

# CertRL: Formalizing Convergence Proofs for Value and Policy Iteration in Coq

KOUNDINYA VAJJHA, University of Pittsburgh , USA

AVRAHAM SHINNAR, IBM Research , USA

BARRY TRAGER, IBM Research , USA

VASILY PESTUN, IBM Research , USA& IHES

NATHAN FULTON, IBM Research , USA

Reinforcement learning algorithms solve sequential decision-making problems in probabilistic environments by optimizing for long-term reward. The desire to use reinforcement learning in safety-critical settings inspires a recent line of work on formally constrained reinforcement learning; however, these methods place the implementation of the learning algorithm in their Trusted Computing Base. The crucial correctness property of these implementations is a guarantee that the learning algorithm converges to an optimal policy.

This paper begins the work of closing this gap by developing a Coq formalization of two canonical reinforcement learning algorithms: value and policy iteration for finite state Markov decision processes. The central results are a formalization of the Bellman optimality principle and its proof, which uses a contraction property of Bellman optimality operator to establish that a sequence converges in the infinite horizon limit. The CertRL development exemplifies how the Giry monad and mechanized metric coinduction streamline optimality proofs for reinforcement learning algorithms. The CertRL library provides a general framework for proving properties about Markov decision processes and reinforcement learning algorithms, paving the way for further work on formalization of reinforcement learning algorithms.

Additional Key Words and Phrases: Formal Verification, Policy Iteration, Value Iteration, Reinforcement Learning, Coinduction

## 1 INTRODUCTION

Reinforcement learning (RL) algorithms solve sequential decision making problems in which the goal is to choose actions that maximize a quantitative utility function [Bel54, How60, Put94, SB98]. Recent high-profile applications of reinforcement learning include beating the world's best players at Go [SHM<sup>+</sup>16], competing against top professionals in Dota [Ope18], improving protein structure prediction [SEJ<sup>+</sup>20], and automatically controlling complex robots [GHLL16]. These successes motivate the use of reinforcement learning in safety-critical and correctness-critical settings.

Reinforcement learning algorithms produce, at a minimum, a *policy* that specifies which action(s) should be taken in a given state. The primary correctness property for reinforcement learning algorithms is *convergence*: in the limit, a reinforcement learning algorithm should converge to a policy that optimizes for the expected future-discounted value of the reward signal.

This paper contributes CertRL, a formal proof of convergence for value iteration and policy iteration two canonical reinforcement learning algorithms [Bel54, How60, Put94]. They are often taught as the first reinforcement learning methods in machine learning courses because the algorithms are relatively simple but their convergence proofs contain the main ingredients of a typical convergence argument for a reinforcement learning algorithm.

There is a cornucopia of presentations of these iterative algorithms and an equally diverse variety of proof techniques for establishing convergence. Many presentations state but do not prove the fact that the optimal policy of an infinite-horizon Markov decision process with  $\gamma$ -discounted reward is a *stationary policy*; i.e., the optimal decision in a given state does not depend on the time step at which the state is encountered. Following this convention, this paper contributes the first formal proof that policy and value iteration converge in the limit to the optimal policy in the space of stationary policies for infinite-horizon Markov decision processes. In addition to establishing convergence results for the classical iterative algorithms under classical infinitary and

stationarity assumptions, we also formalize an optimality result about  $n$ -step iterations of value iteration without a stationarity assumption. The former formalization matches the standard theoretical treatment, while the latter is closer to real-world implementations. We shall refer to the former case – where the set of time steps is an infinite set – as *infinite-horizon* and the latter case as *finite-horizon*.

In all cases, the convergence argument for policy/value iteration proceeds by proving that a contractive mapping converges to a fixed point and that this fixed point is an optimum. This is typical of convergence proofs for reinforcement learning algorithms. CertRL is intentionally designed for ongoing reinforcement learning formalization efforts.

Formalizing the convergence proof directly would require complicated and tedious  $\epsilon$ -hacking as well as long proofs involving large matrices. CertRL obviates these challenges using a combination of the Giriy monad [Gir82, Jac18] and a proof technique called *Metric coinduction* [Koz07].

Metric coinduction was first identified by Kozen and Ruoizzi as a way to streamline and simplify proofs of theorems about streams and stochastic processes [KR09]. Our convergence proofs use a specialized version of metric coinduction called contraction coinduction [FHM18] to reason about order statements concerning fixed points of contractive maps. Identifying a *coinduction hypothesis* allows us to automatically infer that a given (closed) property holds in the limit whenever it holds *ab initio*. The coinduction hypothesis guarantees that this property is a limiting invariant. This is significant because the low level  $\epsilon - \delta$  arguments – typically needed to show that a given property holds of the limit – are now neatly subsumed by a single proof rule, allowing reasoning at a higher level of abstraction.

The *finitary Giriy monad* is a monad structure on the space of all finitely supported probability mass functions on a set. Function composition in the Kleisli category of this monad recovers the Chapman-Kolmogorov formula [Per19, Jac18]. Using this fact, our formalization recasts iteration of a stochastic matrix in a Markov decision process as iterated Kleisli composites of the Giriy monad, starting at an initial state. Again, this makes the presentation cleaner since we identify and reason about the basic operations of bind and ret, thus bypassing the need to define matrices and matrix multiplication and substantially simplifying convergence proofs.

This paper shows how these two basic building blocks – the finitary Giriy monad and metric coinduction – provide a compelling foundation for formalizing reinforcement learning theory. CertRL develops the basic concepts in reinforcement learning theory and demonstrates the usefulness of this library by proving several results about value and policy iteration. CertRL contains a proof of the Bellman optimality principle, an inductive relation on the optimal value and policy over the horizon length of the Markov decision process.

In practice, reinforcement learning algorithms almost always run in finite time by either fixing a run time cutoff (e.g., number training steps) or by stopping iteration after the value/policy changes become smaller than a fixed threshold. Therefore, our development also formalizes a proof that  $n$ -step value iteration satisfies a finite time analogue of our convergence results.

To summarize, the CertRL library contains:

- (1) a formalization of Markov decision processes and their long-term values in terms of the finitary Giriy monad,
- (2) a formalization of optimal value functions and the Bellman operator,
- (3) a formal proof of convergence for value iteration and a formalization of the policy improvement theorem in the case of stationary policies, and
- (4) a formal proof that the optimal value function for finitary sequences satisfies the finite time analogue of the Bellman equation.

Throughout the text which follows, hyperlinks to theorems, definitions and lemmas which have formal equivalents in the Coq [Tea04] development are indicated by a  $\clubsuit$ .<sup>1</sup>

CertRL is part of a larger project for verifying machine learning theory with applications to program synthesis. The entire development is available online at the following URL: <https://github.com/IBM/FormalML>.

## 2 BACKGROUND

We provide a brief introduction to value/policy iteration and to the mathematical structures upon which our formalization is built: contractive metric spaces, metric coinduction, the Giry monad and Kleisli composition.

### 2.1 Reinforcement Learning

This section gently introduces the basics of reinforcement learning with complete information about the stochastic reward and transition functions. In this simplified situation the focus of the algorithm is on optimal exploitation of reward. This framework is also known as the stochastic optimal control problem.

We give an informal definition of Markov decision processes, trajectories, long-term values, and dynamic programming algorithms for solving Markov decision processes. Many of these concepts will be stated later in a more formal type-theoretic style; here, we focus on providing an intuitive introduction to the field.

The basic mathematical object in reinforcement learning theory is the Markov decision process. A Markov decision process is a 4-tuple  $(S, A, R, T)$  where  $S$  is a set of states,  $A$  is a set of actions,  $R : S \times A \times S \rightarrow \mathbb{R}$  is a *reward function*, and  $T$  is a *transition relation* on states and actions mapping each  $(s, a, s') \in S \times A \times S$  to the probability that taking action  $a$  in state  $s$  results in a transition to  $s'$ . Markov decision processes are so-called because they characterize a sequential decision-making process (each action is a decision) in which the transition structure on states and actions depends only on the current state.

*Example 1 (CeRL the turtle  $\clubsuit$ ).* Consider a simple grid world environment in which a turtle can move in cardinal directions throughout a 2D grid. The turtle receives +1 point for collecting stars, -10 for visiting red squares, and +2 for arriving at the green square. The turtle chooses which direction to move, but with probability  $\frac{1}{4}$  will move in the opposite direction. For example, if the turtle takes action `left` then it will go `left` with probability  $\frac{3}{4}$  and `right` with probability  $\frac{1}{4}$ . The game ends when the turtle arrives at the green square.

This environment is formulated as a Markov decision process as follows:

- The set of states  $S$  are the coordinates of each box  $\clubsuit$ :

$$\{(x, y) \mid 1 \leq x \leq 5 \text{ and } 1 \leq y \leq 5\}$$

so that  $(1, 1)$  is the top-left corner and  $(5, 5)$  is the bottom-right corner.

- The set of actions  $A$  are  $\{\text{up, down, left, right}\} \clubsuit$ .

---

<sup>1</sup>We recommend MacOS users view this document in Adobe, Firefox, or Chrome, as Preview and Safari parse the URLs linked to by  $\clubsuit$ 's incorrectly.

