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

Frank Feys, Helle Hvid Hansen, Lawrence S. Moss

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

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

Abstract. This paper studies Markov decision processes (MDPs) from the categorical perspective of coalgebra and algebra. Probabilistic systems, similar to MDPs but without rewards, have been extensively studied, also coalgebraically, from the perspective of program semantics. In this paper, we focus on the role of MDPs as models in optimal planning, where the reward structure is central. The main contributions of this paper are (i) to give a coinductive explanation of policy improvement using a new proof principle, based on Banach's Fixpoint Theorem, that we call contraction coinduction, and (ii) to show that the long-term value function of a policy with respect to discounted sums can be obtained via a generalized notion of corecursive algebra, which is designed to take boundedness into account. We also explore boundedness features of the Kantorovich lifting of the distribution monad to metric spaces.

Keywords: Markov decision process · long-term value · discounted sum · coalgebra · algebra · corecursive algebra · fixpoint · metric space.

1 Introduction

Markov Decision Processes (MDPs) [23] are a family of probabilistic, state-based models used in planning under uncertainty and reinforcement learning. Informally stated, an MDP models a situation in which an agent (the decision maker) has to make 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. In this paper, we consider a simple version of MDPs known as time-homogeneous, infinite-horizon MDPs in which the set of states and actions are finite, and future rewards are computed according to the discounted summation criterion.

Probabilistic systems of similar type have been studied extensively, also coalgebraically, in the area of program semantics (see for instance [8,9,27,28]). Our focus is not so much on the observable behavior of MDPs viewed as computations, but on their role in solving optimal planning problems.

The classic theory of MDPs with discounting is well-developed (see [23, 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" [4, 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.

The main contributions of this paper are the following. First, we present a coinductive proof of the correctness of a classic iterative procedure known as policy iteration [12]. This leads us to formulate a coinductive proof principle that we have named *contraction (co)induction*, and which is closely related to Kozen's metric coinduction [17]. We believe contraction (co)induction should have applications far beyond the topic of MDPs. Second, we show that long-term values of policies can be obtained from the universal property of a generalized notion of corecursive algebra. The technical challenge here is to encode boundedness information in order to ensure the unique existence of certain fixpoints. This leads us to introduce the notions of b-categories and b-corecursive algebras (bcas). Combining these with well-known techniques from coinductive specification and trace sematics [3,14], we obtain the desired universal maps.

Contents of this paper. In Section 2 we give a brief introduction to MDPs and the classical results that we aim to categorify. In Section 3, we present contraction coinduction and apply it to prove the correctness of policy iteration and related results. In Section 4, we describe the (set-based) coalgebraic and algebraic structures that we use to model MDPs and discounted sums. In Section 5, we move to a category of metric spaces, we introduce b-categories and b-corecursive algebras (bcas), and we show that the long-term value of a policy as well as the optimal value arise as universal arrows. We briefly discuss extensions of our work in Section 6. Finally, we conclude and discuss related and future work in Section 7.

2 Markov Decision Processes

We refer to [23] for a comprehensive overview of MDPs, including numerous applications to planning problems such as inventory management and highway maintenance. Here, we confine ourselves to a brief introduction.

An MDP models a situation in which an agent in each state $s \in S$ chooses to execute an action $a \in Act$, and this choice results in a probabilistic transition to a new state $s' \in S$. That is, for every state s and every 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 $u(s)$. The aim of the agent is to find a policy that will maximize his expected long-term rewards. Letting ΔS denote the set of probability distributions on a finite set S, MDPs and policies are formally defined as follows.

Definition 1. *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 $t\colon S \to (\Delta S)^{Act}$. We often omit S and simply write $m = \langle u, t\rangle$. A* policy *is a function $\sigma\colon S \to Act$.*

More generally, MDPs are considered with respect to a time evolution which may be discrete or continuous, and the transition structure and reward function may depend on the time step. If the time evolution is assumed to end after finitely many steps, the MDP is called *finite-horizon*. In our definition of MDPs, time evolution is implicitly assumed to be discrete, but t and u do not depend on the time step, making them *time-homogeneous*, and the time evolution is not assumed to end, making them *infinite-horizon*.

Similarly, there are more general notions of policy in which the policy may depend on the time step. A policy that does not depend on the time step is called *stationary*. The choices prescribed by a non-stationary policy may depend on the entire history of the system up until the present time step, but if each choice depends only on the current state and not the history, then the policy is called *Markovian* or *memoryless*. Finally, a policy may also be randomized, i.e., of type $S \to \Delta Act$, as opposed to *deterministic*. That means, in this paper we consider *stationary* (and therefore *memoryless*), *deterministic policies*.

Example 1. Consider the MDP m shown in Figure 1, taken from [20]. A startup company can be in one of four states that we abbreviate by PU, PF, RU, and RF. In each state, the company receives an *immediate reward* $u(s)$, and chooses to either *advertise* (**A**) or *save* (**S**). The effect of an action in a state is in general probabilistic, as indicated by the arrows. We take a *discount factor* $\gamma = 0.9$.

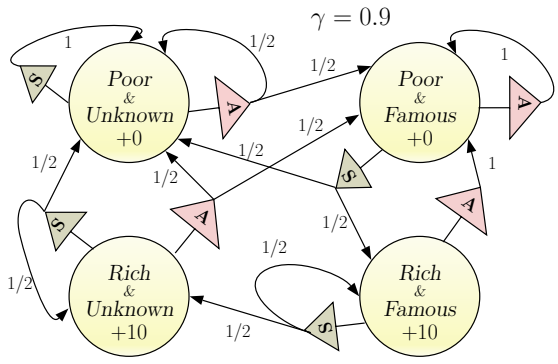

FIG. 1: Example of an MDP modeling of a startup (taken from [20]).

There are several criteria for evaluating the long-term rewards expected by following a given policy. A classic one found in the literature takes the long-term rewards to be the *discounted infinite sum of expected rewards*. The idea is that

rewards collected tomorrow are worth less than rewards collected today. Before we state the definition, we need some notation. Given a probabilistic transition structure $t\colon S \to (\Delta S)^{Act}$ and a policy $\sigma \in Act^S$, we get a map $t_\sigma\colon S \to \Delta S$ by letting $t_\sigma(s) = t(s)(\sigma(s))$. The pair $\langle u, t_\sigma\rangle$ is sometimes called a *Markov reward process*. The map t_σ corresponds to a column-stochastic $|S| \times |S|$-matrix P_σ. Viewing $u \in \mathbb{R}^S$ as a row $|S|$-vector and a start state s as a column $|S|$-vector v_s with 1 in position s and 0 everywhere else, the probability that the agent is in a state s' at time step n is found in position s' of the column-stochastic vector $P_\sigma^n v_s$, and the expected reward $r_n^\sigma(s)$ at time step n is the scalar $uP_\sigma^n v_s$.

Definition 2. *Let γ be a fixed real number with $0 \leq \gamma < 1$. Such a γ is called a* discount factor*. Let an MDP $m = \langle u, t\rangle$ be given. The* long-term value *of a policy σ (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. A policy σ is optimal *if for all $s \in S$, $\mathrm{LTV}_\sigma(s) = \max_{\tau \in Act^S} \mathrm{LTV}_\tau(s)$.*

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 ℓ_σ that takes the expected value of LTV_σ 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 LTV_σ as a column vector in $\mathbb{R}^S$, the equation in (3) represents a linear system, $\mathrm{LTV}_\sigma = u + \gamma P_\sigma \mathrm{LTV}_\sigma$. We find LTV_σ by solving it: $\mathrm{LTV}_\sigma = (I - \gamma P_\sigma)^{-1} u$, where I is the identity matrix.

Equivalently, LTV_σ is defined as the unique fixpoint of the (linear), contractive, monotone (for the pointwise order) operator

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

Note that Ψ_σ is contractive since P_σ is column-stochastic and we multiply with γ, and $\mathbb{R}^S$ is a complete metric space. Hence by the Banach Fixpoint Theorem, the unique fixpoint exists. Moreover, Ψ_σ is monotone, because P_σ has all non-negative entries.

