Rethinking the Discount Factor in Reinforcement Learning:

A Decision Theoretic Approach

Silviu Pitis

University of Toronto, Vector Institute

spitis@cs.toronto.edu https://silviupitis.com

Presentation Outline

- Why we like the fixed discount MDP
- Why the MDP might fail to model preferences
- Characterizing rational preferences using axioms → a state-action dependent discount factor

  → Generalizes the standard MDP!
- Potential applications

Why we like the MDP

- Preferences induced by the (discounted) value function satisfy several notions of consistency
  - E.g., *dynamic consistency*: preferences for actions taken tomorrow do not change come tomorrow

- Fundamental Theorem of Inverse Reinforcement Learning (Ng & Russell 2000)
  - Any arbitrary behavior can be represented as the optimal policy in some MDP

Why the MDP might fail to model preferences

- Human preferences are complex---maybe the agent cannot learn the “optimal policy”
  - Does improving the value function guarantee improvement with respect to modeled human preferences?

- We have good reasons to model preferences with respect to suboptimal policies
  - E.g., in cases where agent ability differs, or when the agent is evaluating policies qua human

- Cliff example in the paper

A universal preference approximator?

- Universal preference approximation is *too* general
- Trivial to show that the MDP cannot model arbitrary preferences
  - e.g., ABAB… > BBBB… > AAAA… (where A & B are fully observed states) cannot be modeled by any MDP

What we really care about is modeling **“rational”** preferences --- can the MDP do that?

"Rational" Preferences

- Rationality is characterized by axioms that we agree preferences should satisfy
  - Whether they do is a different (empirical) question

- Many objects over which preferences can be taken over: actions, states, policies, etc.
  - We will use state, policy pairs: (s, Π) [see paper for why]
  - MDPs induce preferences according to the rule: $(s_1, \Pi_1) > (s_2, \Pi_2)$ iff $V^{\Pi_1}(s_1) > V^{\Pi_2}(s_2)$

- What properties do preferences induced by the typical MDP satisfy?

Von Neumann Axioms...

1) Completeness
   - For all A, B, either we prefer A, prefer B, or are indifferent.
2) Transitivity
3) Independence
   - Roughly, preference between A & B unaffected by C
4) Continuity
   - Roughly, small changes in the probabilities of outcomes → small changes in preference

Von Neumann Axioms… are not enough!

- VNM axioms are stationary / lack a time element
- Can still have ABAB… > BBBB… > AAAA…

Three more axioms (also satisfied by typical MDP)

5) Irrelevance of Unrealizable Actions
   - If two policies differ only when pigs fly → indifference
6) Dynamic Consistency
   - If I plan to do something tomorrow today, I actually do it come tomorrow
7) Impatience
   - Short-term outcomes matter

Axioms are versatile

- E.g., can prove directly from axioms (no value functions / Bellman relation involved):

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

- Sobel (1975) uses a similar axiom set to prove a policy improvement theorem

The main representation theorem

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,

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

A state-action dependent discount factor!

The "discount" can be greater than 1!

- As a result of our impatience axioms, we only require that there be *eventual long-run discounting of future time steps:*

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

- Measure zero trajectories can have undefined (infinite) values.

Other results

- There exists an “Optimizing MDP” whose optimal value (V) and action-value (Q) functions match the state and state-action utilities of the optimal policies.

- Quantify the relationship between the value (according to the Optimizing MDP) and utility of sub-optimal policies

Potential Applications I: Approaches to representing preferences

Approach I: Use both a reward and discount function

- Used by Silver et al.’s Predictron architecture (2017)
- Analyzed theoretically, for discount factors bounded by 1, as part of White’s RL Task Formalism (2017), which proposed the use of a transition-dependent discount

Approach II: Hierarchical RL

- Compose multiple MDPs, or other models, can be used to obtain non-MDP preference structures.
- Maybe it is easier to express consistent preferences at the level of goals.

Potential Applications II: Inverse Reinforcement Learning

- Rather than asking,

  “given the observed behavior, what reward signal is being optimized?” (Russell 1998)

- Ask

  “given the observed behavior, what utility function (parameterized by reward and discount) is being optimized?

The end!

My email: spitis@cs.toronto.edu