Long-Term Values in Markov Decision Processes, (Co)Algebraically

Frank Feys & Helle Hvid Hansen & Larry Moss

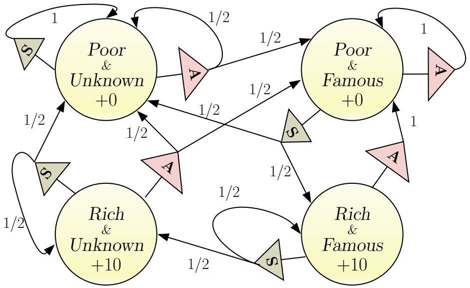
TU Delft TU Delft Indiana University

April 14, 2018

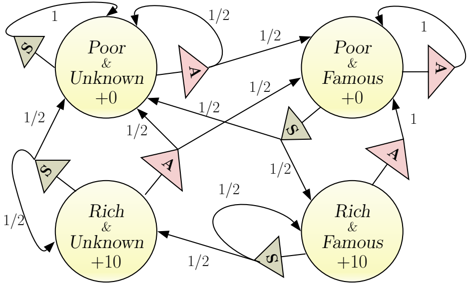
CMCS'18, Thessaloniki, Greece

- Markov decision processes (MDPs) are state-based models of sequential decision-making under uncertainty
- Applications:
  - Planning
  - Reinforcement learning
  - Insurance and finance
  - ...
- We restrict to *finite*, *discrete time-homogeneous*, *infinite-horizon* MDPs, with the *discounting criterion*

Example: A start-up company needs to decide to **A**dvertise or **S**ave money



Example: A start-up company needs to decide to **A**dvertise or **S**ave money

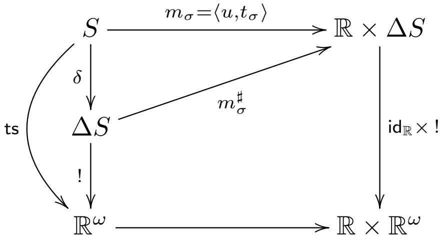


A Markov Decision Process (MDP) is coalgebra $m = \langle u, t \rangle \colon S \to \mathbb{R} \times (\Delta S)^A$ where

- S is finite set of *states* and A is finite set of *actions*
- $u \colon S \to \mathbb{R}$ is *reward* map
- $t \colon S \to (\Delta S)^A$ is *transition* map (where ΔS is set of prob. distr. on S)

A policy σ is a map $\sigma \colon S \to A$

INTRO – Trace Semantics and Long-term Value of a Policy



Given $m = \langle u, t \rangle \colon S \to \mathbb{R} \times (\Delta S)^A$
and policy $\sigma \colon S \to A$

$$t_\sigma \stackrel{\text{def}}{=} t(s)(\sigma(s))$$

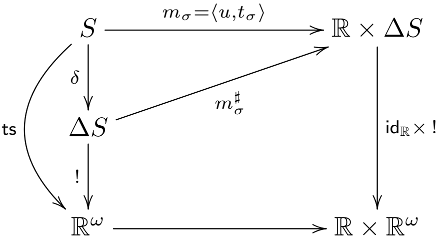
$\rightsquigarrow m_\sigma^\sharp$ by determinization

[Jacobs, Silva, Sokolova]

$\mathsf{ts}(s) = (r_0^\sigma(s), r_1^\sigma(s), r_2^\sigma(s), \cdots)$ is trace semantics of m_σ

($r_n^\sigma(s)$ is expected reward at time n, starting from s)

Intro – Trace Semantics and Long-term Value of a Policy



Given $m = \langle u, t \rangle \colon S \to \mathbb{R} \times (\Delta S)^A$
and policy $\sigma \colon S \to A$

$$t_\sigma \overset{\text{def}}{=} t(s)(\sigma(s))$$

$\rightsquigarrow m_\sigma^\sharp$ by determinization

[Jacobs, Silva, Sokolova]

$\mathsf{ts}(s) = (r_0^\sigma(s)\,, r_1^\sigma(s)\,, r_2^\sigma(s)\,, \cdots)$ is trace semantics of m_σ

($r_n^\sigma(s)$ is expected reward at time n, starting from s)

