Rethinking the Discount Factor in Reinforcement Learning:
A Decision Theoretic Approach

Silviu Pitis

University of Toronto, Vector Institute
Toronto, ON
spitis@cs.toronto.edu

Abstract

Reinforcement learning (RL) agents have traditionally been tasked with maximizing the value function of a Markov decision process (MDP), either in continuous settings, with fixed discount factor $\gamma < 1$, or in episodic settings, with $\gamma = 1$. While this has proven effective for specific tasks with well-defined objectives (e.g., games), it has never been established that fixed discounting is suitable for general purpose use (e.g., as a model of human preferences). This paper characterizes rationality in sequential decision making using a set of seven axioms and arrives at a form of discounting that generalizes traditional fixed discounting. In particular, our framework admits a state-action dependent "discount" factor that is not constrained to be less than 1, so long as there is eventual long run discounting. Although this broadens the range of possible preference structures in continuous settings, we show that there exists a unique "optimizing MDP" with fixed $\gamma < 1$ whose optimal value function matches the true utility of the optimal policy, and we quantify the difference between value and utility for suboptimal policies. Our work can be seen as providing a normative justification for (a slight generalization of) Martha White's RL task formalism (2017) and other recent departures from the traditional RL, and is relevant to task specification in RL, inverse RL and preference-based RL.

1 Introduction

The field of reinforcement learning (RL) studies agents that learn to interact "optimally" with a partially known environment (Sutton and Barto 2018). The common environmental model, underlying almost all analysis in the field, is the Markov decision process (MDP), with optimal behavior defined in terms of the expected sum of future rewards, discounted at a constant rate per time step. The motivation for this approach has historically been a practical one; e.g., to enable us to make "precise theoretical statements" (Sutton and Barto 2018) or because rewards provide "the most succinct, robust, and transferable definition of a task" (Abbeel and Ng 2004). While this has proven effective for specific tasks with well-defined objectives (e.g., games), modern RL is being applied to increasingly general tasks, and it has never been established that the MDP is a suitable model for general purpose use. Recently, researchers have adopted

more flexible value functions that have state, state-action, or transition dependent discount factors (Sutton et al. 2011; Silver et al. 2017; White 2017). Practical considerations aside, which approach, if any, is most sensible from a *normative* perspective?

We address this question by presenting an axiomatic framework to characterize rationality in sequential decision making. After a motivating example in Section 2 and a discussion of related work in Section 3, including an overview of the relevant work in normative decision theory, we proceed in Section 4 to develop our framework. From a set of seven axioms we derive a form of utility function that admits a state-action dependent discount factor that can be greater than 1, so long as there is long run discounting. Although this broadens the range of possible preference structures (vis-à-vis the fixed discount setting), we show that there exists a unique "optimizing MDP" with fixed $\gamma < 1$ whose optimal value function matches the utility of the optimal policy, and we quantify the difference between value and utility for suboptimal policies. Section 5 discusses the implications of our work for task specification in RL, inverse RL and preference-based RL, and highlights an interesting connection to Dayan's successor representation (1993).

2 Motivation

To motivate our work, we show how the traditional fixed γ value function fails to model a set of preferences. Consider an agent that is to walk in a single direction on the side of a cliff forever (the "Cliff Example"). Along the cliff are three parallel paths—low, middle and high—to which we assign utilities of 100, 0 and 50, respectively (Figure 1 (left) a-c). Let us also suppose the agent has the option to jump down one level at a time from a higher path, but is unable to climb back up. Thus the agent has many options. Four of them are shown in Figure 1 (left) d-g with their assigned utilities.

At a glance, there does not appear to be anything irrational about the utility assignments. But it is impossible to represent this utility structure using a 3-state fixed γ MDP with infinite time horizon. To see this, consider that for the Bellman equation to be satisfied with respect to the optimal policy, paths c-g in Figure 1 (left) imply the reward values shown in Figure 1 (right) when one assumes $\gamma = 0.9$. This implies that the utilities of paths a and b are -30 and -20, respectively. Not only is this incorrect, but the order is re-
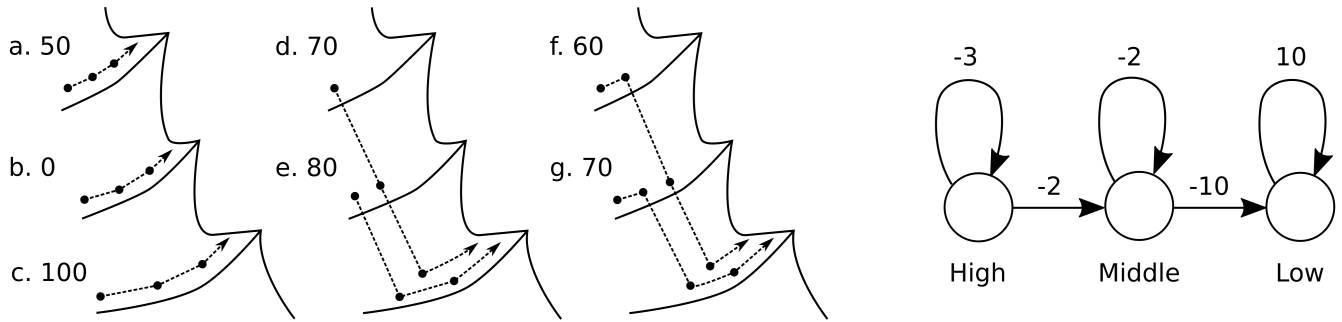
Figure 1: The utilities in the Cliff Example (left) are not compatible with any 3-state MDP; e.g. the only MDP that satisfies the Bellman equation with respect to the optimal policy when $\gamma = 0.9$ (right).

versed! This holds for all $\gamma \in [0, 1)$ (see Supplement for a short proof). It follows that either the utility assignments are irrational, or the MDP structure used is inadequate. This inconsistency motivates a closer look: what characteristics, or axioms, should we demand of "rational" utility assignments? Are they satisfied here, and if so, what does this mean for our default modeling approach (the fixed γ MDP)?

3   Related work

Koopmans (1960) provided the first axiomatic development of discounted additive utility over time. This and several follow-up works are summarized and expanded upon by Koopmans (1972) and Meyer (1976). The applicability of these and other existing discounted additive utility frameworks to general purpose RL is limited in several respects. For instance, as remarked by Sobel (2013), most axiomatic justifications for discounting have assumed deterministic outcomes, and only a handful of analyses address stochasticity (Meyer 1976; Epstein 1983; Sobel 2013). Naively packaging deterministic outcome streams into arbitrary lotteries (as suggested by Meyer (1976), §9.3) is difficult to interpret in case of control (see our commentary in Subsection 4.2 on the resolution of intra-trajectory uncertainty) and entirely hypothetical since the agent never makes such choices—compare our (slightly) less hypothetical "original position" approach in Subsection 4.4.

Existing frameworks have also typically been formulated with a focus on future streams of *consumption* or *income*, which has led to assumptions that do not necessarily apply to sequential decision making. When each unit in a stream is assumed to be scalar, this already rules out "certain types of intertemporal complementarity" (Diamond 1965). This complementarity is also ruled out by mutual independence assumptions (Koopmans 1972) in frameworks admitting vector-valued units (see commentary of Meyer (1976), §9.4, and Frederick, Loewenstein, and O'Donoghue (2002), §3.3). Even seemingly innocuous assumptions like bounded utility over arbitrary outcome streams have important consequences. For instance, Meyer (1976) and Epstein (1983) derive similar utility representations as our Theorem 3, but use the aforementioned assumption to conclude that the discount factor is always less than or equal to 1 (their argument

is discussed in Subsection 4.7). By contrast, our framework admits a state-action dependent "discount" factor that may be greater than 1, *so long as there is eventual long run discounting*—thus, specific, measure zero trajectories may have unbounded reward but the utility of a stochastic process that may produce such trajectories will still exist. This is illustrated in the Supplement.

