A Distributional Analysis of Sampling-Based Reinforcement Learning Algorithms

Philip Amortila[α]   **Doina Precup**[α,β,γ]   **Prakash Panangaden**[α]   **Marc G. Bellemare**[γ,δ]

[α]McGill University; [β]Google DeepMind; [γ]CIFAR Fellow; [δ]Google Research, Brain Team

Abstract

We present a distributional approach to theoretical analyses of reinforcement learning algorithms for constant step-sizes. We demonstrate its effectiveness by presenting simple and unified proofs of convergence for a variety of commonly-used methods. We show that value-based methods such as TD(λ) and Q-Learning have update rules which are contractive in the space of distributions of functions, thus establishing their exponentially fast convergence to a stationary distribution. We demonstrate that the stationary distribution obtained by any algorithm whose target is an expected Bellman update has a mean which is equal to the true value function. Furthermore, we establish that the distributions concentrate around their mean as the step-size shrinks. We further analyse the optimistic policy iteration algorithm, for which the contraction property does not hold, and formulate a probabilistic policy improvement property which entails the convergence of the algorithm.

1 Introduction

Basic results in the theory of Markov decision processes (MDPs) and dynamic programming (DP) rely on the two fundamental properties of the Bellman operator: contraction and monotonicity. For instance, proofs of convergence for value iteration and policy iteration follow immediately from the contractive properties of the Bellman operators and the Banach fixed point theorem. Similarly, convergence of the policy improvement algorithm to an optimal policy follows readily from a monotonicity argument (Puterman, 1994).

However, proving the convergence of sample-based algorithms such as TD-learning (Sutton, 1988) or optimistic policy iteration (Tsitsiklis, 2002) requires substantially more effort. The typical stochastic approximation approach relies on hitting-time or martingale arguments to bound the sequence of value function iterates within progressively smaller regions (see, e.g., Bertsekas and Tsitsiklis, 1996, Section 4.3).

In this work we present a distributional framework for analyzing sample-based reinforcement learning algorithms. Rather than consider the evolution of the random point estimate produced by the learning process, we study the dynamics of the *distribution* of these point estimates. As a concrete example, we view the TD(0) algorithm as defining a sequence of random iterates $(V_n)_{n \in \mathbb{N}}$ whose distributions are recursively defined by the distributional equation

$$V_{n+1}(s) \overset{D}{=} (1-\alpha)V_n(s) + \alpha\left(R(s,A) + \gamma V_n(S')\right), \quad (1)$$

where s is the initial state and (A, R, S') is the random action-reward-next-state transition sampled from the underlying Markov Decision Process.

We study the constant step-size case. Our main contribution is to show that, for a variety of algorithms, the random iterates converge in distribution to a fixed point of the corresponding distributional equation, even though the random point estimate may not converge. We further characterize this fixed point, showing that it depends on both the step-size and the specific Markov Decision Process under consideration. Our framework views the learning process as defining a time-homogeneous Markov chain over the space value functions. We prove convergence by establishing the existence of a stationary distribution for this Markov chain and demonstrating that the sequence of random iterates generated by a sample-based learning rule must converge to this stationary distribution, using tools from optimal transport (Villani, 2008).

We first analyze sample-based algorithms whose corresponding distributional operator is a contraction mapping in the infinity norm, including TD(λ), Q-learning, and double Q-learning. Following a proof technique of (Dieuleveut, Durmus, and Bach, 2017), we lift these stochastic algorithms to the distributional setting. We show that this lifting recovers contractive guarantees, now in the Wasserstein metric using the infinity norm as a cost function. The contraction coefficient depends on the discount factor, as usual, but also on the step-size: updates with smaller step-sizes converge more slowly to their distributional fixed point. TD(0), for example, is a contraction mapping with coefficient $1 - \alpha + \alpha\gamma$.

We also analyze the sample-based equivalent of policy iteration, called optimistic policy iteration (Tsitsiklis, 2002) or Monte Carlo control (Sutton and Barto, 1998). The convergence of policy iteration is not driven by a contraction mapping, but rather by the monotonicity of the policy iteration operator. We derive a similar, weaker property for the sample-based setting which we call probabilistic policy improvement. We use this property to show that optimistic policy improvement also converges to a distributional fixed point.

By recovering the contraction mapping that underlies many dynamic programming algorithms, our distributional analysis significantly simplifies existing proofs of convergence for stochastic algorithms, at least for constant step-sizes. Our approach easily allows us to quantify the limiting behaviour of these algorithms; the same tool even provides us with confidence bounds over the true value function. We believe this type of analysis should prove useful going forward, including for the study of reinforcement learning with function approximation.

2 Background

We write $\mathscr{P}(\mathcal{X})$ for the set of probability distributions on a space $\mathcal{X}$. We consider an agent interacting with an environment modelled as a finite Markov decision process $(\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P}, \gamma)$. As usual, $\mathcal{S}$ is a finite state space, $\mathcal{A}$ is a finite set of actions, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathscr{P}([0, \text{R}_{\text{MAX}}])$ is a bounded reward distribution function, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \mathscr{P}(\mathcal{S})$ is a transition distribution function, and $\gamma \in [0, 1)$ is a discount factor. The strategy of the agent is captured by a policy $\pi : \mathcal{S} \to \mathscr{P}(\mathcal{A})$. The value function $v^\pi : \mathcal{S} \to \mathbb{R}$ of a policy π is the expected discounted sum of rewards observed when starting at state s and following policy π. The value function is the fixed point of the Bellman operator $\mathcal{T}^\pi$

defined by

$$\mathcal{T}^\pi v(s) := \mathbb{E}_{\substack{a \sim \pi(\cdot|s) \\ r \sim \mathcal{R}(\cdot|s,a)}} \left[r + \gamma \mathbb{E}_{s' \sim \mathcal{P}(\cdot|s,a)} \left[v(s') \right] \right]. \quad (2)$$

