A Distributional Analysis of Sampling-Based Reinforcement Learning Algorithms

Philip Amortila[*]
McGill University

Doina Precup
McGill University
Google DeepMind

Prakash Panangaden
McGill University

Marc G. Bellemare
Google Research, Brain Team
CIFAR Fellow

Abstract

We present a distributional approach to theoretical analyses of reinforcement learning algorithms for constant step-sizes. 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. The proof method can be used for many algorithms and simplifies existing arguments for convergence in distribution. For update rules whose expected target is a Bellman update, we further demonstrate that the stationary distribution obtained has a mean which is equal to the true value function in the policy evaluation setting and which is biased in the control setting. Lastly, we establish that the stationary distributions concentrate around their means as the step-size shrinks.

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. However, proving the convergence of sample-based algorithms such as TD-learning (Sutton, 1988) and Q-learning (Watkins and Dayan, 1992) requires substantially more effort. The typical stochastic approximation approach relies on hitting-time or martingale arguments to bound the sequence of value function iterates with 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 $(V_n)_{n\in\mathbb{N}}$ satisfying the distributional equation

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

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

[*]Correspondence to philip.amortila@mail.mcgill.ca

Preprint. Under review.

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 the step-size, the MDP, and the specific update rule under consideration. Following a proof technique of (Dieuleveut, Durmus, and Bach, 2017), we lift these stochastic algorithms to the distributional setting and view the learning process as defining a time-homogeneous Markov process over the space of value functions.

We find that many sampling-based algorithms (e.g. TD(λ), Q-learning, and double Q-learning) induce corresponding distributional operators which are contraction mappings in the Wasserstein metric. In other words, the contraction property of the Bellman operators can be regained by lifting to the space of distributions. 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 $1 - \alpha + \alpha\gamma$.

By recovering the contraction property that underlies many dynamic programming algorithms, our distributional analysis significantly simplifies existing proofs of convergence for stochastic RL 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, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathscr{P}([0, \textsc{Rmax}])$ 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 $\mathcal{T}^\pi v(s) \coloneqq \mathbb{E}\left[r(s, a) + \gamma v(s') \mid a \sim \pi, s' \sim \mathcal{P}(\cdot|s, a)\right]$. The Bellman *optimality* operator is defined by $\mathcal{T}^\star v(s) \coloneqq \max_a \left\{\mathbb{E}\left[r(s, a) + \gamma v(s')|s' \sim \mathcal{P}(\cdot|s, a)\right]\right\}$. 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 $\mathcal{T}^\pi q(s, a)$ and $\mathcal{T}^\star q(s, a)$ are defined analogously to the value function case (we refer the reader to Sutton and Barto, 1998 for more information). 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).

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 $\mu, \nu \in \mathscr{P}(\mathbb{R}^n)$, a pair of random variables (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 $\mathscr{P}(\mathbb{R}^n)$ with the infinity norm as a cost function is defined as:

$$\mathcal{W}(\mu, \nu) = \inf_{(X,Y)\in\Xi(\mu,\nu)} \mathbb{E}\left[\|X - Y\|_\infty\right]. \tag{2}$$

The metric is defined over the set $\mathcal{M}(\mathbb{R}^n) = \left\{\mu \in \mathscr{P}(\mathbb{R}^n) : \int \|x\|_\infty \mu(\mathrm{d}x) < +\infty\right\}$ of measures with finite first moment.

3 Markov Processes 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:

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

We therefore view these methods as inducing Markov processes on the space of value functions. We take their state space 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 $\mathbb{R}^n$, $n \in \mathbb{N}$. Whenever needed, we may also restrict ourselves to the subset of realizable functions $[0, \frac{R_{MAX}}{1-\gamma}]^n \subset \mathbb{R}^n$.

The transition function for an induced Markov process is as follows. Given $f_k \in \mathbb{R}^n$, let f_{k+1} be the random function obtained by a sample-based update rule. For a Borel set $\mathcal{B} \in \text{Borel}(\mathbb{R}^n)$, we define the Markov kernel K as:

$$K(f_k, \mathcal{B}) = \mathbb{P}\left\{f_{k+1} \in \mathcal{B} | f_k\right\}.$$

This Markov kernel describes the probability of transitioning from f_k to some function in the set $\mathcal{B}$ under the update rule. For a given probability measure $\mu \in \mathscr{P}(\mathbb{R}^n)$, the distribution of functions after one transition of the Markov process is given by

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

A probability measure ψ is a *stationary distribution* for a Markov process 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).

Stochastic operators We provide a general formalism for the analysis of stochastic update rules. We call *stochastic operator* any mapping $\widehat{\mathcal{T}} : \mathbb{R}^n \times \Omega \to \mathbb{R}^n$ between functions which depends on a randomly sampled event ω in some probability space Ω. We will write a number of stochastic, value-based algorithms as

$$f_{n+1} = (1-\alpha)f_n + \alpha\widehat{\mathcal{T}}(f_n, \omega), \tag{3}$$

where $f_n, f_{n+1} \in \mathbb{R}^n$ are functions, α is a step-size, and $\widehat{\mathcal{T}}$ is some algorithm-dependent stochastic operator. We say that $\widehat{\mathcal{T}}$ is an *empirical Bellman operator* if it behaves like a Bellman operator in expectation over the samples: $\mathbb{E}_\omega[\widehat{\mathcal{T}}(f, \omega)] = \mathcal{T}^\pi f$. Similarly, $\widehat{\mathcal{T}}$ is an empirical Bellman *optimality* operator if $\mathbb{E}_\omega[\widehat{\mathcal{T}}(f, \omega)] = \mathcal{T}^\star f$.

4 Convergence 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. 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, 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). 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 to a unique stationary distribution $\psi_\alpha \in \mathcal{M}(\mathbb{R}^{|\mathcal{S}|})$.*

Proof. Let $\mu^{(1)}, \mu^{(2)} \in \mathcal{M}(\mathbb{R}^{|\mathcal{S}|})$ be two initial distributions of function estimates, and K_α the Markov kernel induced by (1) for step-size α. Let $V_0^{(1)} \sim \mu^{(1)}, V_0^{(2)} \sim \mu^{(2)}$ be the coupling which minimizes the Wasserstein metric $\mathcal{W}(\mu^{(1)}, \mu^{(2)})$, i.e. such that

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

Such a coupling always exists (Villani, 2008, Theorem 4.1). To show that the map $\mu \mapsto \mu K_\alpha$ is a contraction with respect to $\mathcal{W}$, we couple the updates $(V_1^{(1)}, V_1^{(2)})$ to sample the same transitions for each s:

$$\left.\begin{aligned} V_1^{(1)}(s) &= (1-\alpha)V_0^{(1)}(s) + \alpha\left(r + \gamma V_0^{(1)}(s')\right) \\ V_1^{(2)}(s) &= (1-\alpha)V_0^{(2)}(s) + \alpha\left(r + \gamma V_0^{(2)}(s')\right) \end{aligned}\right\} \text{for the } \underline{\text{same}} \begin{aligned} a &\sim \pi(\cdot|s), \\ r &\sim \mathcal{R}(\cdot|s,a), \\ s' &\sim \mathcal{P}(\cdot|s,a). \end{aligned} \tag{4}$$

