Instituto de Sistemas e Robótica

Pólo de Lisboa

Q-learning with linear function approximation[1]

Francisco S. Melo
M. Isabel Ribeiro

March 2007
RT-602-07

ISR Torre Norte
Av. Rovisco Pais, 1
1049-001 Lisboa
PORTUGAL

[1]This work was partially supported by Programa Operacional Sociedade do Conhecimento (POS_C) that includes FEDER funds. The first author acknowledges the PhD grant SFRH/BD/3074/2000.

Q-learning with linear function approximation

Francisco S. Melo **M. Isabel Ribeiro**
Institute for Systems and Robotics
Instituto Superior Técnico
Av. Rovisco Pais, 1
1049-001 Lisboa,
PORTUGAL
{fmelo,mir}@isr.ist.utl.pt

Abstract

In this paper, we analyze the convergence of Q-learning with linear function approximation. We identify a set of conditions that implies the convergence of this method with probability 1, when a fixed learning policy is used. We discuss the differences and similarities between our results and those obtained in several related works. We also discuss the applicability of this method when a changing policy is used. Finally, we describe the applicability of this approximate method in partially observable scenarios.

1 Introduction

Reinforcement learning addresses the problem of an agent faced with a sequential decision problem and using evaluative feedback as a performance measure. Reinforcement learning methods compute a mapping from the set of states of the agent/environment to the set of possible actions. Such mapping is called a *policy* and it is customary to define a utility-function, or *value-function*, estimating the practical utility of each particular policy. Value-based methods such as TD-learning [34], Q-learning [44], SARSA [32] and many others [5, 12, 21, 36] have been exhaustively covered in the literature and, under mild assumptions, have been proven to converge to the desired solution [6].

However, many such algorithms require explicit representation of the state-space, and it is often the case that the latter is unsuited for explicit representation. Instead, the decision-maker should be able to *generalize* its action-pattern from the collected experience. There are numerous works in the topic of generalization. In many such works, a suitable approximation architecture is proposed and then applied with one's favorite learning method [11, 35]. Encouraging results were reported, perhaps the most spectacular of which by Tesauro's Gammon player [38, 39]. Several other works provided formal analysis of convergence when RL algorithms are combined with function approximation. We refer the early works by Singh et al. [33], Gordon [17] and Tsitsiklis and Van Roy [40]. A few other works further extended the applicability/performance of these methods [2, 12, 27, 30, 37, 41].

In this paper, we analyze the convergence of Q-learning with linear function approximation. Our approach is closely related to interpolation-based Q-learning [37] and the learning algorithm by Borkar [8]. We identify conditions that ensure convergence with probability 1 (w.p.1). We also interpret the obtained approximation and discuss the error bounds in the obtained approximation. We conclude the paper by addressing the applicability of our methods to partially observable scenarios.

2 The Framework of Markov Decision Process

Let $\mathcal{X}$ be a compact subspace of $\mathbb{R}^p$ and $\{X_t\}$ a $\mathcal{X}$-valued controlled Markov chain.[1] The transition probabilities for the chain are given by a the kernel

$$\mathbb{P}\left[X_{t+1} \in U \mid X_t = x, A_t = a\right] = \mathsf{P}_a(x, U),$$

where U is any measurable subset of $\mathcal{X}$. The $\mathcal{A}$-valued process $\{A_t\}$ represents the control process: A_t is the control action at time instant t and $\mathcal{A}$ is the finite set of possible actions. The agent aims to choose the control process $\{A_t\}$ so as to maximize the expected total discounted reward, *i.e.*,

$$V(\{A_t\}, x) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(X_t, A_t) \mid X_0 = x\right],$$

where $0 < \gamma < 1$ is a discount-factor and $R(x, a)$ represents a random "reward" received for taking action $a \in \mathcal{A}$ in state $x \in \mathcal{X}$.

We assume throughout this paper that there is a deterministic function $r : \mathcal{X} \times \mathcal{A} \times \mathcal{X} \longrightarrow \mathbb{R}$ assigning a reward $r(x, a, y)$ every time a transition from x to y occurs after taking action a and that

$$\mathbb{E}\left[R(x, a)\right] = \int_{\mathcal{X}} r(x, a, y)\mathsf{P}_a(x, dy).$$

This simplifies the notation without introducing a great loss in generality. We further assume that there is a constant $\mathcal{R} \in \mathbb{R}$ such that $|r(x, a, y)| < \mathcal{R}$ for all $x, y \in \mathcal{X}$ and all $a \in \mathcal{A}$.[2] We refer to the 5-tuple $(\mathcal{X}, \mathcal{A}, \mathsf{P}, r, \gamma)$ as a *Markov decision process* (MDP).

Given the MDP $(\mathcal{X}, \mathcal{A}, \mathsf{P}, r, \gamma)$, the *optimal value function* V^* is defined for each state $x \in \mathcal{X}$ as

$$V^*(x) = \max_{\{A_t\}} V(\{A_t\}, x) = \max_{\{A_t\}} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(X_t, A_t) \mid X_0 = x\right]$$

and verifies

$$V^*(x) = \max_{a \in \mathcal{A}} \int_{\mathcal{X}} \big[r(x, a, y) + \gamma V^*(y)\big]\mathsf{P}_a(x, dy),$$

which is a form of the Bellman optimality equation. From the optimal value function, the optimal Q-values, $Q^*(x, a)$, are defined for each state-action pair $(x, a) \in \mathcal{X} \times \mathcal{A}$ as

$$Q^*(x, a) = \int_{\mathcal{X}} \big[r(x, a, y) + \gamma V^*(y)\big]\mathsf{P}_a(x, dy). \tag{2.1}$$

From Q^* it is also possible to define a mapping $\pi^* : \mathcal{X} \longrightarrow \mathcal{A}$ as

$$\pi^*(x) = \arg\max_{a \in \mathcal{A}} Q^*(x, a), \qquad \text{for all } x \in \mathcal{X},$$

and the control process $\{A_t\}$ defined by $A_t = \pi^*(X_t)$ is optimal in the sense that

$$V(\{A_t\}, x) = V^*(x),$$

for all $x \in \mathcal{X}$. The mapping π^* is an *optimal policy* for the MDP $(\mathcal{X}, \mathcal{A}, \mathsf{P}, r, \gamma)$.

More generally, a *policy* is a mapping $\pi_t : \mathcal{X} \times \mathcal{A} \longrightarrow [0, 1]$ that generates a control process $\{A_t\}$ verifying

$$\mathbb{P}\left[A_t = a \mid X_t = x\right] = \pi_t(x, a),$$

for all t. Clearly, since $\pi_t(x, \cdot)$ is a probability distribution over $\mathcal{A}$, it must satisfy $\sum_{a \in \mathcal{A}} \pi_t(x, a) = 1$, for all $x \in \mathcal{X}$. A *stationary policy* is a policy δ that does not depend on t. A *deterministic*

[1] We refer to a *subspace* in the topological sense.

[2] This assumption is tantamount to the standard requirement that the rewards $R(x, a)$ have uniformly bounded variance.

policy is a policy assigning probability 1 to a single action in each state. We denote such policy as a function $\pi_t : \mathcal{X} \longrightarrow \mathcal{A}$, that generates a control process $\{A_t\}$ verifying $A_t = \pi_t(X_t)$ for all t. We write $V^{\pi_t}(x)$ instead of $V(\{A_t\}, x)$ if the control process $\{A_t\}$ is generated by a policy π_t.

The optimal control process can be obtained from the optimal (stationary, deterministic) policy π^*, which can in turn be obtained from Q^*. Therefore, the optimal control problem is solved once the function Q^* is known for all pairs $(x, a) \in \mathcal{X} \times \mathcal{A}$.

