\pi}(h_{<t}) \cdot \boldsymbol{d}^k := \sum_{i=1}^{\infty} R^{\pi}(h_{<t})_i d^k_i = \sum_{i=t}^{\infty} R^{\pi}(h_{<t})_i d^k_i.$$

The sum can be taken to start from t since $R^{\pi}(h_{<t})_i = 0$ for $i < t$. This means that the value of d^k_t for $t < k$ is unimportant, and never will be for any result in this paper. As the scalar product is linear, a scaling of a discount vector has no affect on the ordering of the policies. Formally, if $V^{\pi_1}_{\boldsymbol{d}^k}(h_{<t}) \geq V^{\pi_2}_{\boldsymbol{d}^k}(h_{<t})$ then $V^{\pi_1}_{\alpha\boldsymbol{d}^k}(h_{<t}) \geq V^{\pi_2}_{\alpha\boldsymbol{d}^k}(h_{<t})$ for all $\alpha > 0$.

Definition 5 (Optimal Policy/Value). *In general, our agent will try to choose a policy $\pi^*_{\boldsymbol{d}^k}$ to maximise $V^{\pi}_{\boldsymbol{d}^k}(h_{<t})$. This is defined as follows.*

$$\pi^*_{\boldsymbol{d}^k}(h_{<t}) := \arg\max_{\pi} V^{\pi}_{\boldsymbol{d}^k}(h_{<t}), \qquad \boldsymbol{R}^*_{\boldsymbol{d}^k}(h_{<t}) := \boldsymbol{R}^{\pi^*_{\boldsymbol{d}^k}}(h_{<t}),$$
$$V^*_{\boldsymbol{d}^k}(h_{<t}) := V^{\pi^*_{\boldsymbol{d}^k}}_{\boldsymbol{d}^k}(h_{<t}).$$

If multiple policies are optimal then $\pi^*_{\boldsymbol{d}^k}$ is chosen using some arbitrary rule. Unfortunately, $\pi^*_{\boldsymbol{d}^k}$ need not exist without one further assumption.

Assumption 2. *For all π and $k \geq 1$, $\lim_{t\to\infty} \sum_{h_{<t}} P(h_{<t}|\pi) V^{\pi}_{\boldsymbol{d}^k}(h_{<t}) = 0$.*

Assumption 2 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 frequently 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. This restriction is necessary to guarantee the existence of the value function.

From now on, including in theorem statements, we only consider environments/discount vectors satisfying Assumptions 1 and 2. The following theorem then guarantees the existence of $\pi^*_{\boldsymbol{d}^k}$.

Theorem 6 (Existence of Optimal Policy). *$\pi^*_{\boldsymbol{d}^k}$ exists for any environment and discount vector $\boldsymbol{d}^k$ satisfying Assumptions 1 and 2.*

The proof of the existence theorem is in the appendix.

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

Definition 7 (Discount Matrix). *A discount matrix $\mathbf{d}$ 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 $\mathbf{d}$ (written bold), a discount vector $\boldsymbol{d}^k$ (bold and italics), and a particular value in a discount vector d^k_t (just italics).

Definition 8 (Sliding Discount Matrix). *A discount matrix* $\mathbf{d}$ *is sliding if* $d^k_{k+t} = d^1_{t+1}$ *for all* $k, t \geq 1$.

Definition 9 (Mixed Policy). *The mixed policy is the policy where at each time step* t*, the agent acts according to the possibly different policy* $\pi^*_{\boldsymbol{d}^t}$.

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

We do not denote the mixed policy by $\pi^*_{\mathbf{d}}$ 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 $\pi_{\mathbf{d}}$. All theorems are proved with respect to any choice $\pi_{\mathbf{d}}$.

Definition 10 (Time Consistency). *A discount matrix* $\mathbf{d}$ *is time consistent if and only if for all environments* $\pi^*_{\boldsymbol{d}^k}(h_{<t}) = \pi^*_{\boldsymbol{d}^j}(h_{<t})$*, for all* $h_{<t}$ *where* $t \geq k, j$.