Since the marginal distributions of $V_1^{(1)}$ and $V_1^{(2)}$ correspond to the distributions $\mu^{(1)}K_\alpha$ and $\mu^{(2)}K_\alpha$, respectively, this is a valid coupling. We upper bound $\mathcal{W}(\mu^{(1)}K_\alpha, \mu^{(2)}K_\alpha)$ with the coupling above. We write $\widehat{\mathcal{T}}^\pi(V, (a_s, r_s, s'_s))(s) \coloneqq r_s + \gamma V(s'_s)$ for the empirical Bellman update, where the subscript emphasizes the dependence on s. Then:

$$\begin{aligned} \mathcal{W}(\mu^{(1)}K_\alpha, \mu^{(2)}K_\alpha) &\leq \mathbb{E}\left[\|V_1^{(1)} - V_1^{(2)}\|\right] \\ &\leq (1-\alpha)\mathbb{E}\left[\|V_0^{(1)} - V_0^{(2)}\|\right] + \alpha\mathbb{E}\left[\|\widehat{\mathcal{T}}^\pi(V_0^{(1)}) - \widehat{\mathcal{T}}^\pi(V_0^{(2)})\|\right]. \end{aligned} \tag{5}$$

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 for each s. By our coupling construction,

$$\begin{aligned} \mathbb{E}\left[\|\widehat{\mathcal{T}}^\pi(V_0^{(1)}) - \widehat{\mathcal{T}}^\pi(V_0^{(2)})\|\right] &= \mathbb{E}\left[\max_s |(r_s - r_s) + \gamma\big(V_0^{(1)}(s'_s) - V_0^{(2)}(s'_s)\big)|\right] \\ &= \gamma\mathbb{E}\left[\max_s |V_0^{(1)}(s'_s) - V_0^{(2)}(s'_s)|\right] \\ &\leq \gamma\mathbb{E}\left[\max_s |V_0^{(1)}(s) - V_0^{(2)}(s)|\right] = \mathcal{W}(\mu^{(1)}, \mu^{(2)}) \end{aligned} \tag{6}$$

Using Equation (6) in Equation (5) 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. Lastly, $\mathcal{M}(\mathbb{R}^{|\mathcal{S}|})$ 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 $(\mu K_\alpha^n)_{n\geq 0}$ converges to a unique fixed point $\psi_\alpha^{\text{TD(0)}}$ for any initial distribution $\mu \in \mathcal{M}(\mathbb{R}^{|\mathcal{S}|})$. 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 expected 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 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 Appendix A.

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

Table 1: Different update rules which are contractive in $\mathcal{W}$ over distributions of value functions. We provide the corresponding contraction factor. All algorithms converge for any $\alpha \in (0, 1]$. Acronyms: Monte Carlo (MC), Q-Learning (QL).

5 Characterizing the Stationary Distributions

In this section, we characterize the stationary distributions which are attained by any algorithm whose target is, in expectation, a Bellman operator or Bellman optimality operator. As before, we write the discussion in terms of $\mathbb{R}^n$ 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 functions v^π or q^π. Second, they concentrate around this mean in inverse proportion to the step-size α. Hence, 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). Proofs for this section are provided in Appendix B.

Sample-based Evaluation Algorithms The following characterization will hold for any algorithm which converges and performs Bellman updates in expectation.

Theorem 5.1. *Suppose $\widehat{\mathcal{T}}^\pi$ is such that the updates* (3) *with step-size α converge to a stationary distribution ψ_α. Assume that $\widehat{\mathcal{T}}^\pi$ is an empirical Bellman operator for some policy π, and let f^π is the fixed point of $\mathcal{T}^\pi$. Then*

$$\mathbb{E}_{f_\alpha \sim \psi_\alpha}[f_\alpha] = f^\pi.$$

The effects of the specific update rules will be reflected in the higher moments of the stationary distribution. Thus, we next derive a closed-form expression for the covariance matrix of the stationary distribution.

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* (3) *with step-size α converge to a stationary distribution ψ_α. Define $\xi_\omega(f) = \widehat{\mathcal{T}}^\pi(f, \omega) - \mathcal{T}^\pi f$ to be the zero-mean noise term for a given function f and $\mathcal{C}(f) := \mathbb{E}_\omega[\xi_\omega(f)\xi_\omega(f)^\mathsf{T}]$ to be its covariance. The covariance of $f_\alpha \sim \psi_\alpha$ is given by*

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

The integral in the final line corresponds to the expected covariance of the empirical Bellman operator when sampling from the MDP under the sampling distribution. As a corollary, we show that the distribution concentrates around its mean when α is close to 0. We write $\|A\|_{\text{op}} = \sup\{\|Av\| : \|v\| \leq 1, v \in \mathbb{R}^n\}$ for the operator norm of a matrix A.

Corollary 5.2.1. *Assume that the state space of the Markov process is bounded. Let $C := (\frac{2R_{\text{MAX}}}{1-\gamma})^2$. 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:*

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

We remark that the boundedness of the state space (e.g. by $[0, \frac{\textsc{Rmax}}{1-\gamma}]^n \subset \mathbb{R}^n$) is easily satisfied in the presence of a bounded reward function in the MDP.

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).

Theorem 5.3. *Suppose $\widehat{\mathcal{T}}^\star$ is such that the updates* (3) *with step-size α converge to a stationary distribution $\psi_\alpha^\star$. Assume that $\widehat{\mathcal{T}}^\star$ is an empirical Bellman optimality operator and let $f^\star$ be the fixed point of $\mathcal{T}^\star$. Then*

$$\mathbb{E}_{f_\alpha \sim \psi_\alpha^\star}[f_\alpha] \geq f^\star,$$

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

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, bringing fresh evidence about the algorithm's well-known overestimation problem.

6 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, Liapunov functions, and the martingale method (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 Beck and Srikant (2012), Yu (2016), Lakshminarayanan and Szepesvári (2017), and Chen et al. (2019). However, our results and methods are different. With the exception of Yu's work, the above references do not cast the algorithms under consideration as Markov processes or discuss convergence to a stationary distribution. Furthermore, as far as we are aware, the result that RL algorithms are contractions on the space of distributions of functions is novel. The resulting proofs of convergence in distribution using said contraction 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, our proofs in Section 4 are inspired from the proof of their Proposition 2, and those of Section 5 follow the methods of their Proposition 3.

7 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. 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. 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.

Our work opens a number of interesting avenues for future research. First, it would be valuable to further characterize the stationary distributions obtained by control algorithms. 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 type algorithms. 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.

References

Beck, Carolyn L and Rayadurgam Srikant (2012). "Error bounds for constant step-size Q-learning". In: *Systems & Control Letters* 61.12, pp. 1203–1208.

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.

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*.

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.

Munos, Rémi et al. (2016). "Safe and efficient off-policy reinforcement learning". In: *Advances in Neural Information Processing Systems*, pp. 1054–1062.

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.