The value function of the optimal policy $\pi^\star$ is also the fixed point of the Bellman optimality operator $\mathcal{T}^\star$, defined by

$$\mathcal{T}^\star v(s) := \max_a \left\{ \mathbb{E}_{\substack{r \sim \mathcal{R}(\cdot|s,a) \\ s' \sim \mathcal{P}(\cdot|s,a)}} \left[r + \gamma v(s') \right] \right\}. \quad (3)$$

A closely-related object is the *action-value function q^π*, the expected discounted return of first taking action a and thereafter following policy π. The action-value function satisfies the Bellman equations $q^\pi(s,a) = \mathcal{T}^\pi q^\pi(s,a)$ and $q^\star(s,a) = \mathcal{T}^\star q^\star(s,a)$, where $\mathcal{T}^\pi$ and $\mathcal{T}^\star$ are defined analogously to Equations (2) and (3) (Sutton and Barto, 1998). The Bellman operators for value functions (resp. action-value functions) are contractions on $\mathbb{R}^{|\mathcal{S}|}$ (resp. $\mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}$) with respect to the infinity norm $\|v\| := \|v\|_\infty = \max_i |v_i|$ (Puterman, 1994). A policy π is called greedy with respect to an action-value function $Q \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}$ if $\pi(s) \in \text{argmax}_a Q(s,a)$ for each $s \in \mathcal{S}$.

2.1 Couplings and the Wasserstein Metric

To establish convergence in distribution, we will use the Wasserstein metric $\mathcal{W}$ between distributions (Villani, 2008). As a cost function, we use the infinity norm. For two distributions $\mu, \nu \in \mathscr{P}(\mathbb{R}^d)$, a pair of random vectors (X, Y) is a coupling of (μ, ν) if $X \sim \mu$ and $Y \sim \nu$. We write $\Xi(\mu, \nu)$ for the set of such couplings. The Wasserstein metric on $\mathscr{P}(\mathbb{R}^d)$ with the infinity norm as a cost function is defined as:

$$\mathcal{W}(\mu, \nu) = \inf_{(X,Y) \in \Xi(\mu,\nu)} \mathbb{E}\left[\|X - Y\|_\infty \right]. \quad (4)$$

The metric is defined over the set $\mathcal{M}(\mathbb{R}^d) = \left\{ \mu \in \mathscr{P}(\mathbb{R}^d) : \int \|x\|_\infty \mu(dx) < +\infty \right\}$ of measures with finite first moment. Assuming a bounded reward function, we will always be dealing with finite-moment measures. The Wasserstein metric characterizes the *weak convergence* of measures (Villani, 2008, Theorem 6.9), or equivalently the convergence in distribution.

3 Markov Chains on the Space of Functions

With many value-based RL algorithms, the stochasticity of the algorithm depends only on the sampled transition and the random current estimate. For example, recalling the update rule for TD(0) (Equation (1)), the value of $V_{n+1}(s)$ for a particular state is fully

determined by knowledge of V_n and the action, reward, and successor state which was sampled from s:

$$\mathbb{P}\{V_{n+1} \mid V_n, V_{n-1}, ..., V_1, V_0\} = \mathbb{P}\{V_{n+1} \mid V_n\}.$$

We therefore view these methods as inducing Markov chains on the space of value functions. We note that their state space is continuous rather than discrete – we take it to be $\mathbb{R}^{|\mathcal{S}|}$ when modelling value functions or $\mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}$ when modelling action-value functions. When results hold for both cases, we will write the discussion in terms of $\mathbb{R}^d$, $d \in \mathbb{N}$. Whenever needed, we may also restrict ourselves to the subset of realizable functions $[0, \frac{\text{R}_{\text{MAX}}}{1-\gamma}]^d \subset \mathbb{R}^d$.

For a given update rule $\mathcal{U}$ and step-size α, the transition function for an induced Markov chain is as follows. Given $f_k \in \mathbb{R}^d$, let f_{k+1} be the random function obtained by $\mathcal{U}$ and α. For a Borel set $\mathcal{B} \in \text{Borel}(\mathbb{R}^d)$, we define the Markov kernel $K_{\mathcal{U},\alpha}$ as:

$$K_{\mathcal{U},\alpha}(f_k, \mathcal{B}) = \mathbb{P}\{f_{k+1} \in \mathcal{B}|f_k\}.$$

This Markov kernel describes the probability of transitioning from f_k to some function in the set $\mathcal{B}$ under the update rule. In the sequel, we omit the subscripts on the kernel when the update rule is clear from context. For a given probability measure $\mu \in \mathscr{P}(\mathbb{R}^d)$, the distribution of functions after one transition of the Markov chain is given by

$$\mu K(\mathcal{B}) = \int_{\mathbb{R}^d} K(\theta, \mathcal{B})\mu(\text{d}\theta).$$

The distribution of functions after n transitions is given by K^n, which is defined inductively as:

$$K^n(\theta, \mathcal{B}) = \int_{\mathbb{R}^d} K(\theta', \mathcal{B})K^{n-1}(\theta, \text{d}\theta').$$

A probability measure ψ is a *stationary distribution* for a Markov chain with kernel K if

$$\psi = \psi K.$$

An algorithm updates synchronously when all states or state-actions pairs are updated at every iteration. In the regime of constant step-sizes and synchronous updates, the Markov kernels are *time-homogeneous* (or time-independent). Thus, the law $\mu_n(\mathcal{B}) = \mathbb{P}\{f_n \in \mathcal{B}\}$ of the random variable f_n is given by:

$$\mu_n \leftarrow \mu_0(K)^n.$$

3.1 Stochastic operators

In this section, we introduce the notion of a stochastic operator and provide a general formalism for the analysis of stochastic update rules. We will distinguish two classes of stochastic operators which will require different analyses.

We model the sampling space as a probability space $(\Omega, \mathcal{F}, \eta)$. A stochastic operator is a map between functions which depends on a randomly sampled event $\omega \in \Omega$.