This means that a time-consistent agent taking action $\pi^*_{\boldsymbol{d}^t}(h_{<t})$ at each time 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 $h_{<t}$ ($\pi^*_{\boldsymbol{d}^0}(h_{<t}) = a$). 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 behavior.

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 behavior 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 it requires you to know the lifetime of the agent beforehand and also makes asymptotic analysis impossible.

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 behavior by suggesting that we discount hyperbolically [14]. The hyperbolic discount matrix is not summable, so may be replaced by the following (similar to [5]), which has similar properties for β close to 1.

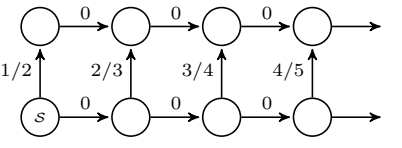
$$d^k_t = 1/(1 + \kappa(t-k))^\beta \text{ with } \beta > 1.$$

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

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

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.

No Discounting. $d_t^k = 1$, for all k, t. [8] and [7] 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. Of course, this "discount matrix" is not summable, but can be made to work by insisting that all environments satisfy Assumption 2. 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.



To illustrate the potential consequences of time-inconsistent discount matrices we consider the policies of several agents acting in the following 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 $2/3$. However, once in second time-step, it 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. $V_{\boldsymbol{d}^k}^{\pi_t}(h_{<1}) = \gamma^t \frac{(t+1)}{(t+2)}$. This value does not depend on k and so the agent will move right until $t = \arg\max \left\{ \gamma^t \frac{(t+1)}{t+2} \right\} < \infty$ when it will move up and receive a reward.

The actions of agent A are an example of the worst possible behavior 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 behavior 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 11 (Characterisation). *Let* $\mathbf{d}$ *be a discount matrix, then the following are equivalent.*

1. $\mathbf{d}$ *is time-consistent (Definition 10)*
2. *For each* k *there exists an* $\alpha_k \in \mathbb{R}$ *such that* $d_t^k = \alpha_k d_t^1$ *for all* $t \geq k \in \mathbb{N}$.

Recall that a discount matrix is sliding if $d_t^k = d_{t-k+1}^1$. Theorem 11 can be used to show that if a sliding discount matrix is used as in [13] then the only time-consistent discount matrix is geometric. Let $\mathbf{d}$ be a time-consistent sliding discount matrix. By Theorem 11 and the definition of sliding, $\alpha_1 d_{t+1}^1 = d_{t+1}^2 = d_t^1$. Therefore $\frac{1}{\alpha_1} d_2^1 = d_1^1$ and $d_3^1 = \frac{1}{\alpha_1} d_2^1 = \left(\frac{1}{\alpha_1}\right)^2 d_1^1$ and similarly, $d_t^1 = \left(\frac{1}{\alpha_1}\right)^{t-1} d_1^1 \propto \gamma^t$ with $\gamma = 1/\alpha_1$, which is geometric discounting. This is the analogue to the results of [13] 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 $d_t^k := d_t^1$ for all $t \geq k$ will always be time-consistent, for example, the *fixed lifetime* discount matrix with $d_t^k = 1$ if $t \leq H$ for some horizon H. Indeed, all time-consistent discount rates can be constructed in this way (up to scaling).

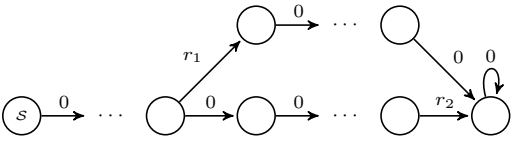
Proof (Theorem 11). $2 \implies 1$: This direction follows easily from linearity of the scalar product.

$$\pi^*_{\boldsymbol{d}^k}(h_{<t}) \equiv \arg\max_\pi V^\pi_{\boldsymbol{d}^k}(h_{<t}) \equiv \arg\max_\pi \boldsymbol{R}^\pi(h_{<t}) \cdot \boldsymbol{d}^k = \arg\max_\pi \boldsymbol{R}^\pi(h_{<t}) \cdot \alpha_k \boldsymbol{d}^1 \tag{2}$$

$$= \arg\max_\pi \alpha_k \boldsymbol{R}^\pi(h_{<t}) \cdot \boldsymbol{d}^1 = \arg\max_\pi \boldsymbol{R}^\pi(h_{<t}) \cdot \boldsymbol{d}^1 \equiv \pi^*_{\boldsymbol{d}^1}(h_{<t})$$

as required. The last equality of (2) follows from the assumption that $d_t^k = \alpha_k d_t^1$ for all $t \geq k$ and because $\boldsymbol{R}^\pi(h_{<t})_i = 0$ for all $i < t$.