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

Watkins, Christopher JCH and Peter Dayan (1992). "Q-learning". In: *Machine learning* 8.3-4, pp. 279–292.

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.

Appendices

Appendix A Laundry list of other algorithms

We outline the general proof recipe, which will be re-using for the following examples.

Proof strategy

- **(P1)** Let $\mu^{(1)}, \mu^{(2)}$ be initial distributions and $(f_0^{(1)}, f_0^{(2)})$ be the optimal coupling which minimizes $\mathcal{W}(\mu^{(1)}, \mu^{(2)})$;
- **(P2)** Define an appropriate coupling $f_1^{(1)} \sim \mu^{(1)}K, f_1^{(2)} \sim \mu^{(2)}K$ – e.g. by defining them to follow the same trajectories if the updates sample from the same distributions;
- **(P3)** Use the upper bound $\mathcal{W}(\mu^{(1)}K, \mu^{(2)}K) \leq \mathbb{E}\left[\|f_1^{(1)} - f_2^{(2)}\|\right]$ and bound $\mathbb{E}\left[\|f_1^{(1)} - f_1^{(2)}\|\right] \leq \rho\mathbb{E}\left[\|f_0^{(1)} - f_0^{(2)}\|\right]$ for some $\rho < 1$ (usually follows from the recursive nature of the updates) to show that $\mu \mapsto \mu K$ is a contraction.

A.1 Convergence of synchronous Monte Carlo Evaluation with constant step-sizes

We prove that Monte Carlo Evaluation with synchronous updates & constant step-size converges to a stationary distribution. The algorithm aims to evaluate the value function of a given policy π using Monte Carlo returns. The update rule is given by:

$$\forall\, s \in \mathcal{S}: \quad V_{n+1}(s) = (1-\alpha)V_n(s) + \alpha\mathcal{G}_n^\pi(s) \tag{MCE}$$

where $\mathcal{G}_n^\pi(s) = \sum_{n\geq 0}\gamma^n r_n(s_n, a_n)$ is the return of a random trajectory $(s_n, a_n, r_n)_{n\geq 0}$ starting from s, following $a_n \sim \pi(\cdot|s_n), r_n \sim \mathcal{R}(\cdot|s_n, a_n)$, and $s_{n+1} \sim \mathcal{P}(\cdot|s_n, a_n)$.

Theorem A.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 (MCE) converges in distribution to a unique stationary distribution $\varphi_\alpha \in \mathcal{M}(\mathbb{R}^{|\mathcal{S}|})$.*

Proof. Following the proof strategy outlined above, we skip to step **(P2)** of the proof. We define the coupling of the updates $(V_1^{(1)}, V_1^{(2)})$ to sample the same trajectories:

$$\left.\begin{aligned} V_1^{(1)}(s) &= (1-\alpha)V_0^{(1)}(s) + \alpha\mathcal{G}_k^\pi(s) \\ V_1^{(2)}(s) &= (1-\alpha)V_0^{(2)}(s) + \alpha\mathcal{G}_k^\pi(s). \end{aligned}\right\} \text{for the } \underline{\text{same}}\ \mathcal{G}_k^\pi(s) \tag{7}$$

Note that this is a valid coupling of $(\mu^{(1)}K_\alpha, \mu^{(2)}K_\alpha)$, since $V_1^{(1)}(s)$ and $V_1^{(2)}(s)$ have access to the same sampling distributions. We upper bound $\mathcal{W}(\mu^{(1)}K_\alpha, \mu^{(2)}K_\alpha)$ by the coupling defined in Equation (7). This gives:

$$\begin{aligned} \mathcal{W}(\mu^{(1)}K_\alpha, \mu^{(2)}K_\alpha) &\leq \mathbb{E}\left[\left\|V_1^{(1)} - V_1^{(2)}\right\|\right] \\ &= \mathbb{E}\left[\left\|(1-\alpha)V_0^{(1)} + \alpha\mathcal{G}_1^\pi - \left((1-\alpha)V_0^{(2)} + \alpha\mathcal{G}_1^\pi\right)\right\|\right] \\ &= \mathbb{E}\left[\left\|(1-\alpha)(V_0^{(1)} - V_0^{(2)})\right\|\right] \\ &= (1-\alpha)\mathbb{E}\left[\left\|V_0^{(1)} - V_0^{(2)}\right\|\right] = (1-\alpha)\mathcal{W}(\mu^{(1)}, \mu^{(2)}) \end{aligned}$$

Since $1-\alpha < 1$, K_α is a contraction mapping and we are done. □

A.2 Convergence of synchronous Q-Learning with constant step-sizes

We prove that Q-Learning with synchronous updates & constant step-sizes converges to a stationary distribution. The algorithm aims to learn the optimal action-value function $Q^\star$.

The updates are given by:

$$\forall\,(s,a)\in\mathcal{S}\times\mathcal{A}:\quad Q_{n+1}(s,a)=(1-\alpha)Q_n(s,a)+\alpha\left(r+\gamma\max_{a'}Q_n(s',a')\right),\qquad \text{(QL)}$$

where $r\sim\mathcal{R}(\cdot|s,a), s'\sim\mathcal{P}(\cdot|s,a)$, and $\alpha>0$.

Theorem A.2. *For any constant step size $0<\alpha\leq 1$ 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 (QL) converges in distribution to a unique stationary distribution $\xi_\alpha\in\mathcal{M}(\mathbb{R}^{|\mathcal{S}|})$.*

Proof. We use the proof outline given above, and jump straight to step **(P2)**. We witness the same-sampling coupling again:

$$\left.\begin{aligned}Q_1^{(1)}(s,a)&=(1-\alpha)Q_0^{(1)}(s,a)+\alpha\left(r+\gamma\max_{a'}Q_0^{(1)}(s',a')\right)\\ Q_1^{(2)}(s.a)&=(1-\alpha)Q_0^{(2)}(s,a)+\alpha\left(r+\gamma\max_{a'}Q_0^{(2)}(s',a')\right)\end{aligned}\right\}\text{ for the }\underline{\text{same}}\ \begin{matrix}r\\ s'\end{matrix}\sim\begin{matrix}\mathcal{R}(s,a),\\ \mathcal{P}(\cdot|s,a)\end{matrix}$$

The bound follows similarly, but with one additional step. Again we write $\widehat{\mathcal{T}}(Q)(s,a)=r+\gamma\max_{a'}Q(s'_{(s,a)},a')$ for the empirical Bellman (optimality) operator.

$$\begin{aligned}\mathbb{E}\left[\left\|\widehat{\mathcal{T}}(Q^{(1)})-\widehat{\mathcal{T}}(Q^{(2)})\right\|\right]&=\mathbb{E}\left[\max_{s,a}\left|r-r+\gamma\left(\max_{a'}Q^{(1)}(s'_{(s,a)},a')-\max_{a'}Q^{(2)}(s'_{(s,a)},a')\right)\right|\right]\\&=\gamma\mathbb{E}\left[\max_{s,a}\left|\max_{a'}Q^{(1)}(s'_{(s,a)},a')-\max_{a'}Q^{(2)}(s'_{(s,a)},a')\right|\right]\\&\leq\gamma\mathbb{E}\left[\max_{s,a}\max_{a'}\left|Q^{(1)}(s'_{(s,a)},a')-Q^{(2)}(s'_{(s,a)},a')\right|\right]\\&\leq\gamma\mathbb{E}\left[\max_{s,a}\left|Q^{(1)}(s,a)-Q^{(2)}(s,a)\right|\right]=\gamma\mathbb{E}\left[\left\|Q^{(1)}-Q^{(2)}\right\|\right]\end{aligned}$$

□

The first inequality follows from $|\max_{a'}Q_1(s,a')-\max_{a'}Q_2(s,a')|\leq\max_{a'}|Q_1(s,a')-Q_2(s,a')|$, and the second inequality follows since $Q^{(1)}$ and $Q^{(2)}$ sampled the same s'. Concluding the proof as before we see that the kernel is contractive with Lipschitz constant $1+\alpha-\alpha\gamma<1$, and we are done.

