Challenging the MDP Status Quo: An Axiomatic Approach to Rationality for Reinforcement Learning Agents

Silviu Pitis [1 2]

Abstract

The status quo for objective function design in reinforcement learning (RL) is to use the value function of a Markov decision process (MDP). But this prescribes RL agents with an additive utility function, which is not obviously suitable for general purpose use. This paper presents a minimal axiomatic framework for rationality in sequential decision making and shows that the implied cardinal utility function is of a more general form than the discounted additive utility function of an MDP. In particular, our framework allows for a state-action dependent "discount" factor that is not constrained to be less than 1 (so long as there is eventual long run discounting). We show that although the MDP is not sufficiently expressive to model all rational preference structures (as defined by our framework), there exists a unique "optimizing MDP" whose optimal value function matches the utility of the optimal policy. The relation between the value and utility of suboptimal policies is quantified and the implications for objective function design in RL are discussed.

1. Introduction

Should we seek to use reinforcement learning (RL) as a foundation for developing general purpose agents, it behooves us to ensure that its framework is theoretically able to represent, or at least approximate, any "rational" set of preferences. The concept of rationality is not uncontroversial, and one could simply define rational preferences to be those representable by the value function of a Markov decision process (MDP). Indeed, this is the predominant approach for objective function design in RL (Sutton & Barto, 2018). Though this "MDP status quo" may be suitable for specific tasks with well-defined objectives (e.g., many games), it is unclear that it is suitable in general. For example, Christiano et al. (2017) propose to model *human* preferences using an MDP—does this make sense?

This paper challenges the MDP status quo by presenting a minimal axiomatic framework for rationality in sequential decision making and showing that the implied cardinal utility function has a more general form than the value function of an MDP. After a motivating example in Section 2 and a discussion of related work in Section 3, we proceed in Section 4 to develop our framework and analyse the value function as an approximate model of utility. Our analysis suggests that a single MDP cannot model certain preference structures, and that we should be thinking about ways to coordinate the use of multiple MDPs. This and other implications of our work are discussed in Section 5.

2. Motivation

To motivate our work, we show how additive utility may fail using a simple "Cliff Example". Consider an agent that is to walk in a single direction on the side of a cliff forever. Along the cliff are three parallel paths. The agent assigns cardinal[1] utilities of 100, 0 and 50 to walking along the low, middle and high paths, 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 associated utilities.

At a glance, there does not appear to be anything irrational about the utility assignments. But try as we might, we cannot force this utility structure into a 3-state discounted 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 a discount factor γ of 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 reversed! This is true for all $\gamma \in [0, 1)$ (a simple proof of this fact is found in the Supplement). It follows that either the utility assignments

[1]University of Toronto [2]Vector Institute. Correspondence to: Silviu Pitis <spitis@cs.toronto.edu>.

Accepted at the 1st Workshop on Goal Specifications for Reinforcement Learning, FAIM 2018, Stockholm, Sweden, 2018.

[1]Cardinal (as opposed to ordinal) means that relative differences in utility values indicate degree of preference.

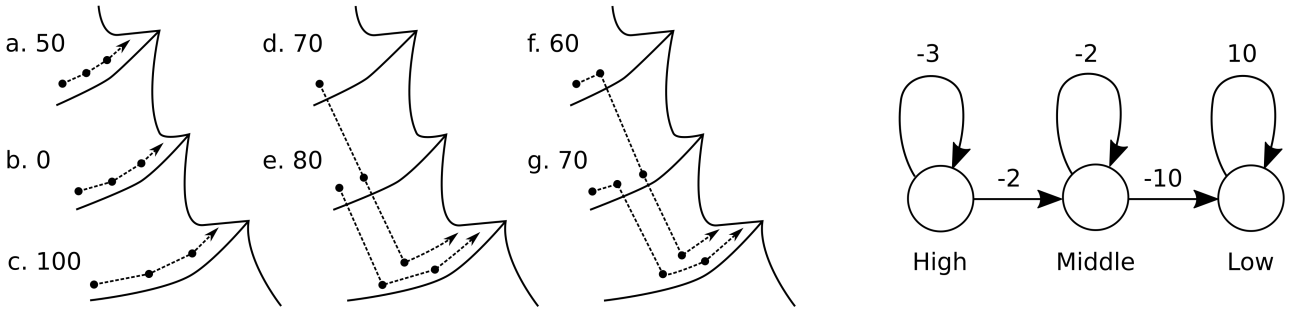

Figure 1. Utilities over paths in Cliff Example (left); MDP implied by optimal policy when $\gamma = 0.9$ (right).

are irrational, or the MDP structure used is inadequate.