$1 \implies 2$: Let $\boldsymbol{d}^0$ and $\boldsymbol{d}^k$ be the discount vectors used at times 0 and k respectively. Now let $k \leq t_1 < t_2 < \cdots$ and consider the deterministic environment below where the agent has a choice between earning reward r_1 at time t_1 or r_2 at time t_2. In this environment there are only two policies, π_1 and π_2, where $\boldsymbol{R}^{\pi_1}(h_{<k}) = r_1 \boldsymbol{e}_{t_1}$ and $\boldsymbol{R}^{\pi_2}(h_{<k}) = r_2 \boldsymbol{e}_{t_2}$ with $\boldsymbol{e}_i$ the infinite vector with all components zero except the ith, which is 1.



Since $\mathbf{d}$ is time-consistent, for all $r_1, r_2 \in \mathcal{R}$ and $k \in \mathbb{N}$ we have:

$$\arg\max_\pi V^\pi_{\boldsymbol{d}^1}(h_{<k}) \equiv \arg\max_\pi \boldsymbol{R}^\pi(h_{<k}) \cdot \boldsymbol{d}^1 \tag{3}$$

$$= \arg\max_\pi \boldsymbol{R}^\pi(h_{<k}) \cdot \boldsymbol{d}^k \equiv \arg\max_\pi V^\pi_{\boldsymbol{d}^k}(h_{<k}). \tag{4}$$

Now $V^{\pi_1}_{\boldsymbol{d}^k} \geq V^{\pi_2}_{\boldsymbol{d}^k}$ if and only if $\boldsymbol{d}^k \cdot [\boldsymbol{R}^{\pi_1}(h_{<k}) - \boldsymbol{R}^{\pi_2}(h_{<k})] = [d_{t_1}^k, d_{t_2}^k] \cdot [r_1, -r_2] \geq 0$. Therefore we have that,

$$[d_{t_1}^1, d_{t_2}^1] \cdot [r_1, -r_2] \geq 0 \Leftrightarrow [d_{t_1}^k, d_{t_2}^k] \cdot [r_1, -r_2] \geq 0. \tag{5}$$

Letting $\cos\theta_k$ be the cosine of the angle between $[d^k_{t_1}, d^k_{t_2}]$ and $[r_1, -r_2]$ then Equation (5) becomes $\cos\theta_0 \geq 0 \Leftrightarrow \cos\theta_k \geq 0$. Choosing $[r_1, -r_2] \propto [d^1_{t_2}, -d^1_{t_1}]$ implies that $\cos\theta_0 = 0$ and so $\cos\theta_k = 0$. Therefore there exists $\alpha_k \in \mathbb{R}$ such that

$$[d^k_{t_1}, d^k_{t_2}] = \alpha_k [d^1_{t_1}, d^1_{t_2}]. \tag{6}$$

Let $k \leq t_1 < t_2 < t_3 < \cdots$ be a sequence for which $d^1_{t_j} > 0$. By the previous argument we have that, $[d^k_{t_i}, d^k_{t_{i+1}}] = \alpha_k [d^1_{t_i}, d^1_{t_{i+1}}]$ and $[d^k_{t_{i+1}}, d^k_{t_{i+2}}] = \tilde{\alpha}_k [d^1_{t_{i+1}}, d^1_{t_{i+2}}]$. Therefore $\alpha_k = \tilde{\alpha}_k$, and by induction, $d^k_{t_i} = \alpha_k d^1_{t_i}$ for all i. Now if $t \geq k$ and $d^1_t = 0$ then $d^k_t = 0$ by equation (6). By symmetry, $d^k_t = 0 \implies d^1_t = 0$. Therefore $d^k_t = \alpha_k d^1_t$ for all $t \geq k$ as required. □