Definition 3.1 (Stochastic operator). A stochastic operator is a function $\widehat{\mathcal{T}} : \mathbb{R}^d \times \Omega \to \mathbb{R}^d$.

When operating on functions, a stochastic operator $\widehat{\mathcal{T}}$ outputs a random function. We will write a number of stochastic value-based algorithms as

$$f_{n+1} = (1-\alpha)f_n + \alpha\widehat{\mathcal{T}}(f_n, \omega), \quad (5)$$

where $f_n, f_{n+1} \in \mathbb{R}^d$ are functions, α is a step-size, and $\widehat{\mathcal{T}}$ is some algorithm-dependent stochastic operator. In this notation, the operator $\widehat{\mathcal{T}}$ is the *target* of the algorithm. We say that $\widehat{\mathcal{T}}$ is an *empirical Bellman operator* if it behaves like a Bellman operator in expectation.

Definition 3.2 (Empirical Bellman Operator). The stochastic operator $\widehat{\mathcal{T}}$ is an empirical Bellman operator for a policy π if

$$\mathbb{E}_{\omega \sim \eta}[\widehat{\mathcal{T}}(f, \omega)] = \mathcal{T}^\pi f \quad \forall f \in \mathbb{R}^d.$$

Similarly, $\widehat{\mathcal{T}}$ is an empirical Bellman *optimality* operator if $\mathbb{E}_{\omega \sim \eta}[\widehat{\mathcal{T}}(f, \omega)] = \mathcal{T}^\star f$.

In general, the sampling distribution of the stochastic operator may depend on the function which it is acting on. Two examples of methods for which the sampling distribution is independent of the current function estimates are TD(0), which applies an empirical Bellman operator, and Q-Learning, which applies an empirical Bellman optimality operator. TD(0) is defined by the stochastic operator

$$\widehat{\mathcal{T}}(V, (a_s, r_s, s'_s)_{s \in \mathcal{S}})(s) = r_s + \gamma V(s'_s),$$

where (a_s, r_s, s'_s) is a transition sampled for every state, and Q-Learning is defined by the operator

$$\widehat{\mathcal{T}}(Q, (r_{s,a}, s'_{s,a})_{(s,a)})(s, a) = r_{s,a} + \gamma \max_{a'} Q(s'_{s,a}, a'),$$

where $(r_{s,a}, s'_{s,a})$ is a transition sampled for every state-action pair. Convergence of these methods is covered in Section 4. On the other hand, methods for which the sampling of the update rule depends on the function being updated are more akin to policy iteration, which applies Bellman operators that depend on the current greedy policy. We will see an example of such a method in Section 6.

4 Convergence via Contraction to a Stationary Distribution

In this section we demonstrate that common value-based algorithms converge to a stationary distribution when updated synchronously and with constant step-sizes. The convergence follows by showing that their Markov kernels are contractive with respect to the Wasserstein metric. To illustrate our approach, we provide a proof of convergence for TD(0). With the same proof method, we also establish convergence and give convergence rates for Monte Carlo evaluation, Q-Learning, TD(λ), SARSA, Expected SARSA (Van Seijen et al., 2009), and Double Q-Learning (Hasselt, 2010). The proofs for these other algorithms are given in Appendix A.

Recall the update rule of the synchronous TD(0) algorithm given by Equation (1).

Proposition 4.1. *For any step size $0 < \alpha \leq 1$, the TD(0) algorithm has a contractive Markov kernel K_α:*

$$\mathcal{W}(\mu K_\alpha, \nu K_\alpha) \leq (1 - \alpha + \alpha\gamma)\mathcal{W}(\mu, \nu),$$

for all $\mu, \nu \in \mathcal{M}(\mathbb{R}^{|\mathcal{S}|})$.

Proof. Let $\mu^{(1)}, \mu^{(2)} \in \mathcal{M}(\mathbb{R}^{|\mathcal{S}|})$ be two distributions of function estimates. Let $V_0^{(1)} \sim \mu^{(1)}, V_0^{(2)} \sim \mu^{(2)}$ be the coupling which minimizes the Wasserstein metric, i.e.:

$$\mathcal{W}(\mu^{(1)}, \mu^{(2)}) = \inf_{(X,Y)} \mathbb{E}[\|X - Y\|] = \mathbb{E}\left[\|V_0^{(1)} - V_0^{(2)}\|\right].$$

Such an optimal coupling always exists (Villani, 2008, Theorem 4.1). We couple the updates $(V_1^{(1)}, V_1^{(2)})$ to sample the same transitions at each state:

$$
\begin{aligned}
V_1^{(1)}(s) &= (1 - \alpha)V_0^{(1)}(s) + \alpha\left(r_s + \gamma V_0^{(1)}(s'_s)\right) \\
V_1^{(2)}(s) &= (1 - \alpha)V_0^{(2)}(s) + \alpha\left(r_s + \gamma V_0^{(2)}(s'_s)\right),
\end{aligned}
\tag{6}
$$

for the *same* $a \sim \pi(\cdot|s)$, $r_s \sim \mathcal{R}(\cdot|s,a)$, and $s'_s \sim \mathcal{P}(\cdot|s,a)$. Note that this is a valid coupling of $(\mu^{(1)}K_\alpha, \mu^{(2)}K_\alpha)$ since $V_1^{(1)}$ and $V_1^{(2)}$ sample transitions from the same distributions. We upper-bound $\mathcal{W}(\mu^{(1)}K_\alpha, \mu^{(2)}K_\alpha)$ with the coupling above.

$$
\begin{aligned}
\mathcal{W}(\mu^{(1)}K_\alpha, \mu^{(2)}K_\alpha) &\leq \mathbb{E}\left[\|V_1^{(1)} - V_1^{(2)}\|\right] \\
&\leq (1 - \alpha)\mathbb{E}\left[\|V_0^{(1)} - V_0^{(2)}\|\right] \\
+ \alpha\mathbb{E}\left[\max_s \left|(r_s - r_s) + \gamma\left(V_0^{(1)}(s'_s) - V_0^{(2)}(s'_s)\right)\right|\right]
\end{aligned}
\tag{7}
$$

We note that the expectation is over the pair $(V_0^{(1)}, V_0^{(2)})$ as well as the random samples a_s, r_s, s'_s.