A.3 TD(λ)

We prove that TD(λ) with synchronous updates & constant step-size converges to a stationary distribution. The algorithm aims to evaluate the value function of a given policy π using a convex combination of n-step returns. The update rule is given by:

$$\forall s:\ V_{n+1}(s)=(1-\alpha)V_n(s,a)+\alpha(1-\lambda)\sum_{k=1}^{\infty}\lambda^{k-1}\left(\sum_{i=0}^{k}\gamma^i r(s_i,a_i)+\gamma^k V_n(s_k)\right)\quad\text{(TD}(\lambda)\text{)}$$

where each n-step trajectory is sampled starting from s and following policy π.

Theorem A.3. *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 (TD(λ)) converges in distribution to a unique stationary distribution $\zeta_\alpha\in\mathcal{M}(\mathbb{R}^{|\mathcal{S}|})$.*

Proof. Again, we jump straight to step **(P2)** of the template given above. We couple every n-step trajectory to sample the same n rewards, actions, and successors states.

$$\left.\begin{aligned}V_{k+1}^{(1)}(s)&=(1-\alpha)V_k^{(1)}(s)+\alpha(1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}\left(\sum_{i=0}^{n-1}\gamma^i r_i(s_i,a_i)+\gamma^n V_k^{(1)}(s_n)\right)\\ V_{k+1}^{(2)}(s)&=(1-\alpha)V_k^{(2)}(s)+\alpha(1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}\left(\sum_{i=0}^{n-1}\gamma^i r_i(s_i,a_i)+\gamma^n V_k^{(2)}(s_n)\right)\end{aligned}\right\}\begin{matrix}\underline{\text{same}}\\ (s_i,a_i,r_i)_{i=0}^{n}\\ \forall n\end{matrix}$$

By the coupling, the reward terms will cancel in every n-step trajectory. We write $R_n^{(i)} = \sum_{i=0}^{n-1} \gamma^i r_i(s_i, a_i) + \gamma^n V_k^{(i)}(s_n)$ for the n-step return and $\hat{\mathcal{T}}(V)(s) = \sum_{k=1}^{\infty} \lambda^{k-1} \left(\sum_{i=0}^{k} \gamma^i r(s_i, a_i) + \gamma^k V_n(s_k) \right)$ for the empirical Bellman operator of TD(λ).

$$
\begin{aligned}
\mathbb{E}\left[\left\|\hat{\mathcal{T}}(V^{(1)}) - \hat{\mathcal{T}}(V^{(2)})\right\|\right] &= \mathbb{E}\left[\max_s \left| \sum_{n=1}^{\infty} \lambda^{n-1} R_n^{(1)} - \sum_{n=1}^{\infty} \lambda^{n-1} R_n^{(2)} \right|\right] \\
&= \mathbb{E}\left[\max_s \left| \sum_{n=1}^{\infty} \lambda^{n-1} \left(R_n^{(1)} - R_n^{(2)} \right) \right|\right] \\
&= \mathbb{E}\left[\max_s \left| \sum_{n=1}^{\infty} \lambda^{n-1} \gamma^n \left(V^{(1)}(s_n) - V^{(2)}(s_n) \right) \right|\right] \\
&\qquad \text{(reward terms cancel)} \\
&\leq \mathbb{E}\left[\sum_{n=1}^{\infty} \lambda^{n-1} \gamma^n \max_s \left| \left(V^{(1)}(s_n) - V^{(2)}(s_n) \right) \right|\right] \\
&\qquad \text{(triangle inequality)} \\
&\leq \sum_{n=1}^{\infty} \lambda^{n-1} \gamma^n \mathbb{E}\left[\max_s \left| V^{(1)}(s) - V^{(2)}(s) \right|\right] \qquad \text{(by the coupling)} \\
&= \sum_{n=1}^{\infty} \lambda^{n-1} \gamma^n \mathbb{E}\left[\left\| V^{(1)} - V^{(2)} \right\|\right] = \gamma \frac{1}{1-\lambda\gamma} \mathbb{E}\left[\left\| V^{(1)} - V^{(2)} \right\|\right]
\end{aligned}
$$

Concluding the proof as before, we have $\mathcal{W}(\mu^{(1)}K, \mu^{(2)}K) \leq (1 - \alpha + \alpha\gamma\frac{1-\lambda}{1-\lambda\gamma})\mathcal{W}(\mu^{(1)}, \mu^{(2)})$. Since $1 - \alpha + \alpha\gamma\frac{1-\lambda}{1-\lambda\gamma} < 1$ we are done. □

A.4 SARSA with ε-greedy policies

In this example we will example the use of ε-greedy policies for control. In particular, we examine SARSA updates with ε-greedy policies. Let $\pi(\cdot|s)$ be some base policy. The updates are as follow:

$$
Q_{k+1}(s,a) = \begin{cases} (1-\alpha)Q_k(s,a) + \alpha\left(r(s,a) + \gamma Q_k(s', a')\right) & \text{w.p. } \varepsilon \\ (1-\alpha)Q_k(s,a) + \alpha\left(r(s,a) + \gamma \max_{a'} Q_k(s', a')\right) & \text{w.p. } 1-\varepsilon \end{cases} \tag{SARSA}
$$

where $r \sim \mathcal{R}(\cdot|s,a)$ and $s' \sim \mathcal{P}(\cdot|s,a)$ in both cases and $a' \sim \pi(\cdot|s')$ in the first case.

Theorem A.4. *For any constant step size $0 < \alpha \leq 1$ 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 (SARSA) converges in distribution to a unique stationary distribution $\theta_\alpha \in \mathcal{M}(\mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|})$.*

Proof. We jump straight to step **(P2)** of the proof template. We use the same-sampling coupling, where $Q_1^{(1)}$ takes the greedy action if and only if $Q_1^{(2)}$ does. In the non-greedy case, they sample the same $a' \sim \pi(\cdot|s')$. In all cases, both functions sample the same $r(s,a)$ and s'.

