Rethinking the Discount Factor in Reinforcement Learning

A Decision Theoretic Approach

Silviu Pitis • University of Toronto • Vector Institute • spitis@cs.toronto.edu

Can all "rational" preferences be represented using a fixed discount factor MDP?

This is an important question, especially as agents become more general purpose, because it is commonly assumed that arbitrary preferences can be modeled using fixed discount factors. E.g., Christiano et al. (2017) model human preferences as an MDP – does this make sense?

This paper derives a generalization of the MDP reward structure from axioms. The derived reward structure has a state-action dependent "discount" factor that is not constrained to be less than 1. Instead of the standard Bellman equation, the derived model uses the equation:

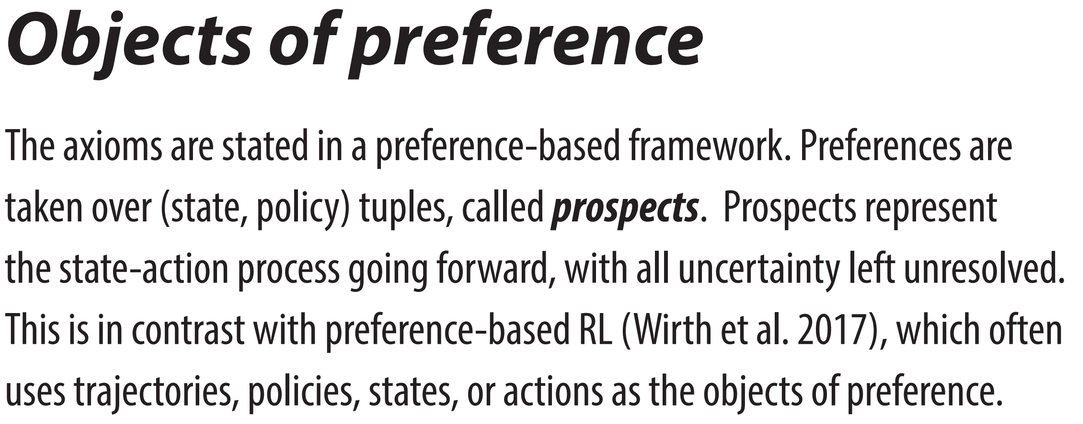
$$Q(s, a) = R(s, a) + \Gamma(s, a)\, E[Q(s', a')].$$

Objects of preference

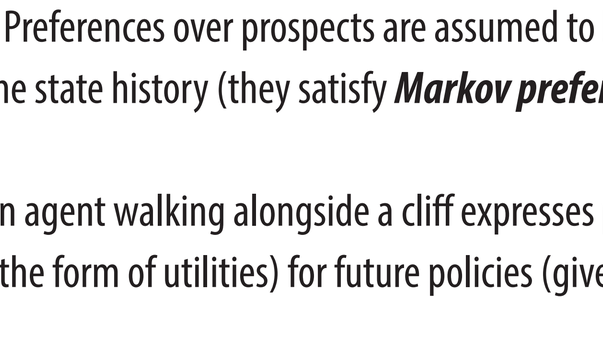
The axioms are stated in a preference-based framework. Preferences are taken over (state, policy) tuples, called ***prospects***. Prospects represent the state-action process going forward, with all uncertainty left unresolved. This is in contrast with preference-based RL (Wirth et al. 2017), which often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

Strict preference is denoted by $\succ$. The set of lotteries of over prospect set P is denoted $\mathcal{L}(\mathrm{P})$. Preferences over prospects are assumed to be independent of the state history (they satisfy ***Markov preference***).

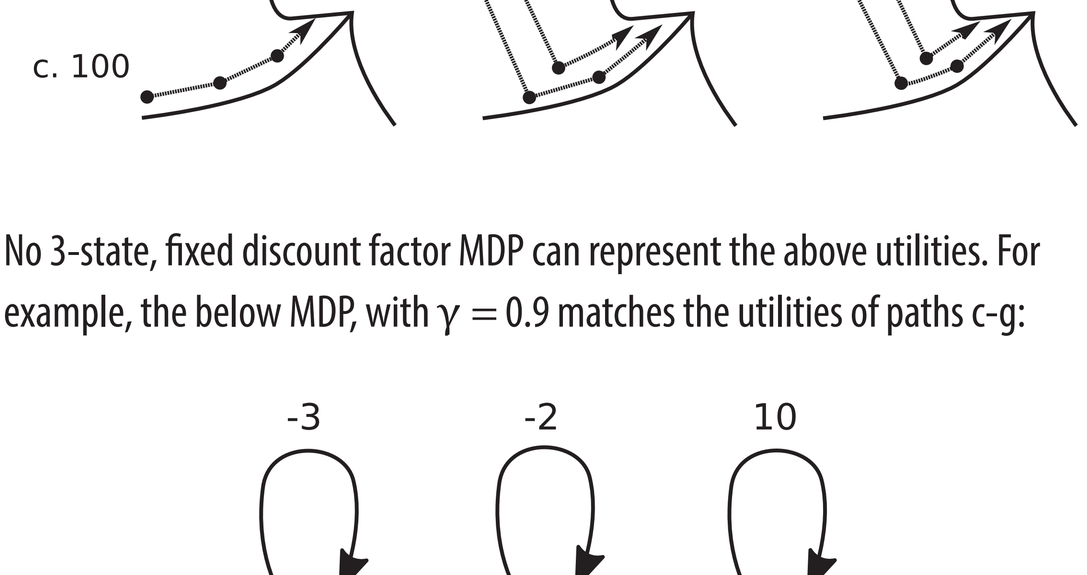
Cliff example: an agent walking alongside a cliff expresses preferences (shown below in the form of utilities) for future policies (given a start state):



No 3-state, fixed discount factor MDP can represent the above utilities. For example, the below MDP, with $\gamma = 0.9$ matches the utilities of paths c-g:



but implies the following utilities (a and b are reversed!):



Axioms

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 $\alpha\tilde{p} + (1-\alpha)\tilde{r} \succ \alpha\tilde{q} + (1-\alpha)\tilde{r}$.*

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

Dynamic Rationality

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 $U(s, \Pi)$.*

Results

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

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.*

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 $|S|$) :*

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

Implications and future work

- fixed-discount MDP may not be sufficient to model general preferences
- should consider more general models (MDP-Γ or composition of MDPs)
- is it possible (practical) to learn Γ, or $(I - \Gamma T)^{-1}$, from data?
- should investigate Γ empirically in inverse RL or preference-based RL
  e.g., does using a state-dependent discount improve IRL results?

VECTOR INSTITUTE

UNIVERSITY OF TORONTO