Example 2. We continue with Example 1. An example of a policy is the "miserly" σ given by $\sigma(s) = \mathbf{S}$ for all states s, i.e., the startup chooses to save in each state. The equations that describe the probabilistic system m_σ resulting from following σ are given in the equations on the left below. To compute LTV_σ, the expression from (1) for this policy may be rewritten to the equations on the right in (5) below. (Recall that the discount factor is $\gamma = 0.9$.)

$$\begin{array}{l|l} m_\sigma(\mathrm{PU}) = (0, 1 \cdot \mathrm{PU}) & \mathrm{LTV}_\sigma(\mathrm{PU}) = 0 + \gamma \cdot \mathrm{LTV}_\sigma(\mathrm{PU}) \\ m_\sigma(\mathrm{PF}) = (0, \frac{1}{2} \cdot \mathrm{PU} + \frac{1}{2} \cdot \mathrm{RF}) & \mathrm{LTV}_\sigma(\mathrm{PF}) = 0 + \gamma \cdot (\frac{1}{2} \cdot \mathrm{LTV}_\sigma(\mathrm{PU}) + \frac{1}{2} \cdot \mathrm{LTV}_\sigma(\mathrm{RF})) \\ m_\sigma(\mathrm{RU}) = (10, \frac{1}{2} \cdot \mathrm{PU} + \frac{1}{2} \cdot \mathrm{RU}) & \mathrm{LTV}_\sigma(\mathrm{RU}) = 10 + \gamma \cdot (\frac{1}{2} \cdot \mathrm{LTV}_\sigma(\mathrm{PU}) + \frac{1}{2} \cdot \mathrm{LTV}_\sigma(\mathrm{RU})) \\ m_\sigma(\mathrm{RF}) = (10, \frac{1}{2} \cdot \mathrm{RU} + \frac{1}{2} \cdot \mathrm{RF}) & \mathrm{LTV}_\sigma(\mathrm{RF}) = 10 + \gamma \cdot (\frac{1}{2} \cdot \mathrm{LTV}_\sigma(\mathrm{RU}) + \frac{1}{2} \cdot \mathrm{LTV}_\sigma(\mathrm{RF})) \end{array} \tag{5}$$

Solving this linear system, we get $\mathrm{LTV}_\sigma(\mathrm{PU}) = 0$, $\mathrm{LTV}_\sigma(\mathrm{PF}) = 14.876$, $\mathrm{LTV}_\sigma(\mathrm{RU}) = 18.182$, and $\mathrm{LTV}_\sigma(\mathrm{RF}) = 33.058$.

The long-term value induces an ordering on policies: $\sigma \leq \tau$ if $\mathrm{LTV}_\sigma \leq \mathrm{LTV}_\tau$ in the pointwise order on $\mathbb{R}^S$. It is a classic result [23, Theorem 6.2.7] that for our simple model of MDPs with discounting, the best stationary, memoryless, deterministic policy is as good as any policy. In other words, one cannot do better by allowing time-dependence, memory, or randomization in policies. This result is also the theoretical basic for finding optimal policies via policy iteration, as we describe further below.

Before we move on to policy iteration, we recall the notion of optimal value function. 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 [23]:

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

We note that a transition structure $t\colon S \to (\Delta S)^{Act}$ corresponds to an Act-indexed set of maps $t_a\colon S \to \Delta S$, $a \in Act$, each of which in turn corresponds to a column-stochastic $|S| \times |S|$-matrix. It is an important classic result that V^* is the unique (bounded) map $v\colon S \to \mathbb{R}$ that satisfies *Bellman's optimality equation* [4,23]:

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

In other words, V^* is a fixpoint of the (non-linear) contractive, monotone *Bellman operator*, given by

$$\Psi^*\colon \mathbb{R}^S \to \mathbb{R}^S \qquad \Psi^*(v) = u + \gamma \cdot \max_{a \in Act} \{t_a v\},$$

where the maximum is taken in the pointwise order on $\mathbb{R}^S$.

3 Policy Improvement via Contraction Coinduction

3.1 Policy Iteration

The optimality equation together with the abovementioned result that an optimal policy may be found among the stationary, deterministic policies is the basis for an effective algorithm for finding optimal policies, known as *policy iteration* [12]. The algorithm starts from any policy $\sigma \in Act^S$, and iteratively improves σ to some τ 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 policy improvement step of the algorithm is obtained via the following defintion.

Definition 3 (Improved Policy). *A policy τ is called an* improvement *of a policy σ if for all $s \in S$ it holds that*

$$\tau(s) = argmax_{a \in Act}\{\ell_\sigma(t_a(s))\}. \tag{6}$$

Informally, $\tau(s)$ is an action a that maximizes the expected future rewards obtained by doing a now, and then continuing with σ. However, it is not prima facie clear that τ is an improvement, since following τ means to also "continue with τ" (not with σ). Proving that $\sigma \leq \tau$ is the content of the Policy Improvement Theorem (Theorem 2) which we prove below.

Note that improved policies need not be unique, because there could be different actions a that maximize $\ell_\sigma(t_a(s))$, but (6) describes a procedure for improving a policy σ assuming that LTV_σ has been computed (e.g., by solving the associated linear system).

Example 3. We return to Example 1 and to the "miserly" policy σ in Example 2. To determine $\tau(\mathrm{PF})$ where τ is an improved policy, we compare

$$\ell_\sigma(t(\mathrm{PF})(\mathbf{S})) = \mathrm{E} \circ \Delta\, \mathrm{LTV}_\sigma \left(\frac{1}{2}\,\mathrm{PU} + \frac{1}{2}\,\mathrm{RF}\right) = \left(\frac{1}{2} \cdot 0\right) + \left(\frac{1}{2} \cdot 33.058\right)$$

and

$$\ell_\sigma(t(\mathrm{PF})(\mathbf{A})) = \mathrm{E} \circ \Delta\, \mathrm{LTV}_\sigma(1 \cdot \mathrm{PF}) = 1 \cdot 14.876.$$

Since the latter is smaller, we have $\tau(\mathrm{PF}) = \mathbf{S}$.

Clasically, policy improvement is proved [12,23] using that $(I - \gamma P_\sigma)^{-1}$ is a monotone operator. This in turn follows from the matrix $(I - \gamma P_\sigma)^{-1}$ having only non-negative entries, a property which we show in Example 5 below using contraction coinduction.

3.2 The Contraction Coinduction Principle

We now introduce the contraction coinduction principle. We only assume basic knowledge of metric spaces, as can be found in, e.g., [21]. Here we just recall a few basic definitions and fix notation. A *metric space* (X, d_X) is a set X equipped

with a metric $d_X \colon X \to \mathbb{R}$. Sometimes the metric is left implicit and we simply refer to the metric space X. We always assume the standard Euclidean metric on the set of real numbers $\mathbb{R}$. Any set X can be equipped with the discrete metric, given by $d_X(x, y) = 1$ if $x \neq y$, and $d_X(x, y) = 0$ if $x = y$, for all $x, y \in X$.

A function $f \colon X \to Y$ between metric spaces is *bounded* if there is a real number C such that for all $x, y \in X$, it holds that $d_Y(f(x), f(y)) \leq C$. We write $B(X, Y)$ for the set of all bounded $f \colon X \to Y$. The set $B(X, Y)$ can be equipped with the *supremum metric*: for all $f, g \in B(X, Y)$,

$$d(f, g) = \sup\{d_Y(f(x), g(x)) \mid x \in X\}. \tag{7}$$

When Y is a complete space (for example, $\mathbb{R}$), so is $B(X, Y)$. We recall that a function $f \colon X \to X$ on a metric space X is *contractive* if there is a $C < 1$ such that for all $x_1, x_2 \in X$, we have $d_X(f(x_1), f(x_2)) \leq C \cdot d_X(x_1, x_2)$. A *fixpoint* of f is an element x^* such that $f(x^*) = x^*$.

The contraction coinduction principle is a variation of the classic Banach Fixpoint Theorem, asserting that any contractive mapping has a unique fixpoint. We need a version of this theorem which, in addition to a complete metric, also has an order. For this reason we introduce the following definition.

Definition 4. *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.*[3]

Example 4. For any set X, $B(X, \mathbb{R})$ with the pointwise order (and supremum metric, as in (7) above) is a complete ordered metric space.

We can now state contraction (co)induction. It will lead to elegant proofs of order statements concerning fixpoints, as we shall see below.

Theorem 1 (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 $x \leq x^*$).*

Proof. We only verify the first assertion; the second is similar. Suppose that $f(x) \leq x$. By induction on n and monotonicity of f, we have for all $n \geq 0$, $f^n(x) \leq x$. Since f is contractive, the proof of the Banach Fixpoint Theorem shows that $\{f^n(x)\}_n$ is a convergent sequence. But $\{z \mid z \leq x\}$ is closed and contains this sequence, so $\lim_n f^n(x) \leq x$. The proof of the Banach Fixpoint Theorem also shows that $\lim_n f^n(x)$ equals the fixpoint x^*. Thus, $x^* \leq x$. □

[3] We could weaken the partial order in the definition of an ordered metric space to a transitive relation. However, our aim is not the highest level of generality. Rather, we see contraction (co)induction as an instance of metric coinduction (see Remark 1) that suffices to prove interesting results about MDPs.

Remark 1. Theorem 1 follows from the Metric Coinduction Principle [17,24]. E.g., to derive contraction induction, let x be such that $f(x) \leq x$. The set $A = \{y \in X \mid y \leq x\}$ is non-empty, since $x \in A$, and closed by our assumption that X is an ordered metric space. Moreover, $f[A] \subseteq A$ by monotonicity. Hence by metric coinduction, $x^* \in A$.