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 type of 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 et al. (2002), §3.3). Even seemingly innocuous assumptions like bounded utility over deterministic outcome streams have important consequences. For instance, Meyer (1976) and Epstein (1983) derive similar utility representations as our Theorem 3, but use their assumption that utility over outcome streams is bounded to conclude that the discount factor is always less than or equal to 1 (see the discussions following equation 9.22 in the former and equation 17 in the latter). By contrast, our framework admits the possibility of a state-action dependent "discount" factor 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 might produce such trajectories will still exist. This is illustrated in the Supplement.

Frederick et al. (2002) provides a comprehensive empirical review of the discounted additive utility model as it pertains to human behavior and concludes 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 & 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 & Morgenstern (1953), Kreps & 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 traditionally been viewed as a task-specific model. For example, the MDP in Bellman's classic study (1957) arose "in connection with an equipment replacement problem".

The reinforcement learning problem, framed generally, involves an agent learning to interact "optimally" with a partially known environment (Sutton & Barto, 2018). The most common model for this problem is the MDP, and it is often assumed that human preferences can be well represented by the reward and value functions of an MDP (Abbeel & Ng, 2004; Christiano et al., 2017). To the authors' knowledge, the best (indeed, *only*) theoretical justification for this assumption is found in Ng & Russell (2000). Its connec-

tion to our work is discussed in Subsection 4.7. Further connections to the RL literature 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—the set of which is denoted $\mathcal{L}(S)$ and has generic element $\tilde{s}$—and $T_0 \in \mathcal{L}(S)$ is the distribution from which initial state s_0 is chosen.

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. $\Pi_{[i]}$ 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 $\mathbf{\Pi}$.

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. Note that $V^{a\Pi}$ is well-defined regardless of $\pi(z)$ for $z \neq s$.

4.2. Preferences over prospects

We would like to apply the machinery of expected utility theory (Von Neumann & 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, bare trajectories and policies both fail to satisfy asymmetry.

Suppose that preferences were defined over bare trajectories (or segments thereof), 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 & 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 & 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, however, we will apply a "veil of ignorance" argument (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 $y_t \in \cup_t Y_t$ 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: $y_i, y_j \in \cup_t Y_t$ 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 $\mathcal{L}(\mathcal{P})$. Let us call this state the "original position" and say that the choice is being made from behind the "veil of ignorance", after Rawls (2009). In our case, this is not entirely hypothetical: the agent's designer, from whom preferences are derived, is faced with a similar choice problem. The designer can theoretically instantiate the agent's internal state, which encapsulates the agent's subjective belief about the history, to be anything (albeit, the external world state is generally 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 the following four *axioms of static rationality*:

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 $M_\alpha(\tilde{p}, \tilde{r}) \succ M_\alpha(\tilde{q}, \tilde{r})$.*

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

The notation $M_\alpha(x, y)$ represents the *mixture* $\alpha x + (1-\alpha)y$. Mixtures of prospect lotteries are themselves prospect lotteries, where $M_\alpha(\tilde{p}, \tilde{q})$ represents an $\alpha\%$ chance of lottery $\tilde{p}$ and a $(1-\alpha)\%$ chance of lottery $\tilde{q}$. Note that every prospect is a (degenerate) prospect lottery.

We may now apply Theorem 5.15 of Kreps (1988), 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 $U(\tilde{p}) = \sum_z \tilde{p}(z)U(z)$. This gives us the corollary (cf. equation 5.13 of Kreps (1988)):

Corollary (Mixture of prospects)**.** *For $\tilde{p}, \tilde{q} \in \mathcal{L}(\mathcal{P})$ and $\alpha \in [0, 1]$, $U(M_\alpha(\tilde{p}, \tilde{q})) = M_\alpha(U(\tilde{p}), U(\tilde{q}))$.*

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 $U^\Pi(s, a) = U(s, a\Pi)$, 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. Implicit in this definition is the irrelevance of unrealizable actions axiom proposed below.