In Section 3 we saw an example where time-inconsistency led to very bad behavior. 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 behavior? For example, could rounding errors when using a geometric discount matrix seriously affect the agent's behavior? The following Theorem shows that this is not possible. First we require a measure of the cost of time-inconsistent behavior. The regret experienced by the agent at time zero from following policy $\pi_{\mathbf{d}}$ rather than $\pi^*_{\boldsymbol{d}^1}$ is $V^*_{\boldsymbol{d}^1}(h_{<1}) - V^{\pi_{\mathbf{d}}}_{\boldsymbol{d}^1}(h_{<1})$. We also need a distance measure on the space of discount vectors.

Definition 12 (Distance Measure). *Let $\boldsymbol{d}^k, \boldsymbol{d}^j$ be discount vectors then define a distance measure D by*

$$D(\boldsymbol{d}^k, \boldsymbol{d}^j) := \sum_{i=\max\{k,j\}}^{\infty} |d^k_i - d^j_i|.$$

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

Theorem 13 (Continuity). *Suppose $\epsilon \geq 0$ and $D_{k,j} := D(\boldsymbol{d}^k, \boldsymbol{d}^j)$ then*

$$V^*_{\boldsymbol{d}^1}(h_{<1}) - V^{\pi_{\mathbf{d}}}_{\boldsymbol{d}^1}(h_{<1}) \leq \epsilon + D_{1,t} + \sum_{k=1}^{t-1} D_{k,k+1}$$

with $t = \min\left\{t : \sum_{h_{<t}} P(h_{<t}|\pi^*_{\boldsymbol{d}^1}) V^*_{\boldsymbol{d}^1}(h_{<t}) \leq \epsilon\right\}$, which for $\epsilon > 0$ is guaranteed to exist by Assumption 2.

Theorem 13 implies that the regret of the agent at time zero 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 behavior.

The proof is omitted due to limitations in space. It relies on proving the result for finite horizon environments and showing that this extends to the infinite case by using the horizon, t, after which the actions of the agent are no longer important. The bound in Theorem 13 is tight in the following sense.

Theorem 14. *For $\delta > 0$ and $t \in \mathbb{N}$ and any sufficiently small $\epsilon > 0$ there exists an environment and discount matrix such that*

$$(t-2)(1-\epsilon)\delta < V^*_{\boldsymbol{d}^1}(h_{<1}) - V^{\pi_{\mathbf{d}}}_{\boldsymbol{d}^1}(h_{<1}) < (t+1)\delta$$
$$\equiv D_{1,t} + \sum_{i=1}^{t-1} D_{i,i+1}$$

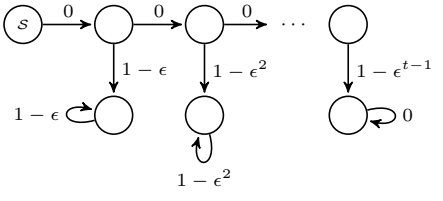
where $t = \min\left\{t : \sum_{h_{<t}} P(h_{<t}|\pi^*_{\boldsymbol{d}^1}) V^*_{\boldsymbol{d}^1}(h_{<t}) = 0\right\} < \infty$ *and where* $D(\boldsymbol{d}^k, \boldsymbol{d}^j) \equiv D_{k,j} = \delta$ *for all* k, j.

Note that t in the statement above is the same as that in the statement of Theorem 13. Theorem 14 shows that there exists a discount matrix, environment and $\epsilon > 0$ where the regret due to time-inconsistency is nearly equal to the bound given by Theorem 13.

Proof (Theorem 14). Define $\mathbf{d}$ by

$$d^k_i = \begin{cases} \delta & \text{if } k < i < t \\ 0 & \text{otherwise} \end{cases}$$

Observe that $D(\boldsymbol{d}^k, \boldsymbol{d}^j) = \delta$ for all $k < j < t$ since $d^j_i = d^k_i$ for all i except $i = j$. Now consider the environment below.