Example 5. We recover a fact that comes up frequently in the area (e.g., [12] uses it to justify policy improvement): if P is a column-stochastic $n \times n$ matrix, then $(I - \gamma P)^{-1}$ has all non-negative entries. We do not really need this fact, and we mainly mention it to point out that contraction coinduction might streamline the proofs of known results. To see this, let $M = \mathbb{R}^{n \times n}$. We order M pointwise, and as metric we use $d(X, Y) = \|X - Y\|$, where $\|X\| = \max_{i,j} |x_{i,j}|$ (so that $\|PX\| \leq \|X\|$). This gives a complete ordered metric space. Let $f \colon M \to M$ be $f(X) = I + (\gamma P)X$. Easily, f is a monotone contraction, and its fixpoint is $(I - \gamma P)^{-1}$. Note that $f(0) \geq 0$, where 0 is the zero matrix. By contraction coinduction, we conclude that $(I - \gamma P)^{-1} \geq 0$.

We now give our proof of policy improvement using contraction coinduction.

Theorem 2 (Policy Improvement). *Let an MDP be given by $t \colon S \to (\Delta S)^{Act}$ and $u \colon S \to \mathbb{R}$. Let σ and τ 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 $\mathrm{LTV}_\tau \leq \mathrm{LTV}_\sigma$.*

Proof. Assuming that $\ell_\sigma \circ t_\tau \geq \ell_\sigma \circ t_\sigma$, we have for all $s \in S$,

$$u(s) + \gamma \sum_{s' \in S} t_\sigma(s)(s') \cdot \mathrm{LTV}_\sigma(s') \leq u(s) + \gamma \sum_{s' \in S} t_\tau(s)(s') \cdot \mathrm{LTV}_\sigma(s').$$

Since LTV_σ and LTV_τ are the fixpoints of the contractive, monotone operators Ψ_σ and Ψ_τ, respectively (cf. (4)), the above inequality may be recast to say that $\Psi_\tau(\mathrm{LTV}_\sigma) \geq \Psi_\sigma(\mathrm{LTV}_\sigma) = \mathrm{LTV}_\sigma$. By contraction coinduction (Theorem 1), $\mathrm{LTV}_\tau \geq \mathrm{LTV}_\sigma$. This completes the proof of the first assertion. The second one is proved similarly. □

Next, we use contraction coinduction to show the classic result that V^* is an upper bound for the long-term value of all policies, and that a so-called *greedy policy* is optimal [4,23]. Lemma 1 below is standard, and essentially the same proof as ours appears as Lemma 5.2 in Kozen and Ruozzi [24].

Lemma 1. *For all policies σ, $\mathrm{LTV}_\sigma \leq V^*$.*

Proof. A straightforward calculation and monotonicity argument shows that for all $f \in B(S, \mathbb{R})$, $\Psi_\sigma(f) \leq \Psi^*(f)$. In particular, $\mathrm{LTV}_\sigma = \Psi_\sigma(\mathrm{LTV}_\sigma) \leq \Psi^*(\mathrm{LTV}_\sigma)$. By contraction coinduction we conclude that $\mathrm{LTV}_\sigma \leq V^*$. □

Define the *greedy policy* $\sigma^* \colon S \to Act$ by

$$\sigma^*(s) = \mathrm{argmax}_{a \in Act} \left\{ \sum_{s' \in S} t_a(s)(s') \cdot V^*(s') \right\}. \tag{8}$$

Due to ties, this policy is strictly speaking not unique. But following standard usage, we speak of "the" greedy policy when we mean "a" greedy one.

Proposition 1. *The greedy policy is optimal. That is,* $\mathrm{LTV}_{\sigma^*} = V^*$.

Proof. Observe that $\Psi_{\sigma^*}(V^*) \geq V^*$ (in fact, equality holds). By contraction coinduction, $V^* \leq \mathrm{LTV}_{\sigma^*}$. The other direction follows from Lemma 1. □

It is a direct consequence of Proposition 1 that the optimal value is attained by some policy.

4 Coalgebras and Algebras for MDPs

In this section, we present the coalgebraic and algebraic structures that we use to model MDPs and their long-term values. We assume the reader is familiar with basic notions in coalgebra [25] and category theory [18], but we briefly recall some definitions and results (see, e.g., [3,13,16]) related to monads and distributive laws that are needed for this paper.

4.1 Algebras, Monads, and Distributive Laws

Given a functor $T\colon \mathsf{C} \to \mathsf{C}$ on a category C, a *T-algebra* is a pair (A, α) where A is a C-object and $\alpha\colon TA \to A$ is a C-arrow. A *T-algebra homomorphism* from (A, α) to (B, β) is a C-arrow $f\colon A \to B$ such that $f \circ \alpha = \beta \circ Tf$.

A *monad* (on C) is a triple (T, η, μ) where T is a C-endofunctor, and $\eta\colon \mathrm{Id} \Rightarrow T$ and $\mu\colon TT \Rightarrow T$ are natural transformations such that $\mu \circ T\eta = \mathsf{id} = \mu \circ \eta_T$ and $\mu \circ \mu_T = \mu \circ T\mu$. Given a monad (T, η, μ), an *Eilenberg-Moore T-algebra* is a T-algebra (A, ω) such that $\omega \circ \eta_A = \mathsf{id}$ and $\omega \circ \mu_A = \omega \circ T\omega$. We denote the category of Eilenberg-Moore T-algebras and T-algebra homomorphisms by $\mathcal{EM}(T)$. Note that (TX, μ_X) is an Eilenberg-Moore T-algebra.

Example 6. The well-known *distribution monad* is the discrete variant of the *Giry monad* [11]. The functor part $\Delta\colon \mathsf{Set} \to \mathsf{Set}$ maps a set X to the finitely supported probability distributions on X:

$$\begin{aligned} \Delta X &= \{\varphi\colon X \to [0,1] \mid \mathrm{supp}(\varphi) \text{ is finite and } \textstyle\sum_x \varphi(x) = 1\}, \\ (\Delta f)(\varphi)(y) &= \textstyle\sum_{x \in f^{-1}(y)} \varphi(x) \qquad \text{for all } f\colon X \to Y. \end{aligned}$$

It is sometimes convenient to write an element φ of ΔX as a formal linear combination $\varphi = r_1 x_1 + \cdots + r_n x_n$, where $\mathrm{supp}(\varphi) = \{x_1, \ldots, x_n\}$ and $\varphi(x_i) = r_i$, or also $\varphi = \sum_{x \in X} \varphi(x)\, x$. In this notation, $(\Delta f)(\varphi) = r_1 f(x_1) + \cdots + r_n f(x_n)$ for $f\colon X \to Y$, where coefficients of identical $f(x_i)$-values are summed implicitly. Equivalently stated, we have

$$(\Delta f)(\varphi) = \sum_{x \in X} f(x)\varphi(x). \tag{9}$$

The unit $\delta\colon \mathrm{Id} \Rightarrow \Delta$ is $\delta_X(x) = 1x$ (the Dirac distribution at x), and the multiplication $\mu\colon \Delta\Delta \Rightarrow \Delta$ is given as follows. For $\psi = r_1\varphi_1 + \cdots + r_n\varphi_n \in \Delta\Delta X$, we have $\mu_X(\psi)(x) = \sum_{i=1}^n r_i \varphi_i(x)$, i.e., $\mu_X(\psi) = \sum_{x \in X} \big(\sum_{\varphi \in \Delta X} \psi(\varphi) \cdot \varphi(x)\big)x$.

The category $\mathcal{EM}(\Delta)$ is also known as the category of convex sets and affine (or linear) maps, since an Eilenberg-Moore Δ-algebra can be seen as a set X in which convex combinations $r_1x_1 + \cdots + r_nx_n$ can be evaluated.