By our coupling construction,

$$
\begin{aligned}
&\mathbb{E}\left[\max_s \left|(r_s - r_s) + \gamma\left(V_0^{(1)}(s'_s) - V_0^{(2)}(s'_s)\right)\right|\right] \\
&= \gamma\mathbb{E}\left[\max_s |V_0^{(1)}(s'_s) - V_0^{(2)}(s'_s)|\right] \\
&\leq \gamma\mathbb{E}\left[\max_s |V_0^{(1)}(s) - V_0^{(2)}(s)|\right] = \gamma\mathcal{W}(\mu^{(1)}, \mu^{(2)})
\end{aligned}
\tag{8}
$$

The inequality follows since $V_0^{(1)}$ and $V_0^{(2)}$ sample the same set of successor states: the maximum is the same if each s samples a different s'_s and is lesser otherwise. Using Equation (8) in Equation (7) gives:

$$
\begin{aligned}
&\mathcal{W}(\mu^{(1)}K_\alpha, \mu^{(2)}K_\alpha) \\
&\leq \mathbb{E}\left[(1 - \alpha)\|V_0^{(1)} - V_0^{(2)}\| + \alpha\gamma\|V_0^{(1)} - V_0^{(2)}\|\right] \\
&= (1 - \alpha + \alpha\gamma)\mathcal{W}(\mu^{(1)}, \mu^{(2)}).
\end{aligned}
$$

Since $1 - \alpha + \alpha\gamma < 1$, the kernel K_α is a contraction mapping. $\square$

The contraction property readily entails the convergence to a stationary distribution. We initialize with any V_0 drawn from an arbitrary distribution of finite first moment.

Theorem 4.1. *For any constant step size $0 < \alpha \leq 1$ and initialization $V_0 \sim \mu_0 \in \mathcal{M}(\mathbb{R}^{|\mathcal{S}|})$, the sequence of random variables $(V_n)_{n \geq 0}$ defined by the recursion (1) converges in the Wasserstein metric to a unique stationary distribution $\psi_\alpha \in \mathcal{M}(\mathbb{R}^{|\mathcal{S}|})$.*

Proof. The space of probability measures $\mathcal{M}(\mathbb{R}^{|\mathcal{S}|})$ metrized with $\mathcal{W}$ is a complete metric space (Villani, 2008, Theorem 6.16), and therefore it follows from Banach's fixed point theorem that $(\mu_0 K_\alpha^n)_{n \geq 0}$ converges exponentially quickly to a unique fixed point $\psi_\alpha^{\text{TD}(0)}$. The distribution $\psi_\alpha^{\text{TD}(0)}$ is a stationary distribution by the fixed point property:

$$\psi_\alpha^{\text{TD}(0)} K_\alpha = \psi_\alpha^{\text{TD}(0)}. \quad \square$$

As evidenced by the above, lifting the analysis to distributions over value functions greatly simplifies the proof. The key is in the choice of a proper coupling. The same technique extends to a broad class of algorithms, with relatively few modifications. This avoids, for example, the additional hurdles caused by the greedy probability kernel in Q-learning (Tsitsiklis, 1994). We further note some surprising connections with distributional reinforcement learning (Bellemare, Dabney, and Munos, 2017). For $\alpha = 1$, the fixed point of TD(0) is in fact Bellemare, Dabney, and Munos's return distribution. The same coupling, which forces two processes to sample the same transitions, has also been implicitly used to study the

	MC Evaluation	TD(λ)	SARSA	Expected SARSA	QL	Double QL
Contraction factor	$1 - \alpha$	$1 - \alpha + \alpha\gamma\frac{1-\lambda}{1-\lambda\gamma}$	$1 - \alpha + \alpha\gamma$	$1 - \alpha + \alpha\gamma$	$1 - \alpha + \alpha\gamma$	$\frac{1}{2}(2 - \alpha + \alpha\gamma)$

Table 1: Different sample-based algorithms which imply a contraction mapping in the Wasserstein metric over distributions on value functions. For each method, we also provide the corresponding contraction factor. Acronyms: Monte Carlo (MC), Q-Learning (QL).

behaviour of distributional algorithms (Lyle, Castro, and Bellemare, 2019).

To demonstrate the power of the approach, we summarize in Table 1 a series of results regarding common sampling-based RL algorithms. Under similar conditions to Theorem 4.1, each algorithm listed in Table 1 converges to a stationary distribution (which is in general different for different algorithms, as we show in the next section). Each proof only requires small adjustments to the basic proof template, for example an extended state space (Double Q-Learning). Full details, along with the proof template, are given in the appendix.

5 The Stationary Distributions

In this section, we characterize the stationary distributions which are attained by any algorithm whose target is a Bellman operator or Bellman optimality operator in expectation. In our notation, these algorithms are defined in terms of *empirical* Bellman (optimality) operators. As before, we write the discussion in terms of $\mathbb{R}^d$ since results will hold for both value functions and action-value functions.

What do these distributions look like? We first consider the case of policy evaluation algorithms, which have as expected operator $\mathcal{T}^\pi$. In that case, their mean corresponds to the fixed point of $\mathcal{T}^\pi$, i.e. the value functions v^π or q^π. Second, they concentrate around this mean in inverse proportion to the step-size α. Hence, as expected, small step sizes lead to a more accurate distribution at the cost of a larger contraction factor. The full distributions are not symmetric or easily described, however; as a simple example, take $\alpha = 1$ in TD(0), corresponding to the return distribution (Bellemare, Dabney, and Munos, 2017). In the case of optimality operators, we show that the mean of the stationary distributions is in fact greater than the fixed points $v^\star$ or $q^\star$.