For sufficiently small ϵ, the agent at time zero will plan to move right and then down leading to $\boldsymbol{R}^*_{\boldsymbol{d}^1}(h_{<1}) = [0, 1-\epsilon, 1-\epsilon, \cdots]$ and $V^*_{\boldsymbol{d}^1}(h_{<1}) = (t-1)\delta(1-\epsilon)$.

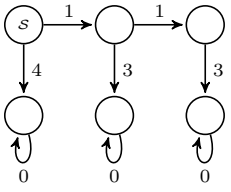
To compute $\boldsymbol{R}_{\mathbf{d}}$ note that $d^k_k = 0$ for all k. Therefore the agent in time-step k doesn't care about the next instantaneous reward, so prefers to move right with the intention of moving down in the next time-step when the rewards are slightly better. This leads to $\boldsymbol{R}_{\mathbf{d}}(h_{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

$$V^*_{\boldsymbol{d}^1}(h_{<1}) - V^{\pi_{\mathbf{d}}}_{\boldsymbol{d}^1}(h_{<1}) = (t-1)\delta(1-\epsilon) - (1-\epsilon^{t-1})\delta \geq (t-2)\delta(1-\epsilon)$$

as required. □

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 $\boldsymbol{d}^1 = [2,1,2,0,0,\cdots]$ and $\boldsymbol{d}^2 = [*,3,1,0,0,\cdots]$ and consider the environment on the right. Initially, the agent has two choices. It can either move down to guarantee a reward sequence of $\boldsymbol{r} = [4,0,0,\cdots]$ which has utility of $\boldsymbol{d}^1 \cdot [4,0,0,\cdots] = 8$ or it can move right in which case it will receive a reward sequence of either $\boldsymbol{r}' = [1,3,0,0,\cdots]$ with utility 5 or $\boldsymbol{r}'' = [1,1,3,0,0,\cdots]$ 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 $[1,3,0,0,\cdots]$ 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 which is used to find sub-game perfect equilibria in finite extensive games. A variant of Kuhn's theorem proves that backwards induction can be used to find such equilibria in finite extensive games [9]. For arbitrary extensive games (possibly infinite) 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 [9].

Definition 15 (Sub-game Perfect Equilibria). *A policy $\pi^*_{\mathbf{d}}$ is a sub-game perfect equilibrium policy if and only if for each t $V^{\pi^*_{\mathbf{d}}}_{\boldsymbol{d}^t}(h_{<t}) \geq V^{\tilde{\pi}}_{\boldsymbol{d}^t}(h_{<t})$, for all $h_{<t}$, where $\tilde{\pi}$ is any policy satisfying $\tilde{\pi}(h_{<i}) = \pi^*_{\mathbf{d}}(h_{<i}) \forall h_{<i}$ where $i \neq t$.*

Theorem 16 (Existence of Sub-game Perfect Equilibrium Policy). *For all environments and discount matrices $\mathbf{d}$ satisfying Assumptions 1 and 2 there exists at least one sub-game perfect equilibrium policy $\pi^*_{\mathbf{d}}$.*

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 2). The closest related work appears to be that of Drew Fudenberg in [2], but our proof (see appendix) 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 [9] a sub-game perfect equilibrium policy must exist in each of these finite games. However the space of policies is compact (Lemma 21) and so this sequence of sub-game perfect equilibrium policies contains a convergent

sub-sequence converging to policy π. It is not then hard to show that π is a sub-game prefect equilibrium policy in the original environment.

Proof (Theorem 16). Add an action a^{death} to $\mathcal{A}$ and μ such that if a^{death} is taken at any time in $h_{<t}$ then μ returns zero reward. Essentially, once in the agent takes action a^{death}, the agent receives zero reward forever. Now if $\pi^*_{\mathbf{d}}$ is a sub-game perfect equilibrium policy in this modified environment then it is a sub-game perfect equilibrium policy in the original one.