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 $U^{\tilde{s}} : \mathbf{\Pi} \to \mathbb{R}$ as $U^{\tilde{s}}(\Pi) = U(\tilde{s}, \Pi)$.

4.5. Rational planning

The axioms of static rationality 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. The main point of this paper—that the MDP framework may not be sufficiently versatile to represent all rational preference structures—would stand even if the axioms were weakened, for that would broaden the range of rational preference structures; indeed, the MDP optimality criteria implies the axioms (see Theorem 5).

Rather than weaken this foundation, it is prudent to strengthen it 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 can infer very little about future behavior from

past behavior and learning would be impossible; see, e.g., the "No Free Lunch" theorem for inverse reinforcement learning (Armstrong & Mindermann, 2017). The axioms and results in this subsection, provide a minimal characterization of rational planning, thereby permitting a more detailed discussion of "rationality" and allowing us to better understand the role of the MDP framework. 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 the $a\Pi$ notation introduced in the previous subsection is sensible: $U(s, a\Pi)$ is constant for all 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, $U|_X : \mathbf{\Pi} \to \mathbb{R}$, 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 & 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). Impatience can be understood as the desire to make the finite-term behaviors and consequences meaningful in light of an infinite time horizon (Koopmans, 1960). In the statement below, we use $\Pi_n\Omega$ to refer to 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. This would be slightly orthogonal to our main point, and we restrict our analysis to three immediately relevant results. 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 $(s, \Pi_1\Pi) \succeq (s, \Pi)$. □

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 $\pi \succeq \Pi$. □

The next result is somewhat similar Lemma 4 of Kreps & 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., $\tilde{p}, \tilde{q} \in X \implies M_\alpha(\tilde{p}, \tilde{q}) \in X$, $\forall \alpha \in [0, 1]$), and the axioms of static rationality 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. □

The final result of this subsection, which builds upon Theorem 3, will be used to prove the value-utility relation of Subsection 4.8. Suppose $|S|$ is finite, and 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 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 $T(s_i, \pi(s_i))(s_j)$. We have:

Theorem 4 (Generalized successor representation)**.** *For finite* $|S|$*,* $\lim_{n\to\infty}(\mathbf{\Gamma}^\pi\mathbf{T}^\pi)^n = \mathbf{0}$*, so that the matrix* $(\mathbf{I} - \mathbf{\Gamma}^\pi\mathbf{T}^\pi)^{-1} = \mathbf{I} + (\mathbf{\Gamma}\mathbf{T})^1 + (\mathbf{\Gamma}\mathbf{T})^2 + \dots$ *is invertible.*

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

$$\begin{aligned}\mathbf{u}^{\pi_n\omega} &= \mathbf{r}^\pi + \mathbf{\Gamma}^\pi\mathbf{T}^\pi\mathbf{u}^{\pi_{n-1}\omega} \\ &= \mathbf{r}^\pi(\mathbf{I} + (\mathbf{\Gamma}\mathbf{T})^1 + \dots + (\mathbf{\Gamma}\mathbf{T})^{n-1}) + (\mathbf{\Gamma}\mathbf{T})^n\mathbf{u}^\omega\end{aligned}$$

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

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

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. We've made no assumptions about the action space or ω, and can augment any MDP with more actions so that $\mathbf{u}^\omega$ (and hence $(\mathbf{u}^\omega - \mathbf{u}^\pi)$) is arbitrary without affecting $\mathbf{u}^\pi, \mathbf{r}^\pi, \mathbf{\Gamma}^\pi$ and $\mathbf{T}^\pi$. It follows that $(\mathbf{\Gamma}^\pi\mathbf{T}^\pi)^n \to \mathbf{0}$.