Discounting criterion: letting $0 \le \gamma < 1$ be a *discount factor*, the long-term value of policy σ is $\mathrm{V}^\sigma \colon S \to \mathbb{R}$:

$$\mathrm{V}^\sigma(s) = r_0^\sigma(s) + \gamma \cdot r_1^\sigma(s) + \cdots + \gamma^n \cdot r_n^\sigma(s) + \cdots$$

The optimal value function $\mathrm{V}^*\colon S \to \mathbb{R}$ of m in state s is given by

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

The optimal value function $\mathrm{V}^*\colon S \to \mathbb{R}$ of m in state s is given by

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

We define:

- $\sigma \geq \sigma'$ if for all s, $\mathrm{V}^{\sigma}(s) \geq \mathrm{V}^{\sigma'}(s)$
- σ is optimal iff $\sigma \geq \tau$ for all τ

The optimal value function $\mathrm{V}^*\colon S \to \mathbb{R}$ of m in state s is given by

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

We define:

- $\sigma \geq \sigma'$ if for all s, $\mathrm{V}^{\sigma}(s) \geq \mathrm{V}^{\sigma'}(s)$
- σ is optimal iff $\sigma \geq \tau$ for all τ

Classical facts (cf. (Puterman, 2014)):

- Optimal policy always exists
- Optimal policies need not be unique
- If σ is optimal, then $\mathrm{V}^{\sigma} = \mathrm{V}^*$
- Stationary (memoryless), deterministic policies suffice

Observations:

- Classic theory uses low-level, analytic methods
- MDPs are coalgebras

Goal: to develop coalgebraic methods for reasoning about LTVs

The main **contributions**:

Part 1: Value function V^{σ} from b-corecursive algebras

Part 2: Coinductive proof of policy improvement theorem

PART 1:

Value Function Arises from a Universal Property

- Following Bellman, the value function has a natural recursive structure:

$$(\mathrm{V}^\sigma \text{ from today}) = \text{reward today} + \gamma \cdot (\mathrm{V}^\sigma \text{ from tomorrow})$$

$$\boxed{\mathrm{V}^\sigma = u + \gamma P_\sigma \mathrm{V}^\sigma} \tag{1}$$

- Following Bellman, the value function has a natural recursive structure:

$$(\mathrm{V}^\sigma \text{ from today}) = \text{reward today} + \gamma \cdot (\mathrm{V}^\sigma \text{ from tomorrow})$$

$$\boxed{\mathrm{V}^\sigma = u + \gamma P_\sigma \mathrm{V}^\sigma} \tag{1}$$

- So, V^σ arises as a fixpoint of the operator $\Psi_\sigma \colon \mathbb{R}^S \to \mathbb{R}^S$ given by

$$\Psi_\sigma(v) = u + \gamma P_\sigma v$$

- Following Bellman, the value function has a natural recursive structure:

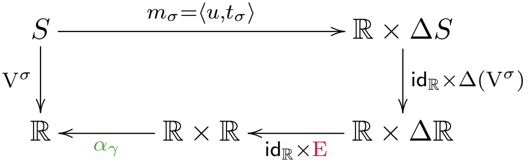
$$(V^\sigma \text{ from today}) = \text{reward today} + \gamma \cdot (V^\sigma \text{ from tomorrow})$$

$$\boxed{V^\sigma = u + \gamma P_\sigma V^\sigma} \tag{1}$$

- So, V^σ arises as a fixpoint of the operator $\Psi_\sigma \colon \mathbb{R}^S \to \mathbb{R}^S$ given by

$$\Psi_\sigma(v) = u + \gamma P_\sigma v$$

- **Our observation**: we can re-express (1) as V^σ being a coalgebra-to-algebra morphism, as in



where $\alpha_\gamma \colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is $\alpha_\gamma(x, y) = x + \gamma \cdot y$ and $\mathrm{E} \colon \Delta\mathbb{R} \to \mathbb{R}$ is EV

Part 1 – Universality from Corecursive Algebra?