We write $\hat{\mathcal{T}}(Q)(s,a) = \begin{cases} r + \gamma Q(s', a') \text{ w.p. } \varepsilon \\ r + \gamma \max_{a'} Q(s', a') \text{ w.p. } 1-\varepsilon \end{cases}$

The bound follows similarly to the examples of Q-learning and TD(0). We omit the subscripts

on the Q-functions.

$$\begin{aligned}\mathbb{E}\left[\left\|\hat{\mathcal{T}}(Q^{(1)}) - \hat{\mathcal{T}}(Q^{(2)})\right\|\right] &= \mathbb{P}\left\{\text{greedy action chosen}\right\}\mathbb{E}\left[\max_{s,a}\gamma|(\max_{a'}Q^{(1)}(s',a') - \max_{a'}Q^{(2)}(s',a')|\right] \\ &\quad + \mathbb{P}\left\{\text{non-greedy action chosen}\right\}\mathbb{E}\left[\max_{s,a}|\gamma(Q^{(1)}(s',a') - Q^{(2)}(s',a'))|\right] \\ &\leq \varepsilon\gamma\mathbb{E}\left[\left\|Q^{(1)} - Q^{(2)}\right\|\right] + (1-\varepsilon)\gamma\mathbb{E}\left[\left\|Q^{(1)} - Q^{(2)}\right\|\right] \\ &= \gamma\mathbb{E}\left[\|Q^{(1)} - Q^{(2)}\|\right]\end{aligned}$$

The bound $\mathbb{E}\left[\max_{s,a}\gamma|(\max_{a'}Q^{(1)}(s',a') - \max_{a'}Q^{(2)}(s',a')|\right] \leq \gamma\mathbb{E}\left[\left\|Q^{(1)} - Q^{(2)}\right\|\right]$ follows from $|\max_{a'}Q_1(s,a') - \max_{a'}Q_2(s,a')| \leq \max_{a'}|Q_1(s,a') - Q_2(s,a')|$, and since $Q^{(1)}$ and $Q^{(2)}$ sampled the same s' in the greedy case. The bound $\mathbb{E}\left[\max_{s,a}|\gamma(Q^{(1)}(s',a') - Q^{(2)}(s',a'))|\right] \leq \mathbb{E}\left[\left\|Q^{(1)} - Q^{(2)}\right\|\right]$ follows since $Q^{(1)}$ and $Q^{(2)}$ sampled the same state-action pair in the non-greedy case. Concluding the proof as before, we have that $\mathbb{E}\left[\|Q_1^{(1)} - Q_1^{(2)}\|\right] \leq (1-\alpha+\alpha\gamma)\mathbb{E}\left[\|Q_0^{(1)} - Q_0^{(2)}\|\right]$, and thus the kernel is a contraction. □

A.5 Expected SARSA with ε-greedy policies

In this example we examine the Expected SARSA updates with ε-greedy policies. Let $\pi(\cdot|s)$ be some base policy. Define $\pi_\varepsilon(\cdot|s)$ as the ε-greedy policy which takes the greedy action with probability 1-ε and π otherwise. The updates are as follow:

$$Q_{k+1}(s,a) = (1-\alpha)Q_k(s,a) + \alpha\left(r(s,a) + \gamma\sum_{a'}\pi_\varepsilon(a'|s)Q_k(s',a')\right) \quad \text{(Expected-SARSA)}$$

where $r \sim \mathcal{R}(\cdot|s,a)$ and $s' \sim \mathcal{P}(\cdot|s,a)$ in both cases and $a' \sim \pi(\cdot|s')$ in the first case.

Theorem A.5. *For any constant step size $0 < \alpha \leq 1$ 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 (Expected-SARSA) converges in distribution to a unique stationary distribution $\beta_\alpha \in \mathcal{M}(\mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|})$.*

Proof. We jump straight to step **(P2)** of the proof template. We use the same-sampling coupling.

We write $\hat{\mathcal{T}}(Q)(s,a) = r + \gamma\sum_{a'}\pi(a'|s)Q(s',a')$. The bound follows similarly to the examples of Q-learning and TD(0). We omit the subscripts on the Q-functions.

$$\begin{aligned}\mathbb{E}\left[\left\|\hat{\mathcal{T}}(Q^{(1)}) - \hat{\mathcal{T}}(Q^{(2)})\right\|\right] &= \mathbb{E}\left[\max_{s,a}\gamma|\sum_{a'}\pi_\varepsilon(a')Q^{(1)}(s',a') - \sum_{a'}\pi_\varepsilon(a')Q^{(2)}(s',a')|\right] \\ &\leq \mathbb{E}\left[\max_{s,a}\gamma\sum_{a'}\pi_\varepsilon(a')|Q^{(1)}(s',a') - Q^{(2)}(s',a')|\right] \\ &\leq \mathbb{E}\left[\max_{s,a}\gamma\sum_{a'}\pi_\varepsilon(a')\left\|Q^{(1)}(s',a') - Q^{(2)}(s',a')\right\|\right] \\ &\leq \gamma\mathbb{E}\left[\|Q^{(1)} - Q^{(2)}\|\right]\end{aligned}$$

Concluding the proof as before, we have that $\mathbb{E}\left[\|Q_1^{(1)} - Q_1^{(2)}\|\right] \leq (1-\alpha+\alpha\gamma)\mathbb{E}\left[\|Q_0^{(1)} - Q_0^{(2)}\|\right]$, and thus the kernel is a contraction. □

A.6 Double Q-Learning

In this example we will have to modify our state-space and introduce a new metric on pairs of Q-functions. The Double Q-Learning algorithm (Hasselt, 2010)[2] maintains two random estimates (Q^A, Q^B) and updates Q^A with probability p and Q^B with probability $1-p$. Should Q^A be chosen to be updated, the update is:

$$Q^A_{n+1}(s,a) = (1-\alpha)Q^A_n(s,a) + \alpha\left(r(s,a) + \gamma Q^B_n(s, \operatorname{argmax}_{a'} Q^A_n(s',a'))\right).$$

Analogously, the update for Q^B is:

$$Q^B_{n+1}(s,a) = (1-\alpha)Q^B_n(s,a) + \alpha\left(r(s,a) + \gamma Q^A_n(s, \operatorname{argmax}_{a'} Q^B_n(s',a'))\right).$$

In both cases, we have $s' \sim \mathcal{P}(\cdot|s,a)$. For this algorithm, the updates are Markovian on *pairs* of action-value functions. Thus we set the state space to be $\mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|} \times \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$. We choose the product metric defined by $d_1((Q^A,Q^B),(R^A,R^B)) = \left\|Q^A - R^A\right\| + \left\|Q^B - R^B\right\|$.