5.1 Sample-based Evaluation Algorithms

Theorem 5.1. *Suppose $\widehat{\mathcal{T}}^\pi$ is an empirical Bellman operator for some policy π and that the updates* (5) *with step-size α converge to a stationary distribution ψ_α. Let $f_\alpha \sim \psi_\alpha$ and f^π be the fixed point of $\mathcal{T}^\pi$. Then $\mathbb{E}[f_\alpha] = f^\pi$.*

Proof. Let f_0 be distributed according to ψ_α. By stationarity,

$$f_1 = (1 - \alpha)f_0 + \alpha\widehat{\mathcal{T}}^\pi(f_0, \omega) \qquad (9)$$

is also distributed according to ψ_α. We write $\overline{f_\alpha} := \mathbb{E}[f_0]$. Taking expectations on both sides, and using that $\mathbb{E}_\omega[\widehat{\mathcal{T}}^\pi(f, \omega)] = \mathcal{T}^\pi(f)$ for any f:

$$\overline{f_\alpha} = (1 - \alpha)\overline{f_\alpha} + \alpha\mathbb{E}_{\omega, \psi_\alpha}[\widehat{\mathcal{T}}^\pi(f_0, \omega)]$$
$$\overline{f_\alpha} = \mathbb{E}_{\psi_\alpha}[\mathcal{T}^\pi f_0]$$
$$\overline{f_\alpha} = \mathcal{T}^\pi\mathbb{E}_{\psi_\alpha}[f_0] = \mathcal{T}^\pi\overline{f_\alpha}$$

And therefore $\overline{f_\alpha} = f^\pi$ since it is the unique fixed point of $\mathcal{T}^\pi$. $\qquad \square$

We remark again that this characterization will hold for any algorithm which converges and performs Bellman updates in expectation. Although they have the same mean, the stationary distributions will depend on the update rule. These differences will be reflected in their higher moments. To this effect, we next derive a closed-form expression for the covariance of the stationary distribution. We write A^T for the transpose of a matrix A. The outer product of two vectors $x, y \in \mathbb{R}^d$ is the matrix $xy^\mathsf{T} \in \mathbb{R}^{d\times d}$ defined by $(xy^\mathsf{T})_{i,j} = x_iy_j$. Thus, $\mathbb{E}[(\vec{X} - \vec{\mu})(\vec{X} - \vec{\mu})^\mathsf{T}]$ is the covariance of a random vector $\vec{X}$ with mean $\vec{\mu}$. The proof of the following result is provided in Appendix B.

Theorem 5.2. *Let $\widehat{\mathcal{T}}^\pi$ be an empirical Bellman operator for some policy π. Suppose $\widehat{\mathcal{T}}^\pi$ is such that the updates* (5) *with step-size α converge to a stationary distribution ψ_α. Define $\xi_\omega(f) = \widehat{\mathcal{T}}^\pi(f, \omega) - \mathcal{T}^\pi f$, and*

$$\mathcal{C}(f) := \mathbb{E}_\omega[\xi_\omega(f)\xi_\omega(f)^\mathsf{T}]$$

to be the covariance of the zero-mean noise term $\xi_\omega(f)$ for a given function f. Define $C = (1 - (1 - \alpha))^2$. The covariance of $f_\alpha \sim \psi_\alpha$ is given by

$$C\mathbb{E}\left[(f_\alpha - f^\pi)(f_\alpha - f^\pi)^\mathsf{T}\right] =$$
$$\quad \alpha^2(\gamma\mathcal{P}^\pi)\mathbb{E}\left[(f_\alpha - f^\pi)(f_\alpha - f^\pi)^\mathsf{T}\right](\gamma\mathcal{P}^\pi)^\mathsf{T}$$
$$\quad + \alpha(1 - \alpha)(\gamma\mathcal{P}^\pi)\mathbb{E}\left[(f_0 - f^\pi)(f_0 - f^\pi)^\mathsf{T}\right]$$
$$\quad + \alpha(1 - \alpha)\mathbb{E}\left[(f_\alpha - f^\pi)(f_\alpha - f^\pi)^\mathsf{T}\right](\gamma\mathcal{P}^\pi)^\mathsf{T}$$
$$\quad + \alpha^2\int\mathcal{C}(f)\psi_\alpha(\mathrm{d}f).$$

Theorem 5.2 provides a recursive definition for the covariance of the stationary distribution ψ_α. The integral in the final line corresponds to the expected covariance of the empirical Bellman operator when sampling from the distribution. Under the assumption that states are updated independently, this "one-step" covariance is diagonal. More generally, the covariance matrix is scaled by α, suggesting that the distribution concentrates around its mean when α is close to 0. The following makes this precise. We write $\|A\|_{\mathrm{op}} = \sup\{\|Av\| : \|v\| \leq 1, v \in \mathbb{R}^d\}$ for the operator norm of a matrix A.

Corollary 5.2.1. *Assume that the state space of the Markov chain is bounded. Let $C := (\frac{2R_{\mathrm{MAX}}}{1-\gamma})^2$. Then, we have that $\left\|\mathbb{E}\left[(f_\alpha - f^\pi)(f_\alpha - f^\pi)^\mathsf{T}\right]\right\|_{op}$ is monotonically decreasing with respect to α. In particular, $\lim_{\alpha \to 0}\left\|\mathbb{E}[(f_\alpha - f^\pi)(f_\alpha - f^\pi)^\mathsf{T}]\right\|_{op} = 0$, and we have that:*

$$\mathbb{P}\left\{\min_i |f_\alpha(i) - f^\pi(i)| \geq \varepsilon\right\} \leq \frac{C}{\mathsf{d}\varepsilon^2}\frac{\alpha^2}{1 - (1 - \alpha + \alpha\gamma)^2}$$
$$\xrightarrow{\alpha \to 0} 0.$$

We remark that the boundedness of the state space (e.g. by $[0, \frac{R_{\mathrm{MAX}}}{1-\gamma}]^d \subset \mathbb{R}^d$) is easily satisfied in the presence of bounded rewards in the MDP. Furthermore, the above results about the mean and covariance can easily be extended beyond Bellman operators to any operator which has a unique fixed point and commutes with expectation.