Frederick, Loewenstein, and O'Donoghue (2002) provide a comprehensive empirical review of the discounted additive utility model as it pertains to human behavior and conclude that it "has little empirical support." While this is consistent with our normative position, it does not invalidate discounting, as humans are known to exhibit regular violations of rationality (Tversky and Kahneman 1986). Such violations are not surprising, but rather a necessary result of bounded rationality (Simon 1972). Our work is in a similar vein to Russell (2014) in that it argues that this boundedness necessitates new research directions.

Several papers in economics—including Kreps (1977) (and sequels), Jaquette (1976) (and prequels), Porteus (1975) (and sequel)—examine sequential decision processes that do not use an additive utility model. Of particular relevance to our work are Von Neumann and Morgenstern (1953), Kreps and Porteus (1978) and Sobel (1975), on which our axiomatic framework is based. To the authors' knowledge, no study, including this one, has ever provided a direct axiomatic justification of the MDP as a model for general rationality. This is not so surprising given that the MDP has historically been applied as a task-specific model. For example, the MDP in Bellman's classic study (1957) arose "in connection with an equipment replacement problem".

The MDP is the predominant model used in RL research. It is commonly assumed that complex preference structures, including arbitrary human preferences, can be well represented by traditional MDP value functions (Abbeel and Ng 2004; Christiano et al. 2017). To the authors' knowledge, the best (indeed, *only*) theoretical justification for this assumption is due to Ng and Russell (2000), discussed in Subsection 4.8. Some RL researchers have proposed generalizations of the traditional value functions that include a transition dependent discount factor (White 2017; Silver et al. 2017). Our work can be understood as a normative justification for (a slight generalization of) these approaches. This, and further

connections to the RL literature, including to inverse RL and preference-based RL, are discussed in Section 5.

4 Theory

4.1 Preliminaries

Our basic environment is a sequential decision process (SDP) with infinite time horizon, formally defined as the tuple (S, A, T, T_0) where S is the state space, A is the action space, $T : S \times A \to \mathcal{L}(S)$ is the transition function mapping state-action pairs to *lotteries* (i.e., probability distributions with finite support) over next states, and $T_0 \in \mathcal{L}(S)$ is the distribution from which initial state s_0 is chosen. To accommodate episodic (terminating) processes, one may simply introduce a terminal state that repeats ad infinitum.

A *trajectory* is an infinite sequence of states and actions, $(s_t, a_t, s_{t+1}, a_{t+1}, \dots)$. For each non-negative integer t, $y_t \in Y_t$ denotes the *history* from time $t = 0$ through the state at time t; e.g., $(s_0, a_0, s_1, a_1, s_2) \in Y_2$. $y_{t[i]}$ indexes the i-th state in y_t; e.g., $(s_0, a_0, s_1, a_1, s_2)_{[2]} = s_2$.

A (stochastic) *stationary policy* $\pi : S \to \mathcal{L}(A)$ (or ω when a second generic is needed) maps states to lotteries over actions. A *non-stationary policy* from time t, $\Pi_t = (\pi_t, \pi_{t+1} \mid y_{t+1}, \pi_{t+2} \mid y_{t+2}, \dots)$ (or just Π, or Ω when a second generic is needed) is a conditional sequence of stationary policies where the choice of π_t may depend on y_t. $\Pi_{t[i]}$ indexes Π's i-th element (e.g., $\Pi_{t[1]} = \pi_{t+1} \mid y_{t+1}$), $\Pi_{[i:]}$ denotes $(\Pi_{[i]}, \Pi_{[i+1]}, \dots)$. $\Pi(s)$ is shorthand for $\Pi_{[0]}(s)$. Note that stationary policy π may be viewed as the non-stationary policy $(\pi, \pi, \dots)$. The space of all non-stationary policies is denoted $\boldsymbol{\Pi}$.

A Markov decision process (MDP) is an SDP together with a tuple (R, γ), where $R : S \times A \to \mathbb{R}$ returns a bounded scalar reward for each transition and $\gamma \in [0, 1)$ is a discount factor. For a given MDP, we define the value function for a policy Π, $V^\Pi : S \to \mathbb{R}$, as $V^\Pi(s_t) = \mathbb{E}[\sum_{t=0}^\infty \gamma^t R(s_t, a_t)]$. Further, we define the Q-function for Π, $Q^\Pi : S \times A \to \mathbb{R}$ as $Q(s, a) = V^{a\Pi}(s)$, where $a\Pi = (\pi, \Pi_{[0]}, \Pi_{[1]}, \dots)$ is the non-stationary policy that uses generic policy π with $\pi(s) = a$ in the first step and follows policy Π thereafter.

4.2 Preferences over prospects

As a starting point in our characterization of rationality, we would like to apply the machinery of expected utility theory (Von Neumann and Morgenstern 1953) to preferences over the set of possible futures, or "prospects", $\mathcal{P}$, and more generally, lotteries on prospects, $\mathcal{L}(\mathcal{P})$. Some care is required in defining a prospect so as to satisfy the necessary axioms. In particular, we would like (strict) preference to be *asymmetric*, meaning that between any two prospects $p, q \in \mathcal{P}$, at most one of $p \succ q$ or $q \succ p$ holds, where $\succ$ denotes strict preference (for convenience, we also define weak preference $p \succeq q$ as <u>not</u> $q \succ p$, and indifference $p \sim q$ as $p \succeq q$ and $q \succeq p$). Without additional assumptions, preferences over trajectories or policies fail to satisfy asymmetry.

Suppose that preferences were defined over bare trajectories, as in "preference-based RL" (Wirth et al. 2017). One problem with this is that intra-trajectory uncertainty has already been resolved; e.g., to evaluate a trajectory that risked,

but did not realize, a visit to a difficult region of the state space, one must make an assumption about how the agent would have behaved in that region. More generally, it is unclear whether trajectories should be compared with respect to action quality (what could have happened) or with respect to outcomes (what did happen).

Suppose instead that preferences were defined over bare policies. This is still problematic because preference would depend on the current state distribution (not just T_0). To see this take an SDP with disconnected state trees S_1 and S_2, where $\pi_1 \succ \pi_2$ in S_1 but $\pi_2 \succ \pi_1$ in S_2.

To avoid such difficulties, we define a prospect as a pair (s, Π), where $s \in S$ and Π is an arbitrary non-stationary policy, which represents the stochastic process that results when the agent starts in state s and behaves according to Π thereafter. For this definition to work we need the following assumption, in absence of which different histories leading up to the initial state s could result in preference reversal:

Assumption (Markov preference, MP). *Preferences over prospects are independent of time t and history y_t.*

One might justify MP in several ways. First, *trivially*, one may restrict the scope of inquiry to time $t = 0$. Second, *theoretically*, it is a consequence of certain preference structures (e.g., the structure associated with the standard optimality criteria for MDPs). Third, *practically*, one can view MP as a constraint on agent design. Typical RL agents, like the DQN agent of Mnih et al. (2013) are restricted in this way. Finally, *constructively*, MP may be achieved by using an "internal" SDP that is derived from the environment SDP, as shown in Subsection 4.3. The DRQN agent of Hausknecht and Stone (2015), which augments the DQN agent with a recurrent connection, implements such a construction.