Theorem A.6. *For any constant step size $0 < \alpha \leq 1$ and initialization $(Q^A_0, Q^B_0) \sim \mu_0 \in \mathcal{M}(\mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|} \times \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|})$, the sequence of random variables $(Q^A_n, Q^B_n)_{n\geq 0}$ defined by the Double Q-Learning recursion converges in distribution to a unique stationary distribution $\chi_\alpha \in \mathcal{M}(\mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|} \times \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|})$.*

Proof. As before, let $\mu^{(1)}, \mu^{(2)}\mathcal{M}(\mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}\times\mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|})$ be arbitrary initializations and (Q^A_0, Q^B_0) and (R^A_0, R^B_0) be the optimal coupling of $\mathcal{W}(\mu^{(1)}, \mu^{(2)})$. We couple (Q^A_1, Q^B_1) and (R^A_1, R^B_1) to sample the same function to be updated and the same s'. Assume for a moment that Q^A and R^A are chosen to be updated. Proceeding as in the proof of Q-Learning (cf. Theorem A.2), we find that

$$\mathbb{E}\left[\left\|Q^A_1 - R^A_1\right\|\right] \leq (1-\alpha)\mathbb{E}\left[\left\|Q^A_0 - R^A_0\right\|\right] + \alpha\gamma\mathbb{E}\left[\left\|Q^B_0 - R^B_0\right\|\right].$$

Analogously, if Q^B and R^B are chosen to updated, we have:

$$\mathbb{E}\left[\left\|Q^B_1 - R^B_1\right\|\right] \leq (1-\alpha)\mathbb{E}\left[\left\|Q^B_0 - R^B_0\right\|\right] + \alpha\gamma\mathbb{E}\left[\left\|Q^A_0 - R^A_0\right\|\right].$$

Putting everything together, the full expectation is:

$$\begin{aligned}
\mathbb{E}\left[d((Q^A_1,Q^B_1),(R^A_1,R^B_1))\right] &= \mathbb{E}\left[\left\|Q^A_1 - R^A_1\right\| + \left\|Q^B_1 - R^B_1\right\|\right] \\
&= \mathbb{P}\left\{\text{A is updated}\right\}\mathbb{E}\left[\left\|Q^A_1 - R^A_1\right\| + \left\|Q^B_1 - R^B_1\right\|\right] \\
&\quad + \mathbb{P}\left\{\text{B is updated}\right\}\mathbb{E}\left[\left\|Q^A_1 - R^A_1\right\| + \left\|Q^B_1 - R^B_1\right\|\right] \\
&= p\mathbb{E}\left[\left\|Q^A_1 - R^A_1\right\| + \left\|Q^B_0 - R^B_0\right\|\right] \\
&\quad + (1-p)\mathbb{E}\left[\left\|Q^A_0 - R^A_0\right\| + \left\|Q^B_1 - R^B_1\right\|\right] \\
&\leq p\left((1-\alpha)\mathbb{E}\left[\left\|Q^A_0 - R^A_0\right\|\right] + (1+\alpha\gamma)\mathbb{E}\left[\left\|Q^B_0 - R^B_0\right\|\right]\right) \\
&\quad + (1-p)\left((1+\alpha\gamma)\mathbb{E}\left[\left\|Q^A_0 - R^A_0\right\|\right] + (1-\alpha)\mathbb{E}\left[\left\|Q^B_0 - R^B_0\right\|\right]\right) \\
&\leq \frac{1}{2}(2+\alpha\gamma-\alpha)\left(\mathbb{E}\left[\left\|Q^A_0 - R^A_0\right\|\right] + \mathbb{E}\left[\left\|Q^B_0 - R^B_0\right\|\right]\right) \quad (p=\tfrac{1}{2}) \\
&= \frac{1}{2}(2+\alpha\gamma-\alpha)\mathbb{E}\left[d((Q^A_0,Q^B_0),(R^A_0,R^B_0))\right]
\end{aligned}$$

Since $0 \leq 1/2(2+\alpha\gamma-\alpha) < 1$, so we are done. We note that the first equality only follows since, under the coupling, either A or B is updated for both functions. □

Appendix B Proofs of Section 5

Theorem B.1. *Suppose $\widehat{\mathcal{T}}^\pi$ is such that the updates* (3) *with step-size α converge to a stationary distribution ψ_α. If $\widehat{\mathcal{T}}^\pi$ is an empirical Bellman operator for some policy π, then $\mathbb{E}[f_\alpha] = f^\pi$ where $f_\alpha \sim \psi_\alpha$ and f^π is the fixed point of $\mathcal{T}^\pi$.*

[2]This is the original algorithm, not the deep reinforcement learning version given in (**van2016deep**).

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

$$f_1 = (1-\alpha)f_0 + \alpha\widehat{\mathcal{T}}^\pi(f_0, \omega) \tag{8}$$

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

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

Since $\mathcal{T}^\pi(\cdot) = \mathcal{R}^\pi + \gamma\mathcal{P}^\pi(\cdot)$ is an affine operator it commutes with expectation, thus:

$$f_\alpha = \mathcal{T}^\pi\overline{f_\alpha}$$

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

Theorem B.2. *Suppose $\widehat{\mathcal{T}}^\pi$ is such that the updates* (3) *with step-size α converge to a stationary distribution ψ_α, and that $\widehat{\mathcal{T}}^\pi$ is an empirical Bellman operator for some policy π. Define*

$$\mathcal{C}(f) \coloneqq \mathbb{E}_\omega[(\widehat{\mathcal{T}}^\pi(f,\omega) - \mathcal{T}^\pi f)(\widehat{\mathcal{T}}^\pi(f,\omega) - \mathcal{T}^\pi f)^\mathsf{T}]$$

to be the covariance of the zero-mean noise term $\widehat{\mathcal{T}}^\pi(f,\omega) - \mathcal{T}^\pi f$ for a given function f. Then, the covariance of $f_\alpha \sim \psi_\alpha$ is given by

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

Furthermore, 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, where $\|\cdot\|_{op}$ denotes the operator norm of a matrix. 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:

$$\mathbb{P}\left\{\min_i |f_\alpha(i) - f^\pi(i)| \geq \varepsilon\right\} \overset{\alpha\to 0}{\longrightarrow} 0 \quad \forall\, \varepsilon > 0$$

We preface the proof with some useful identities. We will write the covariance in terms of the tensor product for ease of manipulations

Lemma B.1. *Write $\xi(f) \coloneqq (\widehat{\mathcal{T}}^\pi(f,\omega) - \mathcal{T}^\pi f)$. In the same setup as Theorem 5.2:*

$$\mathbb{E}\left[(f_\alpha - f^\pi)(\mathcal{T}^\pi f_\alpha - f^\pi + \xi(f_0))^\mathsf{T}\right] = \mathbb{E}\left[(f_\alpha - f^\pi)(f_\alpha - f^\pi)^\mathsf{T}\right](\gamma\mathcal{P}^\pi)^\mathsf{T}$$

and

$$\begin{aligned}\mathbb{E}\left[((\mathcal{T}^\pi f_\alpha - f^\pi) + \xi(f_\alpha))((\mathcal{T}^\pi f_\alpha - f^\pi) + \xi(f_\alpha))^\mathsf{T}\right] = {}& (\gamma\mathcal{P}^\pi)\mathbb{E}\left[(f_\alpha - f^\pi)(f_\alpha - f^\pi)^\mathsf{T}\right](\gamma\mathcal{P}^\pi)^\mathsf{T}\\ &+ \int C(v)\psi_\alpha(\mathrm{d}v)\end{aligned}$$