Let (T, η, μ) be a monad and F an endofunctor, both on C. A *distributive law of* (T, η, μ) *over* F is a natural transformation $\lambda\colon TF \Rightarrow FT$ that is compatible with the monad structure, meaning that $\lambda \circ \eta_F = F\eta$ and $\lambda \circ \mu_F = F\mu \circ \lambda_T \circ T\lambda$. We recall (see, e.g., [15,16]) that such a distributive law corresponds to a lifting F_λ of F to the category $\mathcal{EM}(T)$, and equivalently to a lifting T_λ of T to the category $\mathrm{Coalg}_{\mathsf{C}}(F)$ of F-coalgebras. The functors F_λ and T_λ are defined as follows:

$$\begin{aligned} F_\lambda(A, \omega\colon TA \to A) &= (FA, F\omega \circ \lambda_A) & F_\lambda(f) &= Ff, \\ T_\lambda(B, \beta\colon B \to FB) &= (TB, \lambda_B \circ T\beta) & T_\lambda(f) &= Tf. \end{aligned}$$

We also recall (cf. [3,14]) that such a distributive law induces an operation $(-)^\sharp\colon \mathrm{Coalg}_{\mathsf{C}}(FT) \to \mathrm{Coalg}_{\mathcal{EM}(T)}(F_\lambda)$, which is often referred to as an abstract form of *determinization* (cf. [26,14]). For every FT-coalgebra $c\colon X \to FTX$, $c^\sharp$ is defined as

$$c^\sharp = F\mu_X \circ \lambda_{TX} \circ Tc\colon (TX, \mu_X) \to F_\lambda(TX, \mu_X), \text{ and we have } c^\sharp \circ \eta_X = c. \quad (10)$$

Determinization $(-)^\sharp$ is a functor, but we shall not use this fact. Note that the underlying F-coalgebra of $c^\sharp$ is of type $TX \to FTX$.

We write $\mathrm{E}\colon \Delta\mathbb{R} \to \mathbb{R}$ for the map that computes expected value. That is, viewing an element $\varphi \in \Delta\mathbb{R}$ as a formal linear combination, E evaluates φ by interpreting the formal expression in $\mathbb{R}$, i.e., $\mathrm{E}(\varphi) = \sum_{x\in\mathbb{R}} \varphi(x) \cdot x$.

Note that for $f\colon X \to \mathbb{R}$, by (9) we have, for all $\varphi \in \Delta X$, that

$$\mathrm{E}((\Delta f)(\varphi)) = \sum_{x\in X} f(x) \cdot \varphi(x). \quad (11)$$

Lemma 2. *The expected value* $\mathrm{E}\colon \Delta\mathbb{R} \to \mathbb{R}$ *is an Eilenberg-Moore* Δ*-algebra:* $\mathrm{E} \circ \delta_{\mathbb{R}} = \mathsf{id}_{\mathbb{R}}$ *and* $\mathrm{E} \circ \Delta\mathrm{E} = \mathrm{E} \circ \mu_{\mathbb{R}}$.

4.2 Coalgebraic Modeling of MDPs

As we saw in Definition 2, long-term values arise by summing infinite sequences (or streams) of real numbers. It is well-known [25] that such streams form a final coalgebra for the Set-endofunctor $H = \mathbb{R} \times \mathrm{Id}$. The final H-coalgebra structure is given by mapping a stream $x = (x_0, x_1, x_2, \ldots)$ to $(\mathsf{head}(x), \mathsf{tail}(x))$, where $\mathsf{head}(x) = x_0$ and $\mathsf{tail}(x) = (x_1, x_2, \ldots)$.

Given an MDP $m = \langle u, t\rangle$ and a policy $\sigma\colon S \to Act$, the resulting Markov reward process $m_\sigma = \langle u, t_\sigma\rangle$ is easily seen to be an $H\Delta$-coalgebra

$$m_\sigma = \langle u, t_\sigma\rangle\colon S \to \mathbb{R} \times \Delta S,$$

where, as we recall from from Section 2, $t_\sigma(s) = t(s)(\sigma(s))$.

Similarly, it is not hard to see that an MDP $m = \langle u, t\rangle$ is a $K\Delta$-coalgebra $\langle u, t\rangle \colon S \to \mathbb{R} \times (\Delta S)^{Act}$, where $K = H \circ (-)^{Act}$ and $(-)^{Act}$ is the covariant hom-functor.

Since $\mathrm{E} \colon \Delta\mathbb{R} \to \mathbb{R}$ is an Eilenberg-Moore Δ-algebra, there is a distributive law χ of (Δ, δ, μ) over H (cf. [13]) specified by

$$\chi_X \colon \Delta(\mathbb{R} \times X) \xrightarrow{\langle \Delta\pi_1, \Delta\pi_2\rangle} \Delta\mathbb{R} \times \Delta X \xrightarrow{\mathrm{E}\times\mathsf{id}} \mathbb{R} \times \Delta X,$$

i.e.,

$$\chi_X = \langle \mathrm{E} \circ \Delta\pi_1, \Delta\pi_2\rangle. \tag{12}$$

The lifted functor $H_\chi \colon \mathcal{EM}(\Delta) \to \mathcal{EM}(\Delta)$ is concretely given as

$$\begin{aligned} H_\chi(A, \omega) &= (\mathbb{R} \times A, (\mathbb{R} \times \omega) \circ \langle \mathrm{E} \circ \Delta\pi_1, \Delta\pi_2\rangle) \\ &= (\mathbb{R} \times A, \langle \mathrm{E} \circ \Delta\pi_1, \omega \circ \Delta\pi_2\rangle). \end{aligned}$$

Using the distributive law χ from (12), the determinization $m^\sharp_\sigma \colon \Delta S \to \mathbb{R} \times \Delta S$ is given for each $\varphi \in \Delta S$ by

$$\begin{aligned} m^\sharp_\sigma(\varphi) &= ((\mathrm{E} \circ \Delta u)(\varphi), (\mu_S \circ \Delta t_\sigma)(\varphi)) \\ &= (\textstyle\sum_{s\in S} u(s) \cdot \varphi(s), s \mapsto \sum_{s'\in S} t_\sigma(s)(s') \cdot \varphi(s')). \end{aligned}$$

Considering φ as a probabilistic state, the first component of the pair $m^\sharp_\sigma(\varphi)$ is the expected reward given φ, and the second component is the expected next probabilistic state. The morphism $\mu_S \circ \Delta t_\sigma \colon \Delta S \to \Delta S$ is the Kleisli extension of $t_\sigma \colon S \to \Delta S$, which can be seen as a column-stochastic $|S| \times |S|$-matrix. Viewing $u \in \mathbb{R}^S$ as a row $|S|$-vector and a distribution $\varphi \in \Delta S$ as a column-stochastic $|S|$-vector, we have that $m^\sharp_\sigma(\varphi) = \langle u\varphi, t_\sigma\varphi\rangle$, where juxtaposition denotes matrix-vector multiplication. The unique H-coalgebra morphism from $m^\sharp_\sigma$ to the final H-coalgebra of streams maps a distribution φ to the stream of expected rewards $(u\varphi, ut_\sigma\varphi, ut^2_\sigma\varphi, \ldots)$.

The distributive law given by χ is an $\mathcal{EM}$-law in the terminology of [14], where determinization was studied for the purpose of obtaining trace semantics. The trace semantics of $m_\sigma \colon S \to \mathbb{R} \times \Delta S$ is the function that maps a state s to the stream of expected rewards $(r^\sigma_0(s), r^\sigma_1(s), r^\sigma_2(s), \ldots)$ from (1).

4.3 Algebraic Modeling of Discounted Sums

The long-term value of a policy σ in state s is the discounted infinite sum of the stream $\rho(s) = (r^\sigma_0(s), r^\sigma_1(s), r^\sigma_2(s), \ldots)$ of expected rewards. Due to S being finite, the values in this stream are bounded, which ensures that the discounted sum converges. A leading observation of this paper is that we can re-express the recursive equation (3) for LTV_σ by saying that $\mathrm{LTV}_\sigma \colon S \to \mathbb{R}$ makes the following diagram commute:

$$\begin{array}{ccccc} S & & \xrightarrow{\quad m_\sigma \quad} & & \mathbb{R} \times \Delta S \\ {\scriptstyle \mathrm{LTV}_\sigma}\downarrow & & & & \downarrow{\scriptstyle \mathbb{R}\times\Delta(\mathrm{LTV}_\sigma)} \\ \mathbb{R} & \xleftarrow[\alpha_\gamma]{} & \mathbb{R} \times \mathbb{R} & \xleftarrow[\mathbb{R}\times\mathrm{E}]{} & \mathbb{R} \times \Delta\mathbb{R} \end{array} \tag{13}$$

Here, $\alpha_\gamma \colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is the H-algebra

$$\alpha_\gamma \colon H\mathbb{R} \to \mathbb{R} \qquad \alpha_\gamma(x, y) = x + \gamma \cdot y. \tag{14}$$

Notice that LTV_σ 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* [6]: 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? As suggested by the previous discussion, problems can arise if the reward values in f are unbounded. But the question can be framed in an even more basic way. Namely, by [5, Theorem 19], $\alpha_\gamma \circ (\mathbb{R} \times \mathrm{E})$ is a corecursive algebra for $H\Delta$ if and only if α_γ is a corecursive algebra for H. But the latter is not the case. Consider an infinite system of equations

$$x_n = a_n + \gamma \cdot x_{n+1}, \qquad n = 0, 1, \ldots, \tag{15}$$

where a_n are fixed real numbers. Then (15) corresponds uniquely to a H-coalgebra $g \colon X \to \mathbb{R} \times X$. Solutions to (15) in turn correspond to maps $g^\dagger$ such that $g^\dagger = \alpha_\gamma \circ (\mathbb{R} \times g^\dagger) \circ g$, i.e., to coalgebra-to-algebra maps from (X, g) to $(\mathbb{R}, \alpha_\gamma)$. The reason why α_γ is not a corecursive algebra is that (15) always has continuum many solutions. Namely, the solution value for x_0 may be chosen arbitrarily, and the rest are determined from it. Note however, if $(a_n)_n$ is unbounded then all solutions are unbounded. [To see this, let $(b_n)_n$ be a solution. We have: $|a_n| > 2K$ $\Rightarrow |b_{n+1}| = |a_n + \gamma \cdot b_n| \geq |a_n| - \gamma|b_n| \;\Rightarrow\; |b_{n+1}| + \gamma|b_n| > 2K \Rightarrow |b_n| > K$ or $|b_{n+1}| > K$. Hence, for each K there is some i such that $|b_i| > K$.]