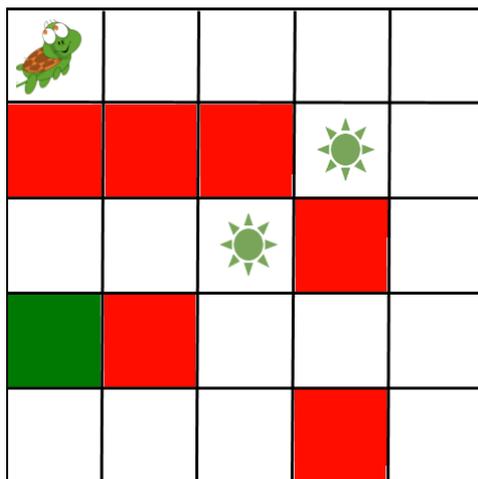


Fig. 1. An example grid-world environment ♣.

- The reward function is defined as ♣:

$$R(1, 4) = 2$$

$$R(4, 2) = 1$$

$$R(3, 3) = 1$$

$$R(\{1, 2, 3\}, 2) = -10$$

$$R(4, 3) = -10$$

$$R(2, 4) = -10$$

$$R(4, 5) = -10$$

$$R(\cdot, \cdot) = 0 \text{ otherwise}$$

- The transition probabilities are as described ♣; e.g.,

$$T((3, 4), \text{up}, (3, 3)) = \frac{3}{4}$$

$$T((3, 4), \text{up}, (3, 5)) = \frac{1}{4}$$

$$T((3, 4), \text{up}, (\cdot, \cdot)) = 0 \text{ otherwise}$$

and so on.

We implement this example in Coq as a proof-of-concept for CertRL. We first define a matrix whose indices are states  $(x, y)$  and whose entries are colors  $\{\text{red, green, star, empty}\}$ . We then define a reward function that maps from matrix entries to a reward depending on the color of the turtle's current state. We also define a transition function that comports with the description given above. At last, we prove that this combination of states, actions, transitions and rewards inhabits our MDP (standing for *Markov decision process*) type. Therefore, all of the theorems developed in this paper apply directly to our Coq implementation of the CertRL Turtle environment.

The goal of reinforcement learning is to find a *policy* ( $\pi : S \rightarrow A$ ) specifying which action the algorithm should take in each state. This policy should maximize the amount of reward obtained by the agent. A policy is *stationary* if it is not a function of time; i.e., if the optimal action in some state  $s \in S$  is always the same and, in particular, independent of the specific time step at which  $s$  is encountered.

Reinforcement learning agents optimize for a *discounted sum* of rewards – placing more emphasis on reward obtained today and less emphasis on reward obtained tomorrow. A constant *discount factor* from the open unit interval, typically denoted by  $\gamma$ , quantitatively discounts future rewards and serves as a crucial hyperparameter to reinforcement learning algorithms.

Value iteration, invented by Bellman [Bel54], is a dynamic programming algorithm that finds optimal policies to reinforcement learning algorithms by iterating a contractive mapping. Value iteration is defined in terms of a *value function*  $V_\pi : S \rightarrow \mathbb{R}$ , where  $V_\pi(s)$  is the expected value of state  $s$  when following policy  $\pi$  from  $s$ .

**Data:**

Markov decision process  $(S, A, T, R)$

Initial value function  $V_0 = 0$

Threshold  $\theta > 0$

Discount factor  $0 < \gamma < 1$

**Result:**  $V^*$ , the value function for an optimal policy.

```

for  $n$  from 0 to  $\infty$  do
  | for each  $s \in S$  do
    |  $V_{n+1}[s] = \max_a \sum_{s'} T(s, a, s')(R(s, a, s') + \gamma V_n[s'])$ 
  | end
  | if  $\forall s | V_{n+1}[s] - V_n | < \theta$  then
    | return  $V_{n+1}$ 
  | end
end
    
```

**Algorithm 1:** Pseudocode for Value Iteration.

The optimal policy  $\pi^*$  is then obtained by

$$\pi^*(a) = \operatorname{argmax}_{a \in A} \sum_{s'} T(s, a, s')(R(s, a, s') + \gamma V_{n+1}[s']).$$

Policy iteration follows a similar iteration scheme, but with a policy estimation function  $Q_\pi : S \times A \rightarrow \mathbb{R}$  where  $Q_\pi(s, a)$  estimates the value of taking action  $a$  in state  $s$  and then following the policy  $\pi$ . In Section 3.4 we will demonstrate a formalized proof that  $V_n$  is the *optimal value* function of a length  $n$  MDP; this algorithm implements the *dynamic programming* principle.

## 2.2 Metric and Contraction Coinduction

Our formalization uses metric coinduction to establish convergence properties for infinite sequences. This section recalls the Banach fixed point theorem and explains how this theorem gives rise to a useful proof technique.

A metric space  $(X, d)$  is a set  $X$  equipped with a function  $d : X \times X \rightarrow \mathbb{R}$  satisfying certain axioms that ensure  $d$  behaves like a measurement of the *distance* between points in  $X$ . A metric space is *complete* if the limit of every Cauchy sequence of elements in  $X$  is also in  $X$ .

Let  $(X, d)$  denote a complete metric space with metric  $d$ . Subsets of  $X$  are modeled by terms of the function type  $\phi : X \rightarrow \text{Prop}$ . Another interpretation is that  $\phi$  denotes all those terms of  $X$  which satisfy a particular property. These *subsets* are also called *Ensembles* in the Coq standard library.

A *Lipschitz map*  $\clubsuit$  is a mapping that is Lipschitz continuous; i.e., a mapping  $F$  from  $(X, d_X)$  into  $(Y, d_Y)$  such that for all  $x_1, x_2 \in X$  there is some  $K \geq 0$  such that

$$d_Y(F(x_1), F(x_2)) \leq K d_X(x_1, x_2).$$

The constant  $K$  is called a Lipschitz constant.

A map  $F : X \rightarrow X$  is called a *contractive map*  $\clubsuit$ , or simply a *contraction*, if there exists a constant  $0 \leq \gamma < 1$  such that

$$d(F(u), F(v)) \leq \gamma d(u, v) \quad \forall u, v \in X.$$

Contractive maps are Lipschitz maps with Lipschitz constant  $\gamma < 1$ .

The Banach fixed point theorem is a standard result of classical analysis which states that contractive maps on complete metric spaces have a unique fixed point.

**THEOREM 2 (BANACH FIXED POINT THEOREM).** *If  $(X, d)$  is a nonempty complete metric space and  $F : X \rightarrow X$  is a contraction, then  $F$  has a unique fixed point; i.e., there exists a point  $x^* \in X$  such that  $F(x^*) = x^*$ . This fixed point is  $x^* = \lim_{n \rightarrow \infty} F^{(n)}(x_0)$  where  $F^{(n)}$  stands for the  $n$ -th iterate of the function  $F$  and  $x_0$  is an arbitrary point in  $X$ .*

The Banach fixed point theorem generalizes to subsets of  $X$ .

**THEOREM 3 (BANACH FIXED POINT THEOREM ON SUBSETS  $\clubsuit$ ).** *Let  $(X, d)$  be a complete metric space and  $\phi$  a closed nonempty subset of  $X$ . Let  $F : X \rightarrow X$  be a contraction and assume that  $F$  preserves  $\phi$ . In other words,*

$$\phi(u) \rightarrow \phi(F(u))$$

*Then  $F$  has a unique fixed point in  $\phi$ ; i.e., a point  $x^* \in X$  such that  $\phi(x^*)$  and  $F(x^*) = x^*$ . The fixed point of  $F$  is given by  $x^* = \lim_{n \rightarrow \infty} F^{(n)}(x_0)$  where  $F^{(n)}$  stands for the  $n$ -th iterate of the function  $F$ .*

Both the Banach fixed point theorem and the more general theorem on subsets were previously formalized in Coq by Boldo et al. [BCF<sup>+</sup>17]. This formalization includes definitions of Lipschitz maps and contractions. We make use of the fact that Boldo et al. prove the above theorem where  $X$  is either a `CompleteSpace` or a `CompleteNormedModule`.

The fixed point of  $F$  in Theorem 3 is unique, but it depends on an initial point  $x_0 \in X$ , which  $F$  then iterates on. Uniqueness of the fixed point implies that different choices of the initial point still give the same fixed point  $\clubsuit$ .

To emphasize how this theorem is used in our formalization, we restate it as an inductive proof rule:

$$\frac{\phi \text{ closed} \quad \exists x_0, \phi(x_0) \quad \phi(u) \rightarrow \phi(F(u))}{\phi(\text{fix } F \ x_0)} \clubsuit \quad (1)$$

This proof rule states that in order to prove some closed  $\phi$  is a property of a fixed point of  $F$ , it suffices to establish the standard inductive assumptions: that  $\phi$  holds for some initial  $x_0$ , and that if  $\phi$  holds at  $u$  then it also holds after a single application of  $F$  to  $u$ . In this form, the Banach fixed point theorem is called *Metric coinduction*. The rule (1) is *coinductive* because it is equivalent to the assertion that a certain coalgebra is final in a category of coalgebras. (Details are given in Section 2.3 of Kozen and Ruozzi [KR09]).