Proof. Let $f_0 \sim \psi_\alpha$, by (3) we have $f_1 = (1-\alpha)f_0 + \alpha(\mathcal{T}^\pi f_0 + \xi(f_0))$ and $f_1 \sim \psi_\alpha$. Furthermore, the distribution of f_0 is independent of the distribution of ω. By independence,

$$\begin{aligned}\mathbb{E}\left[(f_0 - f^\pi)\xi(f_0)^\mathsf{T}\right] &= \mathbb{E}_{f_0}\mathbb{E}_\omega\left[(f_0 - f^\pi)\xi(f_0)^\mathsf{T}\right] && \text{(by independence of } f_0 \text{ and } \xi(\cdot))\\ &= \mathbb{E}_{f_0}\left[(f_0 - f^\pi)(\mathbb{E}_\omega\xi(f_0))^\mathsf{T}\right] = 0 && (\mathbb{E}_\omega[\xi(f)] = 0 \text{ for every } f)\end{aligned}$$

For the first identity, note that

$$\begin{aligned}\mathbb{E}\left[(f_0 - f^\pi)(\mathcal{T}^\pi f_0 - f^\pi))^\mathsf{T}\right] &= \mathbb{E}\left[(f_0 - f^\pi)(\mathcal{R}^\pi + \gamma\mathcal{P}^\pi(f_0) - \mathcal{R}^\pi - \gamma\mathcal{P}^\pi(f^\pi))^\mathsf{T}\right] \\ &= \mathbb{E}\left[(f_0 - f^\pi)(\gamma\mathcal{P}^\pi(f_0 - f^\pi))^\mathsf{T}\right] \\ &= \mathbb{E}\left[(f_0 - f^\pi)(f_0 - f^\pi)^\mathsf{T}(\gamma\mathcal{P}^\pi)^\mathsf{T}\right] \\ &= \mathbb{E}\left[(f_0 - f^\pi)(f_0 - f^\pi)^\mathsf{T}\right](\gamma\mathcal{P}^\pi)^\mathsf{T}\end{aligned}$$

The first identity then follows by using $\mathbb{E}\left[(f_0 - f^\pi)\xi(f_0)^\mathsf{T}\right] = 0$ and linearity of expectations.

For the second identity, expanding the outer product gives:

$$\begin{aligned}\mathbb{E}\left[((\mathcal{T}^\pi f_0 - f^\pi) + \xi(f_0))\,((\mathcal{T}^\pi f_0 - f^\pi) + \xi(f_0))^\mathsf{T}\right] &= \mathbb{E}\left[(\mathcal{T}^\pi f_0 - f^\pi)(\mathcal{T}^\pi f_0 - f^\pi)^\mathsf{T}\right] \\ &\quad + \mathbb{E}\left[(\xi(f_0))(\xi(f_0)))^\mathsf{T}\right] \\ &\quad + \cancel{\mathbb{E}\left[(\mathcal{T}^\pi f_0 - f^\pi)(\xi(f_0))^\mathsf{T}\right]} \\ &\quad + \cancel{\mathbb{E}\left[\xi(f_0)(\mathcal{T}^\pi f_0 - f^\pi)^\mathsf{T}\right]} \\ &= \mathbb{E}\left[(\gamma\mathcal{P}^\pi(f_0 - f^\pi))(\gamma\mathcal{P}^\pi(f_0 - f^\pi))^\mathsf{T}\right] \\ &\quad + \int \mathcal{C}(v)\psi_\alpha(\mathrm{d}v) \\ &= (\gamma P^\pi)\mathbb{E}\left[(f_0 - f^\pi)(f_0 - f^\pi)^\mathsf{T}\right](\gamma P^\pi)^\mathsf{T} \\ &\quad + \int \mathcal{C}(v)\psi_\alpha(\mathrm{d}v)\end{aligned}$$

where we used $\mathbb{E}\left[(\mathcal{T}^\pi f_0 - f^\pi)(\xi(f_0))^\mathsf{T}\right] = 0$. □

Proof (of Theorem 5.2). Again let f_0 be distributed according to ψ_α. Subtracting f^π from equation (8),

$$f_1 - f^\pi = (1-\alpha)(f_0 - f^\pi) + \alpha\left(\mathcal{T}^\pi f_0 - f^\pi + \xi(f_0)\right).$$

and taking outer products:

$$\begin{aligned}(f_1 - f^\pi)\,(f_1 - f^\pi)^\mathsf{T} =& (1-\alpha)^2\,(f_0 - f^\pi)\,(f_0 - f^\pi)^\mathsf{T} \\ &+ \alpha^2\,(\mathcal{T}^\pi f_0 - f^\pi + \xi(f_0))\,(\mathcal{T}^\pi f_0 - f^\pi + \xi(f_0))^\mathsf{T} \\ &+ \alpha(1-\alpha)(f_0 - f^\pi)(\mathcal{T}^\pi f_0 - f^\pi + \xi(f_0))^\mathsf{T} \\ &+ \alpha(1-\alpha)(\mathcal{T}^\pi f_0 - f^\pi + \xi(f_0))(f_0 - f^\pi)^\mathsf{T}.\end{aligned}$$

Taking expectations on both sides, and using Lemma B.1:

$$\begin{aligned}\mathbb{E}\left[(f_1 - f^\pi)(f_1 - f^\pi)^\mathsf{T}\right] =& (1-\alpha)^2\mathbb{E}\left[(f_0 - f^\pi)(f_0 - f^\pi)^\mathsf{T}\right] + \alpha^2(\gamma\mathcal{P}^\pi)\mathbb{E}[(f_0 - f^\pi)](\gamma\mathcal{P}^\pi)^\mathsf{T} \\ &+ \alpha^2\int \mathcal{C}(v)\psi_a(\mathrm{d}v) \\ &+ \alpha(1-\alpha)(\gamma\mathcal{P}^\pi)\mathbb{E}\left[(f_0 - f^\pi)(f_0 - f^\pi)^\mathsf{T}\right] \\ &+ \alpha(1-\alpha)\mathbb{E}\left[(f_0 - f^\pi)(f_0 - f^\pi)^\mathsf{T}\right](\gamma\mathcal{P}^\pi)^\mathsf{T}\end{aligned}$$

Since $\mathbb{E}\left[(f_1 - f^\pi)(f_1 - f^\pi)^\mathsf{T}\right] = \mathbb{E}\left[(f_0 - f^\pi)(f_0 - f^\pi)^\mathsf{T}\right]$ by stationarity, re-arranging to the LHS and factoring gives:

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

For the remainder of the proof we re-write the above expression in terms of tensor products. The tensor product of two vectors x, y is the matrix defined by $x \otimes y = xy^\mathsf{T}$. By extension, the tensor product of two matrices A, B is the operator defined by $(A \otimes B)X = AXB^\mathsf{T}$. Then, the above expression can be re-written as:

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

Factoring the tensor products further gives:

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

We show that the matrix on the LHS is invertible. By (Puterman, 2014, Corollary C.4) it will follow from showing that $\rho\left(((1-\alpha)I + \alpha\gamma P^\pi)^{\otimes 2}\right) < 1$, where $\rho(A)$ is the spectral radius of matrix A. Writing $\|A\|_{\text{op}} = \max_i \sum_j |A(i,j)|$ for the operator norm of a matrix A, and using that $\rho(A) \le \|A\|_{\text{op}}$, $\|A \otimes B\|_{\text{op}} = \|A\|_{\text{op}}\|B\|_{\text{op}}$, and $\|P^\pi\|_{\text{op}} = \|I\|_{\text{op}} = 1$:

$$\left\|((1-\alpha)I + \alpha\gamma P^\pi)^{\otimes 2}\right\|_{\text{op}} = \|(1-\alpha)I + \alpha\gamma P^\pi\|_{\text{op}}^2 \le ((1-\alpha) + \alpha\gamma)^2 < 1, \tag{9}$$

where the last inequality followed since $\gamma < 1$. Finally, for the limit $\alpha \to 0$, we use the following identity: if A is such that $\|I - A\| \le 1$ then $\left\|A^{-1}\right\| \le \frac{1}{1-\|I-A\|}$. We let $A = I - ((1-\alpha)I + \alpha\gamma\mathcal{P}^\pi)^{\otimes 2}$, by the calculation in (9) we have $\|I - A\| < 1$. So we calculate the operator norm of the covariance matrix:

$$\begin{aligned}\left\|\mathbb{E}\left[(f_0 - f^\pi)(f_0 - f^\pi)^\mathsf{T}\right]\right\| &= \alpha^2 \left\|\left[I - ((1-\alpha)I + \alpha\gamma P^\pi)^{\otimes 2}\right]^{-1} \int \mathcal{C}(v)\psi_\alpha(\mathrm{d}v)\right\| \\ &\le \alpha^2 \left\|\left[I - ((1-\alpha)I + \alpha\gamma P^\pi)^{\otimes 2}\right]^{-1}\right\| \left\|\int \mathcal{C}(v)\psi_\alpha(\mathrm{d}v)\right\| \\ &\le \alpha^2 \frac{1}{1 - \left\|I - I + ((1-\alpha)I + \alpha\gamma P^\pi)^{\otimes 2}\right\|} \left\|\int \mathcal{C}(v)\psi_\alpha(\mathrm{d}v)\right\| \\ &= \alpha^2 \frac{1}{1 - \left\|((1-\alpha)I + \alpha\gamma P^\pi)^{\otimes 2}\right\|} \left\|\int \mathcal{C}(v)\psi_\alpha(\mathrm{d}v)\right\| \\ &= \alpha^2 \frac{1}{1 - \|((1-\alpha)I + \alpha\gamma P^\pi)\|^2} \left\|\int \mathcal{C}(v)\psi_\alpha(\mathrm{d}v)\right\| \\ &\le \alpha^2 \frac{1}{1 - (1-\alpha+\alpha\gamma)^2} \left\|\int \mathcal{C}(v)\psi_\alpha(\mathrm{d}v)\right\|\end{aligned}$$

Finally, since $\mathcal{C}(f)$ is bounded for all $f \in \mathbb{R}^n$, we have that $\left\|\int \mathcal{C}(v)\psi_\alpha(\mathrm{d}v)\right\| \le M$ and thus

$$\left\|\mathbb{E}\left[(f_0 - f^\pi)(f_0 - f^\pi)^\mathsf{T}\right]\right\| \le M\frac{\alpha^2}{1 - (1-\alpha+\alpha\gamma)^2} \xrightarrow{\alpha\to 0} 0$$

For the concentration inequality, we will use a multivariate Chebyshev inequality (Marshall and Olkin, 1960, Theorem 3.1), whos statement is as follows:

Theorem B.3. *Let $X = (X_1, ..., X_n)$ be a random vector with $\mathbb{E}X = 0$ and $\mathbb{E}[X^T X] = \Sigma$. Let $T = T_+ \cup \{x : -x \in T_+\}$, where $T_+ \subseteq \mathbb{R}^n$ is a closed, convex set. If $A = \{a \in \mathbb{R}^n : \langle a, x\rangle \ge 1 \ \forall x \in T_+\}$, then*

$$\mathbb{P}\{X \in T\} \le \inf_{a \in A} a^\mathsf{T}\Sigma a$$

Let $\varepsilon > 0$. We first bound $a^\mathsf{T} \Sigma a$ with the operator norm of Σ. Note that

$$\begin{aligned} a^\mathsf{T} \Sigma a &= \sum_i a_i (\Sigma a)_i \\ &\leq \sum_i a_i \left\| \Sigma a \right\| \leq n \left\| \Sigma \right\|_{\text{op}} \left\| a \right\|^2 \end{aligned}$$

We define T_+ to be the intersection of half-planes the $\{x | x_i \geq \varepsilon\}$, so that $T_+ = \{x | x_i \geq \varepsilon \; \forall i\}$. Since the half-planes are closed and convex, T_+ is also closed and convex since it is an intersection of closed and convex sets.Then, $T = T_+ \cup \{x : -x \in T_+\} = \{x | x_i \geq \varepsilon \; \forall i \text{ or } x_i \leq -\varepsilon \; \forall i\}$. Note that $x \in T \iff \min_i |x_i| \geq \varepsilon$. We define $X = f_\alpha - f^\pi$ which has zero-mean. Finally, Theorem B.3 states that

$$\mathbb{P}\left\{X \in T\right\} = \mathbb{P}\left\{f_\alpha - f^\pi \in T\right\} \leq \inf_{a \in A} a^T \Sigma a \leq n \left\| \Sigma \right\|_{\text{op}} \inf_{a \in A} \left\| a \right\|^2 .$$

Note that $\inf_a \left\| a \right\|^2$ is bounded since $a = (\frac{1}{n\varepsilon}, \frac{1}{n\varepsilon}, ..., \frac{1}{n\varepsilon})$ is in A and $\left\| a \right\|^2 = \frac{1}{(n\varepsilon)^2}$. So $n \inf_{a \in A} \left\| a \right\|^2 \leq C$ for some constant C. From the previous result, we can take the limit of $\alpha \to 0$ of $\left\| \Sigma \right\|_{\text{op}} = \left\| \mathbb{E}\left[(f_\alpha - f^\pi)(f_\alpha - f^\pi)^\mathsf{T}\right] \right\|_{\text{op}}$ and obtain:

$$\mathbb{P}\left\{f_\alpha - f^\pi \in T\right\} = \mathbb{P}\left\{\min_i |f_\alpha(i) - f^\pi(i)| \geq \varepsilon\right\} \leq C \cdot \left\| \mathbb{E}\left[(f_\alpha - f^\pi)(f_\alpha - f^\pi)^\mathsf{T}\right] \right\|_{\text{op}} \to 0$$

□

Theorem B.4. *Suppose $\widehat{\mathcal{T}}^\star$ is such that the updates* (3) *with step-size α converge to a stationary distribution $\psi_\alpha^\star$. If $\widehat{\mathcal{T}}^\star$ is an empirical Bellman optimality operator then*

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

where $f_\alpha \sim \psi_\alpha^\star$ and $f^\star$ is the fixed point of $\mathcal{T}^\star$. 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 $\overline{f_\alpha} \coloneqq \mathbb{E}\left[f_\alpha\right]$ gives:

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

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