That $(\mathbf{I} - \mathbf{\Gamma}^\pi\mathbf{T}^\pi)$ is invertible follows from the well known matrix identity (Kemeny & Snell (1976) §1.11.1). □

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 $\sum_{(s,\Pi)} \tilde{p}((s,\Pi))V^\Pi(s) > \sum_{(s,\Pi)} \tilde{q}((s,\Pi))V^\Pi(s)$ where the two sums are over the respective supports of $\tilde{p}$ and $\tilde{q}$. We have that:

Theorem 5. *Preferences induced by the discounted additive value function of an MDP satisfy Axioms 1-7.*

Proof. Axioms 1-4 follow from the necessity part of Theorem 1. Axioms 5 and 6 are obvious. Axiom 7 is true because bounded R and $\gamma < 1$ imply that the total contribution of rewards received after n steps goes to 0 as $n \to \infty$ so that $V^{\Pi_n\Omega}(s) \to V^\Pi(s)$. □

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

What we would have wanted is for the discounted additive preference structure of MDPs to follow from the axioms. Unfortunately, the representation of Theorem 3 is the closest we can get; it is not hard to see that the utilities in the Cliff Example of Section 2 are also consistent with the axioms (this is illustrated in the Supplement).

It follows that Axioms 1-7 allow for more diverse preference structures than those induced by the value function of an MDP. If these axioms are a good characterization of rational planning, then this is problematic, because it seems inconsistent with the common assumption that arbitrarily complex preferences and behaviors can be represented using the MDP framework (Abbeel & Ng, 2004; Christiano et al., 2017). Nevertheless, the results of the next two subsections show that the MDP remains a useful tool in general settings.

4.7. The optimizing MDP

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:

$$\begin{aligned} U^*(s_t) &= R(s_t, \pi^*(s_t)) + \gamma\mathbb{E}[U^*(s_{t+1})] \\ &= \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, \pi^*(s_t))\right] = V^*(s_t) \end{aligned}$$

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. □

A consequence of Theorem 6 is that the inverse reinforcement learning problem (Ng et al., 2000) is solvable (in theory); i.e., there exists a reward function that can *explain* any "rational" set of behavior as being the solution to an MDP. This same consequence follows from Theorem 3 of (Ng et al., 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.

4.8. 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 & 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 $|S|$) :*

$$\begin{aligned}\mathbf{u}^\pi &= \mathbf{u}^* - (\mathbf{I} - \boldsymbol{\Gamma}^\pi\mathbf{T}^\pi)^{-1}(\mathbf{I} - \gamma\mathbf{T}^\pi)(\mathbf{v}^* - \mathbf{v}^\pi)\\ &= \mathbf{v}^\pi - (\mathbf{I} - \boldsymbol{\Gamma}^\pi\mathbf{T}^\pi)^{-1}\boldsymbol{\epsilon}^\pi\mathbf{T}^\pi(\mathbf{v}^* - \mathbf{v}^\pi).\end{aligned}$$

Proof. The Bellman relation provides four equations:

$$\mathbf{u}^\pi = \mathbf{r}^\pi + \boldsymbol{\Gamma}^\pi\mathbf{T}^\pi\mathbf{u}^\pi \quad (1) \qquad \mathbf{v}^\pi = \boldsymbol{r}^\pi + \gamma\mathbf{T}^\pi\mathbf{v}^\pi \quad (3)$$
$$\mathbf{u}^{a*} = \mathbf{r}^\pi + \boldsymbol{\Gamma}^\pi\mathbf{T}^\pi\mathbf{u}^* \quad (2) \qquad \mathbf{u}^{a*} = \boldsymbol{r}^\pi + \gamma\mathbf{T}^\pi\mathbf{v}^* \quad (4)$$

where $\boldsymbol{r}^\pi$ is the vector representing $R(s, \pi(s))$. The first equality of the theorem follows by computing (2) - (4) to obtain an equation for $\mathbf{r}^\pi - \boldsymbol{r}^\pi$, 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 $(\mathbf{I} - \boldsymbol{\Gamma}\mathbf{T})^{-1} = \mathbf{I} + \boldsymbol{\Gamma}\mathbf{T} + (\boldsymbol{\Gamma}\mathbf{T})^2 + \cdots = \mathbf{I} + (\mathbf{I} + \boldsymbol{\Gamma}\mathbf{T} + \dots)\boldsymbol{\Gamma}\mathbf{T} = \mathbf{I} + (\mathbf{I} - \boldsymbol{\Gamma}\mathbf{T})^{-1}\boldsymbol{\Gamma}\mathbf{T}$, and writing:

$$\begin{aligned}\mathbf{u}^\pi &= \mathbf{u}^* - (\mathbf{I} + (\mathbf{I} - \boldsymbol{\Gamma}\mathbf{T})^{-1}\boldsymbol{\Gamma}\mathbf{T})(\mathbf{I} - \gamma\mathbf{T}^\pi)(\mathbf{v}^* - \mathbf{v}^\pi)\\ &= \mathbf{v}^\pi - [(\mathbf{I} - \boldsymbol{\Gamma}\mathbf{T})^{-1}\boldsymbol{\Gamma} - \gamma(\mathbf{I} + (\mathbf{I} - \boldsymbol{\Gamma}\mathbf{T})^{-1}\boldsymbol{\Gamma}\mathbf{T})]\mathbf{T}(\mathbf{v}^* - \mathbf{v}^\pi).\\ &= \mathbf{v}^\pi - [(\mathbf{I} - \boldsymbol{\Gamma}\mathbf{T})^{-1}\boldsymbol{\Gamma} - (\mathbf{I} - \boldsymbol{\Gamma}\mathbf{T})^{-1}\gamma\mathbf{I}]\mathbf{T}(\mathbf{v}^* - \mathbf{v}^\pi).\\ &= \mathbf{v}^\pi - (\mathbf{I} - \boldsymbol{\Gamma}^\pi\mathbf{T}^\pi)^{-1}\boldsymbol{\epsilon}^\pi\mathbf{T}^\pi(\mathbf{v}^* - \mathbf{v}^\pi)\end{aligned}$$

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. □

It is worth examining the factors of the product $(\mathbf{I} - \boldsymbol{\Gamma}^\pi\mathbf{T}^\pi)^{-1}\boldsymbol{\epsilon}^\pi\mathbf{T}^\pi(\mathbf{v}^* - \mathbf{v}^\pi)$. 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 $(\mathbf{I} - \boldsymbol{\Gamma}^\pi\mathbf{T}^\pi)^{-1}$ and $\mathbf{T}^\pi$ also have all non-negative entries (to see this, write the former as an infinite sum). Thus, if $\boldsymbol{\epsilon}^\pi = \boldsymbol{\Gamma}^\pi - \gamma\mathbf{I}$ has mostly positive entries, $\mathbf{v}^\pi$ will tend to be greater than $\mathbf{u}^\pi$ and the optimizing MDP will over-estimate utility.

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 for some (s, a), so long as there is long run net discounting (see Supplement for an illustration).

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. It may be possible to provide a more detailed bound on regret in terms of Γ, but we have yet to obtain one.

4.9. Further Limitations

Besides the lack of a guaranteed pessimistic MDP and a bound on regret caused by the preference reversal problem, our current analysis is limited in several respects. First, we emphasize that constructive MP (Subsection 4.3) may result in uncountable $|S|$, but parts of our analysis assume finite $|S|$ (e.g., Theorem 7). Along these same lines, our analysis is limited to lotteries and we have yet to verify that it generalizes to distributions with uncountable supports.

Second, given our normative focus, we implicitly assumed the existence of so-called "objective" probabilities in a fully-observable environment. Along these same lines, our framework relies on preference over abstract prospects and does not suggest a way for the empirical expression thereof. Further work is required to accommodate subjective probability, partial observability and empirically expressible preference.

5. Discussion

5.1. General reinforcement learning (GRL)

How can we depart from the MDP status quo and create reinforcement learning agents that are capable of representing non-MDP preference structures? Our analysis suggests two possible courses of action. First, one might 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. It is possible that dynamic programming and online reinforcement learning algorithms can be extended to cover this more general model whilst still offering convergence guarantees; cf. Sobel (1975) and Kreps & Porteus (1979). A problem with this approach is that Γ function design seems to be just as hard, if not harder, than reward function design.

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 hierarchical, as in hierarchical reinforcement learning (HRL). The idea is that it is (or 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 hierarchically decompose its preferences into two stages: first pick $g \in G$, and then

optimize a g-specialized MDP, $\boldsymbol{M}_g$, to pick fine-grained actions. Kulkarni et al. (2016) provides an example of how this might be done. Although the goal-selection mechanism used by Kulkarni et al. is modeled as an MDP, in principle it could be modeled as anything (including an MDP+Γ), thereby freeing the agent from the limitations of the MDP. Note that all $\boldsymbol{M}_g$s share world dynamics, so that only R (and optionally γ) change with g. The value function in each $\boldsymbol{M}_g$ is a *general value function* (GVF) (Sutton et al., 2011) and all GVFs (and therefore $\boldsymbol{M}_g$s) might be modelled simultaneously using a *universal value function approximator* (UVFA) (Schaul et al., 2015).