As compared to trajectories, all uncertainty in prospects is left unresolved, which makes them comparable to the "temporal lotteries" of Kreps and Porteus (1978). As a result, it is admittedly difficult to express *empirical* preference over prospects. Indeed, within the bounds of an SDP, an agent only ever chooses between prospects originating in the same state. In Subsection 4.4 we will apply a "veil of ignorance" argument (Harsanyi 1953; Rawls 2009) to enable a general comparison between prospects for normative purposes.

4.3 Constructive Markov preference (MP)

In this section we derive an "internal" SDP from an environment SDP, (S, A, T, T_0), so that Markov preference is satisfied with respect to the internal SDP. First, define a *historical prospect* to be a pair (y_t, Π_t) where $y_t \in \cup_n Y_n$ and Π_t is the policy to be followed when starting in the final state of y_t. One should have little trouble accepting asymmetry with respect to preferences over historical prospects.

Next, define an equivalence relation on the set of all histories as follows: $y_i, y_j \in \cup_n Y_n$ are equivalent if the last states are equal, $y_{i[i]} = y_{j[j]}$, and, for all Π^1 and Π^2, $(y_i, \Pi_i^1) \succ (y_i, \Pi_i^2) \iff (y_j, \Pi_j^1) \succ (y_j, \Pi_j^2)$. Let S' be the set of equivalence classes with generic element $s' = \{y_t\}$. Note that S' may be uncountable even if S is finite.

It follows from our construction that preferences over the prospects (s', Π), where $s' \in S'$ and Π is an arbitrary

non-stationary policy, are independent of time t and history y_t. Therefore, the constructed SDP, (S', A, T', T_0), where $T'(s', a) := T(y_{t[t]}, a)$, satisfies Markov preference.

4.4 Cardinal utility over prospects

We imagine a hypothetical state from which an agent chooses between lotteries (i.e., probability distributions with finite support) of prospects, denoted by $\mathcal{L}(\mathcal{P})$. We might think of this choice being made from behind a "veil of ignorance" (Rawls 2009). In our case, the veil is not entirely hypothetical: the agent's designer, from whom preferences are derived, is faced with a similar choice problem. The designer can instantiate the agent's internal state, which encapsulates the agent's subjective belief about the history, to be anything (albeit, the external state is outside the designer's control). Now there may be some uncertainty about which internal hardware states correspond to which histories, so that the designer is, in a sense, behind a veil of ignorance.

We assume that strict preference ($\succ$) with respect to arbitrary prospect lotteries $\tilde{p}, \tilde{q}, \tilde{r} \in \mathcal{L}(\mathcal{P})$ satisfies:

Axiom 1 (Asymmetry). *If $\tilde{p} \succ \tilde{q}$, then not $\tilde{q} \succ \tilde{p}$.*

Axiom 2 (Negative transitivity). *If not $\tilde{p} \succ \tilde{q}$, and not $\tilde{q} \succ \tilde{r}$, then not $\tilde{p} \succ \tilde{r}$.*

Axiom 3 (Independence). *If $\alpha \in (0, 1]$ and $\tilde{p} \succ \tilde{q}$, then $\alpha\tilde{p} + (1 - \alpha)\tilde{r} \succ \alpha\tilde{q} + (1 - \alpha)\tilde{r}$.*

Axiom 4 (Continuity). *If $\tilde{p} \succ \tilde{q} \succ \tilde{r}$, then $\exists \alpha, \beta \in (0, 1)$ such that $\alpha\tilde{p} + (1 - \alpha)\tilde{r} \succ \tilde{q} \succ \beta\tilde{p} + (1 - \beta)\tilde{r}$.*

The notation $\alpha\tilde{p} + (1 - \alpha)\tilde{q}$ above represents a *mixture* of prospect lotteries, which is itself a prospect lottery with a $\alpha\%$ chance of lottery $\tilde{p}$ and a $(1 - \alpha)\%$ chance of lottery $\tilde{q}$. Note that prospects are (degenerate) prospect lotteries.

We now apply the VNM Expected Utility Theorem (Von Neumann and Morgenstern 1953) to obtain a cardinal utility function over prospects. We use a version of the theorem from Kreps (1988) (Theorem 5.15), restated here without proof and with minor contextual modifications:

Theorem 1 (Expected utility theorem). *The binary relation $\succ$ defined on the set $\mathcal{L}(\mathcal{P})$ satisfies Axioms 1-4 if and only if there exists a function $U : P \to \mathbb{R}$ such that, $\forall \tilde{p}, \tilde{q} \in \mathcal{L}(\mathcal{P})$:*

$$\tilde{p} \succ \tilde{q} \iff \sum_z \tilde{p}(z)U(z) > \sum_z \tilde{q}(z)U(z)$$

where the two sums in the display are over all $z \in \mathcal{P}$ in the respective supports of $\tilde{p}$ and $\tilde{q}$. Moreover, another function U' gives this representation if and only if U' is a positive affine transformation of U.

Applying the theorem produces the cardinal utility function $U : \mathcal{P} \to \mathbb{R}$, as desired. We overload notation and define $U : \mathcal{L}(\mathcal{P}) \to \mathbb{R}$ as $U(\tilde{p}) = \sum_z \tilde{p}(z)U(z)$.

We further define $U^\Pi : S \to \mathbb{R}$ as $U^\Pi(s) = U((s, \Pi)) = U(s, \Pi)$ for policy Π. Similarly, we overload U^Π to define $U^\Pi : S \times A \to \mathbb{R}$ as $U^\Pi(s, a) = U(s, a\Pi)$.

Finally, given a lottery over states, $\tilde{s} \in \mathcal{L}(S)$, we denote the prospect lottery given by $\tilde{s}$ with fixed Π as $(\tilde{s}, \Pi)$, and define preference ordering $\succ_{\tilde{s}}$ over policies induced by $\tilde{s}$ according to the rule $\Pi \succ_{\tilde{s}} \Omega$ if and only if $U(\tilde{s}, \Pi) = \sum_z \tilde{s}(z)U^\Pi(z) > \sum_z \tilde{s}(z)U^\Omega(z) = U(\tilde{s}, \Omega)$. We further define the shorthand $U^{\tilde{s}} : \Pi \to \mathbb{R}$ as $U^{\tilde{s}}(\Pi) = U(\tilde{s}, \Pi)$.

4.5 Rational planning

Axioms 1-4 are classics of decision theory and have been debated extensively over the years (see footnote 1 of Machina (1989) for some initial references). Other than asymmetry, which is natural given MP, we do not wish to argue for their merits. Rather, it is prudent to strengthen this foundation for the sequential context. Without strong normative assumptions about the structure of preferences, such as those implied by the standard optimality criteria of MDPs, one could infer little about future behavior from past experiences and learning would be impossible; see, e.g., the "No Free Lunch" theorem for inverse reinforcement learning (Armstrong and Mindermann 2018). The axioms and results in this subsection provide what we argue is a minimal characterization of rational planning. We begin with:

Axiom 5 (Irrelevance of unrealizable actions). *If the stochastic processes generated by following policies Π and Ω from initial state s are identical, then the agent is indifferent between prospects (s, Π) and (s, Ω).*

A consequence of this axiom is that $U(s, a\Pi)$ is well-defined, since $U(s, a\Pi)$ is constant for all first-step policies π with $\pi(s) = a$.