The following snippet shows how we use the Banach Fixed Point theorem as proven in [BCF<sup>+</sup>17] as a proof rule.

```
Theorem metric_coinduction {phi : X → Prop}
  (nephi : phi init) (Hcphi : closed phi)
  (HFphi : forall x : X, phi x → phi (F x)) :
  phi (fixpt F init).
Proof.
  assert (my_complete phi)
```

```

by (now apply closed_my_complete).
destruct (FixedPoint K F phi fphi (ex_intro _ _ init_phi)
H hF) as [? [Hin [? [? Hsub]]]].
specialize (Hsub init init_phi).
rewrite ← Hsub in Hin.
apply Hin.
Qed.

```

**Definition 4** (Ordered Metric Space). *A metric space  $X$  is called an ordered metric space if the underlying set  $X$  is partially ordered and the sets  $\{z \in X | z \leq y\}$  and  $\{z \in X | y \leq z\}$  are closed sets in the metric topology for every  $y \in X$ .*

For ordered metric spaces, metric coinduction specializes to [FHM18, Theorem 1], which we restate as Theorem 5 below.

**THEOREM 5 (CONTRACTION COINDUCTION).** *Let  $X$  be a non-empty, complete ordered metric space. If  $F : X \rightarrow X$  is a contraction and is order-preserving, then:*

- $\forall x, F(x) \leq x \Rightarrow x^* \leq x$  ✿ and
- $\forall x, x \leq F(x) \Rightarrow x \leq x^*$  ✿

where  $x^*$  is the fixed point of  $F$ .

We will use the above result to reason about Markov decision processes. However, doing so requires first setting up an ordered metric space on the function space  $A \rightarrow \mathbb{R}$  where  $A$  is a finite set ✿.

### 2.3 The function space $A \rightarrow \mathbb{R}$ .

Let  $A$  be a finite set ✿. We endow the function space  $A \rightarrow \mathbb{R}$  with a natural vector space structure and with  $L^\infty$  norm ✿:

$$\|f\|_\infty = \max_{a \in A} |f(a)| \quad (2)$$

Our development establishes several important properties about this function space. The norm (2) is well-defined because  $A$  is finite and furthermore induces a metric that makes  $A \rightarrow \mathbb{R}$  a metric space ✿. With this metric, the space of functions  $A \rightarrow \mathbb{R}$  is also complete ✿. From  $\mathbb{R}$  this metric inherits a pointwise order; viz., for functions  $f, g : A \rightarrow \mathbb{R}$ ,

$$\begin{aligned}
 f \leq g &\iff \forall a \in A, f(a) \leq g(a) \text{ ✿} \\
 f \geq g &\iff \forall a \in A, f(a) \geq g(a) \text{ ✿}
 \end{aligned}$$

We also prove that the sets

$$\{f | f \leq g\} \text{ ✿}$$

and

$$\{f | f \geq g\} \text{ ✿}$$

are closed in the norm topology. Our formalization of the proof of closedness for these sets relies on classical reasoning. Additionally, we rely on functional extensionality to reason about equality between functions.

We now have an ordered metric space structure on the function space  $A \rightarrow \mathbb{R}$  when  $A$  is finite. Constructing a contraction on this space will allow an application of Theorem 5. Once we set up a theory of Markov decision processes we will have natural examples of such a function space and contractions on it. Before doing so, we first introduce the Giriy monad.

## 2.4 (Finitary) Giry Monad

A monad structure on the category of all measurable spaces was first described by Lawvere in [Law62] and was explicitly defined by Giry in [Gir82]. This monad has since been called the Giry monad. While the construction is very general (applying to arbitrary measures on a space), for our purposes it suffices to consider finitely supported probability measures.

The Giry monad for finitely supported probability measures is called the *finitary Giry monad*, although sometimes also goes by the more descriptive names *distribution monad* and *convex combination monad*.

On a set  $A$ , let  $P(A)$  denote the set of all finitely-supported probability measures on  $A$   $\clubsuit$ . An element of  $P(A)$  is a list of elements of  $A$  together with probabilities. The probability assigned to an element  $a : A$  is denoted by  $p(a)$ .

In our development we state this as the record

```
Record Pmf (A : Type) := mkPmf {
  outcomes :> list (nonnegreal * A);
  sum1 : list_fst_sum outcomes = R1
}.
```

where `outcomes` stores all the entries of the type  $A$  along with their atomic probabilities. The field `sum1` ensures that the probabilities sum to 1.

The Giry monad is defined in terms of two basic operations associated to this space:

$$\begin{aligned} \text{ret} &: A \rightarrow P(A) \clubsuit \\ a &\mapsto \lambda x : A, \delta_a(x) \end{aligned}$$

where  $\delta_a(x) = 1$  if  $a = x$  and 0 otherwise. The other basic operation is

$$\begin{aligned} \text{bind} &: P(A) \rightarrow (A \rightarrow P(B)) \rightarrow P(B) \clubsuit \\ \text{bind } p \ f &= \lambda b : B, \sum_{a \in A} f(a)(b) * p(a) \end{aligned}$$

In both cases the resulting output is a probability measure. The above definition is well-defined because we only consider finitely-supported probability measures. A more general case is obtained by replacing sums with integrals.

The definitions of `bind` and `ret` satisfy the following properties:

$$\text{bind } (\text{ret } x) \ f = \text{ret}(f(x)) \clubsuit \tag{3}$$

$$\text{bind } p \ (\lambda x, \delta_x) = p \clubsuit \tag{4}$$

$$\text{bind } (\text{bind } p \ f) \ g = \text{bind } p \ (\lambda x, \text{bind } (f x) \ g) \clubsuit \tag{5}$$

These *monad laws* establish that the triple  $(P, \text{bind}, \text{ret})$  forms a monad.

The Giry monad has been extensively studied and used by various authors because it has several attractive qualities that simplify (especially formal) proofs. First, the Giry monad naturally admits a denotational monadic semantics for certain probabilistic programs [RP02, JP89, ŠGG15, APM09]. Second, it is useful for rigorously formalizing certain informal arguments in probability theory by providing a means to perform *ad hoc* notation overloading [TTV19]. Third, it can simplify certain constructions such as that of the product measure [EHN15].

CertRL uses the Giry monad as a substitute for the stochastic matrix associated to a Markov decision process. This is possible because the Kleisli composition of the Giry monad recovers the Chapman-Kolmogorov formula [Per18, Per19]. The Kleisli composition is the *fish* operator in Haskell parlance.

*2.4.1 Kleisli Composition.* Reasoning about probabilistic *processes* requires composing probabilities. The Chapman-Kolmogorov formula is a classical result in the theory of Markovian processes that states the probability of transition from one state to another through two steps can be obtained by summing up the probability of visiting each intermediate state. This application of the Chapman-Kolmogorov formula plays a fundamental role in the study of Markovian processes, but requires formalizing and reasoning about matrix operations.

Kleisli composition provides an alternative and more elegant mechanism for reasoning about compositions of probabilistic choices. This section defines and provides an intuition for Kleisli composition.

Think of  $P(A)$  as the random elements of  $A$  ([Per18, page 15]). In this paradigm, the set of maps  $A \rightarrow P(B)$  are simply the set of maps with a random outcome. When  $P$  is a monad, such maps are called Kleisli arrows of  $P$ .

In terms of reinforcement learning, a map  $f : A \rightarrow P(B)$  is a rule which takes a state  $a : A$  and gives the probability of transitioning to state  $b : B$ . Suppose now that we have another such rule  $g : B \rightarrow P(C)$ . Kleisli composition puts  $f$  and  $g$  together to give a map  $(f \gg g) : A \rightarrow P(C)$ . It is defined as:

$$f \gg g := \lambda x : A, \text{bind } (f \ x) \ g \tag{6}$$

$$= \lambda x : A, (\lambda c : C, \sum_{b:B} g(b)(c) * f(x)(b)) \tag{7}$$

$$= \lambda(x : A) (c : C), \sum_{b:B} f(x)(b) * g(b)(c) \tag{8}$$

The motivation for (6)–(8) is intuitive. In order to start at  $x : A$  and end up at  $c : C$  by following the rules  $f$  and  $g$ , one must first pass through an intermediate state  $b : B$  in the codomain of  $f$  and the domain of  $g$ . The probability of that point being any particular  $b : B$  is

$$f(x)(b) * g(b)(c).$$

So, to obtain the total probability of transitioning from  $x$  to  $c$ , simply sum over all intermediate states  $b : B$ . This is exactly (8). We thus recover the classical Chapman-Kolmogorov formula, but as a Kleisli composition of the Giry monad. This obviates the need for reasoning about operators on linear vector spaces, thereby substantially simplifying the formalization effort.

Indeed, if we did not use Kleisli composition, we would have to associate a stochastic transition matrix to our Markov process and manually prove various properties about stochastic matrices which can quickly get tedious. With Kleisli composition however, our proofs become more natural and we reason closer to the metal instead of adapting to a particular representation.

### 3 THE CERTRL LIBRARY

CertRL contains a formalization of Markov decision processes, a definition of the Kleisli composition specialized to Markov decision processes, a definition of the long-term value of a Markov decision process, a definition of the Bellman operator, and a formalization of the operator’s main properties.