- Recall that a corecursive algebra (for functor F) is an F-algebra α s.t.

$$
\begin{array}{ccc}
C & \xrightarrow{\forall f} & FC \\
{\scriptstyle \exists!\, f^\dagger}\downarrow & & \downarrow{\scriptstyle F f^\dagger} \\
A & \xleftarrow[\alpha]{} & FA
\end{array}
$$

- Recall that a corecursive algebra (for functor F) is an F-algebra α s.t.

$$\begin{array}{ccc} C & \xrightarrow{\forall f} & FC \\ {\scriptstyle \exists!\, f^\dagger}\downarrow & & \downarrow{\scriptstyle F f^\dagger} \\ A & \xleftarrow[\alpha]{} & FA \end{array}$$

- Recall:

$$\begin{array}{ccc} S & \xrightarrow{m_\sigma = \langle u, t_\sigma \rangle} & \mathbb{R} \times \Delta S \\ {\scriptstyle \mathrm{V}^\sigma}\downarrow & & \downarrow{\scriptstyle \mathrm{id}_{\mathbb{R}} \times \Delta(\mathrm{V}^\sigma)} \\ \mathbb{R} & \xleftarrow[\alpha_\gamma \circ (\mathrm{id}_{\mathbb{R}} \times \mathrm{E})]{} & \mathbb{R} \times \Delta \mathbb{R} \end{array}$$

- Recall that a corecursive algebra (for functor F) is an F-algebra α s.t.

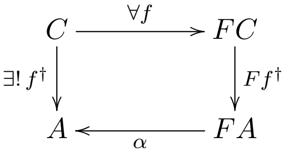
$$
\begin{array}{ccc}
C & \xrightarrow{\ \forall f\ } & FC \\
{\scriptstyle \exists!\, f^\dagger}\downarrow & & \downarrow{\scriptstyle F f^\dagger} \\
A & \xleftarrow[\ \alpha\ ]{} & FA
\end{array}
$$

- Recall:

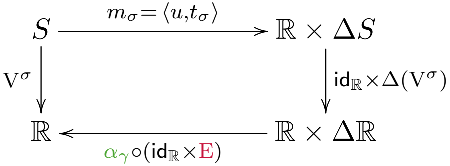
$$
\begin{array}{ccc}
S & \xrightarrow{\ m_\sigma = \langle u, t_\sigma \rangle\ } & \mathbb{R} \times \Delta S \\
{\scriptstyle \mathrm{V}^\sigma}\downarrow & & \downarrow{\scriptstyle \mathsf{id}_\mathbb{R} \times \Delta(\mathrm{V}^\sigma)} \\
\mathbb{R} & \xleftarrow[\ \alpha_\gamma \circ (\mathsf{id}_\mathbb{R} \times \mathrm{E})\ ]{} & \mathbb{R} \times \Delta \mathbb{R}
\end{array}
$$

- Question: is $\alpha_\gamma \circ (\mathsf{id}_\mathbb{R} \times \mathrm{E})$ a corecursive algebra?

- Recall that a corecursive algebra (for functor F) is an F-algebra α s.t.



- Recall:



- Question: is $\alpha_\gamma \circ (\mathrm{id}_{\mathbb{R}} \times \mathrm{E})$ a corecursive algebra?
- Consider a more basic question: is algebra $\alpha_\gamma : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ corecursive?

  By (Capretta et al., 2004), this is **equivalent** with $\alpha_\gamma \circ (\mathrm{id}_{\mathbb{R}} \times \mathrm{E})$ corec.

- Letting $H = \mathbb{R} \times \mathsf{id}$, α_γ is corecursive if

$$\begin{array}{ccc} X & \xrightarrow{\ \forall f\ } & \mathbb{R} \times X \\ {\scriptstyle \exists!\, f^\dagger}\downarrow & & \downarrow{\scriptstyle Hf^\dagger = \mathsf{id}_{\mathbb{R}} \times f^\dagger} \\ \mathbb{R} & \xleftarrow[\ \alpha_\gamma\ ]{} & \mathbb{R} \times \mathbb{R} \end{array}$$