5.2 Sample-Based Control Algorithms

Above we saw that the mean of the stationary distribution of a sample-based method using a fixed policy is the value function for that policy. This no longer holds in the presence of optimality operators, for example in what is called the *control setting* (Sutton, 1988). To conclude this section, we use our distributional approach to highlight behavioural characteristics of control algorithms.

Theorem 5.3. *Suppose $\widehat{\mathcal{T}}^\star$ is an empirical Bellman optimality operator such that the updates* (5) *with step-size α converge to a stationary distribution $\psi_\alpha^\star$. Let $f_\alpha \sim \psi_\alpha^\star$ and $f^\star$ is the fixed point of $\mathcal{T}^\star$. Then*

$$\mathbb{E}[f_\alpha] \geq f^\star.$$

Equality holds if and only if the expectation and the maximum commute, i.e. $\mathbb{E}\widehat{\mathcal{T}}f = \widehat{\mathcal{T}}\mathbb{E}f$

Proof. As before, let f_0 be distributed according to $\psi_\alpha^\star$. Taking expectations on both sides of $f_1 = (1 - \alpha)f_0 +$

$\alpha\widehat{\mathcal{T}}^\star(f_0, \omega)$ and writing $\overline{f_\alpha} := \mathbb{E}[f_\alpha]$ gives:

$$\overline{f_\alpha} = (1 - \alpha)\overline{f_\alpha} + \alpha\mathbb{E}_{\omega, f_0}[\widehat{\mathcal{T}}^\star(f_0, \omega)]$$
$$\overline{f_\alpha} = \mathbb{E}_{f_0}[\max_\pi \mathcal{T}^\pi f_0]$$
$$\overline{f_\alpha} \geq \max_\pi \mathbb{E}_{f_0}[\mathcal{T}^\pi f_0]$$
$$\overline{f_\alpha} \geq \max_\pi \mathcal{T}^\pi \overline{f_\alpha} = \mathcal{T}^\star \overline{f_\alpha}$$

By the linear programming formulation of MDPs (Puterman, 1994, Section 6.9.1), we conclude that $\bar{f}_\alpha \geq f^\star = \min_f\{f \geq \mathcal{T}^\star f\}$. $\qquad\square$

The theorem shows that in general, sample-based control methods such as Q-learning produces a biased (in an expected sense) estimate of the optimal Q-value. This brings fresh evidence about the algorithm's well-known overestimation problem, which double Q-learning seeks to correct.

6 Convergence via Monotonicity: Optimistic Policy Iteration

In a previous section, we showed that a number of sampling-based algorithms induce a contraction mapping in the Wasserstein metric between distributions over value functions. In this section we analyze a non-contractive example, namely the optimistic policy iteration (OPI) algorithm. The OPI algorithm is a sampling-based analogue of the classic policy iteration (PI) algorithm. The latter is driven to convergence by the monotonicity of the greedy policy updates. We show in this section that our Markov chain approach can regain a distributional analogue of the monotonicity property, which we call *probabilistic policy improvement*, and that this property can be used to analyze the algorithm in a restricted setting.

The convergence of optimistic policy iteration is more difficult to prove than that of most sample-based algorithms, and has been previously been established for Robbins-Monro decreasing stepsizes by using monotonicity arguments and assumptions on the sampling distribution (Tsitsiklis, 2002).

Optimistic policy iteration proceeds by constructing a greedy policy from its current value function, sampling one trajectory per state-action pair from this policy, then updating its value function towards the return of these trajectories. We will write

$$\mathcal{G}^\pi(s_0, a_0) = \sum_{t=0}^\infty \gamma^t r_t(s_t, a_t)$$

for a sampled discounted return starting at state s_0, taking first action a_0, and thereafter following policy

π. For any $Q \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}$, we write π_Q for the greedy policy corresponding to Q (assuming a consistent tie-breaking so that this is well-defined). Let Q_0 be some initial estimate and $\pi_0 = \pi_{Q_0}$. The update rule of OPI is as follows ($\alpha \in (0, 1]$):

$$Q_{n+1}(s, a) = (1 - \alpha)Q_n(s, a) + \alpha \mathcal{G}^{\pi_n}(s, a)$$
$$\pi_{n+1} = \pi_{Q_{n+1}}. \qquad (10)$$

Analyzing optimistic policy iteration in the distributional setting poses a few challenges. First, the distribution of sampled trajectories depends on the exact value function. Informally, the greedy mapping from value functions to policies induces a *greedy partition* (Bertsekas and Tsitsiklis, 1996, Figure 6.9), with a different empirical Bellman operator corresponding to each region of this partition. This rules out a simple coupling argument, as functions with different greedy policies may have arbitrarily different return distributions. Bertsekas and Tsitsiklis point out that optimistic policy iteration can lead to chattering, where the greedy policy fails to converge even the value function converges. For our analysis, we consider the simpler case $\alpha = 1$; we discuss the extension to $\alpha < 1$ at the end of the section.

Theorem 6.1. *For $\alpha = 1$ and initialization $Q_0 \sim \mu_0 \in \mathcal{M}(\mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|})$, the sequence of random variables $(Q_n)_{n \geq 0}$ defined by the recursion (10) converges to a unique stationary distribution $\varphi_1 \in \mathscr{P}(\mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|})$.*

The key lemma is to extend the monotonicity property of the policy iteration operator to the distributional case. In policy iteration, the greedy policy $\pi' = \pi_{Q^\pi}$ with respect to Q^π leads to an improved value function:

$$Q^{\pi'} \geq Q^\pi.$$

The role of Q^π is therefore to provide us with the improved policy π'. We will show that the same holds true for the sampled returns: there is some probability that the greedy policy with respect to $\mathcal{G}^\pi$ is π'. This allows us to argue that there is a chance that optimistic policy iteration follows the correct "greedy path" to π^*. We will call this property *probabilistic policy improvement*.