If, on the other hand, $(a_n)_n$ is bounded, then there is a *unique bounded solution* to (15), namely $x_n = \sum_{i=0}^{\infty} \gamma^i \cdot a_{n+i}$ for all n. Boundedness is used in asserting that the sum converges, and the detailed verification that this solution works and is unique follows from Proposition 6 below. In summary, uniqueness is only obtained by restricting to bounded solutions.

We end this section by noting that α_γ is an algebra for the lifted functor H_χ, essentially because α_γ is affine. We will need this result in Section 5.3.

Lemma 3. *$((\mathbb{R}, \mathrm{E}), \alpha_\gamma)$ is an H_χ-algebra in $\mathcal{EM}(\Delta)$, that is, we have the equality* $\mathrm{E} \circ \Delta\alpha_\gamma = \alpha_\gamma \circ \langle \mathrm{E} \circ \Delta\pi_1, \mathrm{E} \circ \Delta\pi_2 \rangle$.

5 Long-Term Values via *b*-Corecursive Algebras

In this section, we will develop some categorical notions in order to capture boundedness properties, and eventually show that long-term values can be characterized via a universal property of a notion of corecursive algebra for bounded maps.

5.1 MDPs in Metric Spaces

The first step is to identify the appropriate category of metric spaces. There are several types of functions on metric spaces that are of interest. In this

paper, we shall consider the following. Let (X, d_X) and (Y, d_Y) be metric spaces and $f\colon X \to Y$ a function (not necessarily continuous). Then f is said to be *Lipschitz* if $d_Y(f(x_1), f(x_2)) \leq C \cdot d_X(x_1, x_2)$ for all $x_1, x_2 \in X$, for some fixed real number C. A Lipschitz function that satisfies the above inequality for $C = 1$ is called *non-expansive* (or *short*). It is clear that the composition of two Lipschitz functions is again Lipschitz, and the composition of non-expansive functions again non-expansive.

Bounded functions need not be Lipschitz, and vice versa. Although bounded maps are of particular interest to us, we point out the fact that metric spaces with bounded maps do not form a category, since the identity on a space of infinite diameter is not bounded. Our main interest in Lipschitz functions is that if g is bounded and f is Lipschitz, then $f \circ g$ is bounded; also, they are used in the Kantorovich metric just below.

We write Met for the category that has metric spaces as objects and all functions as arrows. Usually, the morphisms of metric spaces are taken to be the non-expanding functions or continuous functions. The reason we take *all* set functions is that we are going to use the metric structure only in connection with *boundedness*, and so our (non-standard) choice will become more sensible. (In Section 6.2, we hint that with additional results we can indeed work with a"real" metric-type category, the Polish metric spaces.)

We lift our Set-endofunctors H and Δ to Met using the maximum and Kantorovich(-Wasserstein) metrics (cf. [29,2]). This last metric is usually defined in the measure-theoretic setting, so discrete probability measures are a special case.

Definition 5 (Product and Kantorovich Metrics). *Let (X, d_X) and (Y, d_Y) be metric spaces.*

- *The product $(X, d_X) \times (Y, d_Y) = (X \times Y, d_X \times d_Y)$ has the maximum metric*
$$(d_X \times d_Y)((x_1, y_1), (x_2, y_2)) = \max\{d_X(x_1, x_2), d_Y(y_1, y_2)\}.$$
- *The Kantorovich lifting of d_X is the metric $d_{\Delta X}$ on ΔX, defined by*
$$d_{\Delta X}(\varphi, \psi) = \sup\{d_{\mathbb{R}}((\mathrm{E}\circ\Delta f)(\varphi), (\mathrm{E}\circ\Delta f)(\psi)) \mid f\colon X \to \mathbb{R} \text{ is non-expansive}\}.$$

Remark 2. See [10] for ten choices for the metric d on ΔX. Incidentally, very little is known concerning the question of whether each d leads to a functor on the category of *all* metric spaces and continuous functions. However, it follows from Theorem 1 of [11] that for the related category of Polish spaces, the Kantorovich lifting does lead to a functor.

We can view a Markow reward process $\langle u, t_\sigma\rangle\colon S \to \mathbb{R} \times \Delta S$ as a coalgebra in Met for the lifted functor $H\Delta$, by equipping the state space S with a metric. (The discrete metric is the canonical choice, but any metric will do.)

The next lemma will be frequently used in Section 5.2 to prove boundedness preservation properties.

Lemma 4. *If $f\colon X \to Y$ is Lipschitz, so is Hf. If $f\colon X \to Y$ is Lipschitz with constant C, then so is Δf.*

5.2 Categorical Structure for Bounded Maps

This section aims at a sparse categorification of boundedness that will permit us to re-work the notion of a corecursive algebra to a *bounded corecursive algebra* in Section 5.3 below. To this aim, we introduce the notion of b-category and related concepts.

Definition 6. *Let* C *be a category and* $\mathcal{B}$ *a class of morphisms in* C*. We call* $\mathcal{B}$ *a* b-structure[4] *on* C *if for all* $f \in \mathcal{B}$ *and all arrows* g *in* C*, if* $f \circ g$ *is defined, then* $f \circ g \in \mathcal{B}$*. A* b-category *is a pair* $(\mathsf{C}, \mathcal{B})$*, where* C *is a category and* $\mathcal{B}$ *is a* b*-structure on* C*. We frequently call a morphism* $f \in \mathcal{B}$ *a* $\mathcal{B}$-morphism*. We denote the collection of all* C*-morphisms* $X \to Y$ *that are also in* $\mathcal{B}$ *by* $\mathcal{B}(X, Y)$*.*

The key feature of Lipschitz and bounded functions for our purposes is captured in the following definition.

Definition 7. *We say that a* C*-arrow* f preserves $\mathcal{B}$ *if whenever* $g \in \mathcal{B}$ *and* $f \circ g$ *is defined, then* $f \circ g \in \mathcal{B}$*.*

It is easy to see that for every category C, $(\mathsf{C}, \mathcal{M})$ is a b-category, where $\mathcal{M}$ is the collection of morphisms of C. If $(\mathsf{C}, \mathcal{B})$ is a b-category, then every morphism in $\mathcal{B}$ preserves $\mathcal{B}$.

Example 7. Our primary example of a b-category is (Met, B), where Met is the category of metric spaces and all functions, and B is the collection of bounded maps of metric spaces. While the metric structure is not used in the Met-morphisms, it figures in the b-structure.

Every Lipschitz function preserves B. For any metric spaces X_1 and X_2, the projections $\pi_i \colon X_1 \times X_2 \to X_i$ preserve B. The algebras $\mathrm{E} \colon \Delta\mathbb{R} \to \mathbb{R}$ and α_γ, from (14), both preserve B.

Next, we formulate definitions of functors and natural transformations which incorporate b-structures. The main motivation for the definitions below are the examples which follow and also the properties that we shall see at the end of this section, in Proposition 5 and Example 8.

Definition 8. *Let* $(\mathsf{C}, \mathcal{B})$ *and* $(\mathsf{C}', \mathcal{B}')$ *be* b*-categories. A functor* $F \colon \mathsf{C} \to \mathsf{C}'$ *is a* b-functor*, written* $F \colon (\mathsf{C}, \mathcal{B}) \to (\mathsf{C}', \mathcal{B}')$*, if whenever* $f \in \mathcal{B}$*, then* Ff *preserves* $\mathcal{B}'$*; and* F *is a* strong b-functor *if whenever* $f \in \mathcal{B}$*, then* $Ff \in \mathcal{B}'$*.*

If $F, G \colon \mathsf{C} \to \mathsf{C}'$ *are functors (not necessarily* b*-functors), then a* b-natural transformation $\sigma \colon F \Rightarrow G$ *is a natural transformation in the usual sense such that every component* σ_X *preserves* $\mathcal{B}'$*.*

Proposition 2. *(1) Constant functors are* b*-functors. (2) The identity on a* b*-category is a* b*-endofunctor. (3) If* F *is a strong* b*-functor, then* F *is a* b*-functor.*

[4] During CMCS 2018, we learned from Henning Urbat that a b-structure is also known as a *sieve*. We currently do not know how to put this fact to use.