Now given any two functions $v : \mathcal{X} \longrightarrow \mathbb{R}$ and $q : \mathcal{X} \times \mathcal{A} \longrightarrow \mathbb{R}$, we can define the operators

$$(\mathbf{T}v)(x) = \max_{a \in \mathcal{A}} \int_{\mathcal{X}} \big[r(x, a, y) + \gamma v(y)\big] \mathsf{P}_a(x, dy)$$

and

$$(\mathbf{H}q)(x, a) = \int_{\mathcal{X}} \big[r(x, a, y) + \gamma \max_{u \in \mathcal{A}} q(y, u)\big] \mathsf{P}_a(x, dy). \tag{2.2}$$

The functions V^* and Q^* introduced above are fixed-points of the operators $\mathbf{T}$ and $\mathbf{H}$, respectively. Each of these operators is a contraction in a corresponding norm and, theoretically, a fixed-point iteration could be used to determine V^* and Q^*.

On the other hand, if P or r (or both) are not known, the *Q-learning algorithm* can be used, defined by the update rule

$$Q_{k+1}(x, a) = (1 - \alpha_k) Q_k(x, a) + \alpha_k \big[R(x, a) + \gamma \max_{u \in \mathcal{A}} Q_k(X(x, a), u)\big], \tag{2.3}$$

where $Q_k(x, a)$ is the kth estimate of $Q^*(x, a)$, $X(x, a)$ is a $\mathcal{X}$-valued random variable obtained according to the probabilities defined by P and $\{\alpha_k\}$ is a step-size sequence. Notice that $R(x, a)$ and $X(x, a)$ can be obtained through some simulation device, not requiring the knowledge of either P or r. The estimates Q_k converge with probability 1 (w.p.1) to Q^* as long as

$$\sum_t \alpha_t = \infty \qquad\qquad \sum_t \alpha_t^2 < \infty.$$

The Q-learning algorithm was first proposed by Watkins in 1989 [44] and its convergence w.p.1 later established by Watkins and Dayan [43] and several other authors [6, 20].

3 Q-learning with linear function approximation

In this section, we establish the convergence properties of Q-learning when using linear function approximation. We identify the conditions ensuring convergence w.p.1 and derive error bounds for the obtained approximation. As will soon become apparent, the results derived herein are deeply related with other approaches described in the literature [8, 17, 33, 37, 40].

3.1 Combining Q-learning with linear function approximation

We previously suggested that a fixed-point iteration could theoretically be used to determine Q^*. However, this implicitly requires that the successive estimates for Q^* can be represented compactly and stored in a computer with finite memory and that the transition kernel P and the reward function r are known.

If $\mathcal{X}$ is finite, Q^* (and hence any corresponding estimates) can be represented as an $|\mathcal{X}| \times |\mathcal{A}|$ matrix. Therefore, the first of the two above conditions is trivially fulfilled. To solve for Q^* we can use the fixed-point iteration proposed in Section 2 or the Q-learning algorithm, if P and r are not known.

However, if $\mathcal{X}$ is an infinite set, it is no longer possible to straightforwardly apply any of the aforementioned methods. For example, the updates in (2.3) explicitly consider the Q-values for each individual state-action pair and there will be infinitely many such pairs if $\mathcal{X}$ is not finite. Therefore, some compact representation of either $\mathcal{X}$ or Q^* is necessary to tackle the infinite nature of $\mathcal{X}$. In our approach, we focus on compact representations for Q^*.

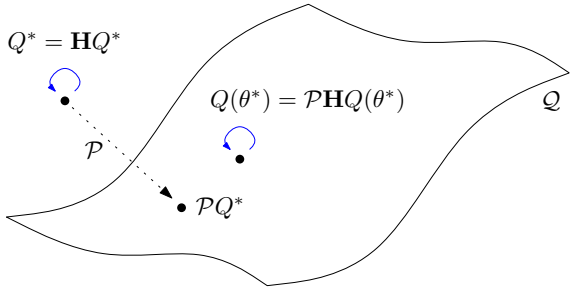

Figure 1: Optimal function $Q^* = \mathbf{H}Q^*$ and the fixed-point of $Q(\theta^*)$ of the combined operator $\mathcal{P}\mathbf{H}$. Notice that, in general $Q(\theta^*) \neq \mathcal{P}Q^*$.

In our pursuit to approximate Q^*, we start by considering a family of functions $\mathcal{Q} = \{Q_\theta\}$ parameterized by a finite-dimensional parameter vector $\theta \in \mathbb{R}^M$. If we replace the iterative procedure to find Q^* by a suitable "equivalent" procedure to find a parameter θ^* so as to best approximate Q^* by a function in $\mathcal{Q}$, we move from a search in an infinite dimensional function space to a search in a finite dimensional space ($\mathbb{R}^M$). This has an immediate implication: unless if $Q^* \in \mathcal{Q}$, we will not be able to determine Q^* exactly. Instead, we will determine the fixed point of a combined operator $\mathcal{P}\mathbf{H}$, where $\mathcal{P}$ is some mapping that "projects" a function q defined in $\mathcal{X} \times \mathcal{A}$ to a point in $\mathcal{Q}$ (see Fig. 1).

In this paper we admit the family $\mathcal{Q}$ to be linear in that if $q_1, q_2 \in \mathcal{Q}$, then so does $\alpha q_1 + q_2$ for any $\alpha \in \mathbb{R}$. $\mathcal{Q}$ is therefore the linear span of some set of linearly independent functions $\xi_i : \mathcal{X} \times \mathcal{A} \longrightarrow \mathbb{R}$, and each $q \in \mathcal{Q}$ can be written as

$$q(x, a) = \sum_{i=1}^{M} \xi_i(x, a)\theta(i),$$

where $\theta(i)$ is the ith component of the vector $\theta \in \mathbb{R}^M$. If $\Xi = \{\xi_1, \ldots, \xi_M\}$ is a set of linearly independent functions, we interchangeably use Q_θ and $Q(\theta)$ to denote the function

$$Q_\theta(x, a) = \sum_{i=1}^{M} \xi_i(x, a)\theta(i) = \xi^\top(x, a)\theta, \tag{3.1}$$

where $\xi(x, a)$ is a vector in $\mathbb{R}^M$ with ith component given by $\xi_i(x, a)$.

We throughout let $\Xi = \{\xi_i, i = 1, \ldots, M\}$ be a set of M bounded, linearly independent functions verifying

$$\sum_i |\xi_i(x, a)| \leq 1 \tag{3.2}$$

for all $(x, a) \in \mathcal{X} \times \mathcal{A}$ and eventually introduce further restrictions on the set Ξ as needed.

3.2 Linear approximation using sample-based projection

We now consider a sample-based approximation model that, while imposing somewhat strict conditions on the set of functions Ξ, will allow us to derive useful error bounds for the obtained approximation Q_{θ^*}. For that we assume that the functions in Ξ verify

$$\|\xi_i\|_\infty = 1. \tag{3.3}$$

We remark that if (3.2) and (3.3) simultaneously hold, linear independence of the functions in Ξ arises as an immediate consequence. To see this, notice that for each function $\xi_i \in \Xi$ there is a point (x, a) such that $|\xi_i(x, a)| = 1$, as $\|\xi_i\|_\infty = 1$. Then, since $\sum_i |\xi_i(x, a)| \leq 1$, $\xi_j(x, a) = 0$ for all $j \neq i$. This, in turn, implies the functions in Ξ are linearly independent. As in the previous subsection, we take the family $\mathcal{Q}$ as the linear span of Ξ.