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 reinforcement learning problem, broadly framed, encompasses GRL.

5.2. Inverse and preference-based RL

Our work leads to a generalization of the inverse reinforcement learning (IRL) question. Rather than asking, "given the observed behavior, what reward signal, if any, is being optimized?" (Russell, 1998), our work suggests asking: *given some preferences, what R and Γ, if any, are consistent with those preferences?* Future work might explore whether equipping an IRL algorithm with the ability to learn variable Γ would produce 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 et al., 2011), over complete trajectories (Wilson et al., 2012), and over partial trajectories (Christiano et al., 2017). Yet, per the discussion in Subsection 4.2, *none* of these objects satisfy the basic requirement of 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.

Another interesting problem has to do with the impact of assuming or inferring an MDP preference structure based on suboptimal preferences. The optimizing MDP of Theorem 6 is designed with respect to V^*, but what if we were to use IRL to design an MDP based on expert behavior that is suboptimal (e.g., because the robotic agent is much more capable than a human expert)? Similarly, in the PBRL setting, human preferences are queried with respect to objects that are almost always suboptimal, since preferences are only queried during learning. Supposing empirical human preferences are better modeled by an MDP-Γ (are they?), how does the assumption that they are induced by an MDP structure (as in, e.g., Christiano et al. (2017)) impact results?

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 $\mathbf{S} = (\mathbf{I} - \boldsymbol{\Gamma}\mathbf{T})^{-1}$ (see Theorem 4), so that $\mathbf{v} = \mathbf{S}\mathbf{r}$. When $\Gamma = \gamma$, $\mathbf{S}$ is the well known successor representation (SR) (Dayan, 1993). The SR is useful for transfer (e.g., when solving $\boldsymbol{M}_g$ for novel g), has been used to measure the distance between policies (Abbeel & Ng, 2004) and is essential to our Theorem 7.

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 a 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

This paper began with the idea that the discounted additive value function of an MDP may not be sufficiently expressive to represent all "rational" sets of preference. This was illustrated concretely with the Cliff Example. In order to analyse this, we drew upon certain classic axioms and results of decision theory to develop an axiomatic framework for rationality in sequential decision making. This framework enabled us to prove several interesting results about cardinal utility over prospects in SDPs and the use of an optimizing MDP as a model thereof. It revealed the limitations of the single MDP, and motivated the use of more expressive GRL architectures, to be explored in future work.

References

Abbeel, Pieter and Ng, Andrew Y. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 1. ACM, 2004.

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

Armstrong, Stuart and Mindermann, Sören. Impossibility of deducing preferences and rationality from human policy. *arXiv preprint arXiv:1712.05812*, 2017.

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

Christiano, Paul F, Leike, Jan, Brown, Tom, Martic, Miljan, Legg, Shane, and Amodei, Dario. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, pp. 4302–4310, 2017.

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

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

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

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

Griffith, Shane, Subramanian, Kaushik, Scholz, Jonathan, Isbell, Charles L, and Thomaz, Andrea L. Policy shaping: Integrating human feedback with reinforcement learning. In *Advances in neural information processing systems*, pp. 2625–2633, 2013.

Hausknecht, Matthew and Stone, Peter. Deep recurrent q-learning for partially observable mdps, 2015.

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

Kemeny, John G and Snell, James Laurie. *Finite markov chains*. Springer-Verlag, second edition, 1976.

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

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

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

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

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

Kreps, David M and Porteus, Evan L. Dynamic choice theory and dynamic programming. *Econometrica: Journal of the Econometric Society*, pp. 91–100, 1979.

Kulkarni, Tejas D, Narasimhan, Karthik, Saeedi, Ardavan, and Tenenbaum, Josh. Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation. In *Advances in neural information processing systems*, pp. 3675–3683, 2016.