We now investigate how the functor H, monad (Δ, δ, μ), and distributive law χ interact with the b-structure B of bounded maps on Met.

Proposition 3. *$H\colon \mathsf{Met} \to \mathsf{Met}$ is a b-endofunctor, but not a strong b-endofunctor, on (Met, B).*

Proposition 4. *$\Delta\colon \mathsf{Met} \to \mathsf{Met}$ is a strong b-endofunctor on (Met, B).*

Lemma 5. *If $f\colon X \to Y$ preserves B, then so does $Hf\colon \mathbb{R} \times X \to \mathbb{R} \times Y$.*

Lemma 6. *For all metric spaces (X, d_X), the following hold.*

1. *δ_X is an isometric embedding.*
2. *μ_X is non-expanding.*
3. *χ_X is Lipschitz.*

It follows that δ, μ, and χ are b-natural transformations in (Met, B).

One crucial observation is that if the $H\Delta$-coalgebra m_σ obtained from an MDP m and a policy σ is bounded, then so is the determinized H-coalgebra $m_\sigma^\sharp$. The following proposition shows that our setup of b-structures ensures that this holds abstractly.

Proposition 5. *Let λ be a distributive law of monad (T, η, μ) over a functor F such that T is a strong b-endofunctor, and $F\mu$ and λ are b-natural transformations. Then $\mathcal{B}$ is closed under $(-)^\sharp$, i.e., if $c \in \mathcal{B}$ then $c^\sharp \in \mathcal{B}$.*

Proof. This follows instantly from $c^\sharp = F\mu_X \circ \lambda_{TX} \circ Tc$ (cf. Equation (10)). □

Example 8. For our running example for MDPs where $F = H$, $T = \Delta$, and $\lambda = \chi$ is given by (12), we have the conclusion of Proposition 5 in the b-category (Met, B). Indeed, by Proposition 4, Δ is a strong b-endofunctor. By Lemma 6, χ is b-natural. Finally, by the second part of Lemma 4 and Lemma 6 (2), $H\mu_X$ is Lipschitz for every X, and thus preserves bounded maps. Therefore, $H\mu$ is b-natural.

5.3 *b*-Corecursive Algebras (bcas)

As we explained in Section 4.3, the long-term value map LTV_σ is a certain coalgebra-to-algebra morphism, i.e., it is a solution to a set of recursive equations, but it is only uniquely defined if we restrict to bounded maps. The following notion of b-corecursive algebra categorifies this observation.

Definition 9. *Let $(\mathsf{C}, \mathcal{B})$ be a b-category, F an endofunctor on C (not necessarily a b-endofunctor), and $\beta\colon FA \to A$ an F-algebra. Then β is a* b-corecursive algebra (bca) *if for every F-coalgebra $f\colon X \to FX$ with $f \in \mathcal{B}$, there is a unique* solution map *$f^\dagger \in \mathcal{B}$ such that the diagram*

$$\begin{array}{ccc} X & \xrightarrow{\ f\ } & FX \\ {\scriptstyle f^\dagger}\downarrow & & \downarrow{\scriptstyle Ff^\dagger} \\ A & \xleftarrow{\ \beta\ } & FA \end{array}$$

commutes, or equivalently stated, such that $f^\dagger$ is the fixed point of the operator $\Phi_{f,\beta}\colon \mathsf{C}(X,A) \to \mathsf{C}(X,A)$, defined for all $j \in \mathsf{C}(X,A)$ by $\Phi_{f,\beta}(j) = \beta \circ Fj \circ f$.

We note that a *corecursive algebra* [6] is a bca with $\mathcal{B}$ the family of *all* morphisms in the underlying category.

Remark 3. A corecursive algebra is a special kind of *completely iterative algebra* (also called cias, see Milius [19]). With the obvious definition, the examples in this paper would be *b-cias*. Alas, we have not found any application of this fact.

The next lemma uses the b-category concepts to give conditions that ensure the operator $\Phi_{f,\beta}$ from Definition 9 restricts to $\mathcal{B}$-morphisms. This is the abstract analogue of showing that the Bellman operator maps bounded maps to bounded maps.

Lemma 7. *Let $(\mathsf{C},\mathcal{B})$ be a b-category. If F is a b-endofunctor on $(\mathsf{C},\mathcal{B})$, and $\beta\colon FA \to A$ is an F-algebra that preserves $\mathcal{B}$, it holds that for every F-coalgebra $f\colon X \to FX$ in $\mathcal{B}$, the operator $\Phi_{f,\beta}\colon \mathsf{C}(X,A) \to \mathsf{C}(X,A)$ from Definition 9 restricts to an operator $\Phi_{f,\beta}\colon \mathcal{B}(X,A) \to \mathcal{B}(X,A)$.*

Proof. Let $j \in \mathcal{B}(X,A)$. Since F is assumed to be an b-endofunctor, Fj preserves $\mathcal{B}$. Thus since $f \in \mathcal{B}$, $Fj \circ f \in \mathcal{B}$ as well. Finally, since β preserves $\mathcal{B}$, it follows that $\Phi_{f,\beta}(j) = \beta \circ Fj \circ f \in \mathcal{B}$. □

The following result is the first step towards obtaining the long-term value map LTV_σ from the universal property of bcas.

Proposition 6. *The H-algebra $\alpha_\gamma\colon H\mathbb{R} \to \mathbb{R}$ is a bca in (Met, B).*

Proof. Fix a bounded $f\colon X \to HX$. Recall from Example 4 that the bounded function space $B(X,\mathbb{R})$ is a complete ordered metric space with the supremum metric. Since H is a b-endofunctor (Lemma 5) and α_γ preserves B (Example 7), the operator

$$\Phi = \Phi_{f,\alpha_\gamma}\colon B(X,\mathbb{R}) \to B(X,\mathbb{R})\colon \Phi(j) = \alpha_\gamma \circ Hj \circ f$$

is well-defined by Lemma 7.

We now show that Φ is a contractive map. So let $j,k \in B(X,\mathbb{R})$, and $x \in X$. We write $f = \langle f_1, f_2\rangle\colon X \to \mathbb{R} \times X$. Then

$$d_{\mathbb{R}}(\Phi(j)(x), \Phi(k)(x)) \leq \gamma \cdot |j(f_2(x)) - k(f_2(x))| \leq \gamma \cdot d(j,k).$$

This holds for all $x \in X$. Since $0 \leq \gamma < 1$, it follows that Φ is contractive. By the Banach Fixpoint Theorem, Φ has a unique fixpoint. This proves that the operator Φ_{f,α_γ} has a unique *bounded* fixpoint, which is what we had to show. □

The second step for obtaining the long-term value function LTV_σ from the universal property of bcas, is to show how to obtain a bca for $H\Delta$ from the bca α_γ for H. The next theorem shows that we can prove this result abstractly using b-structure.

We first note that given a b-structure $(\mathsf{C}, \mathcal{B})$ and a monad (T, η, μ), the category $\mathcal{EM}(T)$ has a b-structure consisting of the T-algebra morphisms φ such that $U\varphi \in \mathcal{B}$ in the base b-structure; we shall write the b-structure on $\mathcal{EM}(T)$ as $\mathcal{B}$ as well.

Theorem 3. *Let $(\mathsf{C}, \mathcal{B})$ be a b-category, F a C-endofunctor, (T, η, μ) a monad on C, and λ a distributive law of (T, η, μ) over F. Assume further that T is a strong b-functor and that λ and $F\mu$ are b-natural in $(\mathsf{C}, \mathcal{B})$.*

1. *If $\beta\colon F_\lambda(A, \omega) \to (A, \omega)$ is an F_λ-algebra in $\mathcal{EM}(T)$ such that the underlying F-algebra $\beta\colon FA \to A$ is a bca for F and ω preserves $\mathcal{B}$, then it holds that $\beta \circ F\omega\colon FTA \to A$ is a bca for FT.*
2. *Let the solution operation for the bca $\beta\colon FA \to A$ be denoted $h \mapsto h^\ddagger$, and the solution operation for the bca $\beta \circ F\omega\colon FTA \to A$ be denoted $h \mapsto h^\dagger$. Then for all $g\colon X \to FTX$ in $\mathcal{B}$, we have $g^\dagger = (g^\sharp)^\ddagger \circ \eta_X$ and $(g^\sharp)^\ddagger = \omega \circ Tg^\dagger$.*

Excluding the "b-considerations", Theorem 3 is formulated and proved in dual form (i.e., for comonads and recursive coalgebras) in [5, Theorem 19]. Our assumptions related to the b-structure ensure that the proof carries over to the case of bcas.