Building on top of its library of results about Markov decision processes, CertRL contains proofs of our main results:

- (1) the (infinite) sequence of value functions obtained by value iteration converges in the limit to a global optimum assuming stationary policies,
- (2) the (infinite) sequence of policies obtained by policy iteration converges in the limit to a global optimum assuming stationary policies, and
- (3) the optimal value function for Markov decision process of length  $n$  is computed inductively by application of Bellman operator, Section 3.4.

The following sections describe the above results more carefully.

### 3.1 Markov Decision Processes

We refer to [Put94] for detailed presentation of the theory of Markov decision processes. Our formalization considers the theory of *infinite-horizon discounted Markov decision processes with deterministic stationary policies*.

We now elaborate on the above definitions and set up relevant notation. Our presentation will be type-theoretic in nature, to reflect the formal development. The exposition (and CertRL formalization) closely follows the work of Frank Feys, Helle Hvid Hansen, and Lawrence Moss [FHM18].

#### 3.1.1 Basic Definitions.

**Definition 6** (Markov Decision Process  $\clubsuit$ ). *A Markov decision process consists of the following data:*

- A nonempty finite type  $S$  called the set of states.<sup>2</sup>
- For each state  $s : S$ , a nonempty finite type  $A(s)$  called the type of actions available at state  $s$ . This is modelled as a dependent type.
- A stochastic transition structure  $T : \prod_{s:S}(A(s) \rightarrow P(S))$ . Here  $P(S)$  stands for the set of all probability measures on  $S$ , as described in Section 2.4.
- A reward function  $r : \prod_{s:S}(A(s) \rightarrow S \rightarrow \mathbb{R})$  where  $r(s, a, s')$  is the reward obtained on transition from state  $s$  to state  $s'$  under action  $a$ .

From these definitions it follows that the rewards are bounded in absolute value: since the state and action spaces are finite, there exists a constant  $D$  such that

$$\forall (s s' : S), (a : A(s)), |r(s, a, s')| \leq D \clubsuit \quad (9)$$

**Definition 7** (Decision Rule / Policy). *Given a Markov decision process with state space  $S$  and action space  $\prod_{s:S} A(s)$ ,*

- A function  $\pi : \prod_{s:S} A(s)$  is called a decision rule  $\clubsuit$ . The decision rule is deterministic<sup>3</sup>.
- A stationary policy is an infinite sequence of decision rules:  $(\pi, \pi, \pi, \dots)$   $\clubsuit$ . Stationary implies that the same decision rule is applied at each step.

This policy  $\pi$  induces a stochastic dynamic process on  $S$  evolving in discrete time steps  $k \in \mathbb{Z}_{\geq 0}$ . In this section we consider only stationary policies, and therefore use the terms *policy* and *decision rule* interchangeably.

**3.1.2 Kleisli Composites in a Markov Decision Process.** Note that for a fixed decision rule  $\pi$ , we get a Kleisli arrow  $T_\pi : S \rightarrow P(S)$  defined as  $T_\pi(s) = T(s)(\pi(s))$ .

Conventionally,  $T_\pi$  is represented as a row-stochastic matrix  $(T_\pi)^s_{s'}$  that acts on the probability co-vectors from the right, so that the row  $s$  of  $T_\pi$  corresponding to state  $s$  encodes the probability distribution of states  $s'$  after a transition from the state  $s$ .

Let  $p_k \in P(S)$  for  $k \in \mathbb{Z}_{\geq 0}$  denote a probability distribution on  $S$  evolving under the policy stochastic map  $T_\pi$  after  $k$  transition steps, so that  $p_0$  is the initial probability distribution on  $S$  (the initial distribution is usually taken to be  $\text{ret } s_0$  for a state  $s_0$ ). These are related by

$$p_k = p_0 T_\pi^k \quad (10)$$

In general (if  $p_0 = \text{ret } s_0$ ) the number  $p_k(s)$  gives the probability that starting out at  $s_0$ , one ends up at  $s$  after  $k$  stages. So, for example, if  $k = 1$ , we recover the stochastic transition structure at the end of the first step  $\clubsuit$ .

<sup>2</sup>There are various definitions of finite. Our mechanization uses surjective finiteness (the existence of a surjection from a bounded set of natural numbers) $\clubsuit$ , and assumes that there is a decidable equality on  $S$ . This pair of assumptions is equivalent to bijective finiteness.

<sup>3</sup>if the decision rule takes a state and returns a probability distribution on actions instead, it is called *stochastic*.

Instead of representing  $T_\pi^k$  as an iterated product of a stochastic matrix in our formalization, we recognize that (10) states that  $p_k$  is the  $k$ -fold iterated Kleisli composite of  $T_\pi$  applied to the initial distribution  $p_0$   $\clubsuit$ .

$$p_k = (p_0 \multimap \underbrace{T_\pi \multimap \dots \multimap T_\pi}_{k \text{ times}}) \quad (11)$$

Thus, we bypass the need to define matrices and matrix multiplication entirely in the formalization.

**3.1.3 Long-Term Value of a Markov Decision Process.** Since the transition from one state to another by an action is governed by a probability distribution  $T$ , there is a notion of expected reward with respect to that distribution.

**Definition 8** (Expected immediate reward). *For a Markov decision process,*

- An expected immediate reward to be obtained in the transition under action  $a$  from state  $s$  to state  $s'$  is a function  $\bar{r} : S \rightarrow A \rightarrow \mathbb{R}$  computed by averaging the reward function over the stochastic transition map to a new state  $s'$

$$\bar{r}(s, a) := \sum_{s' \in S} r(s, a, s') T(s, a)(s') \quad (12)$$

- An expected immediate reward under a decision rule  $\pi$ , denoted  $\bar{r}_\pi : S \rightarrow \mathbb{R}$  is defined to be:

$$\bar{r}_\pi(s) := \bar{r}(s, \pi(s)) \quad \clubsuit \quad (13)$$

That is, we replace the action argument in (12) by the action prescribed by the decision rule  $\pi$ .

- The expected reward at time step  $k$  of a Markov decision process starting at initial state  $s$ , following policy  $\pi$  is defined as the expected value of the reward with respect to the  $k$ -th Kleisli iterate of  $T_\pi$  starting at state  $s$ .

$$r_k^\pi(s) := \mathbb{E}_{T_\pi^k(s)} [\bar{r}_\pi] = \sum_{s' \in S} [\bar{r}_\pi(s') T_\pi^k(s)(s')] \quad \clubsuit$$

The long-term value of a Markov decision process under a policy  $\pi$  is defined as follows:

**Definition 9** (Long-Term Value). *Let  $\gamma \in \mathbb{R}$ ,  $0 \leq \gamma < 1$  be a discount factor, and  $\pi = (\pi, \pi, \dots)$  be a stationary policy. Then  $V_\pi : S \rightarrow \mathbb{R}$  is given by*

$$V_\pi(s) = \sum_{k=0}^{\infty} \gamma^k r_k^\pi(s) \quad \clubsuit \quad (14)$$

The rewards being bounded in absolute value implies that the long-term value function  $V_\pi$  is well-defined for every initial state  $\clubsuit$ .

It can be shown by manipulating the series in (14) that the long-term value satisfies the Bellman equation:

$$V_\pi(s) = \bar{v}(s, \pi(s)) + \gamma \sum_{s' \in S} V_\pi(s') T_\pi(s)(s') \quad \clubsuit \quad (15)$$

$$= \bar{r}_\pi(s) + \gamma \mathbb{E}_{T_\pi(s)} [V_\pi] \quad (16)$$

**Definition 10.** *Given a Markov decision process, we define the Bellman operator as*

$$\mathbf{B}_\pi : (S \rightarrow \mathbb{R}) \rightarrow (S \rightarrow \mathbb{R}) \quad (17)$$

$$W \mapsto \bar{r}_\pi(s) + \gamma \mathbb{E}_{T_\pi(s)} W \quad (18)$$

**THEOREM 11** (PROPERTIES OF THE BELLMAN OPERATOR  $\clubsuit$ ). *The Bellman operator satisfies the following properties:*

- As is evident from (15), the long-term value  $V_\pi$  is the fixed point of the operator  $\mathbf{B}_\pi$   $\clubsuit$ .
- The operator  $\mathbf{B}_\pi$  (called the Bellman operator) is a contraction in the norm (2)  $\clubsuit$ .

- The operator  $\mathbf{B}_\pi$  is a monotone operator. That is,

$$\forall s, W_1(s) \leq W_2(s) \Rightarrow \forall s, \mathbf{B}_\pi(W_1)(s) \leq \mathbf{B}_\pi(W_2)(s)$$

The Banach fixed point theorem now says that  $V_\pi$  is the unique fixed point of this operator.