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

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

Mnih, Volodymyr, Kavukcuoglu, Koray, Silver, David, Graves, Alex, Antonoglou, Ioannis, Wierstra, Daan, and Riedmiller, Martin. Playing atari with deep reinforcement learning. In *NIPS Deep Learning Workshop*. 2013.

Momennejad, Ida, Russek, Evan M, Cheong, Jin H, Botvinick, Matthew M, Daw, ND, and Gershman, Samuel J. The successor representation in human reinforcement learning. *Nature Human Behaviour*, 1(9): 680, 2017.

Ng, Andrew Y, Russell, Stuart J, et al. Algorithms for inverse reinforcement learning. In *Icml*, pp. 663–670, 2000.

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

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

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

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

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

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

Schaul, Tom, Horgan, Daniel, Gregor, Karol, and Silver, David. Universal value function approximators. In *International Conference on Machine Learning*, pp. 1312–1320, 2015.

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

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

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

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

Sutton, Richard S, Modayil, Joseph, Delp, Michael, Degris, Thomas, Pilarski, Patrick M, White, Adam, and Precup, Doina. 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*, pp. 761–768. International Foundation for Autonomous Agents and Multiagent Systems, 2011.

Tversky, Amos and Kahneman, Daniel. Rational choice and the framing of decisions. *Journal of business*, pp. S251–S278, 1986.

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

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

Wirth, Christian, Akrour, Riad, Neumann, Gerhard, and Fürnkranz, Johannes. A survey of preference-based reinforcement learning methods. *Journal of Machine Learning Research*, 18(136):1–46, 2017.

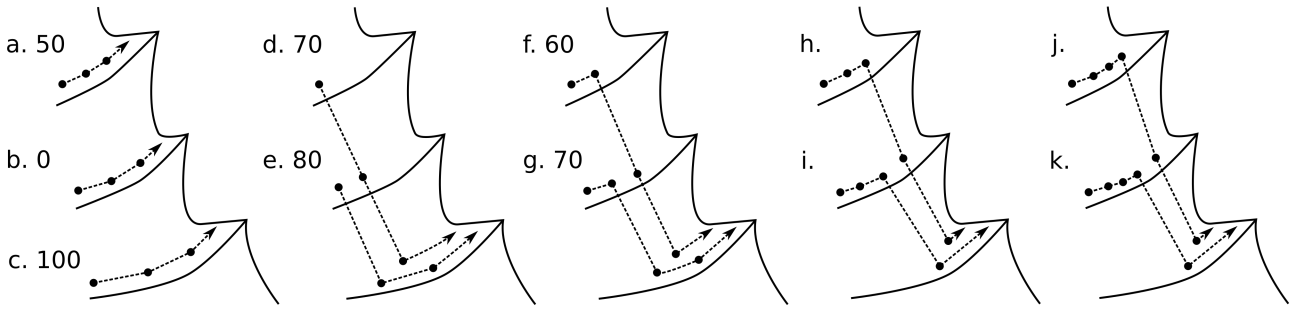

Figure 2. Extension of Figure 1 to illustrate more paths.

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 R and Γ provide one consistent Theorem 3 representation (not unique) and imply the following utilities for paths h-k:

(s, a)	$R(s, a)$	$\Gamma(s, a)$
LL	10	0.9
ML	-10	0.9
MM	0	0.875
HM	-2	0.9
HH	25	0.5

path	U(path)
h	55
i	61.25
j	52.5
k	53.59375

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.

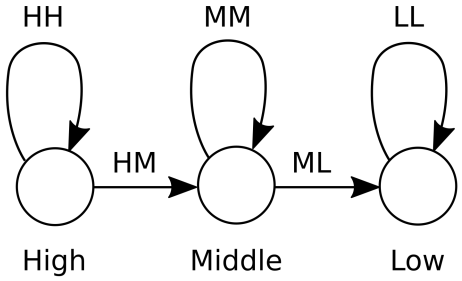

Figure 3. SDP representation of Cliff Example.

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

This claim is made in Section 3 as one of the factors differentiating our framework from those of Meyer (1976) and Epstein (1983). 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 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.

By contrast, Meyer (1976) and Epstein (1983) assume that the utility of all paths, including path a, is bounded, from which it follows that $\Gamma < 1$.