Assuming non-trivial MP, an agent will choose between prospects of the form (s_2, Π_2) at time $t = 2$, where $s_2 \sim T(s_1, a_1)$ and Π_2 is any non-stationary policy. The agent has preferences over *plans* for this choice at $t = 1$, which can be ascertained by restricting the $t = 1$ choice set to the set X of prospects of the form $(s_1, a_1\Pi)$. From a rational agent, we should demand that the restriction of U to X, $U|_X : \Pi \to \mathbb{R}$, represent the same preference ordering over policies as $U^{T(s_1, a_1)}$. We thus assume:

Axiom 6 (Dynamic consistency). *$(s, a\Pi) \succ (s, a\Omega)$ if and only if $(T(s, a), \Pi) \succ (T(s, a), \Omega)$.*

Dynamic consistency is based on the similar axioms of Sobel (1975) and Kreps and Porteus (1978), reflects the general notion of dynamic consistency discussed by Machina (1989), and might be compared to Koopmans' classic "stationarity" axiom (1960). Note that we demand consistency only before and after an action has been chosen, but not before and after environmental uncertainty is resolved. That is, $(T(s, a), \Pi) \succ (T(s, a), \Omega)$ *does not imply* that for all z in the support of $T(s, a)$, $(z, \Pi) \succ (z, \Omega)$.

Finally, we adopt a mild version of "impatience", comparable to Sobel's countable transitivity axiom (1975), but stated here in terms of utility (for clarity). Impatience can be understood as the desire to make the finite-term outcomes meaningful in light of an infinite time horizon (Koopmans 1960). In the statement below, $\Pi_n\Omega$ is the policy that follows Π for the first n steps and Ω thereafter.

Axiom 7 (Horizon continuity). *The sequence $\{U(s, \Pi_n\Omega)\}$ converges with limit $U(s, \Pi)$.*

One might use this basic setup to prove a number of facts about rational behavior in SDPs; e.g., Sobel (1975) uses a similar axiomatic structure to prove a policy improvement theorem alongside the next result. We restrict our analysis to three immediately relevant results (but see our comment on convergence theorems in Subsection 5.1). The first justifies

our later focus on stationary policies. An *optimal policy* is a policy Π for which $(s, \Pi) \succeq (s, \Omega)$ for all s and Ω.

Lemma 1. *If Π is an optimal policy, so too is the policy $\Pi_1 \Pi = (\Pi_{[0]}, \Pi_{[0]}, \Pi_{[1]}, \Pi_{[2]}, \dots)$ formed by delaying Π one step in order to act according to $\Pi_{[0]}$ for that step.*

Proof. Consider any state s. For each state z in the support of $\Pi(s)$, $(z, \Pi) \succeq (z, \Pi_{[1:]})$ (because Π is optimal) so that $(T(s, \Pi(s)), \Pi) \succeq (T(s, \Pi(s)), \Pi_{[1:]})$. By dynamic consistency, this implies $(s, \Pi_1 \Pi) \succeq (s, \Pi)$. $\square$

Theorem 2. *If there exists an optimal policy Π, there exists an optimal stationary policy π.*

Proof. Put $\pi = \Pi_{[0]}$. By repeated application of Lemma 1 we have $\pi_n \Pi \succeq \Pi$ for all $n > 0$. It follows from horizon continuity that $\pi \succeq \Pi$. $\square$

The next result is somewhat similar Lemma 4 of Kreps and Porteus (1978), but with a recursive, affine formulation. Note that the proof does not use horizon continuity.

Theorem 3 (Bellman relation for SDPs). *There exist $\mathcal{R} : S \times A \to \mathbb{R}$ and $\Gamma : S \times A \to \mathbb{R}^+$ such that for all s, a, Π,*

$$U(s, a\Pi) = \mathcal{R}(s, a) + \Gamma(s, a)\mathbb{E}_{s' \sim T(s,a)}[U(s', \Pi)].$$

Proof. Fix s and a. Dynamic consistency ensures that $U^{T(s,a)} = \mathbb{E}_{s' \sim T(s,a)}[U(s', \Pi)]$ represents the same preferences as the restriction of U to the space X of prospects of the form $(s, a\Pi)$. Preferences are cardinal because Π may be stochastic, so that prospects in X are prospect lotteries (i.e., $X \subset \mathcal{L}(\mathcal{P})$), X is closed under mixtures (i.e., $\tilde{p}, \tilde{q} \in X \implies \alpha\tilde{p} + (1 - \alpha)\tilde{q} \in X, \forall \alpha \in [0, 1]$), and Axioms 1-4 apply to prospect lotteries in X. Therefore, by the restriction of Theorem 1 to X, $U|_X$ and $U^{T(s,a)}$ are related by the positive affine transformation $U|_X = \alpha + \beta U^{T(s,a)}$ for some $\alpha \in \mathbb{R}, \beta \in \mathbb{R}^+$. Define $\mathcal{R}(s, a) = \alpha$, $\Gamma(s, a) = \beta$. Since s and a were arbitrary, the result follows. $\square$

The final result of this subsection will be used to prove the value-utility relation of Subsection 4.9 and is useful for analyzing convergence of RL algorithms (see discussion in Subsection 5.1). It uses two new assumptions. First, we assume $|S| = n$ is finite. This allows us to define vectors $\mathbf{u}^{\Pi}$ and $\mathbf{r}^{\pi}$ so that their ith components equal $U(s_i, \Pi)$ and $\mathcal{R}(s_i, \pi(s_i))$, respectively. Further, we define diagonal matrix $\mathbf{\Gamma}^{\pi}$ whose ith diagonal entry is $\Gamma(s_i, \pi(s_i))$ and transition matrix $\mathbf{T}^{\pi}$ whose ijth entry is $T(s_i, \pi(s_i))(s_j)$. Second, we assume the set $\{\mathbf{u}^{\Pi} : \Pi \in \mathbf{\Pi}\}$ spans $\mathbb{R}^n$. This latter assumption is sensible given that $|\{\mathbf{u}^{\Pi}\}| \gg n$ (one can also hypothetically alter S or A to make it true). We have:

Theorem 4 (Generalized successor representation). *If $|S| = n$ and $\text{span}(\{\mathbf{u}^{\Pi}\}) = \mathbb{R}^n$, $\lim_{n \to \infty} (\mathbf{\Gamma}^{\pi} \mathbf{T}^{\pi})^n = \mathbf{0}$, so that $(\mathbf{I} - \mathbf{\Gamma}^{\pi} \mathbf{T}^{\pi})^{-1} = \mathbf{I} + (\mathbf{\Gamma} \mathbf{T})^1 + (\mathbf{\Gamma} \mathbf{T})^2 + \dots$ is invertible.*

Proof. Using $a\Pi = \pi_n \Omega$ in the vector form of Theorem 3, and expanding the recursion $n - 1$ steps gives:

$$\mathbf{u}^{\pi_n \Omega} = \mathbf{r}^{\pi} + \mathbf{\Gamma}^{\pi} \mathbf{T}^{\pi} \mathbf{u}^{\pi_{n-1} \Omega}$$
$$= \mathbf{r}^{\pi}(\mathbf{I} + (\mathbf{\Gamma} \mathbf{T})^1 + \cdots + (\mathbf{\Gamma} \mathbf{T})^{n-1}) + (\mathbf{\Gamma} \mathbf{T})^n \mathbf{u}^{\Omega}$$

where the superscripts on $\mathbf{\Gamma}^{\pi}$ and $\mathbf{T}^{\pi}$ were dropped for convenience. Similarly, using $a\Pi = \pi$, we have:

$$\mathbf{u}^{\pi} = \mathbf{r}^{\pi}(\mathbf{I} + (\mathbf{\Gamma} \mathbf{T})^1 + \cdots + (\mathbf{\Gamma} \mathbf{T})^{n-1}) + (\mathbf{\Gamma} \mathbf{T})^n \mathbf{u}^{\pi}.$$

