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

Silviu Pitis • University of Toronto • Vector Institute • spitis@cs.toronto.edu • *The 1st Workshop on Goal Specifications for Reinforcement Learning, FAIM 2018*, Stockholm, Sweden, 2018.

Can all "rational" preference structures be represented using an MDP?

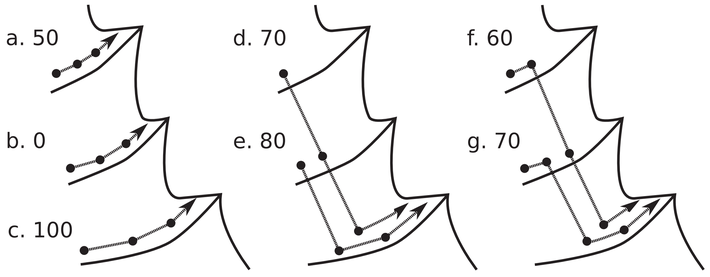
This is an important question, especially as agents become more general purpose, because it is commonly assumed that arbitrary tasks can be modeled as an MDP. E.g., Christiano et al. model *human* preferences as an MDP – does this make sense?

This paper derives a generalization of the MDP reward structure from axioms that has a state-action dependent "discount" factor. Instead of the standard Bellman, the generalized MDP ("MDP-Γ") uses the equation (see Theorem 3):

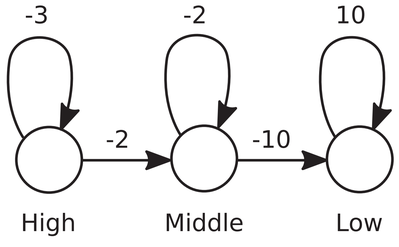
$$Q(s, a) = R(s, a) + \Gamma(s, a)\, \mathbf{E}[Q(s', a')].$$

A motivating example: Walking along a cliff

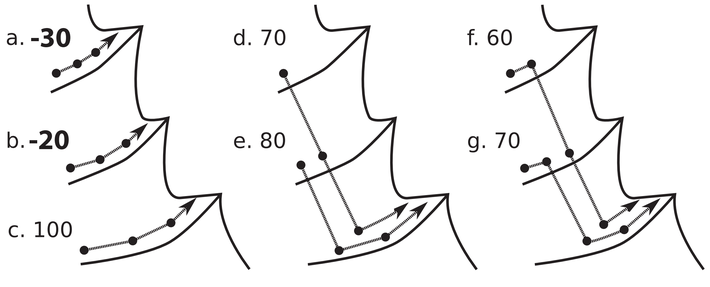
An agent is to walk in a single direction on the side of a cliff forever. The cliff has three paths: high, middle, and low. The agent can jump down, but not up. The agent assigns the following utilities to the paths:



The only discounted 3-state MDP with γ = 0.9 that matches the utilities of paths c-g is:



But this implies the following utilities (the utilities of paths a and b are reversed!):



Either the original utility assignments were irrational, or the MDP structure used is inadequate!

Objects of preference

Preferences are taken over (state, policy) tuples, called ***prospects***. Prospects represent the state-action process going forward, with all uncertainty left unresolved. This is in contrast with preference-based RL (Wirth et al. 2017), which often uses trajectories, policies, states, or actions as the objects of preference. None of these alternatives satisfy the basic requirement of ***asymmetry*** (Axiom 1) .

Strict preference is denoted by $\succ$. Lotteries of over the prospect set P are denoted by $\mathcal{L}(\mathrm{P})$.

Preferences over prospects are assumed to be independent of the state history (they satisfy ***Markov preference***).

Axioms

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

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).*

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

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

Theoretical results

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.

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

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,*

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

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 $(\mathbf{I} - \mathbf{\Gamma}^\pi\mathbf{T}^\pi)^{-1} = \mathbf{I} + (\mathbf{\Gamma T})^1 + (\mathbf{\Gamma T})^2 + \dots$ is invertible.*

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

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 $Q^* = U^*$.*

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

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

Implications and future work

The theoretical analysis suggests that the discounted MDP structure may not be sufficient to model general purpose preference structures. Future work should investigate this empirically, especially for inverse reinforcement learning and preference-based reinforcement learning (does adding a state-dependent discount factor improve results?). In other words, does the state-dependent discount factor allow us to better represent empirical human preferences?