For each function $\xi_i \in \Xi$ take a point (x_i, a_i) in $\mathcal{X} \times \mathcal{A}$ such that $|\xi_i(x_i, a_i)| = 1$, and denote by I the set obtained by gathering M of such points, one for each $\xi_i \in \Xi$. If $\mathcal{B}$ is the set of all (essentially) bounded functions defined on $\mathcal{X} \times \mathcal{A}$ and taking values on $\mathbb{R}$, we define a mapping $\wp : \mathcal{B} \longrightarrow \mathbb{R}^M$ as

$$(\wp f)(i) = f(x_i, a_i), \tag{3.4}$$

where f is an arbitrary function in $\mathcal{B}$, $(\wp f)(i)$ is the ith component of the vector $\wp f$ and (x_i, a_i) is the point in I corresponding to ξ_i. Notice that $\wp f$ is properly defined for every $f \in \mathcal{B}$ and verifies

$$\|\wp f\|_\infty \leq \|f\|_\infty$$

and

$$\wp[\alpha f_1 + f_2] = \alpha \wp f_1 + \wp f_2.$$

Our variant of Q-learning iteratively determines the point $\theta^* \in \mathbb{R}^M$ verifying the fixed-point recursion

$$\theta^* = \wp \mathbf{H} Q(\theta^*), \tag{3.5}$$

where $\mathbf{H}$ is the operator defined in (2.2). Since $\mathbf{H}$ is a contraction in the maximum norm and $\sum_i |\xi_i(x, a)| \leq 1$, the fixed point in (3.5) is properly and uniquely defined.

To derive the expression of the algorithm, we remark that (3.5) can be explicitly written as

$$\theta^*(i) = \int_{\mathcal{X}} \delta_{(x_i,a_i)}(x, a) \int_{\mathcal{X}} \big[r(x, a, y) + \gamma \max_u \xi^\top(y, u)\theta^*\big] \mathsf{P}_a(x, dy) d\mu(x, a),$$

where μ is some probability measure on $\mathcal{X} \times \mathcal{A}$ and $\delta_{(x_i,a_i)}$ is the Dirac delta centered around (x_i, a_i).

We are now in position to describe our algorithm. Let g_ε be a smooth Dirac approximation,[3] such that

$$\int g_\varepsilon(x, a; y, u) d\mu(y, u) = 1$$

$$\lim_{\varepsilon \to 0} \int g_\varepsilon(x, a; y, u) f(y, u) d\mu(y, u) = f(x, a).$$

Let π be a stochastic stationary policy and suppose that $\{x_t\}$, $\{a_t\}$ and $\{r_t\}$ are sampled trajectories from the MDP $(\mathcal{X}, \mathcal{A}, \mathsf{P}, r, \gamma)$ using policy π. Then, given any initial estimate θ_0, we generate a sequence $\{\theta_t\}$ according to the update rule

$$\theta_{t+1}(i) = \theta_t(i) + \alpha_t g_{\varepsilon_t}(x_i, a_i; x_t, a_t)\big[r_t + \gamma \max_{u \in \mathcal{A}} \xi^\top(x_{t+1}, u)\theta_t - \xi^\top(x_t, a_t)\theta_t\big],$$

where $\{\varepsilon_t\}$ is a sequence verifying

$$\varepsilon_{t+1} = (1 - \beta_t)\varepsilon_t.$$

More generally, we can have

$$\varepsilon_{t+1} = \varepsilon_t + \beta_t h(\varepsilon_t),$$

where h is chosen so that the ODE $\dot{x}_t = h(x_t)$ has a globally asymptotically stable equilibrium in the origin.

Under some regularity assumptions on the Markov chain $(\mathcal{X}, \mathsf{P}_\pi)$ obtained using the policy π and on the step-sizes α_t and β_t, the trajectories of the algorithm closely follow those of an associated ODE with a globally asymptotically stable equilibrium point θ^*. Therefore, the sequence $\{\theta_t\}$ will converge w.p.1 to the equilibrium point θ^* of the ODE.

We now state our main convergence result. Given a MDP $(\mathcal{X}, \mathcal{A}, \mathsf{P}, r, \gamma)$, let π be a stationary stochastic policy and $(\mathcal{X}, \mathsf{P}_\pi)$ the corresponding Markov chain with invariant probability measure

[3] There are several common smooth Dirac approximations, *e.g.*,

$$g_\varepsilon(x; y) = \frac{1}{\varepsilon\sqrt{\pi}} e^{-\|x-y\|^2/\varepsilon^2}.$$

μ_X. Denote by $\mathbb{E}_\pi[\cdot]$ the expectation w.r.t. the probability measure μ_π defined for every set $Z \times U \subset \mathcal{X} \times \mathcal{A}$ as

$$\mu_\pi(Z \times U) = \int_Z \sum_{a \in U} \pi(x,a)\mu_X(dx).$$

Also, define $\hat{\alpha}_t(i)$ as

$$\hat{\alpha}_t(i) = \alpha_t g_{\varepsilon_t}(x_i, a_i; x_t, a_t).$$

Theorem 3.1. *Let $(\mathcal{X}, \mathcal{A}, \mathsf{P}, r, \gamma)$ be a Markov decision process and assume the Markov chain $(\mathcal{X}, \mathsf{P}_\pi)$ to be geometrically ergodic with invariant probability measure μ_X. Suppose that $\pi(x,a) > 0$ for all $a \in \mathcal{A}$ and μ_X-almost all $x \in \mathcal{X}$.*

Let $\Xi = \{\xi_i, i = 1, \ldots, M\}$ be a set of M functions defined on $\mathcal{X} \times \mathcal{A}$ and taking values in $\mathbb{R}$. In particular, admit the functions in Ξ to verify $\|\xi_i\|_\infty = 1$ and $\sum_i |\xi_i(x,a)| \leq 1$.

Then, the following hold:

1. ***Convergence:*** *For any initial condition $\theta_0 \in \mathbb{R}^M$, the algorithm*

$$\theta_{t+1}(i) = \theta_t(i) + \alpha_t g_{\varepsilon_t}(x_i, a_i; x_t, a_t)\big[r_t + \gamma \max_{u \in \mathcal{A}} \xi^\top(x_{t+1}, u)\theta_t - \xi^\top(x_t, a_t)\theta_t\big], \tag{3.6a}$$

$$\varepsilon_{t+1} = (1 - \beta_t)\varepsilon_t. \tag{3.6b}$$

converges w.p.1 as long as the step-size sequences $\{\alpha_t\}, \{\beta_t\}$ are such that

$$\sum_t \alpha_t = \infty; \qquad \sum_t \alpha_t^2 < \infty; \tag{3.7a}$$

$$\sum_t \beta_t = \infty; \qquad \sum_t \beta_t^2 < \infty, \tag{3.7b}$$

$\beta_t = \mathrm{o}(\alpha_t)$ and α_t is built so that $\min_i \sum_t \hat{\alpha}_t(i) = \infty$.

2. ***Limit of convergence:*** *Under these conditions, the limit function $Q(\theta^*)$ of (3.6) verifies*

$$Q_{\theta^*}(x,a) = (\mathcal{P}\mathbf{H}Q_{\theta^*})(x,a), \tag{3.8}$$

where $\mathcal{P} : \mathcal{B} \to \mathcal{Q}$ denotes the operator given by

$$(\mathcal{P}Q)(x,a) = \xi^\top(x,a)\wp Q.$$

3. ***Error bounds:*** *Under these conditions, the limit function Q_{θ^*} verifies the bound*

$$\|Q(\theta^*) - Q^*\|_\infty \leq \frac{1}{1-\gamma}\|\mathcal{P}Q^* - Q^*\|_\infty. \tag{3.9}$$

Proof See Appendix A. □

3.3 Discussion

Before concluding this section, we briefly discuss the conditions of Theorem 3.1 and compare our results with several related works in the literature.

3.3.1 Convergence conditions:

In Theorem 3.1 we identified several conditions that guarantee convergence w.p.1 of the algorithm defined by the update rule in (3.6). These conditions can be classified in two main groups: *conditions on the problem* and *conditions on the algorithm*.

The fundamental condition on the model is that of *geometric ergodicity of the Markov chain* $(\mathcal{X}, \mathsf{P}_\pi)$. Geometric ergodicity ensures that the chain converges exponentially fast to stationarity and, as such, its steady-state behavior is properly captured by the sample trajectories used in the

updates. The analysis of convergence of our algorithm can be conducted in terms of a stationary "version" of it.