We analyze the case $\alpha = 1$ by considering a Markov chain over policies. Formally, $\Pi = \{\pi : \mathcal{S} \to \mathcal{A}\}$ will be our state space, with the Markov kernel:

$$K(\pi, \pi') := \mathbb{P}\{\pi' = \pi_{\mathcal{G}^\pi}\}$$
$$= \mathbb{P}\{\pi' \text{ is greedy for } \mathcal{G}^\pi\}.$$

This Markov chain reflects the OPI process since, at every step, the greedy policy π_n corresponding to Q_n is sufficient to determine the distribution of Q_{n+1}. Since the set Π of deterministic policies is finite, K is a discrete Markov chain.

Lemma 6.1 (Probabilistic policy improvement). *Suppose $\pi' = \pi_{Q^\pi}$. Then $K(\pi, \pi') > 0$.*

The proof of Lemma 6.1 is given in Appendix C. This shows that there is a nonzero probability that the chain improves on the current policy. This implies that there is some probability that OPI applied from π^* produces π^* as a greedy policy.

Lemma 6.2 (π^* is aperiodic). *The optimal policy $\pi^\star$ is aperiodic. In particular: $K(\pi^\star, \pi^\star) > 0$.*

Proof. Since the optimal policy $\pi^\star$ is greedy with respect to $Q^\star$, from Lemma 6.1 we conclude that $K(\pi^\star, \pi^\star) > 0$. $\qquad \square$

All that remains to show is that the optimal policy $\pi^\star$ is *reachable* from any other policy with positive probability.

Lemma 6.3 (π^* is reachable from any initial π_0). *For every $\pi_0 \in \Pi$, there exists an $n(\pi_0) \in \mathbb{N}$ such that $K^{n(\pi_0)}(\pi_0, \pi^\star) > 0$.*

Proof. Let π_0 be an initial policy. Let $Q^{\pi_0}, Q^{\pi_1}, ..., Q^{\pi^\star}$ be the sequence of action-value functions obtained from classical PI. Since PI converges in a finite number of steps (say n_{π_0}) and is a deterministic process, this sequence is well-defined. For every $i \in \{1, .., n(\pi_0)\}$, we have that $K(\pi_i, \pi_{i+1}) > 0$ by Lemma 6.1 (since π_{i+1} is greedy with respect to Q^{π_i} by construction). Thus we have that $K(\pi_0, \pi_1)K(\pi_1, \pi_2) \cdots K(\pi_{n(\pi_0)-1}, \pi^\star) > 0$ and in particular $K^{n(\pi_0)}(\pi_0, \pi^\star) > 0$. $\qquad \square$

Finally, the reachability and aperiodicity of $\pi^\star$ allow us to apply the ergodic theorem for finite Markov chains.

Proof (*of Theorem 6.1*). The policy $\pi^\star$ must be contained in a communicating class $\mathscr{C}^\star$ of policies (perhaps consisting of only $\pi^\star$) which is aperiodic since $\pi^\star$ is. There may be other communicating classes in the Markov chain, but by Lemma 6.3 they must all be transient since they can reach $\pi^\star$. By the Markov chain convergence theorem (Levin and Peres, 2017, Theorem 4.3), any initial distribution converges to a stationary distribution $\varphi_1 \in \mathscr{P}(\Pi)$ with support over $\mathscr{C}^\star$. $\qquad \square$

Our result shows that optimistic policy iteration, applied with a step-size of $\alpha = 1$, converges to a stationary distribution φ_1 over aperiodic policies (and thus to a stationary distribution over value functions through the possible returns of these policies). We note that the convergence happens in the Total Variation (TV) metric, as opposed to previous results Since

$K(\pi^*, \pi^*) < 1$ in general, we know that this distribution has support on suboptimal policies; in fact, we know that

$$\varphi_1(\pi^*) = \frac{1}{1 - K(\pi^*, \pi^*)} \sum_{\pi \neq \pi^*} \varphi_1(\pi) K(\pi, \pi^*).$$

By "continuity", this suggests that the algorithm should also converge for the general case $\alpha \in [0, 1)$. Unfortunately, our proof technique does not immediately carry over. The issue is that, for $\alpha < 1$, we no longer have a Markov chain over policies: the greedy policy depends on the history of past policies, through the value function. One path forward may be to study the Markov chain over value functions, but the known brittleness of optimistic policy iteration suggests that its distributional behaviour may be quite complex. We leave as open questions whether the algorithm does converge, and to which distribution.

7 Related Work

In the constant step-size case, convergence in distribution results are typically derived using tools common to stochastic approximation theory such as the mean ODE method and Lyapunov functions (see, e.g., Kushner and Yin (2003, Chapter 8) and Borkar (2009, Chapter 9)). In RL, examples of constant step-size analyses which feature these methods include Srikant and Ying (2019), Chen et al. (2019), and Lakshminarayanan and Szepesvári (2017).

The Markov chain perspective has previously been used in the stochastic approximation literature (Borkar and Meyn, 2000; Yu, 2016). The convergence of Markov chains in continuous spaces has mainly been studied with respect to the Total Variation metric (e.g. Meyn and Tweedie, 2012, Chapters 13-16). In applications to the analysis of RL algorithms, this type of convergence (stronger than the Wasserstein convergence) does not hold without restrictive assumptions (see Appendix D for a simple counterexample). For example, Assumption (2.6) in (Borkar and Meyn, 2000) fails to hold in the case of deterministic rewards. On the other hand, results about weak convergence of RL algorithms (Yu, 2016) have established the convergence of the averaged iterates rather than the full sequence of distributions. The methods are also different, and rely on the weak Feller property (Meyn and Tweedie, 2012) amongst other stochastic approximation techniques (Kushner and Yin, 2003). As far as we are aware, the use of the Wasserstein metric and the result that RL algorithms are contractive with respect to this metric are novel. The resulting proofs of convergence in distribution using said properties are therefore simpler than the existing literature.