Let  $V_{\pi,n} : S \rightarrow \mathbb{R}$  be the  $n$ -th iterate of the Bellman operator  $\mathbf{B}_\pi$ . It can be computed by the recursion relation

$$\begin{aligned} V_{\pi,0}(s_0) &= 0 \\ V_{\pi,n+1}(s_0) &= \bar{r}_\pi(s_0) + \gamma \mathbb{E}_{T_\pi(s_0)} V_{\pi,n} \quad n \in \mathbb{Z}_{\geq 0} \end{aligned} \quad (19)$$

where  $s_0$  is an arbitrary initial state. The first term in the reward function  $V_{\pi,n+1}$  for the process of length  $n+1$  is the sum of the reward collected in the first step (*immediate reward*), and the remaining total reward obtained in the subsequent process of length  $n$  (*discounted future reward*). The  $n$ -th iterate is also seen to be equal to the  $n$ -th partial sum of the series (14)  $\clubsuit$ .

The sequence of iterates  $\{V_{\pi,n}\}_{n=0,1,2,\dots}$  is convergent and equals  $V_\pi$ , by the Banach fixed point theorem.

$$V_\pi = \lim_{n \rightarrow \infty} V_{\pi,n} \quad \clubsuit \quad (20)$$

### 3.2 Convergence of Value Iteration

In the previous subsection we defined the long-term value function  $V_\pi$  and showed that it is the fixed point of the Bellman operator. It is also the pointwise limit of the iterates  $V_{\pi,n}$ , which is the expected value of all length  $n$  realizations of the Markov decision process following a fixed stationary policy  $\pi$ .

We note that the value function  $V_\pi$  induces a partial order on the space of all decision rules; with  $\sigma \leq \tau$  if and only if  $V_\sigma \leq V_\tau$   $\clubsuit$ .

The space of all decision rules is finite because the state and action spaces are finite  $\clubsuit$ .

The above facts imply the existence of a decision rule (stationary policy) which maximizes the long-term reward. We call this stationary policy the *optimal policy* and its long-term value the *optimal value function*.

$$V_*(s) = \max_{\pi} \{V_\pi(s)\} \quad \clubsuit \quad (21)$$

The aim of reinforcement learning, as we remarked in the introduction, is to have tractable algorithms to find the optimal policy and the optimal value function corresponding to the optimal policy.

Bellman's *value iteration* algorithm is such an algorithm, which is known to converge asymptotically to the optimal value function. In this section we describe this algorithm and formally prove this convergence property.

**Definition 12.** Given a Markov decision process we define the Bellman optimality operator as:

$$\begin{aligned} \hat{\mathbf{B}} : (S \rightarrow \mathbb{R}) &\rightarrow (S \rightarrow \mathbb{R}) \\ W &\mapsto \lambda s, \max_{a \in A(s)} (\bar{r}(s, a) + \gamma \mathbb{E}_{T(s,a)} [W]) \quad \clubsuit \end{aligned}$$

**THEOREM 13.** The Bellman optimality operator  $\hat{\mathbf{B}}$  satisfies the following properties:

- The operator  $\hat{\mathbf{B}}$  is a contraction with respect to the  $L^\infty$  norm (2)  $\clubsuit$ .
- The operator  $\hat{\mathbf{B}}$  is a monotone operator. That is,

$$\forall s, W_1(s) \leq W_2(s) \Rightarrow \forall s, \hat{\mathbf{B}}(W_1)(s) \leq \hat{\mathbf{B}}(W_2)(s) \quad \clubsuit$$

Now we move on to proving the most important property of  $\hat{\mathbf{B}}$ : the optimal value function  $V_*$  is a fixed point of  $\hat{\mathbf{B}}$ .

By Theorem 13 and the Banach fixed point theorem, we know that the fixed point of  $\hat{\mathbf{B}}$  exists. Let us denote it  $\hat{V}$ . Then we have:

**THEOREM 14 (LEMMA 1 OF [FHM18] ✿).** *For every decision rule  $\sigma$ , we have  $V_\sigma \leq \hat{V}$ .*

**PROOF.** Fix a policy  $\sigma$ . Note that for every  $f : S \rightarrow \mathbb{R}$ , we have  $\mathbf{B}_\sigma(f) \leq \hat{\mathbf{B}}(f)$  ✿. In particular, applying this to  $f = V_\sigma$  and using Theorem 11, we get that  $V_\sigma = \mathbf{B}_\sigma(V_\sigma) \leq \hat{\mathbf{B}}(V_\sigma)$ . Now by contraction coinduction (Theorem 5 with  $F = \hat{\mathbf{B}}$  along with Theorem 13) we get that  $V_\sigma \leq \hat{V}$ .  $\square$

Theorem 14 immediately implies that  $V_* \leq \hat{V}$ .

To go the other way, we introduce the following policy, called the *greedy* decision rule.

$$\sigma_*(s) := \operatorname{argmax}_{a \in A(s)} \left( \bar{r}(a, s) + \gamma \mathbb{E}_{T(s,a)}[\hat{V}] \right) \quad \text{✿} \quad (22)$$

We now have the following theorem:

**THEOREM 15 (PROPOSITION 1 OF [FHM18] ✿).** *The greedy policy is the policy whose long-term value is the fixed point of  $\hat{\mathbf{B}}$ :*

$$V_{\sigma_*} = \hat{V}$$

**PROOF.** We observe that  $\mathbf{B}_{\sigma_*}(\hat{V}) = \hat{V}$  ✿. Thus,  $\hat{V} \leq \mathbf{B}_{\sigma_*}(\hat{V})$ . Note that we have  $V_{\sigma_*}$  is the fixed point of  $B_{\sigma_*}$  by Theorem 11. Now applying contraction coinduction with  $F = \mathbf{B}_{\sigma_*}$ , we get  $\hat{V} \leq V_{\sigma_*}$ . From Theorem 14 we get that  $V_{\sigma_*} \leq \hat{V}$ .  $\square$

Theorem 15 implies that  $V_* \geq \hat{V}$  and so we conclude that  $V_* = \hat{V}$  ✿.

Thus, the fixed point of the optimal Bellman operator  $\hat{\mathbf{B}}$  exists and is equal to the optimal value function. Stated fully, value iteration proceeds by:

- (1) Initialize a value function  $V_0 : S \rightarrow \mathbb{R}$ .
- (2) Define  $V_{n+1} = \hat{\mathbf{B}}V_n$  for  $n \geq 0$ . At each stage, the following policy is computed

$$\pi_n(s) \in \operatorname{argmax}_{a \in A(s)} \left( \bar{r}(s, a) + \gamma \mathbb{E}_{T(s,a)}[V_n] \right)$$

By the Banach Fixed Point Theorem, the sequence  $\{V_n\}$  converges to the optimal value function  $V_*$  ✿. In practice, one repeats this iteration as many times as needed until a fixed threshold is breached.

In Section 3.4 we explain and provide a formalized proof of the *dynamic programming principle*: the value function  $V_n$  is equal to the *optimal value function* of a finite-horizon MDP of length  $n$  with a possibly non-stationary optimal policy.

### 3.3 Convergence of Policy Iteration

The convergence of value iteration is asymptotic, which means the iteration is continued until a fixed threshold is breached. Policy iteration is a similar iterative algorithm that benefits from a more definite stopping condition. Define the *Q function* to be:

$$Q_\pi(s, a) := \bar{r}(s, a) + \gamma \mathbb{E}_{T(s,a)}[V_\pi].$$

The policy iteration algorithm proceeds in the following steps:

- (1) Initialize the policy to  $\pi_0$ .
- (2) Policy evaluation: For  $n \geq 0$ , given  $\pi_n$ , compute  $V_{\pi_n}$ .
- (3) Policy improvement: From  $V_{\pi_n}$ , compute the greedy policy:

$$\pi_{n+1}(s) \in \operatorname{argmax}_{a \in A(s)} [Q_{\pi_n}(s, a)]$$

- (4) Check if  $V_{\pi_n} = V_{\pi_{n+1}}$ . If yes, stop.
- (5) If not, repeat (2) and (3).

This algorithm depends on the following results for correctness. We follow the presentation from [FHM18].

**Definition 16** (Improved policy  $\clubsuit$ ). *A policy  $\tau$  is called an improvement of a policy  $\sigma$  if for all  $s \in S$  it holds that*

$$\tau(s) = \operatorname{argmax}_{a \in A(s)} [Q_\sigma(s, a)]$$

So, step (2) of the policy iteration algorithm simply constructs an improved policy from the previous policy at each stage.

**THEOREM 17** (POLICY IMPROVEMENT THEOREM). *Let  $\sigma$  and  $\tau$  be two policies.*

- If  $\mathbf{B}_\tau V_\sigma \geq \mathbf{B}_\sigma V_\sigma$  then  $V_\tau \geq V_\sigma$   $\clubsuit$ .
- If  $\mathbf{B}_\tau V_\sigma \leq \mathbf{B}_\sigma V_\sigma$  then  $V_\tau \leq V_\sigma$   $\clubsuit$ .