For each $t \in \mathbb{N}$ choose π_t to be a sub-game perfect equilibrium policy in the further modified environment obtained by setting $r_i = 0$ if $i > t$. That is, the environment which gives zero reward always after time t. We can assume without loss of generality that $\pi_t(h_{<k}) = a^{death}$ for all $k \geq t$. Since Π is compact, the sequence $\pi_1, \pi_2, \cdots$ has a convergent subsequence $\pi_{t_1}, \pi_{t_2}, \cdots$ converging to π and satisfying

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

We write $\tilde{V}^{\pi_{t_i}}$ for the value function in the modified environment. It is now shown that π is a sub-game perfect equilibrium policy in the original environment. Fix a $t \in \mathbb{N}$ and let $\tilde{\pi}$ be a 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

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

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

$$\begin{aligned}
V^{\pi}_{\boldsymbol{d}^t}(h_{<t}) &\geq V^{\pi_{t_i}}_{\boldsymbol{d}^t}(h_{<t}) - |V^{\pi}_{\boldsymbol{d}^t}(h_{<t}) - V^{\pi_{t_i}}_{\boldsymbol{d}^t}(h_{<t})| && (7)\\
&\geq \tilde{V}^{\pi_{t_i}}_{\boldsymbol{d}^t}(h_{<t}) - |V^{\pi}_{\boldsymbol{d}^t}(h_{<t}) - V^{\pi_{t_i}}_{\boldsymbol{d}^t}(h_{<t})| && (8)\\
&\geq \tilde{V}^{\tilde{\pi}_{t_i}}_{\boldsymbol{d}^t}(h_{<t}) - |V^{\pi}_{\boldsymbol{d}^t}(h_{<t}) - V^{\pi_{t_i}}_{\boldsymbol{d}^t}(h_{<t})| && (9)\\
&\geq V^{\tilde{\pi}}_{\boldsymbol{d}^t}(h_{<t}) - |V^{\pi}_{\boldsymbol{d}^t}(h_{<t}) - V^{\pi_{t_i}}_{\boldsymbol{d}^t}(h_{<t})| \\
&\quad - |V^{\tilde{\pi}_{t_i}}_{\boldsymbol{d}^t}(h_{<t}) - \tilde{V}^{\tilde{\pi}_{t_i}}_{\boldsymbol{d}^t}(h_{<t})| - |V^{\tilde{\pi}_{t_i}}_{\boldsymbol{d}^t}(h_{<t}) - V^{\tilde{\pi}}_{\boldsymbol{d}^t}(h_{<t})| && (10)
\end{aligned}$$

where (7) follows from arithmetic. (8) since $V \geq \tilde{V}$. (9) since π_{t_i} is a sub-game perfect equilibrium policy. (10) by arithmetic. We now show that the absolute value terms in (10) converge to zero. Since $V^{\pi}(\cdot)$ is continuous in π and $\lim_{i\to\infty} \pi_{t_i} = \pi$ and $\lim_{i\to\infty} \tilde{\pi}_{t_i} = \tilde{\pi}$, we obtain $\lim_{i\to\infty} \left[|V^{\pi}_{\boldsymbol{d}^t}(h_{<t}) - V^{\pi_{t_i}}_{\boldsymbol{d}^t}(h_{<t})| + |V^{\tilde{\pi}_{t_i}}_{\boldsymbol{d}^t}(h_{<t}) - V^{\tilde{\pi}}_{\boldsymbol{d}^t}(h_{<t})| \right] = 0$. Now $\tilde{\pi}_{t_i}(h_{<k}) = a^{death}$ if $k \geq t_i$, so $|V^{\tilde{\pi}_{t_i}}(h_{<t}) - \tilde{V}^{\tilde{\pi}_{t_i}}(h_{<t})| = 0$. Therefore taking the limit as i goes to infinity in (10) shows that $V^{\pi}_{\boldsymbol{d}^t}(h_{<t}) \geq V^{\tilde{\pi}}_{\boldsymbol{d}^t}(h_{<t})$ as required. □

In general, $\pi^*_{\mathbf{d}}$ need not be unique, and different sub-game equilibrium policies can lead to different utilities. This is a normal, but unfortunate, problem with the sub-game equilibrium solution concept. The policy is unique if for all players the value of any two arbitrary policies is different. Also, if $\forall k(V^{\pi_1}_{\boldsymbol{d}^k} = V^{\pi_2}_{\boldsymbol{d}^k} \implies \forall j V^{\pi_1}_{\boldsymbol{d}^j} = V^{\pi_2}_{\boldsymbol{d}^j})$ 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 equilibrium policies.