Some of our methods are similar to the work of Dieuleveut, Durmus, and Bach (2017), which develops the theory of constant step-size stochastic gradient descent (in the context of supervised learning). In particular, the proof method we present in Section 4 is inspired from the proof of their Proposition 2, although simplified and adapted to the RL setting, and the results in Section 5 follow the methods of their Proposition 3.

8 Conclusion and Future Work

We studied the convergence properties of sample-based reinforcement learning algorithms by considering how they induce distributions over value functions. Many of these algorithms are in fact contractive not in the space of functions but in the lifted space of distributions of functions. The proof methods relies on coupling the events sampled by two executions of the algorithm, and can be re-used for many algorithms. Using the same Markov chain approach, we also analyzed a restricted version of optimistic policy iteration, which is not amenable to a contraction mapping-type analysis. One of the key results is to make explicit that constant step-size reinforcement learning algorithms do converge, albeit in the weaker distributional sense. As an upside of using a constant step size, we obtain exponentially fast convergence (as indicated by the presence of a contraction factor). By controlling the step-sizes, the stationary distributions thus obtained can be tailored to yield values close to the true value function with high confidence. In the control setting, this should enable us to better explain the performance of practical reinforcement learning schemes.

Our work opens a number of interesting avenues for future research. First, it would be valuable to fully characterize the stationary distribution of sample-based methods, for example by deriving a closed-form expression for their characteristic functions. A deeper understanding of the distributions obtained by control algorithms is also of interest. Second, we did not analyze the case of decaying step-sizes or online updates, which would correspond to time-inhomogenenous Markov processes. More broadly, the coupling method has historically been invaluable for many applications in probability theory. It would be interesting to see if our approach can be applied to policy-based methods, for example policy gradient or actor critic, which are closer in spirit to optimistic policy iteration. Finally, the simplicity of our analysis suggests that it may be carried to the function approximation setting, perhaps eventually shedding light on the behaviour of reinforcement learning with nonlinear approximation methods such as deep networks.

Acknowledgements

References

Bellemare, Marc G, Will Dabney, and Rémi Munos (2017). "A distributional perspective on reinforcement learning". In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, pp. 449–458.

Bertsekas, Dimitri P and John N Tsitsiklis (1996). *Neuro-dynamic programming*. Vol. 5. Athena Scientific Belmont, MA.

Borkar, Vivek S (2009). *Stochastic approximation: a dynamical systems viewpoint*. Vol. 48. Springer.

Borkar, Vivek S and Sean P Meyn (2000). "The ODE method for convergence of stochastic approximation and reinforcement learning". In: *SIAM Journal on Control and Optimization* 38.2, pp. 447–469.

Chen, Zaiwei et al. (2019). "Finite-Time Analysis of Q-Learning with Linear Function Approximation". In: *arXiv preprint arXiv:1905.11425*.

Dieuleveut, Aymeric, Alain Durmus, and Francis Bach (2017). "Bridging the gap between constant step size stochastic gradient descent and markov chains". In: *arXiv preprint arXiv:1707.06386*.

Hasselt, Hado V (2010). "Double Q-learning". In: *Advances in Neural Information Processing Systems*, pp. 2613–2621.

Kushner, Harold and G George Yin (2003). *Stochastic approximation and recursive algorithms and applications*. Vol. 35. Springer Science & Business Media.

Lakshminarayanan, Chandrashekar and Csaba Szepesvári (2017). "Linear stochastic approximation: Constant step-size and iterate averaging". In: *arXiv preprint arXiv:1709.04073*.

Levin, David A and Yuval Peres (2017). *Markov chains and mixing times*. Vol. 107. American Mathematical Soc.

Lyle, Clare, Pablo Samuel Castro, and Marc G. Bellemare (2019). "A Comparative Analysis of Expected and Distributional Reinforcement Learning". In: *CoRR* abs/1901.11084.

Marshall, Albert W and Ingram Olkin (1960). "Multivariate chebyshev inequalities". In: *The Annals of Mathematical Statistics*, pp. 1001–1014.

Meyn, Sean P and Richard L Tweedie (2012). *Markov chains and stochastic stability*. Springer Science & Business Media.

Puterman, Martin L (1994). "Markov Decision Processes: Discrete Stochastic Dynamic Programming". In:

– (2014). *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

Srikant, R and Lei Ying (2019). "Finite-time error bounds for linear stochastic approximation and TD learning". In: *arXiv preprint arXiv:1902.00923*.

Sutton, Richard S (1988). "Learning to predict by the methods of temporal differences". In: *Machine learning* 3.1, pp. 9–44.

Sutton, Richard S and Andrew G Barto (1998). *Introduction to reinforcement learning*. Vol. 135. MIT press Cambridge.

Tsitsiklis, John N (1994). "Asynchronous stochastic approximation and Q-learning". In: *Machine learning* 16.3, pp. 185–202.

– (2002). "On the convergence of optimistic policy iteration". In: *Journal of Machine Learning Research* 3.Jul, pp. 59–72.

Van Hasselt, Hado, Arthur Guez, and David Silver (2016). "Deep reinforcement learning with double q-learning". In: *Thirtieth AAAI conference on artificial intelligence*.

Van Seijen, Harm et al. (2009). "A theoretical and empirical analysis of Expected Sarsa". In: *2009 ieee symposium on adaptive dynamic programming and reinforcement learning*. IEEE, pp. 177–184.

Villani, Cédric (2008). *Optimal transport: old and new*. Springer-Verlag.

Yu, Huizhen (2016). "Weak convergence properties of constrained emphatic temporal-difference learning with constant and slowly diminishing stepsize". In: *The Journal of Machine Learning Research* 17.1, pp. 7745–7802.