Using Theorem 3, we obtain the bca that will give us the long-term value.

Corollary 1. *The $H\Delta$-algebra $\alpha = \alpha_\gamma \circ (\mathbb{R} \times \mathrm{E})$ is a bca in (Met, B).*

Proof. This result follows from Theorem 3. We check the conditions. First, by Lemma 3, $\alpha_\gamma\colon H\chi(\mathbb{R}, \mathrm{E}) \to (\mathbb{R}, \mathrm{E})$ is a $H\chi$-algebra in $\mathcal{EM}(\Delta)$. Next, by Proposition 6, the underlying H-algebra α_γ is a bca for H, and in Example 7 we saw that E preserves B. Finally, we saw in Example 8 that Δ is a strong b-endofunctor, and χ and $H\mu$ are b-natural. By Theorem 3, we have a bca structure for $H\Delta$ on $\mathbb{R}$, namely $\alpha_\gamma \circ H\mathrm{E} = \alpha_\gamma \circ (\mathbb{R} \times \mathrm{E})$. □

Using that α_γ is a bca for H (Proposition 6), we obtain from the universal property of α_γ a unique bounded map $\ell'_\sigma\colon \Delta S \to \mathbb{R}$ that makes the diagram below on the left commute. Also, we obtain a map $\mathrm{LTV}'_\sigma\colon S \to \mathbb{R}$ from the universal property of $\alpha_\gamma \circ (\mathbb{R} \times \mathrm{E})$ as a bca for $H\Delta$ (Corollary 1). That is, LTV'_σ is the unique bounded map that makes the diagram below on the right commute.

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

Moreover, by Theorem 3(2) we have that

$$\mathrm{LTV}'_\sigma = \ell'_\sigma \circ \delta_S \qquad \text{and} \qquad \ell'_\sigma = \mathrm{E} \circ \Delta\, \mathrm{LTV}'_\sigma\,. \tag{16}$$

In particular, LTV'_σ is the unique fixpoint of the operator

$$\Phi_\sigma = \Phi_{m_\sigma, \alpha_\gamma \circ (\mathbb{R} \times \mathrm{E})}\colon B(S, \mathbb{R}) \to B(S, \mathbb{R}) \qquad \Phi_\sigma(f) = \alpha_\gamma \circ \langle u, \mathrm{E} \circ (\Delta f) \circ t_\sigma \rangle. \tag{17}$$

It is not hard to see that, as expected, $\Phi_\sigma = \Psi_\sigma$ from (4) in Section 2 (note that $B(S,\mathbb{R}) = \mathbb{R}^S$ because S is finite). Hence, by unicity, $\mathrm{LTV}'_\sigma = \mathrm{LTV}_\sigma$. By the definition of ℓ_σ (cf. (2)) and the right-hand side of (16), we also see that $\ell'_\sigma = \ell_\sigma$. Therefore, the equations in (16) express that

$$\mathrm{LTV}_\sigma = \ell_\sigma \circ \delta_S \qquad \text{and} \qquad \ell_\sigma = \mathrm{E} \circ \Delta\, \mathrm{LTV}_\sigma\,. \tag{18}$$

In this way, we re-obtained LTV_σ and ℓ_σ using our categorical perspective. Note however that thanks to Corollary 1, in our novel approach we did not need to show that Φ_σ is contractive in order to get LTV_σ.

5.4 The Optimal Value Function V^*

We recall from Section 2 that the optimal value function V^* is the unique solution to Bellman's optimality equation, which we restate here for convenience:

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

To say that V^* solves this is to say that the diagram below commutes:

$$\begin{array}{ccc} S & \xrightarrow{\quad m=\langle u,t\rangle \quad} & \mathbb{R} \times (\Delta S)^{Act} \\ {\scriptstyle V^*}\downarrow & & \downarrow{\scriptstyle \mathbb{R}\times(\Delta V^*)^{Act}} \\ \mathbb{R} & \xleftarrow[\alpha_\gamma \circ (\mathbb{R} \times \max_{Act} \circ \mathrm{E}^{Act})]{} & \mathbb{R} \times (\Delta\mathbb{R})^{Act} \end{array}$$

This diagram clearly looks like a bca diagram, and it is therefore natural to ask whether we can prove the existence of V^* by generalizing the results for LTV_σ. It turns out that many, but not all, do generalize. We give a brief overview.

The coalgebraic modeling is straightforward. Recall that an MDP is a $K\Delta$-coalgebra, where K is the functor $K = H \circ (-)^{Act}$ where $(-)^{Act}$ is the covariant hom-functor. There is also a distributive law ρ of Δ over K. It uses strength $\mathrm{str}\colon \Delta \circ (-)^{Act} \Rightarrow (-)^{Act} \circ \Delta$; specifically, we have $\rho = \langle \mathrm{E} \circ \Delta\pi_1, \mathrm{str} \circ \Delta\pi_2 \rangle$.

We can lift K to Met by viewing X^{Act} as an Act-fold product, i.e., we use the maximum metric. The metric version of K has the same nice properties as the metric version of H. For example, if f is Lipschitz, then so is Kf (generalizing the second part of Lemma 4), and $K\colon (\mathsf{Met}, B) \to (\mathsf{Met}, B)$ is a b-endofunctor (generalizing Proposition 3).

Moreover, we can show that the K-algebra $\alpha_\gamma \circ H \max_{Act}\colon \mathbb{R} \times \mathbb{R}^{Act} \to \mathbb{R}$ is a bca for K (generalizing Proposition 6). Part of the verification shows that the map $\max_{Act}\colon \mathbb{R}^{Act} \to \mathbb{R}$ preserves B; this uses the simple fact that for all sets A and all $h_1, h_2 \in \mathbb{R}^A$, it holds that $|\max_A h_1 - \max_A h_2| \leq \max_A |h_1 - h_2| = d_{\mathbb{R}^A}(h_1, h_2)$.

Things only go sour when we try to apply Theorem 3 to get a bca for $K\Delta$ from the bca $\beta = \alpha_\gamma \circ H \max_{Act}$ for K. The problem is that in order to do so, we need to show that β is an arrow in $\mathcal{EM}(\Delta)$, which entails that $\max_{Act}$ is

an arrow in $\mathcal{EM}(\Delta)$, and this is not the case since, unlike α_γ, $\max_{Act}$ is not affine as it does not commute with convex linear combinations. Nevertheless, $\alpha_\gamma \circ (\mathbb{R} \times (\max_{Act} \circ \mathrm{E}^{Act}))$ *is* a bca for $K\Delta$, since this is equivalent to the statement that the Bellman operator Φ^* (as a map on $B(S, \mathbb{R})$) has a unique fixpoint. The difference with the situation for H and LTV_σ is that we cannot use Theorem 3 to relate the bca structure for $K\Delta$ to the bca structure for K.

6 Extensions

We briefly discuss some extensions to our work.

6.1 Alternative Treatment of MDPs

In our definition of MDPs, rewards are associated with states. However, often MDPs are presented with rewards associated with transitions, i.e., an MDP is then of type $n \colon S \to (\mathbb{R} \times \Delta S)^{Act}$. The latter is an $H^{Act}\Delta$-coalgebra, where $H^{Act} = (-)^{Act} \circ H$. For the general results, not much changes. We again have a distributive law of H^{Act} over Δ, given by $\langle \mathrm{E} \circ \Delta\pi_1, \Delta\pi_2 \rangle^{Act} \circ \mathrm{str}$, and a policy σ yields an $H\Delta$-coalgebra given by $n_\sigma = \langle u_\sigma, t_\sigma \rangle = n \circ \sigma$. (Compare: $m_\sigma = \langle u, t_\sigma \rangle$.) So we again obtain ℓ_σ and LTV_σ from Proposition 6 together with Corollary 1. Also, (18) holds, just as before. The contractive operator characterizing LTV_σ in n_σ is defined as

$$\Phi_\sigma(f) = \alpha_\gamma \circ \langle u_\sigma, \mathrm{E} \circ \Delta f \circ t_\sigma \rangle.$$

We adapt the definition of improved policies to the current setting by letting

$$\tau(s) = \mathrm{argmax}_{a \in Act}\ \alpha_\gamma(\pi_1(n(s)(a)), \ell_\sigma(\pi_2(n(s)(a)))).$$

We can prove the Policy Improvement Theorem. We expect that with similar adaptations, the results of Section 5.4 also go through as before.

6.2 Changing the Setting to Polish Metric Spaces

The main setting of this paper was the b-category (Met, B) of all metric spaces, where the hom-sets are those of Set and the $\mathcal{B}$-morphisms are the bounded maps. Since it would be more satisfying to have a more "metric" category, we want to sketch how this can be done.

