Long-Term Values in MDPs, Corecursively

Applied Category Theory, 15-16 March 2018, NIST

Helle Hvid Hansen
Delft University of Technology

Introduction

Joint work with Larry Moss (Indiana U.) and Frank Feys (Delft).
(Paper at Coalgebraic Methods for Computer Science, 2018)

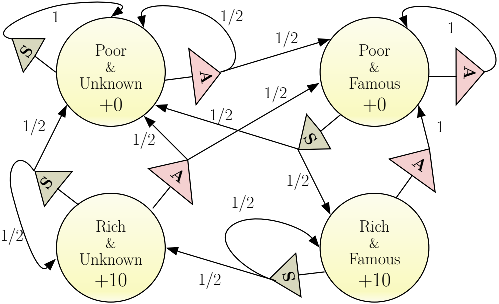
Background & Motivation:

- Coalgebra: categorical theory of systems, observable behaviour, non-wellfounded structures, modal logics.
- Markov Decision Processes (MDPs) are coalgebras.

$\Rightarrow$ Use coalgebraic techniques to reason about MDPs.

Decision-making Under Uncertainty

A startup company has to choose between Saving and Advertising.



- State set S, action set A.
- Probabilistic transitions: $t_a \colon S \to \mathcal{D}S$ for all $a \in A$.
- Reward function: $u \colon S \to \mathbb{R}$.
- MDP is coalgebra $\langle u, t \rangle \colon S \to \mathbb{R} \times (\mathcal{D}S)^A$.

Markov Decision Processes

State-based models of sequential decision-making under uncertainty

- In each state, the agent chooses actions, (but does not have full control over the system), and collects rewards.
- The decision maker wants to find a policy $\sigma \colon S \to A$ that maximizes future rewards
- Applications: maintenance schedules, inventory management, production planning, reinforcement learning, ...
- Classical theory is well-developed (see e.g. Puterman, 2014); uses analytic methods.
- Our motivation: develop high-level, coinductive methods.

Long-Term Value and Optimal Value

Discounting criterion:

Take discounted infinite sum of expected future rewards.

Given an MDP m and a discounting factor $0 \leq \gamma < 1$.

- The long-term value of policy $\sigma \colon S \to A$ in the state s is the discounted infinite sum:

$$V^{\sigma}(s) = \sum_{n=0}^{\infty} \gamma^{n} \cdot r_{n}^{\sigma}(s)$$

  where $r_n^{\sigma}(s)$ = expected reward after n steps, starting from s, following σ.

- The optimal value of m in state s is

$$V^{*}(s) = \max_{\sigma} \{ V^{\sigma}(s) \}$$

Fixpoint Characterisation of V^σ

Given an MDP m and a discounting factor $0 \leq \gamma < 1$.

- The long-term value of policy $\sigma\colon S \to A$ is the unique function $V^\sigma\colon S \to \mathbb{R}$ such that for all $s \in S$:

$$V^\sigma(s) = u(s) + \gamma \sum_{s' \in S} t_{\sigma(s)}(s)(s')V^\sigma(s')$$

- Our observation: This is equivalent to V^σ being coalgebra-to-algebra morphism:

$$\begin{tikzcd} S \arrow[r, "m_\sigma = \langle u, t_\sigma \rangle"] \arrow[d, "V^\sigma"'] & \mathbb{R} \times \mathcal{D}S \arrow[d, "\mathbb{R} \times \mathcal{D}(V^\sigma)"] \\ \mathbb{R} & \mathbb{R} \times \mathcal{D}\mathbb{R} \arrow[l, "\alpha_\gamma \circ (\mathbb{R} \times \mathrm{E})"] \end{tikzcd}$$

fixpt of $\Psi_\sigma(v) = u + \gamma t_\sigma v$

where $t_\sigma(s) = t_{\sigma(s)}(s)$, $\mathrm{E}\colon \mathcal{D}\mathbb{R} \to \mathbb{R}$ computes expected value, and $\alpha_\gamma\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ maps $(x_1, x_2) \mapsto x_1 + \gamma \cdot x_2$.