Subtracting the second from the first gives $\mathbf{u}^{\pi_n \Omega} - \mathbf{u}^{\pi} = (\mathbf{\Gamma}^{\pi} \mathbf{T}^{\pi})^n (\mathbf{u}^{\Omega} - \mathbf{u}^{\pi})$. By horizon continuity, the left side goes to $\mathbf{0}$ as $n \to \infty$ for all π and Ω. Since $\{\mathbf{u}^{\Omega}\}$ spans $\mathbb{R}^n$, so too does $\{\mathbf{u}^{\Omega} - \mathbf{u}^{\pi}\}$. It follows that $(\mathbf{\Gamma}^{\pi} \mathbf{T}^{\pi})^n \to \mathbf{0}$, for otherwise, there exists $\mathbf{x} \neq \mathbf{0}$ for which $(\mathbf{\Gamma}^{\pi} \mathbf{T}^{\pi})^n \mathbf{x} \not\to \mathbf{0}$, but $\mathbf{x}$ can be expressed as a linear combination of the spanning set, which leads to a contradiction. That $(\mathbf{I} - \mathbf{\Gamma}^{\pi} \mathbf{T}^{\pi})$ is invertible follows from the well known matrix identity (Kemeny and Snell (1976) §1.11.1). $\square$

4.6 Preference structure of MDPs

The value function of an MDP induces preferences over $\mathcal{L}(\mathcal{P})$ when treated as a utility function on $\mathcal{P}$; i.e., according to the rule $\tilde{p} \succ \tilde{q}$ if $\sum_{(s,\Pi)} \tilde{p}((s, \Pi))V^{\Pi}(s) > \sum_{(s,\Pi)} \tilde{q}((s, \Pi))V^{\Pi}(s)$. We have that:

Theorem 5. *Preferences induced by the value function of an MDP in continuous settings, with fixed $\gamma < 1$, and in episodic settings, with $\gamma = 1$, satisfy Axioms 1-7.*

Proof. Axioms 1-4 follow from the necessity part of Theorem 1. Axioms 5 and 6 are obvious, as is Axiom 7 in the episodic case. Axiom 7 is true in the continuous case because bounded R and $\gamma < 1$ imply that the total contribution of rewards received after n steps goes to 0 as $n \to \infty$. $\square$

Corollary. *Theorem 2 applies to MDPs (this is well known; see, e.g., Theorem 6.2.9(b) of Puterman (2014)).*

If the converse of Theorem 5 were true (i.e., Axioms 1-7 implied fixed γ), this would validate the use of the MDP as the default model for RL. Unfortunately, Theorem 3 is the closest we can get to such a result; it is not hard to see that the utilities in the Cliff Example of Section 2—which are inconsistent with any fixed γ—are consistent with the axioms (this is illustrated in the Supplement).

4.7 $\Gamma(s, a) > 1$

A key feature of our Theorem 3 representation is that, but for Axiom 7, which demands long run discounting, Γ is unconstrained. As noted in Section 3, this differentiates our result from the otherwise identical recursions of Meyer (1976) and Epstein (1983), which bound the discount factor by 1. It also suggests a more general form of value function than the proposals of White (2017), Silver et al. (2017) and Sutton et al. (2011), which adopt a variable γ, but constrain it to $[0, 1]$.

A natural question is whether our axioms can (and *should*) be extended to provide a normative justification for a constrained discount factor—in particular, should we apply the same argument as Meyer and Epstein? Let us first understand their result. Meyer and Epstein both consider trajectories as primitive objects of preference, so that each trajectory has a well-defined utility. Further, they assume that the utilities of trajectories are bounded, that the domain of preference includes arbitrary trajectories, and that at least one trajectory τ has non-zero utility

(which is true so long as the agent is not indifferent between all trajectories). Now suppose there exists (s, a) for which $\Gamma(s, a) > 1$. It must be that $\mathcal{R}(s, a) = 0$, else $U((s, a, s, a, \dots))$ does not exist. But then the sequence of utilities $\{U((s, a, \tau)), U((s, a, s, a, \tau)), \dots\}$ formed by repeatedly prepending (s, a) to τ diverges. This contradicts the assumptions and it follows that $\Gamma(s, a) \leq 1$ for all (s, a).

We cannot apply their argument directly. Trajectories are not primitive objects in our framework, and can only be valued by successively unrolling the Theorem 3 Bellman relation and summing the "discounted" rewards along the trajectory. However, there is no guarantee that this sum converges (see Supplement for an example). But even if we were to insist that, in addition to Axioms 1-7, all plausible trajectories have bounded utilities, the Meyer-Epstein argument would still not apply and $\Gamma(s, a) > 1$ would still be possible. The reason is that *arbitrary* trajectories like $(s, a, s, a, \dots)$ are not necessarily *plausible*—if these trajectories are not in the support of any prospect, it does not make sense to demand that the agent express preference with respect to them, even hypothetically speaking. For instance, it does not make sense for a chess agent to evaluate the impossible sequence of moves $(\mathrm{e4}, \mathrm{e4}, \mathrm{e4}, \dots)$. For this reason, we argue that the Meyer-Epstein approach to sequential rationality is too restrictive, and that we should allow $\Gamma(s, a) > 1$, so long as there is long run discounting in accordance with Axiom 7.

4.8 The optimizing MDP

Although the fixed γ MDP is not expressive enough to model all preference structures considered rational by our framework, it may still serve as a model thereof. In this subsection we prove the existence of an MDP whose optimal value function equals the optimal utility function. As a preliminary step, let us restate without proof a classic result for MDPs (e.g., Theorem 6.2.6 of Puterman (2014)):

Lemma 2 (Bellman optimality). *Given an MDP, policy π is optimal with respect to V if and only if, $\forall s \in S$,*

$$V^\pi(s) = \arg\max_{a \in A}(R(s, a) + \gamma \mathbb{E}[V^\pi(s')]).$$

We also define $U^* = U^{\pi^*}$, $V^* = V^{\pi^*}$ and $Q^* = Q^{\pi^*}$, and recall that U is overloaded for domains S and $S \times A$.

Theorem 6 (Existence of optimizing MDP). *Given an SDP with cardinal utility U over prospects, and optimal stationary policy π^* with respect to U, for all $\gamma \in [0, 1)$, there exists a unique "optimizing MDP" that extends the SDP with discount factor γ and reward function R such that π^* is optimal with respect to V, and has corresponding optimal $V^* = U^*$ and $Q^* = U^*$.*

Proof. Put $R(s, a) = U^*(s, a) - \gamma \mathbb{E}[U^*(s')]$. Then:

$$U^*(s_t) = R(s_t, \pi^*(s_t)) + \gamma \mathbb{E}[U^*(s_{t+1})]$$

$$= \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, \pi^*(s_t))\right] = V^*(s_t)$$

and:

$$U^*(s, a) = R(s, a) + \gamma \mathbb{E}(V^*(s_{t+1})) = Q^*(s, a).$$

Since $V^*(s) = U^*(s, \pi^*(s_t)) \geq U^*(s, a) = Q^*(s, a)$, $\forall a \in A$, it follows from Lemma 2 that π^* is optimal with respect to V. For uniqueness, suppose that $V^* = U^*$ and $Q^* = U^*$ are optimal; then by definition of Q^*, we have $U^*(s, a) = R(s, a) + \gamma \mathbb{E}[U^*(s')]$, so that $R(s, a) = U^*(s, a) - \gamma \mathbb{E}[U^*(s')]$ as above. $\square$

A consequence of Theorem 6 is that the inverse reinforcement learning problem (Ng and Russell 2000) is solvable (in theory); i.e., there exists a reward function that can *explain* any behavior satisfying Axioms 1-7 as being the solution to an MDP. This same consequence follows from Theorem 3 of (Ng and Russell 2000), which characterizes the *set* of reward functions under which some observed behavior is optimal. Our theorem differs from that of Ng & Russell in that it produces a *unique* solution for completely specified preferences, whereas the theorem of Ng & Russell produces a *set* of solutions for partially specified preferences.