- Letting $H = \mathbb{R} \times \mathsf{id}$, α_γ is corecursive if

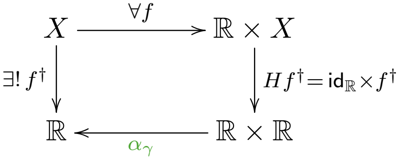
$$\begin{array}{ccc} X & \xrightarrow{\ \forall f\ } & \mathbb{R} \times X \\ {\scriptstyle \exists!\, f^\dagger}\downarrow & & \downarrow{\scriptstyle H f^\dagger = \mathsf{id}_{\mathbb{R}} \times f^\dagger} \\ \mathbb{R} & \xleftarrow[\ \alpha_\gamma\ ]{} & \mathbb{R} \times \mathbb{R} \end{array}$$

- In particular, let X be a set of "variables" $\{x_0, x_1, x_2, \ldots\}$, and fix γ.

$$\left\{ \begin{array}{c} \text{System of lin. eqs.} \\ (3)\ \boxed{x_n = a_n + \gamma \cdot x_{n+1}} \\ (n = 0, 1, 2, \ldots) \end{array} \right\} \quad \leftrightsquigarrow \quad \text{coalgebra } f\colon X \to \mathbb{R} \times X$$

Solutions to (3) $\quad \leftrightsquigarrow \quad f^\dagger$ s.t. $f^\dagger = \alpha_\gamma \cdot (\mathsf{id}_{\mathbb{R}} \times f^\dagger) \cdot f$

- Letting $H = \mathbb{R} \times \mathsf{id}$, α_γ is corecursive if



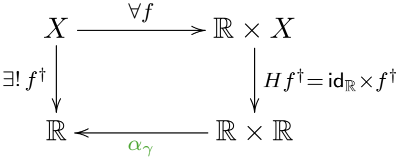
- In particular, let X be a set of "variables" $\{x_0, x_1, x_2, \ldots\}$, and fix γ.

$$\left\{ \begin{array}{c} \text{System of lin. eqs.} \\ (3)\ \boxed{x_n = a_n + \gamma \cdot x_{n+1}} \\ (n = 0, 1, 2, \ldots) \end{array} \right\} \quad \leftrightsquigarrow \quad \text{coalgebra } f\colon X \to \mathbb{R} \times X$$

$$\text{Solutions to (3)} \quad \leftrightsquigarrow \quad f^\dagger \text{ s.t. } f^\dagger = \alpha_\gamma \cdot (\mathsf{id}_{\mathbb{R}} \times f^\dagger) \cdot f$$

- But, (3) has *infinite* number of solutions, even if $(a_n)_n$ is bounded
  $\Rightarrow$ answer to earlier question is NO

- Letting $H = \mathbb{R} \times \mathrm{id}$, α_γ is corecursive if



- In particular, let X be a set of "variables" $\{x_0, x_1, x_2, \ldots\}$, and fix γ.

System of lin. eqs.

$$\left\{ (3) \; \boxed{x_n = a_n + \gamma \cdot x_{n+1}} \; (n = 0, 1, 2, \ldots) \right\} \quad \leftrightsquigarrow \quad \text{coalgebra } f \colon X \to \mathbb{R} \times X$$

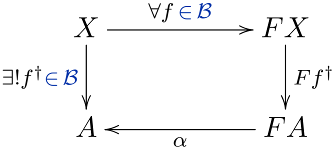
Solutions to (3) $\quad \leftrightsquigarrow \quad f^\dagger$ s.t. $f^\dagger = \alpha_\gamma \cdot (\mathrm{id}_{\mathbb{R}} \times f^\dagger) \cdot f$

- But, (3) has *infinite* number of solutions, even if $(a_n)_n$ is bounded $\Rightarrow$ answer to earlier question is NO
- However, if $(a_n)_n$ is bounded then (3) has a unique **bounded** solution

- To get uniqueness $\Rightarrow$ incorporate boundedness information

- To get uniqueness $\Rightarrow$ incorporate boundedness information
- **Definition** A *b-category* $(\mathsf{C}, \mathcal{B})$ is a category C with a subcollection of "bounded" morphism $\mathcal{B}$ s.t. $(f \in \mathcal{B} \Rightarrow f \circ g \in \mathcal{B})$