Example 17. 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 $\boldsymbol{R}^{\pi_1}(h_{<1}) = [1/2, 0, 0, \cdots]$ and $\boldsymbol{R}^{\pi_2}(h_{<1}) = [0, 2/3, 0, 0, \cdots]$. 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 18. *If* $\mathbf{d}$ *is time-consistent then* $V^*_{\boldsymbol{d}^k} = V^{\pi_{\mathbf{d}}}_{\boldsymbol{d}^k} = V^{\pi^*_{\mathbf{d}}}_{\boldsymbol{d}^k}$ *for all* k *and choices of* $\pi^*_{\boldsymbol{d}^k}$ *and* $\pi_{\mathbf{d}}$ *and* $\pi^*_{\mathbf{d}}$.

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 this (and more general) settings.

Theorem 19. $\exists \mathbf{d}$ *such that* $\pi^*_{\mathbf{d}} \neq \pi^*_{\tilde{\boldsymbol{d}}^0}$, *for all* $\tilde{\boldsymbol{d}}^0$.

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 $\pi^*_{\boldsymbol{d}}$ 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.

6 Discussion

Summary. Theorem 11 gives a characterisation of time-(in)consistent discount matrices and shows that all time-consistent discount matrices follow the simple form of $d_t^k = d_t^1$. Theorem 13 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), $\pi_{\mathbf{d}}$ (mixed policy using $\pi^*_{\boldsymbol{d}^k}$ at each time-step t) and $\pi^*_{\mathbf{d}}$ (sub-game perfect equilibrium policy).

1. A time-consistent agent should play policy $\pi^*_{\boldsymbol{d}^k} = \pi_{\mathbf{d}}$ for any k. In this case, every optimal policy $\pi^*_{\boldsymbol{d}^k}$ is also a sub-game perfect equilibrium policy.
2. $\pi_{\mathbf{d}}$ will be played by an agent that believes it is time-consistent, but may not be. This can lead to very bad behavior as shown in Section 3.
3. An agent may play $\pi^*_{\mathbf{d}}$ 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 2 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 continuous environments should contain all discrete environments. For this reason the proof of Theorem 11 should go through essentially unmodified. The same may not be true for Theorems 13 and 16. 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 [3, 10, 11, 13].

Acknowledgements. We thank reviewers for valuable feedback on earlier drafts.

References

[1] Frederick, S., Oewenstein, G.L., O'Donoghue, T.: Time discounting and time preference: A critical review. Journal of Economic Literature 40(2) (2002)
[2] Fudenberg, D.: Subgame-perfect equilibria of finite and infinite-horizon games. Journal of Economic Theory 31(2) (1983)
[3] Goldman, S.M.: Consistent plans. The Review of Economic Studies 47(3), 533–537 (1980)
[4] Green, L., Fristoe, N., Myerson, J.: Temporal discounting and preference reversals in choice between delayed outcomes. Psychonomic bulletin and review 1(3), 383–389 (1994)
[5] Hutter, M.: Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability. Springer, Berlin (2004)
[6] Hutter, M.: General Discounting Versus Average Reward. In: Balcázar, J.L., Long, P.M., Stephan, F. (eds.) ALT 2006. LNCS (LNAI), vol. 4264, pp. 244–258. Springer, Heidelberg (2006)
[7] Legg, S.: Machine Super Intelligence. PhD thesis, University of Lugano (2008)
[8] Legg, S., Hutter, M.: Universal intelligence: A definition of machine intelligence. Minds & Machines 17(4), 391–444 (2007)
[9] Osborne, M.J., Rubinstein, A.: A Course in Game Theory. The MIT Press, Cambridge (1994)
[10] Peleg, B., Yaari, M.E.: On the existence of a consistent course of action when tastes are changing. The Review of Economic Studies 40(3), 391–401 (1973)
[11] Pollak, R.A.: Consistent planning. The Review of Economic Studies 35(2), 201–208 (1968)
[12] Samuelson, P.A.: A note on measurement of utility. The Review of Economic Studies 4(2), 155–161 (1937)
[13] Strotz, R.H.: Myopia and inconsistency in dynamic utility maximization. The Review of Economic Studies 23(3), 165–180 (1955)
[14] Thaler, R.: Some empirical evidence on dynamic inconsistency. Economics Letters 8(3), 201–207 (1981)