4.9 Relating value to utility

Although $V^* = U^*$ in the optimizing MDP, the Cliff Example tells us that, in general, $V^\pi \neq U^\pi$. This is a potentially serious problem because an agent may *never* find the optimal policy. Indeed, humans are the only general purpose agents we know of, and our bounded rationality is well documented (Simon 1972; Tversky and Kahneman 1986). It is possible that general purpose objective preferences are so complex that *all* intelligent agents—present or future; human, artificial or alien—are so bounded to suboptimality.

Nevertheless, a natural hypothesis is that the closer V^π is to V^*—i.e., the better an agent performs in its approximate model of U—the better off it will tend to be. There is a sense in which this is true, at least for finite $|S|$. Recalling the vector notation defined for Theorem 4, defining $\mathbf{v}^\pi$ accordingly, noting that $\mathbf{u}^* = \mathbf{v}^*$, and setting $\boldsymbol{\epsilon}^\pi = \boldsymbol{\Gamma}^\pi - \gamma \mathbf{I}$, we have:

Theorem 7. *In the optimizing MDP (for finite $|S|$):*

$$\mathbf{u}^\pi = \mathbf{u}^* - (\mathbf{I} - \boldsymbol{\Gamma}^\pi \mathbf{T}^\pi)^{-1}(\mathbf{I} - \gamma \mathbf{T}^\pi)(\mathbf{v}^* - \mathbf{v}^\pi)$$

$$= \mathbf{v}^\pi - (\mathbf{I} - \boldsymbol{\Gamma}^\pi \mathbf{T}^\pi)^{-1} \boldsymbol{\epsilon}^\pi \mathbf{T}^\pi (\mathbf{v}^* - \mathbf{v}^\pi).$$

Proof. The Bellman relation provides four equations:

$$\mathbf{u}^\pi = \mathbf{r}^\pi + \boldsymbol{\Gamma}^\pi \mathbf{T}^\pi \mathbf{u}^\pi \quad (1) \qquad\qquad \mathbf{v}^\pi = \mathbf{r}^\pi + \gamma \mathbf{T}^\pi \mathbf{v}^\pi \quad (3)$$

$$\mathbf{u}^{a*} = \mathbf{r}^\pi + \boldsymbol{\Gamma}^\pi \mathbf{T}^\pi \mathbf{u}^* \quad (2) \qquad\qquad \mathbf{u}^{a*} = \mathbf{r}^\pi + \gamma \mathbf{T}^\pi \mathbf{v}^* \quad (4)$$

where $\mathbf{r}^\pi$ is the vector representing $R(s, \pi(s))$. The first equality of the theorem follows by computing (2) - (4) to obtain an equation for $\mathbf{r}^\pi - \mathbf{r}^\pi$, substituting the result into (1) - (3), adding $\mathbf{v}^* - \mathbf{v}^* (= \mathbf{0})$ to one side, and rearranging, given that $(\mathbf{I} - \boldsymbol{\Gamma}^\pi \mathbf{T}^\pi)$ is invertible by Theorem 4.

The second equality follows after expanding $(\mathbf{I} - \boldsymbol{\Gamma}\mathbf{T})^{-1} = \mathbf{I} + \boldsymbol{\Gamma}\mathbf{T} + (\boldsymbol{\Gamma}\mathbf{T})^2 + \dots = \mathbf{I} + (\mathbf{I} + \boldsymbol{\Gamma}\mathbf{T} + \dots)\boldsymbol{\Gamma}\mathbf{T} = \mathbf{I} + (\mathbf{I} - \boldsymbol{\Gamma}\mathbf{T})^{-1}\boldsymbol{\Gamma}\mathbf{T}$, and writing:

$$\mathbf{u}^\pi = \mathbf{u}^* - (\mathbf{I} + (\mathbf{I} - \boldsymbol{\Gamma}\mathbf{T})^{-1}\boldsymbol{\Gamma}\mathbf{T})(\mathbf{I} - \gamma \mathbf{T}^\pi)(\mathbf{v}^* - \mathbf{v}^\pi)$$

$$= \mathbf{v}^\pi - [(\mathbf{I} - \boldsymbol{\Gamma}\mathbf{T})^{-1}\boldsymbol{\Gamma} - \gamma(\mathbf{I} + (\mathbf{I} - \boldsymbol{\Gamma}\mathbf{T})^{-1}\boldsymbol{\Gamma}\mathbf{T})]\mathbf{T}(\mathbf{v}^* - \mathbf{v}^\pi).$$

$$= \mathbf{v}^\pi - [(\mathbf{I} - \boldsymbol{\Gamma}\mathbf{T})^{-1}\boldsymbol{\Gamma} - (\mathbf{I} - \boldsymbol{\Gamma}\mathbf{T})^{-1}\gamma \mathbf{I}]\mathbf{T}(\mathbf{v}^* - \mathbf{v}^\pi).$$

$$= \mathbf{v}^\pi - (\mathbf{I} - \boldsymbol{\Gamma}^\pi \mathbf{T}^\pi)^{-1} \boldsymbol{\epsilon}^\pi \mathbf{T}^\pi (\mathbf{v}^* - \mathbf{v}^\pi)$$

where we again use the identity $\mathbf{I} + (\mathbf{I} - \boldsymbol{\Gamma}\mathbf{T})^{-1}\boldsymbol{\Gamma}\mathbf{T} = (\mathbf{I} - \boldsymbol{\Gamma}\mathbf{T})^{-1}$ in the third line. $\square$

It is worth examining the factors of the product $(\mathbf{I} - \mathbf{\Gamma}^\pi \mathbf{T}^\pi)^{-1} \boldsymbol{\epsilon}^\pi \mathbf{T}^\pi (\mathbf{v}^* - \mathbf{v}^\pi)$. The final factor $(\mathbf{v}^* - \mathbf{v}^\pi)$ has non-negative entries and tells us that the difference between $\mathbf{u}^\pi$ and $\mathbf{v}^\pi$ is a linear function of the agent's regret in the optimizing MDP, which supports our hypothesis that better approximated performance is correlated with better objective performance. The factors $(\mathbf{I} - \mathbf{\Gamma}^\pi \mathbf{T}^\pi)^{-1}$ and $\mathbf{T}^\pi$ also have all non-negative entries (to see this, write the former as an infinite sum). Thus, the sign of the error depends on the entries of $\boldsymbol{\epsilon}^\pi = \mathbf{\Gamma}^\pi - \gamma\mathbf{I}$. Unfortunately, there is no way to guarantee that $\boldsymbol{\epsilon}^\pi$ has all negative entries, which would guarantee that value estimates in optimizing MDP are pessimistic, since $\Gamma(s, a)$ may be greater than 1 at some (s, a).

It is emphasized that Theorem 7 does not address the preference reversal problem observed in the Cliff Example. Even though $\mathbf{u}^* - \mathbf{u}^\pi$ is related to $\mathbf{v}^* - \mathbf{v}^\pi$ by linear transformation, the relationship is not monotonic (entry-wise, or in norm). Preferences over suboptimal prospects implied by the optimizing MDP's value function may very well be reversed.