- To get uniqueness $\Rightarrow$ incorporate boundedness information
- **Definition** A *b-category* $(\mathsf{C}, \mathcal{B})$ is a category C with a subcollection of "bounded" morphism $\mathcal{B}$ s.t. $(f \in \mathcal{B} \Rightarrow f \circ g \in \mathcal{B})$
- **Definition** Let $(\mathsf{C}, \mathcal{B})$ b-category, F endofunctor on C, $\alpha\colon FA \to A$ an F-algebra. Then α is a *b-corecursive algebra (bca)* if



- To get uniqueness $\Rightarrow$ incorporate boundedness information
- **Definition** A *b-category* $(\mathsf{C}, \mathcal{B})$ is a category C with a subcollection of "bounded" morphism $\mathcal{B}$ s.t. $(f \in \mathcal{B} \Rightarrow f \circ g \in \mathcal{B})$
- **Definition** Let $(\mathsf{C}, \mathcal{B})$ b-category, F endofunctor on C, $\alpha \colon FA \to A$ an F-algebra. Then α is a *b-corecursive algebra (bca)* if

$$\begin{array}{ccc} X & \xrightarrow{\forall f \in \mathcal{B}} & FX \\ {\scriptstyle \exists! f^\dagger \in \mathcal{B}} \downarrow & & \downarrow {\scriptstyle F f^\dagger} \\ A & \xleftarrow[\alpha]{} & FA \end{array}$$

- **Proposition** α_γ is a bca in (Met, B) for H
- Then by b-version of (Capretta et al., 2004) result:

  Proposition $\alpha_\gamma \circ (\mathbb{R} \times \mathrm{E})$ is a bca in (Met, B) for $H \circ \Delta$

Part 1: Value function V^{σ} from b-corecursive algebras ✓

Part 2: Coinductive proof of policy improvement theorem

PART 2:

Correctness of Policy Iteration via Contraction (Co)Induction

- Suppose σ is policy $\rightsquigarrow$ find V^{σ}

- Suppose σ is policy $\rightsquigarrow$ find V^σ
- Now we define a new policy σ' by putting for each state s

$$\sigma'(s) = \mathsf{argmax}_{a \in A} \left\{ \sum_{s' \in S} P(s,a,s')\, \mathrm{V}^\sigma(s') \right\}$$

- Suppose σ is policy $\rightsquigarrow$ find V^σ
- Now we define a new policy σ' by putting for each state s

$$\sigma'(s) = \mathrm{argmax}_{a \in A} \left\{ \sum_{s' \in S} P(s, a, s')\, \mathrm{V}^\sigma(s') \right\}$$

Theorem (Howard, 1960)

- The policy σ' is a better policy than σ, i.e., $\sigma' \geq \sigma$.
- If $\sigma' = \sigma$, then σ is optimal.

- Suppose σ is policy $\rightsquigarrow$ find V^{σ}
- Now we define a new policy σ' by putting for each state s

$$\sigma'(s) = \mathsf{argmax}_{a \in A} \left\{ \sum_{s' \in S} P(s,a,s')\, \mathrm{V}^{\sigma}(s') \right\}$$

Theorem (Howard, 1960)

- The policy σ' is a better policy than σ, i.e., $\sigma' \geq \sigma$.
- If $\sigma' = \sigma$, then σ is optimal.

- Policy Iteration: start with *any* σ, iteratively obtain $\sigma', \sigma'', \sigma''', \ldots$, and continue until there is fixpoint $\Rightarrow$ this outputs an *optimal policy*

Policy Improvement Theorem (Howard, 1960)

$$P_{\sigma'} \cdot \mathrm{V}^{\sigma} \geq P_{\sigma} \cdot \mathrm{V}^{\sigma} \quad \Rightarrow \quad \mathrm{V}^{\sigma'} \geq \mathrm{V}^{\sigma}$$

Policy Improvement Theorem (Howard, 1960)

$P_{\sigma'} \cdot \mathrm{V}^{\sigma} \geq P_{\sigma} \cdot \mathrm{V}^{\sigma} \quad \Rightarrow \quad \mathrm{V}^{\sigma'} \geq \mathrm{V}^{\sigma}$

- Recall

$$\sigma'(s) = \mathsf{argmax}_{a \in A} \left\{ \sum_{s' \in S} P(s, a, s') \, \mathrm{V}^{\sigma}(s') \right\}$$