Using the above theorem, we have:

**THEOREM 18** (POLICY IMPROVEMENT IMPROVES VALUES  $\clubsuit$ ). *If  $\sigma$  and  $\tau$  are two policies and if  $\tau$  is an improvement of  $\sigma$ , then we have  $V_\tau \geq V_\sigma$ .*

**PROOF.** From Theorem 17, it is enough to show  $\mathbf{B}_\tau V_\sigma \geq \mathbf{B}_\sigma V_\sigma$ . We have that  $\tau$  is an improvement of  $\sigma$ .

$$\tau(s) = \operatorname{argmax}_{a \in A(s)} [Q_\sigma(s, a)] \quad (23)$$

$$= \operatorname{argmax}_{a \in A(s)} [\bar{r}(s, a) + \gamma \mathbb{E}_{T(s,a)} [V_\sigma]] \quad (24)$$

Note that

$$\begin{aligned} \mathbf{B}_\tau V_\sigma &= \bar{r}(s, \tau(s)) + \gamma \mathbb{E}_{T(s, \tau(s))} [V_\sigma] \\ &= \max_{a \in A(s)} [\bar{r}(s, a) + \gamma \mathbb{E}_{T(s,a)} [V_\sigma]] \quad \text{by (24)} \\ &\geq \bar{r}(s, \sigma(s)) + \gamma \mathbb{E}_{T(s, \sigma(s))} [V_\sigma] \\ &= \mathbf{B}_\sigma V_\sigma \end{aligned}$$

□

In other words, since  $\pi_{n+1}$  is an improvement of  $\pi_n$  by construction, the above theorem implies that  $V_{\pi_n} \leq V_{\pi_{n+1}}$ . This means that  $\pi_n \leq \pi_{n+1}$ .

Thus, the policy constructed in each stage in the policy iteration algorithm is an improvement of the policy in the previous stage. Since the set of policies is finite  $\clubsuit$ , this policy list must at some point stabilize. Thus, the algorithm is guaranteed to terminate.

In Section 3.4 we will provide formalization of the statement that  $\pi_n$  is actually the *optimal policy* to follow for an MDP process of any finite length at that timestep when  $n$  steps remain towards the end of the process.

### 3.4 Optimal policy for finite time horizon Markov decision processes

All results up to this subsection were stated in terms of the convergences of infinite sequences of states and actions. Stating convergence results in terms of the limits of infinite sequences is not uncommon in texts on reinforcement learning; however, in practice, reinforcement learning algorithms are always run for a finite number of steps. In this section we consider decision processes of finite length and do not impose an assumption that the optimal policy is stationary.

Let  $V_{\bar{\pi}}$  denote the value function of Markov decision process for a finite sequence of policies  $\bar{\pi} = \pi_0 :: \pi_1 :: \pi_2 :: \dots :: \pi_{n-1}$  of length  $n = \operatorname{len}(\bar{\pi})$ . Denote by  $p_0$  the probability distribution over the initial state at the start of the process.

We define the probability measure at step  $k$  in terms of Kleisli iterates for *each decision rule*  $\pi_i$  for  $i$  in  $0 \dots (k-1)$ :

$$p_0 T_{\bar{\pi}[k]} := (p_0 \blacktriangleright T_{\pi_0} \dots \blacktriangleright T_{\pi_{k-1}}) \quad \clubsuit \quad (25)$$

Below, we will use the pairing notation (bra-ket notation)

$$\langle p|V \rangle := \mathbb{E}_p[V] \quad (26)$$

between a probability measure  $p$  on a finite set  $S$  and a function  $V : S \rightarrow \mathbb{R}$ , so that  $|V \rangle$  is an element of the vector space of real valued functions on  $S$ ,  $\langle p|$  is a linear form on this vector space associated to a probability measure  $p$  on  $S$ , and  $\langle p|V \rangle$  denotes evaluation of a linear form  $\langle p|$  on a vector  $|V \rangle$ .

**Definition 19** (expectation value function of MDP of length  $n = \text{len}(\bar{\pi})$  over the initial probability distribution  $p_0$ ).

$$\langle p_0|V_{\bar{\pi}} \rangle = \sum_{k=0}^{n-1} \gamma^k \langle p_0 T_{\bar{\pi}[:k]} |\bar{r}_{\pi_k} \rangle \clubsuit \quad (27)$$

Definition 19 implies the recursion relation

$$\langle p_0|V_{\pi_0::\text{tail}} \rangle = \langle p_0|\bar{r}_{\pi_0} + \gamma T_{\pi_0} V_{\text{tail}} \rangle \clubsuit \quad n \in \mathbb{Z}_{\geq 0} \quad (28)$$

where  $\bar{\pi} = \pi_0 :: \text{tail}$ .

Let  $\hat{V}_{*,n}$  be the optimal value function of the Markov decision process of length  $n$  on the space of all policy sequences of length  $n$ :

$$\hat{V}_{*,n} := \sup_{\bar{\pi} | \text{len}(\bar{\pi})=n} V_{\bar{\pi}} \clubsuit \quad (29)$$

Let  $\hat{V}_{\pi_0::*,n+1}$  be the optimal value function of the Markov decision process of length  $n+1$  on the space of all policy sequences of length  $n+1$  whose initial term is  $\pi_0$ . Using the relation (28) and that

$$\sup_{\pi_0::\text{tail}} V_{\pi_0::\text{tail},n+1} = \sup_{\pi_0} \sup_{\text{tail}} V_{\pi_0::\text{tail},n+1} \clubsuit \quad (30)$$

we find

$$\langle p_0|\hat{V}_{*,n+1} \rangle = \sup_{\pi_0 \in \prod_S A(s)} \langle p_0|\bar{r}_{\pi_0} + \gamma T_{\pi_0} \hat{V}_{*,n} \rangle \clubsuit \quad n \in \mathbb{Z}_{>=0} \quad (31)$$

with the initial term of the sequence  $V_{*,0} = 0$ .

The result (31) can be formulated as follows

**THEOREM 20 (BELLMAN'S FINITE-TIME OPTIMAL POLICY THEOREM).** *The optimal value function  $\hat{V}_{*,n+1}$  of a Markov decision process of length  $n+1$  relates to the optimal value function of the same Markov decision process of length  $n$  by the inductive relation*

$$\hat{V}_{*,n+1} = \hat{\mathbf{B}}V_{*,n} \quad (32)$$

where  $\hat{\mathbf{B}}$  is Bellman optimality operator (Definition 12).

The iterative computation of the sequence of optimal value functions  $\{\hat{V}_{*,n}\}_{n \in \mathbb{Z}_{\geq 0}}$  of Markov decision processes of length  $n = 0, 1, 2, \dots$  from the recursion  $\hat{V}_{*,n+1} = \hat{\mathbf{B}}\hat{V}_{*,n}$  is the same algorithm as *value iteration*.

### 3.5 Comments on Formalization

CertRL contributes a formal library for reasoning about Markov decision processes. We demonstrate the effectiveness of this library's building blocks by proving the two most canonical results from reinforcement learning theory. In this section we reflect on the structure of CertRL's formalization, substantiating our claim that CertRL serves as a convenient foundations for a continuing line of work on formalization of reinforcement learning theory.

**3.5.1 Characterizing Optimality.** Most texts on Markov decision processes (for example [Put94, Section 2.1.6]) start out with a probability space on the space of all possible realizations of the Markov decision process. The long-term value for an infinite-horizon Markov decision process is then defined as the expected value over all possible realizations:

$$V_\pi(s) = \mathbb{E}_{(x_1, x_2, \dots)} \left[ \sum_{k=0}^{\infty} \gamma^k v(x_k, \pi(x_k)) \mid x_0 = s; \pi \right] \quad (33)$$

where each  $x_k$  is drawn from the distribution  $T(x_{k-1}, \pi(x_{k-1}))$ . This definition is hard to work with because, as [Put94] notes, it ignores the dynamics of the problem. Fortunately, it is also unnecessary since statements about the totality of all realizations are rarely made.

In our setup, following [FHM18], we only consider the probability space over the finite set of states of the Markov decision process. By identifying the basic operation of Kleisli composition, we generate more realizations (and their expected rewards) on the fly as and when needed.

Implementations of reinforcement learning algorithms often compute the long-term value using matrix operators for efficiency reasons. The observation that clean theoretical tools do not necessarily entail efficient implementations is not a new observation; both Puterman [Put94] and Hölzl [Hoe17a] make similar remarks. Fortunately, the design of our library provides a clean interface for future work on formalizing efficiency improvements. Extending CertRL with correctness theorems for algorithms that use matrix operations requires nothing more than a proof that the relevant matrix operations satisfy the definition of Kleisli composition.

**3.5.2 Comparison of English and Coq Proofs.** Comparing Theorem 14 and Theorem 15 with the the equivalent results from Puterman [Put94, Theorem 6.2.2] demonstrates that CertRL avoids reasoning about low-level  $\epsilon - \delta$  details through strategic use of coinduction.

The usefulness of contraction coinduction is reflected in the formalization, sometimes resulting in Coq proofs whose length is almost the same as the English text.

