# (Co)Algebraic Techniques for Markov Decision Processes

Frank M. V. Feys[1], Helle Hvid Hansen[1], and Lawrence S. Moss[2]

[1] Department of Engineering Systems and Services, TPM, Delft University of Technology, Delft, The Netherlands {f.m.v.feys, h.h.hansen}@tudelft.nl
[2] Department of Mathematics, Indiana University, Bloomington IN, 47405 USA lsm@cs.indiana.edu

## 1 Introduction

Markov Decision Processes (MDPs) [11] are a family of probabilistic, state-based models used in planning under uncertainty and reinforcement learning. Informally, an MDP models a situation in which an agent (the decision maker) makes choices at each state of a process, and each choice leads to some reward and a probabilistic transition to a next state. The aim of the agent is to find an optimal policy, i.e., a way of choosing actions that maximizes future expected rewards.

The classic theory of MDPs with discounting is well-developed (see [11, Chapter 6]), and indeed we do not prove any new results about MDPs as such. Our work is inspired by Bellman's principle of optimality, which states the following: "An optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision" [2, Chapter III.3]. This principle has clear coinductive overtones, and our aim is to situate it in a body of mathematics that is also concerned with infinite behavior and coinductive proof principles, i.e., in coalgebra.

Probabilistic systems of similar type have been studied extensively, also coalgebraically, in the area of program semantics (see for instance [5, 6, 14, 15]). Our focus is not so much on the observable behavior of MDPs viewed as computations, but on their role in solving optimal planning problems.

This abstract is based on [7] to which we refer for a more detailed account.

## 2 Markov Decision Processes

We briefly introduce the relevant basic concepts from the classic theory MDPs[11]. Letting $\Delta S$ denote the set of probability distributions with finite support on the set $S$, we define MDPs[1] and policies as follows.

**Definition 2.1 (MDP, Policy)** *Let $Act$ be a finite set of* actions. *A* Markov decision process (MDP) $m = \langle S, u, t\rangle$ *consists of a finite set $S$ of* states*, a reward function $u\colon S \to \mathbb{R}$, and a probabilistic transition structure $t\colon S \to (\Delta S)^{Act}$. A* policy *is a function $\sigma\colon S \to Act$.*

That is, in state $s$ when the agent chooses action $a$, there is a probability distribution $t(s)(a)$ over states. Furthermore, in each state $s$, the agent collects a reward (or utility) specified by a real number $u(s)$.

We shall often leave $S$ implicit and simply write $m = \langle u, t\rangle$. Given a probabilistic transition structure $t\colon S \to (\Delta S)^{Act}$ and a policy $\sigma \in Act^S$, we write $t_\sigma\colon S \to \Delta S$ for the map defined by $t_\sigma(s) = t(s)(\sigma(s))$, and $t_a$ when $\sigma$ is constant equal to $a$.

[1] Our simple type of MDPs is known as time-homogeneous, infinite-horizon MDPs with finite state and action spaces, and our policies as stationary, memoryless deterministic policies.

There are several criteria for evaluating the long-term rewards expected by following a given policy. A classic criterion uses discounting. The idea is that rewards collected tomorrow are worth less than rewards collected today.

**Definition 2.2** *Let $\gamma$ be a fixed real number with $0 \leq \gamma < 1$. Such a $\gamma$ is called a* discount factor. *Let an MDP $m = \langle u, t \rangle$ be given. The* long-term value *of a policy $\sigma$ (for $m$) according to the* discounted sum criterion *is the function $\mathrm{LTV}_\sigma \colon S \to \mathbb{R}$ defined as follows:*

$$\mathrm{LTV}_\sigma(s) = r_0^\sigma(s) + \gamma \cdot r_1^\sigma(s) + \cdots + \gamma^n \cdot r_n^\sigma(s) + \cdots \tag{1}$$

*where $r_n^\sigma(s)$ is the expected reward at time step $n$.*

Note that $r_0^\sigma(s) = u(s)$ for all $s \in S$, and since $S$ is finite, $\max_s r_0^\sigma(s) < \infty$. This boundedness property entails that the infinite sum in (1) is convergent.

It will be convenient to work with the map $\ell_\sigma$ that takes the expected value of $\mathrm{LTV}_\sigma$ relative to some distribution. Formally, $\ell_\sigma \colon \Delta S \to \mathbb{R}$ is defined for all $\varphi \in \Delta S$ by

$$\ell_\sigma(\varphi) = \sum_{s \in S} \varphi(s) \cdot \mathrm{LTV}_\sigma(s). \tag{2}$$

Observe that for each state $s$, $\mathrm{LTV}_\sigma(s)$ is equal to the immediate rewards plus the discounted future expected rewards. Seen this way, (1) may be re-written to the corecursive equation

$$\mathrm{LTV}_\sigma(s) = u(s) + \gamma \cdot \left( \sum_{s' \in S} t_\sigma(s)(s') \cdot \mathrm{LTV}_\sigma(s') \right) = u(s) + \gamma \cdot \ell_\sigma(t_\sigma(s)). \tag{3}$$