A Technical Proofs

Before the proof of Theorem 6 we require a definition and two lemmas.

Definition 20. *Let* $\Pi = \mathcal{A}^{\mathcal{S}}$ *be the set of all policies and define a metric* D *on* Π *by* $T(\pi_1, \pi_2) := \min_{t \in \mathbb{N}} \{t : \exists h_{<t} \ s.t \ \pi_1(h_{<t}) \neq \pi_2(h_{<t})\}$ *or* ∞ *if* $\pi_1 = \pi_2$ *and* $D(\pi_1, \pi_2) := \exp(-T(\pi_1, \pi_2))$.

T is the time-step at which π_1 and π_2 first differ. Now augment Π with the topology induced by the metric $\mathbf{d}$.

Lemma 21. *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 $T(\pi_1, \pi_2) \geq t$. If $\pi \sim \pi'$ then $D(\pi, \pi') \leq \epsilon$. Note that $\Pi/_\sim$ is finite. Now choose a representative from each class 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 about π. Therefore Π is totally bounded.

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

Lemma 22. *When viewed as a function from Π to $\mathbb{R}$, $V^{\pi}_{\boldsymbol{d}^k}(\cdot)$ is continuous. (given Assumption 2)*

Proof. Suppose $D(\pi_1, \pi_2) < \exp(-t)$ then π_1 and π_2 are identical on all histories up to length t. Therefore

$$\begin{aligned} |V^{\pi_1}_{\boldsymbol{d}^k}(h_{<k}) - V^{\pi_2}_{\boldsymbol{d}^k}(h_{<k})| &\leq \boldsymbol{d}^k \cdot [\boldsymbol{R}^{\pi_1}(h_{<k}) + \boldsymbol{R}^{\pi_2}(h_{<k})] \\ &= \sum_{i=k}^{\infty} d_i^k \left(R^{\pi_1}(h_{<k})_i + R_i^{\pi_2}(h_{<k})_i\right). \end{aligned} \tag{11}$$

Since π_1 and π_2 are identical up to time t, (11) becomes

$$\begin{aligned} &\sum_{i=t}^{\infty} d_i^k \left(R^{\pi_1}(h_{<k})_i + R_i^{\pi_2}(h_{<k})_i\right) = \\ &\qquad \sum_{h_{<t}} \left[P(h_{<t}|h_{<k}, \pi_1) V^{\pi_1}_{\boldsymbol{d}^k}(h_{<t}) + P(h_{<t}|h_{<k}, \pi_2) V^{\pi_2}_{\boldsymbol{d}^k}(h_{<t})|\right] \end{aligned} \tag{12}$$

where (12) follows from the definition of the reward and value functions. By Assumption 2, $\lim_{t\to\infty} \sum_{h_{<t}} P(h_{<t}|h_{<k}, \pi_i) V^{\pi_i}_{\boldsymbol{d}^k}(h_{<t}) = 0$ for $i \in \{1, 2\}$ and so, V is continuous. □

Proof (Theorem 6). Let Π be the space of all policies with the metric of Definition 20. By Lemmas 21/22 Π is compact and V is continuous. Therefore $\arg\max_{\pi} V^{\pi}_{\boldsymbol{d}^k}(h_{<1})$ exists by the extreme value theorem. □