TU Delft

Fixpoint Characterisation of V^*

Similarly,

- The optimal value of m is the unique function $V^*: S \to \mathbb{R}$ that satisfies the Bellman Equation:

$$V^*(s) = u(s) + \gamma \max_A \sum_{s' \in S} t(s)(a)(s') V^*(s')$$

- Our observation: This is equivalent to V^* being coalgebra-to-algebra morphism:

$$\begin{tikzcd} S \arrow[r, "\langle u,t \rangle"] \arrow[d, "V^*"'] & \mathbb{R} \times (\mathcal{D}S)^A \arrow[d, "\mathbb{R} \times (\mathcal{D}V^*)^A"] \\ \mathbb{R} & \mathbb{R} \times (\mathcal{D}\mathbb{R})^A \arrow[l, "\alpha_\gamma \circ (\mathbb{R} \times \max_A \circ \mathrm{E}^A)"'] \end{tikzcd}$$

$$\Psi^*(V^*) = u + \gamma \max_{a \in A} t_a V^*$$

where $\max_A : \mathbb{R}^A \to \mathbb{R}$.

Universal Property as Definition Principle

- $\alpha\colon F(Y) \to Y$ is a corecursive algebra (for functor F)

$$\begin{tikzcd} X \arrow[r, "\forall f"] \arrow[d, dotted, "\exists ! f^\dagger"'] & F(X) \arrow[d, "F(f^\dagger)"] \\ Y & F(Y) \arrow[l, "\alpha"] \end{tikzcd}$$

- Our algebras are corecursive only for a subclass of $f\colon X \to F(X)$ (unique only among bounded maps).
- We give categorical conditions for how to obtain V^σ and V^* from a universal property (axiomatise properties of bounded maps).

Coinductive Reasoning About Optimal Policies

We say $\sigma \geq \tau$ if $V^{\sigma} \geq V^{\tau}$ (pointwise).
A policy σ is optimal if for all policies τ, $\sigma \geq \tau$.

Some basic facts, see e.g. (Puterman, 2014)

- If σ is optimal, then $V^{\sigma} = V^{*}$.
- Optimal policies need not be unique.
- Stationary (memory-free), deterministic policies suffice.
- Several algorithms for computing optimal policy:
    - policy iteration
    - value iteration
    - linear programming
    - (plus variations)

TUDelft

Policy Improvement

1. Initialise σ_0 to any policy.
2. Compute V^{σ_k} (e.g. by solving system of linear equations).
3. Define σ_{k+1} by

$$\sigma_{k+1}(s) \quad := \quad \mathrm{argmax}_{a \in A} \sum_{s' \in S} t(s,a,s') V^{\sigma_k}(s')$$

4. If $\sigma_{k+1} = \sigma_k$ then stop, else go to step 2.

Why is $\sigma_k \leq \sigma_{k+1}$?

Policy Improvement Lemma:

$$t_\sigma V^\sigma \leq t_\tau V^\sigma \quad \Rightarrow \quad V^\sigma \leq V^\tau$$

TUDelft

Contraction Coinduction Principle

Theorem Let $(M, d, \leq)$ be a non-empty, complete ordered metric space. If $f\colon M \to M$ is contractive and order-preserving, then the fixpoint x^* of f is a least pre-fixpoint (if $f(x) \leq x$, then $x^* \leq x$), and also a greatest post-fixpoint (if $x \leq f(x)$, then $x \leq x^*$).

Proof of policy improvement:
Apply to contractive and order-preserving

$$\Psi_\sigma \colon \mathbb{R}^S \to \mathbb{R}^S \qquad \Psi_\sigma(v) = u + \gamma T_\sigma v.$$

$$t_\tau V^\sigma \geq t_\sigma V^\sigma \quad \Rightarrow \quad \Psi_\tau(V^\sigma) \geq \Psi_\sigma(V^\sigma) = V^\sigma \quad \Rightarrow \quad V^\tau \geq V^\sigma$$

Concluding

We have:

- identified coalgebraic and algebraic structure in the theory of MDPs
- given coinductive proof of policy improvement.

Related work

- Equilibria in infinite games without discounting (Abramsky & Winschel)
- Semantics of equilibria (Pavlovic)
- Open games (Hedges, Ghani,...)