Moreover, geometric ergodicity also ensures that all "interesting" regions of the state-space are visited infinitely often [26].[4] The condition that $\pi(x,a) > 0$ for all $a \in \mathcal{A}$ and μ_X-almost every $x \in \mathcal{X}$ ensures that, in these "interesting" regions of the state-space, every action is tried infinitely often. Therefore, geometric ergodicity and the requirement that $\pi(x,a) > 0$ for all $a \in \mathcal{A}$ and μ_X-almost all $x \in \mathcal{X}$ can be interpreted as a continuous counterpart to the usual condition that all state-action pairs are visited infinitely often.

The conditions on the algorithm are those concerning the basis functions used and those concerning the step-size sequences ($\{\alpha_t\}$ and $\{\beta_t\}$). With respect to the former, we require that the functions are linearly independent. This is a simple way of guaranteeing (in a rather conservative way) that no two functions ξ_i lead to "colliding updates" as happens in the known counter-example presented by Baird [3]. Furthermore, by requiring that $\sum |\xi_i(x,a)| \leq 1$ for all $(x,a) \in \mathcal{X} \times \mathcal{A}$, we ensure that $\|Q(\theta)\|_\infty \leq \|\theta\|_\infty$, thus making $\mathbf{H}Q(\theta)$ a contraction in θ (in the sup-norm). This fact plays an important role in establishing our main convergence result.

To clarify the conditions on the step-size sequences, we start by remarking that, if ε is held fixed, the algorithm will converge to a *neighborhood of the desired point in parameter space*. We could then proceed as follows. As soon as the estimates were "sufficiently close" to this neighborhood, we could decrease ε and wait for the estimates to, once again, approach a new, smaller neighborhood of the desired point. We would then decrease ε once again, etc.

This "gross" version of our algorithm illustrates the fact that ε cannot go to zero arbitrarily fast. In particular, it is necessary to ensure that each component of the estimate vector θ_t is "sufficiently" updated as ε is decreased. This clearly depends on the smooth Dirac approximation chosen. The relation between the two referred entities (g_ε and the rate of convergence of ε_t) is formalized in the relations (3.7) and the conditions on α_t.

Such condition on the step-sizes $\{\alpha_t\}$ can be ensured in different ways (for example, defining α_t from the c-cuts of g_ε as in [37]). As one final note, we remark that the use of "broader" Dirac approximations will probably allow faster convergence of ε_t while "narrower" Dirac approximations will probably lead to slower convergence of ε_t.

Finally, since the space $\mathcal{B}$ of essentially bounded functions with the sup-norm is a Banach space (with no orthogonal projection defined), we defined a projection operator $\mathcal{P}$ that is non-expansive in the sup-norm, thus making the combined operator $\mathcal{P}\mathbf{H}$ a contraction in this norm.

3.3.2 Related work:

The early works by Gordon [17] and Tsitsiklis and Van Roy [40] provide convergence analysis for several RL methods using function approximation. The two referred papers portray similar results, although with a slightly different setting and focus on variations of dynamic programming using function approximation. There is also a brief discussion on how stochastic variations of these algorithms (closer in spirit to the Q-learning algorithm) can be used. These stochastic variations are essentially equivalent to the Q-learning algorithm with soft-state aggregation portrayed by Singh et al. [33], as pointed out by Bertsekas and Tsitsiklis [6].

Soft-state aggregation is extensively studied in [33]. In this work, the authors propose the use of a "soft"-partition of the state-space (each state x belongs to region i with a probability $p_i(x)$) and an "average" Q-value $Q(i,a)$ is defined for each region-action pair. Each of these regions is then treated as a "hyper-state" and the method uses standard Q-learning updates to determine the average Q-values for each region. The function Q^* is then approximated for a state-action pair (x,a) as $Q^*(x,a) \approx \sum_i p_i(x)Q(i,a)$.

In a different work, Tsitsiklis and Van Roy [41] provide a detailed analysis of temporal difference methods for policy evaluation. Given a stationary policy π whose value function V^π is to be estimated, a parameterized linear family $\mathcal{V}$ of functions is used to approximate V^π. The authors

[4] In this context, "interesting" regions are those with positive μ_X measure.

analyze the sequence of parameters $\{\theta_t\}$ obtained with the update rule

$$\theta_{t+1} = \theta_t + \alpha_t \xi(x_t)\big(r_t + \gamma V_{\theta_t}(x_{t+1}) - V_{\theta_t}(x_t)\big) \tag{3.10}$$

and establish that the trajectories of this sequence closely follow those of an associated globally asymptotically stable ODE.[5] This implies that $\{\theta_t\}$ converges w.p.1 to the unique equilibrium point of such ODE. The authors provide an interpretation of the obtained limit point as a fixed point of a composite operator $\mathcal{P}\mathbf{T}^{(\lambda)}$, where $\mathcal{P}$ is the orthogonal projection into $\mathcal{V}$ and $\mathbf{T}^{(\lambda)}$ is the TD operator. The authors also provide error bounds on the obtained approximation. Several authors later proposed variations of the fundamental method [7, 12, 13] that include off-policy policy evaluation algorithms [30].

Szepesvári and Smart [37] proposed a version of Q-learning that approximates the optimal Q-values at a given set of sample points $\{(x_i, a_i), i = 1, \ldots, N\}$ and then uses interpolation to estimate Q^* at any query point. This method, dubbed *interpolation-based Q-learning* (IBQL) uses the update rule

$$\theta_{t+1}(i) = \theta_t(i) + \alpha_t(i) g_\varepsilon(x_i, a_i; x_t, a_t)\big(r_t + \max_{u \in \mathcal{A}} Q_{\theta_t}(x_{t+1}, b) - \theta_t(i)\big). \tag{3.11}$$

This update rule uses a *spreading function* g_ε as in multi-state Q-learning [31]. The authors establish convergence w.p.1 of the algorithm and provide an interpretation of the limit point as the fixed-point of a composite operator $\mathcal{P}\hat{\mathbf{H}}$, where $\mathcal{P}$ is a projection-like operator and $\hat{\mathbf{H}}$ can be interpreted as a modified Bellman operator.

We emphasize the similarity between the update rules in (3.11) and (3.6). The fundamental difference between these two methods lies on the fact that IBQL only makes use of the estimated Q-function to predict the value of the next state, as seen in (3.11). Therefore, the updates of IBQL rely on a vector $\hat{d}_t$ of modified temporal differences with ith component given by

$$\begin{aligned}\hat{d}_t(i) &= r_t + \gamma \max_{u \in \mathcal{A}} Q_{\theta_t}(x_{t+1}, u) - \theta_t(i) = \\ &= r_t + \gamma \max_{u \in \mathcal{A}} Q_{\theta_t}(x_{t+1}, u) - Q_{\theta_t}(x_i, a_i).\end{aligned}$$

Notice that each $\hat{d}_t(i)$ is not a temporal-difference in the strict sense, since it does not provide a one-step estimation "error". This means that the information provided by $\hat{d}_t(i)$ may lead to "misleading" updates. Although not affecting the convergence of IBQL in the long-run, IBQL may exhibit slower convergence because of this. On the other hand, if IBQL is used with a vanishing ε, the effect of these misleading updates will vanish as $t \to \infty$. In the experimental results portrayed by Szepesvári and Smart [37], a vanishing ε was used. Nevertheless, IBQL exhibited initially slower convergence than of other methods, probably because of this reported effect.

We also remark that, in [37], the convergence result requires the underlying Markov chain to be positive Harris and aperiodic. These conditions are actually weaker than the geometric ergodicity required by our result. However, in many practical situations, the former conditions will actually imply the latter.[6] This means that the conditions on the problem required in Theorem 3.1 are essentially similar to those in [37] placing the results of both papers in a common line of work and, basically, leading to concordant conclusions.

Finally, we also refer the close relation between the method in Subsection 3.2 and the algorithm described by Borkar [8]. In the aforementioned work, Borkar provides a convergence analysis of what we may refer to as *functional Q-learning*. This functional Q-learning can be seen as an extension of classical Q-learning to functional spaces, and arises from the approach proposed by Baker [4] to stochastic approximation in function spaces. The update equation for this method is fundamentally similar to (3.6). The main difference is that, while we consider only a fixed, finite set of points $I = \{(x_1, a_1), \ldots, (x_M, a_M)\}$, the algorithm by Borkar [8] maintains a *complete representation of*

[5] Actually, the update rule featured in [41] is more general than the one in (3.10), as it includes the use of *eligibility traces* to speed convergence. We considered the simpler version in (3.10) to ease the presentation.