We compare in Table 1 the Coq proof of Theorem 15 to an English proof of the same. The two proofs are roughly equivalent in length and, crucially, also make essentially the same argument at the same level of abstraction. Note that what we compare is not *exactly* the proof from Feys et al. [FHM18, Proposition 1], but is as close as possible to a restatement of their Proposition 1 and Lemma 1 with the proof of Lemma 1 inlined and the construction restated in terms of our development. The full proof from [FHM18], with Lemma 1 inlined, reads as follows:

**Proposition 1:** The greedy policy is optimal. That is,  $LTV_{\sigma^*} = V^*$ .

- (1) Observe that  $\Psi_{\sigma^*} \geq V^*$  (in fact, equality holds).
- (2) By contraction coinduction,  $V^* \leq LTV_{\sigma^*}$ .
- (3) **Lemma 1:** For all policies  $\sigma$ ,  $LTV_\sigma \leq V^*$ .
- (4) A straightforward calculation and monotonicity argument shows that for all  $f \in B(S, \mathbb{R})$ ,  $\Psi_\sigma(f) \leq \Psi^*(f)$ .
- (5) In particular,  $LTV_\sigma = \Psi_\sigma(LTV_\sigma) \leq \Psi^*(LTV_\sigma)$ .
- (6) By contraction coinduction we conclude that  $LTV_\sigma \leq V^*$ .

Table 1 compares two coinductive proofs – one in English and the other in Coq. Another important comparison is between a Coq coinductive proof and an English non-coinductive proof. The Coq proof of the policy improvement theorem provides one such point of comparison. Recall that theorem states that a particular closed property (the set  $\{x \mid x \leq y\}$ ) holds of the fixed point of a particular contractive map (the Bellman operator). The most common argument – presented in the most common textbook on reinforcement learning – proves this theorem by expanding the infinite sum in multiple steps [SB98, Section 4.2]. We reproduce this below:

**THEOREM 15 (PROPOSITION 1 OF [FHM18] ♣).** *The greedy policy is the policy whose long-term value is the fixed point of  $\hat{\mathbf{B}}$ :*

$$V_{\sigma_*} = \hat{V}$$

PROOF.

- (1)  $V_{\sigma_*} \leq \hat{V}$  follows by Theorem 14.
- (2) Now we have to show  $\hat{V} \leq V_{\sigma_*}$ . Note that we have  $V_{\sigma_*}$  is the fixed point of  $B_{\sigma_*}$  by Theorem 11.
- (3) We can now apply contraction coinduction with  $F = \mathbf{B}_{\sigma_*}$ .
- (4) The hypotheses are satisfied since by Theorem 11, the  $\mathbf{B}_{\sigma_*}$  is a contraction and it is a monotone operator.
- (5) The only hypothesis left to show is  $\hat{V} \leq \mathbf{B}_{\sigma_*} \hat{V}$ .
- (6) But in fact, we have  $\mathbf{B}_{\sigma_*}(\hat{V}) = \hat{V}$  by the definition of  $\sigma_*$ .

□

(a) English proof adapted from [FHM18].

```

1 Lemma exists_fixpt_policy
  : forall init,
2   let V' := fixpt (bellman_max_op) in
3   let pi' := greedy init in
4   ltv gamma pi' = V' init.
5 Proof.
6 intros init V' pi';
7 eapply Rfct_le_antisym; split.
8 - eapply ltv_Rfct_le_fixpt.
9 - rewrite (ltv_bellman_op_fixpt _ init).
10  apply contraction_coinduction_Rfct_ge'.
11  + apply is_contraction_bellman_op.
12  + apply bellman_op_monotone_ge.
13  + unfold V', pi'.
14    now rewrite greedy_argmax_is_max.
15 Qed.
    
```

(b) Coq proof ♣

Table 1. Comparison of English and Coq proofs of Theorem 15.

**THEOREM 21 (POLICY IMPROVEMENT THEOREM [SB98]).** *Let  $\pi, \pi'$  be a pair of deterministic policies such that, for all states  $s$ ,*

$$Q_{\pi}(s, \pi'(s)) \geq V_{\pi}(s) \tag{34}$$

then  $V_{\pi'}(s) \geq V_{\pi}(s)$ .

PROOF. Starting with (34) we keep expanding the  $Q_{\pi}$  side and reapplying (34) until we get  $V_{\pi'}(s)$ .

$$\begin{aligned}
 V_{\pi}(s) &\leq Q_{\pi}(s, \pi'(s)) \\
 &= \mathbb{E}_{\pi'} \{r_{t+1} + \gamma V_{\pi}(s_{t+1}) \mid s_t = s\} \\
 &\leq \mathbb{E}_{\pi'} \{r_{t+1} + \gamma Q_{\pi}(s_{t+1}, \pi'(s_{t+1})) \mid s_t = s\} \\
 &= \mathbb{E}_{\pi'} \{r_{t+1} + \gamma \mathbb{E}_{\pi'} \{r_{t+2} + \gamma V_{\pi}(s_{t+2})\} \mid s_t = s\} \\
 &= \mathbb{E}_{\pi'} \{r_{t+1} + \gamma r_{t+2} + \gamma^2 V_{\pi}(s_{t+2}) \mid s_t = s\} \\
 &\leq \mathbb{E}_{\pi'} \{r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \gamma^3 V_{\pi}(s_{t+3}) \mid s_t = s\} \\
 &\vdots \\
 &\leq \mathbb{E}_{\pi'} \{r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \gamma^3 r_{t+4} \dots \mid s_t = s\} \\
 &= V_{\pi'}(s)
 \end{aligned}$$

□

At a high level, this proof proceeds by showing that the closed property  $\{x|x \leq y\}$  holds of each partial sum of the infinite series  $V_\pi(s)$ . By completeness and using the fact that partial sums converge to the full series  $V_\pi$ , this property is also shown to hold of the fixed point  $V_\pi$ .

However, the coinductive version of this proof (Theorem 17) is simpler because it exploits the fact that this construction has already been done once in the proof of the fixed point theorem: the iterates of the contraction operator were already proven to converge to the fixed point and so there is no reason to repeat the construction again. Thus, the proof is reduced to simply establishing the “base case” of the (co)induction.

The power of this method goes beyond simplifying proofs for Markov decision processes. See [KR09] for other examples.

#### 4 RELATED AND FUTURE WORK

To our knowledge, CertRL is the first formal proof of convergence for value iteration or policy iteration. Related work falls into three categories:

- (1) libraries that CertRL builds upon,
- (2) formalizations of results from probability and machine learning, and
- (3) work at the intersection of formal verification and reinforcement learning.

*Dependencies.* CertRL builds on the Coquelicot [BLM14] library for real analysis. Our main results are statements about fixed points of contractive maps in complete normed modules. CertRL therefore builds on the formal development of the Lax-Milgram theorem and, in particular, Boldo et al.’s formal proof of the Banach fixed point theorem [BCF<sup>+</sup>17]. CertRL also makes extensive use of some utilities from the Q\*cert project [AHM<sup>+</sup>17]. CertRL includes a bespoke implementation of some basic results and constructions from probability theory and also an implementation of the Giry monad. Our use of the monad for reasoning about probabilistic processes, as well as the design of our library, is highly motivated by the design of the Polaris library [TH19]. Many of the thorough formalizations of probabilities in Coq – such as the Polaris [TH19], Infotheo [AH12], and Alea [APM09] – also contain these results. Refactoring CertRL to build on top of one or more of these formalizations might allow future work on certified reinforcement learning to leverage future improvements to these libraries.

Building on these other foundations, CertRL demonstrates how existing work on formalization enables formalization of key results in reinforcement learning theory.

*Related Formalizations.* There is a growing body of work on formalization of machine learning theory [TTV19, TTV<sup>+</sup>20, Hoe17b, SLD17, BS19, BBK19].

Johannes Hölzl’s Isabelle/HOL development of Markov processes is most related to our own work [Hoe17b, Hoe17a]. Hölzl builds on the probability theory libraries of Isabelle/HOL to develop continuous-time Markov chains. Many of Hölzl’s basic design choices are similar to ours; for example, he also uses the Giry monad to place a monadic structure on probability spaces and also uses coinductive methods. CertRL focuses instead on formalization of convergence proofs for dynamic programming algorithms that solve Markov decision processes. In the future, we plan to extend our formalization to include convergence proofs for model-free methods, in which a fixed Markov decision process is not known *a priori*.

The CertiGrad formalization by Selsam et al. contains a Lean proof that the gradients sampled by a stochastic computation graph are unbiased estimators of the true mathematical function [SLD17]. This result, together with our development of a library for proving convergence of reinforcement learning algorithms, provides a path toward a formal proof of correctness for deep reinforcement learning.

*Formal Methods for RL.* The likelihood that reinforcement learning algorithms will be deployed in safety-critical settings during the coming decades motivates a growing body of work on formal methods for safe reinforcement learning. This approach – variously called formally constrained reinforcement learning [HAK18],

shielding[ABE<sup>+</sup>18], or verifiably safe reinforcement learning [HFM<sup>+</sup>20] – uses temporal or dynamic logics to specify constraints on the behavior of RL algorithms.

Global convergence is a fundamental theoretical property of classical reinforcement learning algorithms, and in practice at least local convergence is an important property for any useful reinforcement learning algorithm. However, the formal proofs underlying these methods typically establish the correctness of a safety constraint but do not formalize any convergence properties. In future work, we plan to establish an end-to-end proof that constrained reinforcement learning safely converges by combining our current development with the safe RL approach of Fulton et al. [FP18] and the VeriPhy pipeline of Bohrer et al. [BTM<sup>+</sup>18].