5 Discussion

5.1 General reinforcement learning (GRL)

Our analysis suggests that, from a normative perspective, the MDP may not be sufficiently expressive. Two approaches for representing non-MDP preference structures come to mind. First, one can adopt the Theorem 3 representation and use an "MDP+Γ" model of the environment that defines both an external reward signal R and an external anticipation function Γ. This is the approach of White's RL task formalism (2017), which, motivated by practical considerations, proposes to specify tasks using a transition-based discount factor. This is also the approach taken by Silver et al. (2017). The theoretical analysis of White (2017) covers convergence of standard algorithms for the case of $\Gamma \leq 1$. Although we defer a rigorous analysis of convergence in our more general case, we note that Theorem 4, which guarantees that $\mathbf{\Gamma}^\pi \mathbf{T}^\pi$ (corresponding to White's $\mathbf{P}_{\pi,\gamma}$) has maximum eigenvalue less than 1, makes our setting easier to analyse (we get White's Lemma 3 for free).

Second, one might avoid the use of a single, monolithic MDP to model global preferences, and instead think about ways of coordinating many specialized MDPs. One way to do so is hierarchically, as in hierarchical RL. The idea is that it is, or at least should be, easier to express accurate preference at the level of goals (i.e., without incurring preference reversal between suboptimal goals) than at the level of fine-grained prospects. So given a set of goals G, an agent can decompose its preferences into two stages: first pick $g \in G$, and then optimize a g-specialized MDP, M_g, to pick fine-grained actions. Kulkarni et al. (2016) provides an example of how this might be done. Note that all M_gs share world dynamics, so that only R (and optionally γ) change with g.

As these approaches allow an agent to represent more general preference structures than the standard RL framework, one might term them *general reinforcement learning* (GRL). This is a bit of a misnomer, however, as the RL problem, broadly framed, includes GRL.

5.2 Inverse and preference-based RL

Our work leads to a generalization of the inverse reinforcement learning (IRL) problem. Rather than asking, "given the observed behavior, what reward signal, if any, is being optimized?" (Russell 1998), our work suggests asking: *given the observed behavior, what utility function (parameterized by R and Γ), if any, is being optimized?* Future work should explore whether this produces empirical improvements.

Preference-based RL (PBRL) (Wirth et al. 2017) side steps reward function design by learning directly from human preferences. Although a wide variety of algorithms fall under this general umbrella, PBRL "still lacks a coherent framework" (Wirth et al. 2017). In particular, the *object* of preference has varied widely between different algorithms. For instance, studies in PBRL have seen preferences expressed over actions given a state (Griffith et al. 2013), over entire policies (Akrour, Schoenauer, and Sebag 2011), over complete trajectories (Wilson, Fern, and Tadepalli 2012), and over partial trajectories (Christiano et al. 2017). Yet, per the discussion in Subsection 4.2, none of these objects satisfy Axiom 1 (asymmetry) without additional assumptions, which are not always explicitly stated or analyzed. Our work is relevant to PBRL in that it proposes a basic object—the prospect (s, Π)—over which (objective) preferences are asymmetric (given MP). Our axiomatic framework falls well short of the coherent foundation sought by Wirth, however, as there is no obvious way in which preferences over abstract prospects can be empirically expressed.

5.3 Generalized successor representation

An interesting connection between the MDP and the MDP-Γ is that, in both, a policy-specific value function can be decomposed into the product of "discounted expected future state occupancies" and rewards. In the finite case, the former factor is represented by the matrix $\mathbf{S} = (\mathbf{I} - \mathbf{\Gamma T})^{-1}$ (see Theorem 4), so that $\mathbf{v} = \mathbf{Sr}$. When $\Gamma = \gamma$, $\mathbf{S}$ is the well known successor representation (SR) (Dayan 1993).

What makes the SR interesting here is that it seems to solve some of the problems posed by the abstract anticipation function Γ. First, Γ is sensitive to the discretization of time (for the same reason annual interest rates are larger than monthly ones). Second, small changes in the average Γ can result in large changes in value (increasing γ from 0.990 to 0.991 increases the value of a constant positive perpetuity by over 10%). By contrast, the entries of $\mathbf{S}$—despite its opaque formula—provide an interpretable and time-scale invariant measure of causality (Pitis 2018). Changes in $\mathbf{S}$ impact $\mathbf{v}$ in proportion to $\mathbf{r}$, and there is even evidence to suggest that the humans utilize the SR to cache multi-step predictions (Momennejad et al. 2017). For these reasons, it may be easier and more effective to elicit and use values of $\mathbf{S}$, representing long run accumulations of the Γ function, than individual values of Γ, whether for GRL, IRL or PBRL.

6 Conclusion

We have provided normative justification for a departure from the traditional MDP setting that allows for a state-action dependent discount factor that can be greater than 1.

Future work should empirically test whether our proposal has practical merit and confirm that classic convergence results extend to our more general setting.

References

Abbeel, P., and Ng, A. Y. 2004. Apprenticeship learning via inverse reinforcement learning. In *The Twenty-first International Conference on Machine learning*, 1. ACM.

Akrour, R.; Schoenauer, M.; and Sebag, M. 2011. Preference-based policy learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 12–27. Springer.

Armstrong, S., and Mindermann, S. 2018. Occams razor is insufficient to infer the preferences of irrational agents. In *Advances in Neural Information Processing Systems*.

Bellman, R. 1957. A markovian decision process. *Journal of Mathematics and Mechanics* 679–684.

Christiano, P. F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*.

Dayan, P. 1993. Improving generalization for temporal difference learning: The successor representation. *Neural Computation* 5(4):613–624.

Diamond, P. A. 1965. The evaluation of infinite utility streams. *Econometrica: Journal of the Econometric Society* 170–177.

Epstein, L. G. 1983. Stationary cardinal utility and optimal growth under uncertainty. *Journal of Economic Theory* 31(1):133–152.

Frederick, S.; Loewenstein, G.; and O'Donoghue, T. 2002. Time discounting and time preference: A critical review. *Journal of economic literature* 40(2):351–401.

Griffith, S.; Subramanian, K.; Scholz, J.; Isbell, C. L.; and Thomaz, A. L. 2013. Policy shaping: Integrating human feedback with reinforcement learning. In *Advances in neural information processing systems*, 2625–2633.

Harsanyi, J. C. 1953. Cardinal utility in welfare economics and in the theory of risk-taking. *Journal of Political Economy*.

Hausknecht, M., and Stone, P. 2015. Deep recurrent q-learning for partially observable MDPs.

Jaquette, S. C. 1976. A utility criterion for markov decision processes. *Management Science* 23(1):43–49.

Kemeny, J. G., and Snell, J. L. 1976. *Finite markov chains*. Springer-Verlag, second edition.

Koopmans, T. C. 1960. Stationary ordinal utility and impatience. *Econometrica: Journal of the Econometric Society* 287–309.

Koopmans, T. C. 1972. Representation of preference orderings over time. *Decision and organization* 1(1).

Kreps, D. M., and Porteus, E. L. 1978. Temporal resolution of uncertainty and dynamic choice theory. *Econometrica: journal of the Econometric Society* 185–200.

Kreps, D. M. 1977. Decision problems with expected utility criteria, I: upper and lower convergent utility. *Mathematics of Operations Research* 2(1):45–53.

Kreps, D. 1988. *Notes on the Theory of Choice*. Westview press.

Kulkarni, T. D.; Narasimhan, K.; Saeedi, A.; and Tenenbaum, J. 2016. Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation. In *Advances in neural information processing systems*, 3675–3683.

Machina, M. J. 1989. Dynamic consistency and non-expected utility models of choice under uncertainty. *Journal of Economic Literature* 27(4):1622–1668.

Meyer, R. F. 1976. Preferences over time. *Decisions with multiple objectives* 473–89.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Graves, A.; Antonoglou, I.; Wierstra, D.; and Riedmiller, M. 2013. Playing atari with deep reinforcement learning. In *NIPS Deep Learning Workshop*.

Momennejad, I.; Russek, E. M.; Cheong, J. H.; Botvinick, M. M.; Daw, N.; and Gershman, S. J. 2017. The successor representation in human reinforcement learning. *Nature Human Behaviour* 1(9):680.