[6] An aperiodic, positive Harris chain is geometrically ergodic as long as supp μ_X has non-empty interior.

Q^*, each component of which is updated at each iteration. Clearly, maintaining such a representation of Q^* is computationally impossible. Therefore, the algorithm by Borkar [8] boils down to maintaining a complete record of the history of past events $\mathcal{H} = \{(x_0, a_0), \ldots, (x_t, a_t), \ldots\}$ and of the estimates Q_t at each of these points. Then, the value of Q^* at a generic point $(x, a) \in \mathcal{X} \times \mathcal{A}$ is estimated as

$$Q_{t+1}(x,a) = Q_0(x,a) + \sum_{k=0}^{t} \alpha_k g_{\varepsilon_k}(x_k, a_k; x, a)\big[r_k + \gamma \max_{u \in \mathcal{A}} Q_t(x_{k+1}, u) - Q_t(x_k, a_k)\big].$$

4 Partially observable Markov decision processes

Recall that, in a Markov decision process $(\mathcal{X}, \mathcal{A}, \mathsf{P}, r, \gamma)$, an agent acts at each time instant based on the current state of the environment and so as to maximize its expected total discounted reward. However, if the current state is unknown and the agent has available only a noisy observation of it, the elegant theory and effective algorithms developed for Markov decision processes are in general not applicable, even in the simpler case of finite $\mathcal{X}$.

Partially observable Markov decision processes (POMDPs) present a complex challenge due to the remarkable complications arising from the "simple" consideration of partial state observability. Exact solution methods for POMDPs generally consist on dynamic-programming based iterative procedures and have been found computationally too expensive for systems with more than a few dozen states [24, 28]. This has led many researchers to focus on developing approximate methods using a variety of approaches. We refer to [1, 14] for good surveys on POMDP exact and approximate methods.

Some approximate solution methods rely on value-based reinforcement learning algorithms such as Q-learning. Examples include the Linear-Q algorithm [23], the SPOVA-RL algorithm [29] or the Fast-RL algorithm [18]. A thorough analysis of several such methods can also be found in [14].

In this section we discuss how our results from the previous section can be applied to POMDPs. We identify a set of conditions on POMDPs that ensure the applicability of the method in Section 3. As a side-note, we remark that the Linear-Q algorithm referred above can be cast as a simple variation of the method described in Section 3. Our analysis in this section can easily be adapted to provide a formal proof of the convergence of this algorithm.

4.1 Partial observability and internal state

Let $(\mathcal{X}, \mathsf{P})$ be a finite state-space Markov chain. Let $\mathcal{Z}$ be a finite set of possible observations and suppose that, at each time instant, the the state X_t of the chain is unaccessible. Instead, a random measurement Z_t is "observed" which depends on the state X_t according to an observation probability given by

$$\mathbb{P}\left[Z_t = z \mid X_t = x\right] = \mathsf{O}(x, z), \tag{4.1}$$

A partially observable Markov chain is a 4-tuple $(\mathcal{X}, \mathcal{Z}, \mathsf{P}, \mathsf{O})$, where $\mathcal{X}$ and $\mathcal{Z}$ are, respectively, the state and observation spaces (both considered finite) and P and O are the transition and observation probability matrices.

Let b_t be a discrete probability measure on $\mathcal{X}$ conveying the probability distribution of the state X_t over the set $\mathcal{X}$ at time instant t. Since $\mathcal{X}$ is assumed finite, b_t is a vector with xth component

$$b_t(x) = \mathbb{P}\left[X_t = x \mid \mathcal{F}_t\right], \tag{4.2}$$

where $\mathcal{F}_t$ is the history up to time t. Suppose that at time instant t the chain is in state $x \in \mathcal{X}$ with probability $b_t(x)$ and a transition occurs, with an observation $Z_{t+1} = z$ made at instant $t+1$. Then it holds that

$$b_{t+1}(y) = \frac{\sum_{x \in \mathcal{X}} b_t(x)\mathsf{P}(x,y)\mathsf{O}(y,z)}{\sum_{x,w \in \mathcal{X}} b_t(x)\mathsf{P}(x,w)\mathsf{O}(w,x)}. \tag{4.3}$$

It is clear from (4.3) that b_{t+1} *is Markovian in its dependence of the past history.* Therefore, we define from b_t a sequence $\{B_t\}$ of random variables, each taking the value $B_t = b_t$ at time instant

t. Since each b_t is a probability vector with, say, n components, B_t lies in the n-dimensional probability simplex $\mathbb{S}^n$.

Summarizing, for any partially observable Markov chain $(\mathcal{X}, \mathcal{Z}, \mathsf{P}, \mathsf{O})$ there is an equivalent fully-observable Markov chain $(\mathbb{S}^n, \hat{\mathsf{P}})$, where the kernel $\hat{\mathsf{P}}$ is given, for any $b \in \mathbb{S}^n$ and any measurable set $U \subset \mathbb{S}^n$, by

$$\hat{\mathsf{P}}(b, U) = \sum_z \sum_{x,y} b(x)\mathsf{P}(x, y)\mathsf{O}(y, z)\mathbb{I}_U(B(b, z)),$$

where $B(b, z)$ is the vector obtained from b using (4.3) with observation z and $\mathbb{I}_U$ is the indicator function for the set U. Notice that the xth coordinate of vector B_t describes the *belief* that the underlying state of the chain is $X_t = x$, and it is common to refer to the b vectors as *belief-states*.

Notice that, by considering the chain $(\mathbb{S}^n, \hat{\mathsf{P}})$ of beliefs instead of the partially observable chain $(\mathcal{X}, \mathcal{Z}, \mathsf{P}, \mathsf{O})$ we move from a finite, partially observable Markov chain with state-space $\mathcal{X}$ to an infinite, fully observable Markov chain with state-space $\mathbb{S}^n$. We now identify conditions on P and/or O that ensure the chain $(\mathbb{S}^n, \hat{\mathsf{P}})$ to be uniformly ergodic.

Theorem 4.1. *Let $(\mathcal{X}, \mathcal{Z}, \mathsf{P}, \mathsf{O})$ be a partially observable Markov chain, where the chain $(\mathcal{X}, \mathsf{P})$ is irreducible and aperiodic. Suppose that there is an observation $z \in \mathcal{Z}$ and a state $x^* \in \mathcal{X}$ such that, for all $y \in \mathcal{X}$, $\mathsf{O}(y, z) = \delta(x^*, y)$. Then, the Markov chain $(\mathbb{S}^n, \hat{\mathsf{P}})$ is uniformly ergodic.*

Proof See Section B. □

4.2 POMDPs and associated MDPs

A tuple $(\mathcal{X}, \mathcal{A}, \mathcal{Z}, \mathsf{P}, \mathsf{O}, r, \gamma)$ is a *partially Observable Markov Decision Process* (POMDP), where $\mathcal{X}, \mathcal{A}, \mathsf{P}, r$ and γ are as defined in Section 2, $\mathcal{Z}$ is the observation-space and O represents the (action-dependent) observation probabilities. We consider $\mathcal{X}$, $\mathcal{A}$ and $\mathcal{Z}$ to be finite sets.