## 5 CONCLUSIONS

Reinforcement learning algorithms are an important class of machine learning algorithms that are now being deployed in safety-critical settings. Ensuring the correctness of these algorithms is societally important, but proving properties about stochastic processes presents several challenges. In this paper we show how a combination of metric coinduction and the Giry monad provides a convenient setting for formalizing convergence proofs for reinforcement learning algorithms.

## ACKNOWLEDGMENTS

We thank Larry Moss, Sylvie Boldo and Mark Squillante for discussions related to this paper. Some of the work described in this paper was performed while Koundinya Vajjha was an intern at IBM Research. Vajjha was additionally supported by the Alfred P. Sloan Foundation under grant number G-2018-10067.

## REFERENCES

- [ABE<sup>+</sup>18] Mohammed Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Könighofer, Scott Niekum, and Ufuk Topcu. Safe reinforcement learning via shielding. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI 2018)*. AAAI Press, 2018.
- [AH12] Reynald Affeldt and Manabu Hagiwara. Formalization of Shannon’s theorems in SSReflect-Coq. In *3rd Conference on Interactive Theorem Proving (ITP 2012)*, Princeton, New Jersey, USA, August 13–15, 2012, volume 7406 of *Lecture Notes in Computer Science*, pages 233–249. Springer, Aug 2012.
- [AHM<sup>+</sup>17] Joshua S. Auerbach, Martin Hirzel, Louis Mandel, Avraham Shinnar, and Jérôme Siméon. Q\*cert: A platform for implementing and verifying query compilers. In Semih Salihoglu, Wencho Zhou, Rada Chirkova, Jun Yang, and Dan Suciu, editors, *Proceedings of the 2017 ACM International Conference on Management of Data, SIGMOD Conference 2017, Chicago, IL, USA, May 14–19, 2017*, pages 1703–1706. ACM, 2017.
- [APM09] Philippe Audebaud and Christine Paulin-Mohring. Proofs of randomized algorithms in Coq. *Science of Computer Programming*, 74(8):568–589, 2009.
- [BBK19] Alexander Bentkamp, Jasmin Christian Blanchette, and Dietrich Klakow. A formal proof of the expressiveness of deep learning. *J. Autom. Reason.*, 63(2):347–368, 2019.
- [BCF<sup>+</sup>17] Sylvie Boldo, François Clément, Florian Faissole, Vincent Martin, and Micaela Mayero. A Coq formal proof of the Lax–Milgram theorem. In *6th ACM SIGPLAN Conference on Certified Programs and Proofs*, Paris, France, January 2017.
- [Bel54] Richard Bellman. The theory of dynamic programming. *Bull. Amer. Math. Soc.*, 60(6):503–515, 11 1954.
- [BLM14] Sylvie Boldo, Catherine Lelay, and Guillaume Melquiond. Coquelicot: A user-friendly library of real analysis for Coq. *Mathematics in Computer Science*, 9, 03 2014.
- [BS19] Alexander Bagnall and Gordon Stewart. Certifying the true error: Machine learning in Coq with verified generalization guarantees. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019*, pages 2662–2669. AAAI Press, 2019.
- [BTM<sup>+</sup>18] Brandon Bohrer, Yong Kiam Tan, Stefan Mitsch, Magnus O. Myreen, and André Platzer. VeriPhy: Verified controller executables from verified cyber-physical system models. In Dan Grossman, editor, *Proceedings of the 39th ACM SIGPLAN Conference on Programming Language Design and Implementation (PLDI 2018)*, pages 617–630. ACM, 2018.
- [EHN15] Manuel Eberl, Johannes Hölzl, and Tobias Nipkow. A verified compiler for probability density functions. In Jan Vitek, editor, *ESOP 2015*, volume 9032 of *LNCS*, pages 80–104. Springer, 2015.
- [FHM18] Frank MV Feys, Helle Hvid Hansen, and Lawrence S Moss. Long-term values in Markov decision processes, (co)algebraically. In *International Workshop on Coalgebraic Methods in Computer Science*, pages 78–99. Springer, 2018.

- [FP18] Nathan Fulton and André Platzer. Safe reinforcement learning via formal methods: Toward safe control through proof and learning. In Sheila McIlraith and Kilian Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI 2018)*, pages 6485–6492. AAAI Press, 2018.
- [GHLL16] Shixiang Gu, Ethan Holly, Timothy P. Lillicrap, and Sergey Levine. Deep reinforcement learning for robotic manipulation. *CoRR*, abs/1610.00633, 2016.
- [Gir82] Michèle Giry. A categorical approach to probability theory. In B. Banaschewski, editor, *Categorical Aspects of Topology and Analysis*, pages 68–85, Berlin, Heidelberg, 1982. Springer Berlin Heidelberg.
- [HAK18] Mohammadhosein Hasanbeig, Alessandro Abate, and Daniel Kroening. Logically-correct reinforcement learning. *CoRR*, abs/1801.08099, 2018.
- [HFM<sup>+</sup>20] Nathan Hunt, N. Fulton, Sara Magliacane, N. Hoàng, Subhro Das, and Armando Solar-Lezama. Verifiably safe exploration for end-to-end reinforcement learning. *ArXiv*, abs/2007.01223, 2020.
- [Hoe17a] Johannes Hoelzl. Markov chains and Markov decision processes in Isabelle/HOL. *Journal of Automated Reasoning*, 2017.
- [Hoe17b] Johannes Hoelzl. Markov processes in Isabelle/HOL. In *Proceedings of the 6th ACM SIGPLAN Conference on Certified Programs and Proofs, CPP 2017*, page 100–111, New York, NY, USA, 2017. Association for Computing Machinery.
- [How60] R.A. Howard. *Dynamic Programming and Markov Processes*. Technology Press of Massachusetts Institute of Technology, 1960.
- [Jac18] Bart Jacobs. From probability monads to commutative effectuses. *Journal of Logical and Algebraic Methods in Programming*, 94:200 – 237, 2018.
- [JP89] C. Jones and Gordon D. Plotkin. A probabilistic powerdomain of evaluations. In *Proceedings of the Fourth Annual Symposium on Logic in Computer Science (LICS '89), Pacific Grove, California, USA, June 5-8, 1989*, pages 186–195. IEEE Computer Society, 1989.
- [Koz07] Dexter Kozen. Coinductive proof principles for stochastic processes. *CoRR*, abs/0711.0194, 2007.
- [KR09] Dexter Kozen and Nicholas Ruozi. Applications of metric coinduction. *Log. Methods Comput. Sci.*, 5(3), 2009.
- [Law62] F William Lawvere. The category of probabilistic mappings. *preprint*, 1962.
- [Ope18] OpenAI. OpenAI five. <https://blog.openai.com/openai-five/>, 2018.
- [Per18] Paolo Perrone. *Categorical Probability and Stochastic Dominance in Metric Spaces*. PhD thesis, University of Leipzig, 2018.
- [Per19] Paolo Perrone. Notes on category theory with examples from basic mathematics, 2019.
- [Put94] Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley and Sons, Inc., USA, 1st edition, 1994.
- [RP02] Norman Ramsey and Avi Pfeffer. Stochastic lambda calculus and monads of probability distributions. In John Launchbury and John C. Mitchell, editors, *Conference Record of POPL 2002: The 29th SIGPLAN-SIGACT Symposium on Principles of Programming Languages, Portland, OR, USA, January 16-18, 2002*, pages 154–165. ACM, 2002.
- [SB98] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.
- [SEJ<sup>+</sup>20] Andrew Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Židek, Alexander Nelson, Alex Bridgland, Hugo Penedones, Stig Petersen, Karen Simonyan, Steve Crossan, Pushmeet Kohli, David Jones, David Silver, Koray Kavukcuoglu, and Demis Hassabis. Improved protein structure prediction using potentials from deep learning. *Nature*, 577:1–5, 01 2020.
- [ŠGG15] Adam Ścibior, Zoubin Ghahramani, and Andrew D. Gordon. Practical probabilistic programming with monads. In Ben Lippmeier, editor, *Proceedings of the 8th ACM SIGPLAN Symposium on Haskell, Haskell 2015, Vancouver, BC, Canada, September 3-4, 2015*, pages 165–176. ACM, 2015.
- [SHM<sup>+</sup>16] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, January 2016.
- [SLD17] Daniel Selsam, Percy Liang, and David L. Dill. Developing bug-free machine learning systems with formal mathematics. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3047–3056. PMLR, 2017.
- [Tea04] The Coq Development Team. *The Coq Proof Assistant Reference Manual*. LogiCal Project, 2004. Version 8.0.
- [TH19] Joseph Tassarotti and Robert Harper. A separation logic for concurrent randomized programs. *Proceedings of the ACM on Programming Languages*, 3(POPL):1–30, 2019.
- [TTV19] Joseph Tassarotti, Jean-Baptiste Tristan, and Koundinya Vajjha. A formal proof of PAC learnability for decision stumps. *CoRR*, abs/1911.00385, 2019.
- [TTV<sup>+</sup>20] Jean-Baptiste Tristan, Joseph Tassarotti, Koundinya Vajjha, Michael L. Wick, and Anindya Banerjee. Verification of ML systems via reparameterization. *CoRR*, abs/2007.06776, 2020.