$\Rightarrow$ antecedent holds

Policy Improvement Theorem (Howard, 1960)

$$P_{\sigma'} \cdot \mathrm{V}^{\sigma} \geq P_{\sigma} \cdot \mathrm{V}^{\sigma} \quad \Rightarrow \quad \mathrm{V}^{\sigma'} \geq \mathrm{V}^{\sigma}$$

- Recall

$$\sigma'(s) = \mathsf{argmax}_{a \in A} \left\{ \sum_{s' \in S} P(s, a, s')\, \mathrm{V}^{\sigma}(s') \right\}$$

$\Rightarrow$ antecedent holds

- New proof using Contraction (Co)Induction

Contraction (Co)Induction Theorem

Let M be a non-empty, complete, ordered (i.e., there is partial order $\leq$ s.t. $\uparrow x = \{y \mid x \leq y\}$ and $\downarrow x = \{y \mid y \leq x\}$ are closed) metric space. If $f \colon M \to M$ is contractive and order-preserving, then the (unique) fixpoint x^* is

- least pre-fixpoint (if $f(x) \leq x$, then $x^* \leq x$),
- greatest post-fixpoint (if $x \leq f(x)$, then $x \leq x^*$).

Cf. *Metric Coinduction* (Kozen & Ruozzi, 2009) and (Denardo, 1967).

Contraction (Co)Induction Theorem

Let M be a non-empty, complete, ordered (i.e., there is partial order $\leq$ s.t. $\uparrow x = \{y \mid x \leq y\}$ and $\downarrow x = \{y \mid y \leq x\}$ are closed) metric space. If $f\colon M \to M$ is contractive and order-preserving, then the (unique) fixpoint x^* is

- least pre-fixpoint (if $f(x) \leq x$, then $x^* \leq x$),
- greatest post-fixpoint (if $x \leq f(x)$, then $x \leq x^*$).

Cf. *Metric Coinduction* (Kozen & Ruozzi, 2009) and (Denardo, 1967).

Proof of Policy Improvement (i.e., $P_{\sigma'} \cdot \mathrm{V}^{\sigma} \geq P_{\sigma} \cdot \mathrm{V}^{\sigma} \Rightarrow \mathrm{V}^{\sigma'} \geq \mathrm{V}^{\sigma}$).
Apply theorem to $\Psi_\pi\colon \mathbb{R}^S \to \mathbb{R}^S$ (contractive and order-preserving ✓)

$$\Psi_\pi(v) = u + \gamma P_\pi v, \qquad \text{and } \mathrm{V}^\pi \text{ is its fixpoint.}$$

Thus $P_{\sigma'} \cdot \mathrm{V}^{\sigma} \geq P_{\sigma} \cdot \mathrm{V}^{\sigma} \Rightarrow \Psi_{\sigma'}(\mathrm{V}^{\sigma}) \geq \Psi_{\sigma}(\mathrm{V}^{\sigma}) = \mathrm{V}^{\sigma} \Rightarrow \mathrm{V}^{\sigma'} \geq \mathrm{V}^{\sigma}$. ■

Contributions:

- Value functions V^σ and V^* from b-corecursive algebras
- Coinductive proof of policy improvement theorem

Future work:

- Generalize the setting (e.g., to stochastic games)
- Make connections with related literature:
  Combining semantics of computation and game theory (Pavlovic, 2009)
  Coalgebraic formulation of infinite games (Abramsky & Winschel, 2017)
  Open games (Hedges, Ghani, Winschel, Zahn, 2018)
- Investigate contraction coinduction further and look for other applications (e.g., in social choice)

Contributions:

- Value functions V^{σ} and V^{*} from b-corecursive algebras
- Coinductive proof of policy improvement theorem

Future work:

- Generalize the setting (e.g., to stochastic games)
- Make connections with related literature:
  - Combining semantics of computation and game theory (Pavlovic, 2009)
  - Coalgebraic formulation of infinite games (Abramsky & Winschel, 2017)
  - Open games (Hedges, Ghani, Winschel, Zahn, 2018)
- Investigate contraction coinduction further and look for other applications (e.g., in social choice)

Thank you!