Using a development entirely similar to the one presented in the previous subsection, given a POMDP $(\mathcal{X}, \mathcal{A}, \mathcal{Z}, \mathsf{P}, \mathsf{O}, r, \gamma)$ we can derive a fully observable MDP $(\mathbb{S}^n, \mathcal{A}, \hat{\mathsf{P}}, \hat{r}, \gamma)$, where, for each $a \in \mathcal{A}$, $\hat{\mathsf{P}}$ and $\hat{r}$ are defined as

$$\hat{\mathsf{P}}_a(b, U) = \sum_z \sum_{x,y} b(x)\mathsf{P}_a(x, y)\mathsf{O}_a(y, z)\mathbb{I}_U(B(b, a, z));$$

$$\hat{r}(b, a, b') = \sum_{x,y} b(x)\mathsf{P}_a(x, y)r(x, a, y),$$

where $B(b, a, z)$ is the updated probability vector given action a and observation z with yth component given by

$$B(b, a, z)_y = \frac{\sum_{x \in \mathcal{X}} b_t(x)\mathsf{P}_a(x, y)\mathsf{O}_a(y, z)}{\sum_{x,w \in \mathcal{X}} b_t(x)\mathsf{P}_a(x, w)\mathsf{O}_a(w, x)}.$$

Notice that the reward $\hat{r}(b, a, b')$ corresponds to the expected immediate reward for being in each state x with probability $b(x)$ and taking action a. As expected, it does not depend on b'.[7]

This new MDP is an infinite state-space counterpart to the partially observable Markov decision process $(\mathcal{X}, \mathcal{A}, \mathcal{Z}, \mathsf{P}, \mathsf{O}, r, \gamma)$ and we are interested in applying the methods from the previous section to this continuous-state MDP.

Notice that, even if the complete POMDP model is known, the use of a simulation-based solution may still be preferable to the computationally heavier, exact methods. On the other hand, it may happen that the reward r is unknown and, therefore, recurring to simulation-based methods is the only alternative available. Finally, we emphasize that, in order to use the methods from the previous section, the MDP $(\mathbb{S}^n, \mathcal{A}, \hat{\mathsf{P}}, \hat{r}, \gamma)$ needs to be fully observable, *i.e.*, the beliefs b_t must be computable at every time step t. This means that the agent must know the model parameters P and O.

[7] Notice that the rewards do not depend on the observations and the belief b' is a function of the current belief, action and observation, so it is natural that $\hat{r}$ is independent of b'.

In the new MDP $(\mathbb{S}^n, \mathcal{A}, \hat{\mathsf{P}}, \hat{r}, \gamma)$, it is straightforward to define the optimal value function $V^* : \mathbb{S}^n \longrightarrow \mathbb{R}$, verifying

$$V^*(b) = \max_{a \in \mathcal{A}} \mathbb{E}\left[\hat{r}(b, a, b') + \gamma V^{\delta}(b')\right],$$

and the optimal Q-function, verifying

$$Q^*(b, a) = \mathbb{E}\left[r(b, a, b') + \gamma \max_{u \in \mathcal{A}} Q^*(b', u)\right].$$

More intuitive and well-known expressions for these functions can readily be obtained by replacing $\hat{\mathsf{P}}$ and $\hat{r}$ by the corresponding definitions, yielding

$$V^*(b) = \max_{a \in \mathcal{A}} \sum_{x,y \in \mathcal{X}} b(x)\mathsf{P}_a(x, y)\left[r(x, a, y) + \gamma \sum_{z \in \mathcal{Z}} \mathsf{O}_a(y, z)V^*(b_z)\right];$$

$$Q^*(b, a) = \sum_{x,y \in \mathcal{X}} b(x)\mathsf{P}_a(x, y)\left[r(x, a, y) + \gamma \sum_{z \in \mathcal{Z}} \mathsf{O}_a(y, z) \max_{b \in \mathcal{A}} Q^*(b_z, b)\right].$$

To apply the method from Section 3 to the MDP $M = (\mathbb{S}^n, \mathcal{A}, \hat{\mathsf{P}}, \hat{r}, \gamma)$ with guaranteed convergence, we need to check if M verifies all conditions *on the problem* required in Theorem 3.1. This condition is concerned with the geometric ergodicity of the chain obtained with the learning policy. Combining Theorem 3.1 with Theorem (4.1), it is immediate that the Q-learning algorithm with linear function approximation analyzed in Section 3 can be applied to POMDPs with guaranteed convergence, as long as the underlying MDP is ergodic and there is a *distinguishable* state $x^* \in \mathcal{X}$. We note that ergodicity of the underlying MDP is a standard assumption in classical RL methods and, therefore, partial observability simply requires the single additional condition of a distinguishable state.

5 Conclusions and future work

In this paper we have analyzed the convergence of Q-learning with linear function approximation. Given a linear family $\mathcal{Q}$ of functions, we defined an update rule that "relies" on a projection operator $\mathcal{P}$ defined in the space of (essentially) bounded functions. For the algorithm thus obtained we identified the conditions under which convergence w.p.1 is guaranteed. We also showed the limit function to verify the fixed-point recursion

$$Q_{\theta^*}(x, a) = (\mathcal{P}\mathbf{H}Q_{\theta^*})(x, a)$$

and discussed the relation between the method and results in this paper and those in related works such as [8, 37]. Finally, we showed that partially observable Markov decision processes can be addressed by reinforcement learning algorithms using function approximation as long as the typical convergence conditions are verified for the underlying Markov decision process and there is, at least, one observable state.

Several important remarks are in order. First of all, the error bound in Theorem 3.1 is given as a function of the quantity $\|\mathcal{P}Q^* - Q^*\|$. Notice that the function $\mathcal{P}Q^*$ can be interpreted as the "best" representation of Q^* in $\mathcal{Q}$. The error bound in Theorem 3.1 means that the obtained approximation is, at most, "almost as good" as $\mathcal{P}Q^*$. It also means that, this approximation may be of little use, if the space $\mathcal{Q}$ poorly represents the desired function: the closest function in $\mathcal{Q}$ will still be a poor approximation, and there are no guarantees on its practical usefulness (in terms of the corresponding greedy policy). Notice nevertheless that, if $Q^* \in \mathcal{Q}$, the method will deliver the optimal function Q^*. Therefore, when using function approximation, the space $\mathcal{Q}$ should be chosen so as to include all available information regarding the true function to be estimated. The problem of how to choose the basis functions is currently the target of intense research in the RL community. Some work has been done in this area [16, 22, 25, 42], but a lot more can be done.

A second remark concerns the usefulness of the algorithm in Section 3 if a fixed policy must be used during learning (instead of a policy that depends on the estimates Q_t). Although the result described in the paper considers a fixed learning policy, it is possible to extend this result to encompass the use of a policy π_θ that depends continuously on θ. In particular, if the following condition holds for every $(x, a) \in \mathcal{X} \times \mathcal{A}$

$$|\pi_\theta(x, a) - \pi_{\theta'}(x, a)| \leq C \|\theta - \theta'\|,$$

with $C > 0$, it is possible to extend the conclusions of Theorem 3.1 to algorithms using θ-dependent policies. Further work can explore results on the stability of perturbed ODEs to extend the fundamental ideas in this paper to address the convergence of on-policy learning algorithm (*e.g.*, SARSA).

Also, the methods proposed make no use of eligibility traces. It seems likely that the results in this paper can be modified so as to accommodate eligibility traces and thus improve their overall performance.

Thirdly, we comment on the results presented in Section 3. In this section, we described the use of the algorithm in Section 3 to POMDPs by considering equivalent, fully observable MDPs. Recall that tracking the state of an associated MDP consists in tracking the belief-state b_t of the original POMDP. As already stated, this implies that the agent must know the parameters P and O of the POMDP. This is less general than the approach adopted in many RL methods, where no model of the system is assumed. However, in several practical applications (*e.g.*, robotic applications) this is a reasonable assumption.

Finally, notice that the overall conditions required to ensure convergence of the methods in Section 3 in partially observable scenarios are similar to the requirements for convergence in fully observable scenarios. Convergence in partially observable scenarios simply requires one extra condition: that at least one state is identifiable. If we consider that, in many situations, *the reinforcement function provides additional information on the underlying state of the system*, the existence of a distinguishable state may be a less stringent condition than it appears at first sight. Nevertheless, it is likely that results on the ergodic behavior of the posterior probabilities of hidden Markov models may be adapted so as to alleviate this condition.