Viewing $\mathrm{LTV}_\sigma$ as a column vector in $\mathbb{R}^S$ and $t_\sigma$ as a column-stochastic matrix $P_\sigma$, the equation in (3) shows that $\mathrm{LTV}_\sigma$ is a fixpoint of the (linear) operator

$$\Psi_\sigma \colon \mathbb{R}^S \to \mathbb{R}^S \qquad \Psi_\sigma(v) = u + \gamma P_\sigma v. \tag{4}$$

By the Banach Fixpoint Theorem, this fixpoint unique, since $\Psi_\sigma$ is contractive (due to $0 \leq \gamma < 1$), and $\mathbb{R}^S$ is a complete metric space. The long-term value induces a preorder on policies: $\sigma \leq \tau$ if $\mathrm{LTV}_\sigma \leq \mathrm{LTV}_\tau$ in the pointwise order on $\mathbb{R}^S$. A policy $\sigma$ is *optimal* if for all policies $\tau$, we have $\tau \leq \sigma$.

Given an MDP $m$, the *optimal value of $m$* is the map $V^* \colon S \to \mathbb{R}$ that for each state gives the best long-term value that can be obtained for any policy [11]:

$$V^*(s) = \max_{\sigma \in Act^S} \{\mathrm{LTV}_\sigma(s)\}.$$

It is an important classic result that $V^*$ is the unique (bounded) map that satisfies *Bellman's optimality equation* [2, 11]:

$$V^*(s) = u(s) + \gamma \cdot \max_{a \in Act} \left\{ \sum_{s' \in S} t_a(s)(s') \cdot V^*(s') \right\}.$$

# 3 Main Contributions

## 3.1 Policy Improvement via Contraction Coinduction

For our simple model of MDPs with discounting, it is known that the simple type of policies that we consider here, are sufficient. In other words, an optimal policy can always be found among

stationary, memoryless, deterministic policies [11, Theorem 6.2.7]. This result together with the optimality equation forms the basis for an effective algorithm for finding optimal policies, known as *policy iteration* [8]. The algorithm starts from any policy $\sigma \in Act^S$, and iteratively improves $\sigma$ to some $\tau$ such that $\sigma \leq \tau$. This leads to an increasing sequence of policies in the preorder of all policies $(S^{Act}, \leq)$. Since this preorder is finite, this process will at some point stabilize. The correctness of policy iteration follows from the following theorem.

**Theorem 3.1 (Policy Improvement)** *Let an MDP be given by $t\colon S \to (\Delta S)^{Act}$ and $u\colon S \to \mathbb{R}$. Let $\sigma$ and $\tau$ be policies. If $\ell_\sigma \circ t_\tau \geq \ell_\sigma \circ t_\sigma$, then $\mathrm{LTV}_\tau \geq \mathrm{LTV}_\sigma$. Similarly, if $\ell_\sigma \circ t_\tau \leq \ell_\sigma \circ t_\sigma$, then $\mathrm{LTV}_\tau \leq \mathrm{LTV}_\sigma$.*

We present a coinductive proof of the Policy Improvement theorem. This leads us to formulate a coinductive proof principle that we have named *contraction (co)induction*. The contraction coinduction principle is a variation of the classic Banach Fixpoint Theorem, asserting that any contractive mapping on a complete metric space has a unique fixpoint. We need a version of this theorem which, in addition to a complete metric, also has an order.

**Definition 3.2** *An* ordered metric space *is a structure $(M, d, \leq)$ such that $d$ is a metric on $M$ and $\leq$ is a partial order on $M$, satisfying the extra property that for all $y \in M$, $\{z \mid z \leq y\}$ and $\{z \mid y \leq z\}$ are closed sets in the metric topology. This space is said to be* complete *if it is complete as a metric space.*

**Theorem 3.3 (Contraction (Co)Induction)** *Let $M$ be a non-empty, complete ordered metric space. If $f\colon M \to M$ is both 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 $x \leq x^*$).*

Theorem 3.3 follows from the Metric Coinduction Principle [10, 12]. Our aim is not the highest level of generality. Rather, we see contraction (co)induction as a particular instance of Metric Coinduction that suffices to prove interesting results about MDPs. We also believe contraction (co)induction should have applications far beyond the topic of MDPs.