The first way is to restrict the morphisms between metric spaces to be non-expansive maps. We call the resulting category Met_1. Taking B to be the bounded, non-expansive maps, (Met_1, B) is a b-category. For our work, the problem with Met_1 is that when products are given the maximum metric, α_γ is not a morphism in Met_1. Using the sum metric on products, we get a b-category, and even bca structures for H, $H\Delta$, K and $K\Delta$. However, other technical problems arise that suggest that this is not a worthwhile approach.

A more fruitful direction is to work with the category of *Polish metric spaces* (complete separable spaces) and continuous functions, called PolMet. Let us write

$\widehat{P}\colon \mathsf{Met} \to \mathsf{Met}$ for the endofunctor which takes a space M to the metric space of all Borel probability measures on M, using (for concreteness) the Kantorovich metric, defined using integrals instead of sums. The resulting topology is the weak topology. Giry [11] proved that $\widehat{P}$ restricts to an endofunctor $P\colon \mathsf{PolMet} \to \mathsf{PolMet}$. For every Polish space X, the set ΔX of discrete probability distributions is a dense subset of $\widehat{P}X$ (see [29, Theorem 6.18]). For every map $k\colon X \to Y$ in Met, $\widehat{P}k$ and Δk both work the same way, by "pushing forward" a distribution. The appropriate version of E is $\mathrm{E}(\mu) = \int x \,\mathrm{d}\mu$. All of the results in Sections 4 and 5 adapt to this setting, mutatis mutandis. The upshot is that we get a b-category (PolMet, B), where B is the class of bounded continuous functions. Furthermore, the policy improvement theorem can be done in that setting.

7 Conclusion

Our main goal has been to show that the value functions LTV_σ and V^* arise from a universal property of sorts, and to re-prove the correctness of policy improvement using a coinductive argument. The universal property was explained in terms of bca-structures, and for this we needed the notion of a b-category. The main examples led us to study boundedness preservation properties of the liftings of the stream functor and the distribution monad to metric spaces.

The coinductive analysis of policy improvement went by means of a new contraction coinduction principle. In essence, contraction coinduction allows one to infer qualitative relationships (e.g., policy improvement) without a detour into quantitative results. We would like to think this principle has many other uses.

We have a few comments on earlier work in the same general area.

Kozen and Ruozzi [24] surely had the intuition that aspects of the theory of MDPs should be understood coinductively. Their paper has a very interesting coinductive proof of the fact that the optimal policies in MDPs may be taken to be deterministic. They were not concerned with policy improvement, our target for coinduction. As for ourselves, we formulated contraction coinduction; this is an easy consequence of the metric coinduction principle from [17,24], and it seems to do the work one would want for inequalities as one finds in policy improvement.

One should go back to Shapley games and other infinite games to see if the metric coinduction principle from [17,24] could simplify the (subtle) positive results in the area. Also, the metric coinduction principle was used by Abramsky and Winschel [1] to establish a predicate coinduction principle. They use that result in connection with subgame perfect equilibria in infinite games such as the dollar auction. Pavlovic [22] shares some programmatic features with our work, even though the formal work appears different. There are connections to be made with all of these papers.

Denardo [7] is concerned with some of the same issues that we address. In some ways his work is more abstract than ours, as he does not assume a particular system type, and in some ways less. His work does not use categorical notions, so it does not directly compare with our work, but assumptions pertaining to

contraction mappings and to order-preservation are prominent in the paper. Our contraction coinduction principle simplifies several of the proofs in [7].

A related point: Denardo assumes that (his version of) Φ maps bounded functions to bounded functions. Our notions of b-functor and b-preservation give us a compositional account of this fact. This was put to use in Proposition 6. On the other hand, to show that Φ is a contraction, our general machinery was not useful. So there certainly is more work to be done on that.

This paper 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 paper is devoted to this extra information. We hope that our work will be useful in settings beyond MDPs. We have some pilot results in this direction, but for lack of space these do not appear in this paper.

Acknowledgments

We would like to thank Tarmo Uustalu for pointing us to [5, Theorem 19], thereby improving the paper. We also thank Jasmine Blanchette, Wan Fokkink and Ana Sokolova for useful comments.

References

1. Abramsky, S., Winschel, V.: Coalgebraic Analysis of Subgame-Perfect Equilibria in Infinite Games without Discounting. Mathematical Structures in Computer Science 27(5), 751–761 (2017)
2. Baldan, P., Bonchi, F., Kerstan, H., König, B.: Behavioral Metrics via Functor Lifting. In: 34th International Conference on Foundation of Software Technology and Theoretical Computer Science, FSTTCS 2014. pp. 403–415 (2014)
3. Bartels, F.: On Generalised Coinduction and Probabilistic Specification Formats. Ph.D. thesis, Vrije Universiteit Amsterdam (2004)
4. Bellman, R.: Dynamic Programming. Princeton University Press, Princeton, NJ, USA, 1 edn. (1957)
5. Capretta, V., Uustalu, T., Vene, V.: Recursive Coalgebras from Comonads. Inf. Comp. 204, 437468 (2006)
6. Capretta, V., Uustalu, T., Vene, V.: Corecursive Algebras: A Study of General Structured Corecursion. In: Formal Methods: Foundations and Applications. pp. 84–100 (2009)
7. Denardo, E.V.: Contraction Mappings in the Theory Underlying Dynamic Programming. SIAM Review 9(2), 165–177 (1967)
8. Desharnais, J., Edalat, A., Panangaden, P.: Bisimulation for Labelled Markov Processes. Information and Computation 179(2), 163–193 (2002)
9. Ferns, N., Panangaden, P., Precup, D.: Metrics for Finite Markov Decision Processes. In: Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence. pp. 162–169. UAI '04, AUAI Press, Arlington, Virginia, United States (2004), http://dl.acm.org/citation.cfm?id=1036843.1036863

10. Gibbs, A.L., Su, F.E.: On Choosing and Bounding Probability Metrics. International Statistical Review/Revue Internationale de Statistique 70(3), 419–435 (2002)
11. Giry, M.: A Categorical Approach to Probability Theory. In: Categorical Aspects of Topology and Analysis, vol. 915, pp. 68–85. Springer (1982)
12. Howard, R.A.: Dynamic Programming and Markov Processes. The M.I.T. Press (1960)
13. Jacobs, B.: Distributive Laws for the Coinductive Solution of Recursive Equations. Information and Computation 204(4), 561–587 (2006)
14. Jacobs, B., Silva, A., Sokolova, A.: 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)
15. Johnstone, P.T.: Adjoint Lifting Theorems for Categories of Algebras. Bulletin of the London Mathematical Society 7, 294–297 (1975)
16. Klin, B.: Bialgebras for Structural Operational Semantics: An Introduction. Theoretical Computer Science 412(38), 5043–5069 (2011)
17. Kozen, D.: Coinductive Proof Principles for Stochastic Processes. Logical Methods in Computer Science 5, 1–19 (2009)
18. Mac Lane, S.: Categories for the Working Mathematician, vol. 5. Springer Science & Business Media (2013)
19. Milius, S.: Completely Iterative Algebras and Completely Iterative Monads. Information and Computation 196(1), 1–41 (2005)
20. Moore, A.W.: Markov Systems, Markov Decision Processes, and Dynamic Programming (2002), lecture slides available at https://www.autonlab.org/tutorials
21. Ó'Searcóid, M.: Metric Spaces. Springer Science & Business Media (2006)
22. Pavlovic, D.: A Semantical Approach to Equilibria and Rationality. In: Proceedings, Algebra and Coalgebra in Computer Science, CALCO 2009. pp. 317–334 (2009)
23. Puterman, M.L.: Markov Decision Processes: Discrete Stochastic Dynamic Programming. John Wiley & Sons (2014)
24. Ruozzi, N., Kozen, D.: Applications of Metric Coinduction. Logical Methods in Computer Science 5 (2009)
25. Rutten, J.: Universal Coalgebra: A Theory of Systems. Theoretical Computer Science 249(1), 3–80 (2000)
26. Silva, A., Bonchi, F., Bonsangue, M., Rutten, J.: Generalizing Determinization from Automata to Coalgebras. Logical Methods in Computer Science 9, 1–27 (2013)
27. Silva, A., Sokolova, A.: Sound and Complete Axiomatization of Trace Semantics for Probabilistic Systems. Electronic Notes in Theoretical Computer Science 276, 291–311 (2011), https://doi.org/10.1016/j.entcs.2011.09.027
28. Sokolova, A.: Probabilistic Systems Coalgebraically. Theoretical Computer Science 412(38), 5095–5110 (September 2011), http://dx.doi.org/10.1016/j.tcs.2011.05.008
29. Villani, C.: Optimal Transport, Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol. 338. Springer-Verlag, Berlin (2009)