Acknowledgements

The authors would like to acknowledge the helpful discussions with Prof. João Xavier and the useful comments from the anonymous reviewers that helped to greatly improve the paper.

References

[1] Douglas A. Aberdeen. A (revised) survey of approximate methods for solving partially observable Markov decision processes. Technical report, National ICT Australia, Canberra, Australia, 2003.

[2] Christopher G. Atkeson, Andrew W. Moore, and Stefan Schaal. Locally weighted learning for control. *Artificial Intelligence Review*, 11(1-5):75–113, 1997.

[3] Leemon C. Baird. Residual algorithms: Reinforcement learning with function approximation. In *Proceedings of the 12th International Conference on Machine Learning (ICML'95)*, pages 30–37, San Francisco, CA, 1995. Morgan Kaufman Publishers.

[4] W. Baker. Learning via stochastic approximation in function space. PhD Thesis, 1997.

[5] Andrew G. Barto, Steven J. Bradtke, and Satinder P. Singh. Learning to act using real-time dynamic programming. Technical Report UM-CS-1993-002, Department of Computer Science, University of Massachusetts at Amherst, 1993.

[6] Dimitri P. Bertsekas and John N. Tsitsiklis. *Neuro-Dynamic Programming*. Optimization and Neural Computation Series. Athena Scientific, Belmont, Massachusetts, 1996.

[7] Dimitri P. Bertsekas, Vivek S. Borkar, and Angelia Nedić. *Improved temporal difference methods with linear function approximation*, chapter 9, pages 235–260. Wiley Publishers, 2004.

[8] Vivek S. Borkar. A learning algorithm for discrete-time stochastic control. *Probability in the Engineering and Informational Sciences*, 14:243–258, 2000.

[9] Vivek S. Borkar. Stochastic approximation with two time scales. *Systems & Control Letters*, 29(5): 291–294, 1997.

[10] Vivek S. Borkar and Sean P. Meyn. The O.D.E. method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization*, 38(2):447–469, 2000.

[11] Justin Boyan and Andrew Moore. Generalization in reinforcement learning: Safely approximating the value function. In G. Tesauro, D.S. Touretzky, and T.K. Lee, editors, *Neural Information Processing Systems 7*, pages 369–376, Cambridge, MA, 1995. The MIT Press.

[12] Justin A. Boyan. Least-squares temporal difference learning. In *Proceedings of the 16th International Conference on Machine Learning (ICML'99)*, pages 49–56, San Francisco, CA, 1999. Morgan Kaufmann.

[13] Justin A. Boyan. Technical update: Least-squares temporal difference learning. *Machine Learning*, 49:233–246, 2002.

[14] Anthony R. Cassandra. *Exact and approximate algorithms for partially observable Markov decision processes*. PhD thesis, Brown University, May 1998.

[15] Bernard Delyon. General results on the convergence of stochastic algorithms. *IEEE Transactions on Automatic Control*, AC-41(9):1245–1256, 1996.

[16] Robert Glaubius and William D. Smart. Manifold representations for value-function approximation in reinforcement learning. Technical Report 05-19, Department of Computer Science and Engineering, Washington University in St. Louis, 2005.

[17] Geoffrey J. Gordon. Stable function approximation in dynamic programming. Technical Report CMU-CS-95-103, School of Computer Science, Carnegie Mellon University, 1995.

[18] Qiming He and Mark A. Shayman. Solving POMDPs by on-policy linear approximate learning algorithm. In *Proceedings of the Conference on Information Sciences and Systems*. Princeton University, 2000.

[19] Morris W. Hirsch. Convergent activation dynamics in continuous time networks. *Neural Networks*, 2: 331–349, 1989.

[20] Tommi Jaakkola, Michael I. Jordan, and Satinder P. Singh. On the convergence of stochastic iterative dynamic programming algorithms. *Neural Computation*, 6(6):1185–1201, 1994.

[21] Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. In *Proceedings of the 15th International Conference on Machine Learning (ICML'98)*, pages 260–268, San Francisco, CA, 1998. Morgan Kaufmann Publishers.

[22] Philipp W. Keller, Shie Mannor, and Doina Precup. Automatic basis function construction for approximate dynamic programming and reinforcement learning. In *Proceedings of the 23rd International Conference on Machine Learning (ICML'06)*, pages 449–456, New York, NY, 2006. ACM Press.

[23] Michael L. Littman, Anthony R. Cassandra, and Leslie P. Kaelbling. Learning policies for partially observable environments: Scaling up. In Armand Prieditis and Stuart Russell, editors, *Proceedings of the 12th International Conference on Machine Learning (ICML'95)*, pages 362–370, San Francisco, CA, 1995. Morgan Kaufmann Publishers.

[24] Christopher Lusena, Judy Goldsmith, and Martin Mundhenk. Nonapproximability results for partially observable Markov decision processes. *Journal of Artificial Intelligence Research*, 14:83–103, 2001.

[25] Ishai Menache, Shie Mannor, and Nahum Shimkin. Basis function adaptation in temporal difference reinforcement learning. *Annals of Operations Research*, 134(1):215–238, February 2005.

[26] Sean P. Meyn and Richard L. Tweedie. *Markov Chains and Stochastic Stability*. Communications and Control Engineering Series. Springer-Verlag, New York, 1993.

[27] Dirk Ormoneit and Śaunak Sen. Kernel-based reinforcement learning. *Machine Learning*, 49:161–178, 2002.

[28] Christos H. Papadimitriou and John N. Tsitsiklis. The complexity of Markov chain decision processes. *Mathematics of Operations Research*, 12(3):441–450, 1987.

[29] Ronald Parr and Stuart Russell. Approximating optimal policies for partially observable stochastic domains. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 1088–1094, 1995.

[30] Doina Precup, Richard S. Sutton, and Sanjoy Dasgupta. Off-policy temporal-difference learning with function approximation. In *Proceedings of the 18th International Conference on Machine Learning (ICML'01)*, pages 417–424, San Francisco, CA, 2001. Morgan Kaufmann.

[31] Carlos Ribeiro and Csaba Szepesvári. Q-learning combined with spreading: Convergence and results. In *Proceedings of the ISRF-IEE International Conference: Intelligent and Cognitive Systems (Neural Networks Symposium)*, pages 32–36, 1996.

[32] Gavin A. Rummery and Mahesan Niranjan. On-line Q-learning using connectionist systems. Technical Report CUED/F-INFENG/TR 166, Cambridge University Engineering Department, 1994.

[33] Satinder P. Singh, Tommi Jaakkola, and Michael I. Jordan. Reinforcement learning with soft state aggregation. In *Advances in Neural Information Processing Systems*, volume 7, pages 361–368. 1994.

[34] Richard S. Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 3: 9–44, 1988.

[35] Richard S. Sutton. Generalization in reinforcement learning: Successful examples using sparse coarse coding. *Advances in Neural Information Processing Systems*, 8:1038–1044, 1996.

[36] Richard S. Sutton. DYNA, an integrated architecture for learning, planning, and reacting. *ACM SIGART Bulletin*, 2(4):160–163, 1991.

[37] Csaba Szepesvári and William D. Smart. Interpolation-based Q-learning. In *Proceedings of the 21st International Conference on Machine learning (ICML'04)*, pages 100–107, New York, USA, July 2004. ACM Press.

[38] Gerald Tesauro. TD-Gammon, a self-teaching backgammon program, achieves master-level play. *Neural Computation*, 6(2):215–219, 1994.

[39] Gerald Tesauro. Temporal difference learning and TD-Gammon. *Communications of the ACM*, 38 (3):58–68, 1995.

[40] John N. Tsitsiklis and Benjamin Van Roy. Feature-based methods for large scale dynamic programming. *Machine Learning*, 22:59–94, 1996.

[41] John N. Tsitsiklis and Benjamin Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690, May 1996.

[42] Benjamin Van Roy. *Learning and value function approximation in complex decision processes.* PhD thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, June 1998.