Ng, A. Y., and Russell, S. J. 2000. Algorithms for inverse reinforcement learning. In *The Seventeenth International Conference on Machine Learning*, 663–670.

Pitis, S. 2018. Source traces for temporal difference learning. In *The Thirty-Second AAAI Conference on Artificial Intelligence*.

Porteus, E. L. 1975. On the optimality of structured policies in countable stage decision processes. *Management Science* 22(2):148–157.

Puterman, M. L. 2014. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

Rawls, J. 2009. *A theory of justice: Revised edition*. Harvard university press.

Russell, S. 1998. Learning agents for uncertain environments. In *Proceedings of the eleventh annual conference on Computational learning theory*, 101–103. ACM.

Russell, S. 2014. Rationality and intelligence: A brief update. *Vincent C. Mller (ed.), Fundamental Issues of Artificial Intelligence (Synthese Library)*.

Silver, D.; van Hasselt, H.; Hessel, M.; Schaul, T.; Guez, A.; Harley, T.; Dulac-Arnold, G.; Reichert, D. P.; Rabinowitz, N. C.; Barreto, A.; and Degris, T. 2017. The predictron: End-to-end learning and planning. In *The Thirty-fourth International Conference on Machine Learning*.

Simon, H. A. 1972. Theories of bounded rationality. *Decision and organization* 1(1):161–176.

Sobel, M. J. 1975. Ordinal dynamic programming. *Management science* 21(9):967–975.

Sobel, M. J. 2013. Discounting axioms imply risk neutrality. *Annals of Operations Research* 208(1):417–432.

Sutton, R. S., and Barto, A. G. 2018. *Reinforcement Learning: An Introduction (in preparation)*. MIT Press.

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

Tversky, A., and Kahneman, D. 1986. Rational choice and the framing of decisions. *Journal of business* S251–S278.

Von Neumann, J., and Morgenstern, O. 1953. *Theory of games and economic behavior*. Princeton university press.

White, M. 2017. Unifying task specification in reinforcement learning. In *The Thirty-fourth International Conference on Machine Learning*.

Wilson, A.; Fern, A.; and Tadepalli, P. 2012. A bayesian approach for policy learning from trajectory preference queries. In *Advances in neural information processing systems*, 1133–1141.

Wirth, C.; Akrour, R.; Neumann, G.; and Fürnkranz, J. 2017. A survey of preference-based reinforcement learning methods. *Journal of Machine Learning Research* 18(136):1–46.
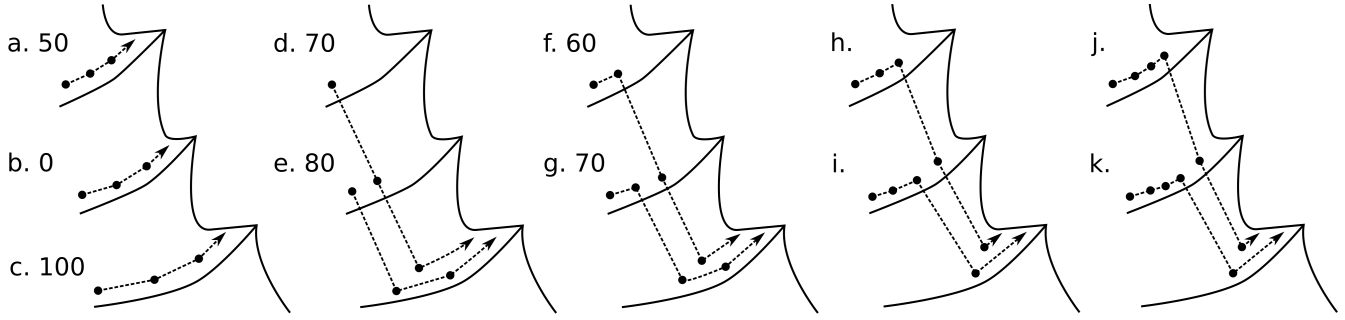
Figure 2: Extension of Figure 1 to illustrate more paths.

Supplement

In this supplement we use the Cliff Example to illustrate concretely three claims made in the paper. The path labels in Figure 2 and the state and state-action labels in Figure 3 are referenced throughout.

Preference reversal occurs for all $\gamma \in [0, 1)$

This claim is made in Section 2. Given the values of paths d, e, f and g in Figure 2, we have:

$$V(g) = R(\text{MM}) + \gamma V(e), \text{ so that } R(\text{MM}) = 70 - \gamma 80,$$

and

$$V(f) = R(\text{HH}) + \gamma V(d), \text{ so that } R(\text{HH}) = 60 - \gamma 70.$$

It follows that $R(\text{MM}) > R(\text{HH})$ for all $\gamma \in [0, 1)$ since $R(\text{MM}) - R(\text{HH}) = 10 - \gamma 10$ is positive.

The Cliff Example satisfies Axioms 1-7

This claim is made in Subsection 4.6. The following $\mathcal{R}$ and Γ provide one consistent Theorem 3 representation (not unique) and imply the following utilities for paths h-k:

(s, a)	$\mathcal{R}(s, a)$	$\Gamma(s, a)$		path	$U(\text{path})$
LL	10	0.9		h	55
ML	-10	0.9		i	61.25
MM	0	0.875		j	52.5
HM	-2	0.9		k	53.59375
HH	25	0.5			

If other paths (not shown) are given utilities consistent with Theorem 3, and we assume the utility of all lotteries (none shown) are computed as expected values, Axioms 1-4 and dynamic consistency hold (by the necessity part of Theorem 1, and by the fact that positive affine transformations represent the same preferences, respectively). Axiom 5 is obviously satisfied. Finally, notice that each step that path d is delayed (paths f, h, j...) brings the utility closer to path a. Similarly, delaying path e (paths g, i, k...) brings the utility closer to that of path b. This is true in general because $\Gamma < 1$, so that the contribution of future rewards to utility goes to zero and horizon continuity is satisfied.
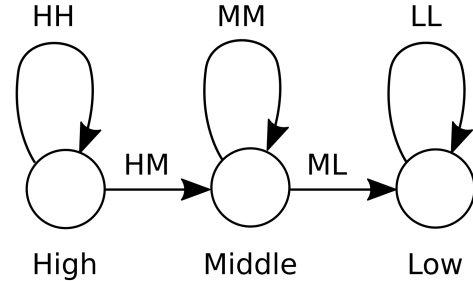

Figure 3: SDP representation of Cliff Example.

$\Gamma(s, a) > 1$ is consistent with Axioms 1-7

This claim is made in Section 3 and Subsection 4.7 as one of the factors differentiating our framework from those of Meyer (1976) and Epstein (1983), in addition to that of White (2017). To see that this is possible, suppose that action HH were stochastic rather than deterministic: although the agent attempts to stay on the high path, half of the time the agent trips and slides down to the middle path. Using the $\mathcal{R}$ and Γ shown in the table to the left, but setting $\Gamma(\text{HH}) = 1.2$, results in an MDP-Γ that has bounded utility over prospects and satisfies Axioms 1-7. Even though the "discounted" reward along path a is unbounded, the agent *almost surely* cannot stay on path a, and the expectation of any policy that tries to stay on path a exists. In particular, an agent that tries to stay on path a, and resorts to the policy that follows path e if it trips (call this prospect X), has utility $U(X) = 25 + 1.2(0.5 \times U(X) + 0.5 \times 80)$, so that $U(X) = 182.5$, which is finite. Since the utilities of all prospects exist, the fact that this revised MDP-Γ is consistent with axioms 1-7 follows by an argument similar to that used for the original Cliff Example.