## 3.2 Long-Term Values via $b$-Corecursive Algebras

We now take a coalgebraic perspective on MDPs and long-term value functions. Let $\Delta$ be the $\mathsf{Set}$-monad of finitely supported probability distributions, and let $H$ be the $\mathsf{Set}$-functor $H = \mathbb{R} \times \mathrm{Id}$. A leading observation of this paper is that we can re-express (3) by saying that $\mathrm{LTV}_\sigma\colon S \to \mathbb{R}$ makes the following diagram commute:

$$\begin{array}{ccccc} S & & \xrightarrow{\quad m_\sigma \quad} & & \mathbb{R} \times \Delta S \\ {\scriptstyle \mathrm{LTV}_\sigma}\downarrow & & & & \downarrow{\scriptstyle \mathbb{R}\times\Delta(\mathrm{LTV}_\sigma)} \\ \mathbb{R} & \xleftarrow[\alpha_\gamma]{} & \mathbb{R} \times \mathbb{R} & \xleftarrow[\mathbb{R}\times \mathrm{E}]{} & \mathbb{R} \times \Delta\mathbb{R} \end{array} \tag{5}$$

Here, $\mathrm{E}\colon \Delta\mathbb{R} \to \mathbb{R}$ is the expected value function and $\alpha_\gamma\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is the $H$-algebra

$$\alpha_\gamma\colon H\mathbb{R} \to \mathbb{R} \qquad \alpha_\gamma(x, y) = x + \gamma \cdot y. \tag{6}$$

This means that $\mathrm{LTV}_\sigma$ is an $H\Delta$-coalgebra-to-algebra map. We naturally wonder whether the $H\Delta$-algebra at the bottom of the diagram is a *corecursive algebra* [4]: for every coalgebra $f\colon X \to H\Delta X$ (where $X$ is possibly infinite), is there a unique map $f^\dagger\colon S \to \mathbb{R}$ making the

diagram commute? This turns out not to be the case for arbitrary state spaces, as problems can arise when reward values are unbounded. To remedy this, we introduce the notions of $b$-categories and $b$-corecursive algebras with which we aim to give a sparse categorification of boundedness. Combining these with techniques from coinductive specification ([3]) and trace sematics [1, 9], we can show that $\alpha_\gamma \circ (\mathbb{R} \times \mathrm{E})$ is a $b$-corecursive algebra, and thereby obtain $\mathrm{LTV}_\sigma$ from its universal property. The optimal value function $V^*$ can be characterised in a similar way via a $b$-corecursive algebra.

This categorical approach emphasizes compositional reasoning about functions and functors. The classical theory of MDPs does not do this; it directly proves properties (such as boundedness) of composites viewed as monolithic entities, instead of deriving them from preservation properties of their constituents. So it neither needs nor uses the extra information that we obtained by working in a categorical setting. Indeed, most of our work is devoted to this extra information, and we hope that it will be useful in settings beyond MDPs. We have some pilot results in this direction for stochastic games [13].

# References


[1] F. Bartels. *On Generalised Coinduction and Probabilistic Specification Formats*. PhD thesis, Vrije Universiteit Amsterdam, 2004.

[2] R. Bellman. *Dynamic Programming*. Princeton University Press, Princeton, NJ, USA, 1 edition, 1957.

[3] V. Capretta, T. Uustalu, and V. Vene. Recursive Coalgebras from Comonads. *Information and Computation*, 204:437468, 2006.

[4] V. Capretta, T. Uustalu, and V. Vene. Corecursive Algebras: A Study of General Structured Corecursion. In *Formal Methods: Foundations and Applications*, pages 84–100, 2009.

[5] J. Desharnais, A. Edalat, and P. Panangaden. Bisimulation for Labelled Markov Processes. *Information and Computation*, 179(2):163–193, 2002.

[6] N. Ferns, P. Panangaden, and D. Precup. Metrics for Finite Markov Decision Processes. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, UAI '04, pages 162–169, Arlington, Virginia, United States, 2004. AUAI Press.

[7] F. M. Feys, H. H. Hansen, and L. S. Moss. Long-term values in Markov decision processes, (co)algebraically. In C. Cîrstea, editor, *Coalgebraic Methods in Computer Science*, volume 11202 of *Lecture Notes in Computer Science*, pages 78–99. Springer, 2018.

[8] R. A. Howard. *Dynamic Programming and Markov Processes*. The M.I.T. Press, 1960.

[9] B. Jacobs, A. Silva, and A. Sokolova. Trace Semantics via Determinization. *Journal of Computer and System Sciences*, 81(5):859 – 879, 2015. 11th International Workshop on Coalgebraic Methods in Computer Science, CMCS 2012 (Selected Papers).

[10] D. Kozen. Coinductive Proof Principles for Stochastic Processes. *Logical Methods in Computer Science*, 5:1–19, 2009.

[11] M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 2014.

[12] N. Ruozzi and D. Kozen. Applications of Metric Coinduction. *Logical Methods in Computer Science*, 5, 2009.

[13] L. S. Shapley. Stochastic Games. *PNAS*, 39(10):1095–1100, 1953.

[14] A. Silva and A. Sokolova. Sound and Complete Axiomatization of Trace Semantics for Probabilistic Systems. *Electronic Notes in Theoretical Computer Science*, 276:291–311, 2011.

[15] A. Sokolova. Probabilistic Systems Coalgebraically. *Theoretical Computer Science*, 412(38):5095–5110, September 2011.