[43] Christopher Watkins and Peter Dayan. Technical note: Q-learning. *Machine Learning*, 8:279–292, 1992.

[44] Christopher J. C. H. Watkins. *Learning from delayed rewards.* PhD thesis, King's College, University of Cambridge, May 1989.

A Proof of Theorem 3.1

We separately establish each of the three assertions of Theorem 3.1. To prove the first assertion, we establish the trajectories $\{\theta_t\}$ generated by algorithm (3.6) to closely follow those of an associated ODE with a globally asymptotically stable equilibrium point. As long as the iterates of the algorithm remain bounded, this will imply the convergence to the equilibrium point of the associated ODE.

To prove the second assertion of Theorem 3.1, we provide an interpretation of the equilibrium point of the associated ODE as the fixed point of a composite operator. This interpretation will then lead to the third assertion of Theorem 3.1.

A.1 Convergence of the iterates

To prove the convergence of the sequence $\{\theta_t\}$ generated by (3.6), we follow a similar argument to that in [8, 9].

Consider a general ODE in $\mathbb{R}^M$,

$$\frac{d}{dt}Z(t) = h(Z(t)), \tag{A.1}$$

for a Lipschitz map $h : \mathbb{R}^p \longrightarrow \mathbb{R}^p$. Further suppose that the ODE (A.1) has a globally asymptotically stable equilibrium Z^*.

Given any $T > 0$ and $\sigma > 0$, a bounded measurable function $z : \mathbb{R}^+ \longrightarrow \mathbb{R}^M$ is a (T, σ)-*perturbation of* (A.1) if there is a sequence T_n of positive real numbers such that $T_0 = 0$, $T_n \to \infty$, with $T_{n+1} - T_n > T$ and the following holds

$$\sup_{t \in [T_n, T_{n+1}]} \|Z^n(t) - z(t)\| \leq \sigma,$$

where $Z^n(t)$ is a solution of (A.1) defined in the interval $[T_n, T_{n+1}]$.

We now introduce the Hirsch lemma [19], whose proof can be found, for example, in [9].

Lemma A.1 (Hirsch Lemma)**.** *Given any $\rho > 0$ and $T > 0$, there is a σ_0 such that, for all $\sigma < \sigma_0$, every (T, σ)-perturbation of* (A.1) *converges to an ρ-neighborhood of Z^*.*

Consider the ODE

$$\frac{d}{dt}\theta_t(i) = (\wp \mathbf{H} Q(\theta_t))_i - \theta_t(i). \tag{A.2}$$

It is not hard to see that this ODE has a globally asymptotically stable equilibrium θ^* verifying

$$\theta^*(i) = (\wp \mathbf{H} Q(\theta^*))_i$$

since, as remarked in Subsection 3.2, $\wp$ is a non-expansion in the sup-norm, $\mathbf{H}$ is a contraction in the sup-norm and $\|\xi_i\|_\infty = 1$ for $i = 1, \ldots, M$.

Consider, on the other hand, the ODE

$$\begin{aligned} \frac{d}{dt}\theta_t^\varepsilon(i) &= h^\varepsilon(\theta_t) = \\ &= \int g_\varepsilon(x_i, a_i; x, a)\big[(\mathbf{H}Q_{\theta_t^\varepsilon})(x, a) - Q_{\theta_t^\varepsilon}(x, a)\big] d\mu_\pi(x, a). \end{aligned} \tag{A.3}$$

By hypothesis, $g_\varepsilon(x_i, a_i; \cdot) \to \delta_{(x_i, a_i)}$ as $\varepsilon \to 0$. By taking $\theta_0^\varepsilon = \theta_0$, a standard argument using the Gronwall inequality leads to the conclusion that the solutions θ_t^ε of (A.3) verify $\theta_t^\varepsilon \to \theta_t$ as $\varepsilon \to 0$, and this convergence holds uniformly in compact time intervals.[8] This implies that, given any $T > 0$ and $\sigma > 0$, θ_t^ε is a $(T, \frac{\sigma}{2})$-perturbation of (A.2) for sufficiently small ε.

[8] In particular, in an interval $[t, t + \tau]$ we have

$$\|\theta_t^\varepsilon - \theta_t\| \leq K_\varepsilon(e^{C\tau} - 1),$$

for some positive constant C and some positive, ε-dependent constant K_ε that goes to 0 with ε.

Our purpose is to establish that, for fixed ε, the trajectories of the algorithm (3.6) closely follow those of (A.3). Thus, for fixed ε, rewrite (3.6) in the form

$$\theta_{t+1} = \theta_t + \alpha_t H^\varepsilon(\theta_t, Y_{t+1}), \tag{A.4}$$

where $Y_{t+1} = (X_t, A_t, X_{t+1})$. Since the chain $\{X_t\}$ is geometrically ergodic and $\pi(x,a) > 0$ for μ_π-almost all $x \in \mathcal{X}$, it follows that so is the chain $\{Y_t\}$.

Using the geometric ergodicity of the chain $\{Y_t\}$ and the Poisson equation involving $H^\varepsilon(\theta, Y)$, we can rewrite (A.4) as

$$\theta_{t+1} = \theta_t + \alpha_t h^\varepsilon(\theta_t) + \alpha_t(e_{t+1} + \eta_{t+1}), \tag{A.5}$$

where, for each $i = 1, \ldots, M$, $\sum_{t=0}^{\infty} \alpha_t e_t(i) < \infty$ and $\|\eta_t\| \to 0$. To see that this is so, recall that the Poisson equation involving $H^\varepsilon(\theta, Y)$ is given by

$$\upsilon_\theta(y) - (\mathsf{P}_\pi \upsilon_\theta)(y) = H^\varepsilon(\theta, y) - h^\varepsilon(\theta),$$

where υ_θ is the solution for a given θ. Under the geometrical ergodicity of $\{Y_t\}$ the solution for this equation always exists [26] and we can rewrite

$$\begin{aligned}
H^\varepsilon(\theta_t, X_{t+1})_i &= h^\varepsilon(\theta_t)_i + \upsilon_{\theta_t}(X_{t+1})_i - (\mathsf{P}_\pi \upsilon_{\theta_t})(X_{t+1})_i = \\
&= h^\varepsilon(\theta_t)_i + \upsilon_{\theta_t}(X_{t+1})_i - (\mathsf{P}_\pi \upsilon_{\theta_t})(X_t)_i + \\
&\quad + (\mathsf{P}_\pi \upsilon_{\theta_t})(X_t)_i - (\mathsf{P}_\pi \upsilon_{\theta_{t+1}})(X_{t+1})_i + \\
&\quad + (\mathsf{P}_\pi \upsilon_{\theta_{t+1}})(X_{t+1})_i - (\mathsf{P}_\pi \upsilon_{\theta_t})(X_{t+1})_i = \\
&= h^\varepsilon(\theta_t)_i + \zeta_{t+1}(i) + (u_t(i) - u_{t+1}(i)) + \eta_{t+1}(i),
\end{aligned}$$

with

$$\begin{aligned}
\zeta_{t+1}(i) &= \upsilon_{\theta_t}(X_{t+1})_i - (\mathsf{P}_\pi \upsilon_{\theta_t})(X_t)_i; \\
u_t(i) &= (\mathsf{P}_\pi \upsilon_{\theta_t})(X_t)_i; \\
u_{t+1}(i) &= (\mathsf{P}_\pi \upsilon_{\theta_{t+1}})(X_{t+1})_i; \\
\eta_{t+1}(i) &= (\mathsf{P}_\pi \upsilon_{\theta_{t+1}})(X_{t+1})_i - (\mathsf{P}_\pi \upsilon_{\theta_t})(X_{t+1})_i.
\end{aligned}$$

Finally, setting $e_{t+1}(i) = \zeta_{t+1}(i) + (u_t(i) - u_{t+1}(i))$ leads to (A.5). The two aforementioned properties of e_t and η_t, namely $\sum_{t=0}^{\infty} \alpha_t e_t(i) < \infty$ and $\|\eta_t\| \to 0$, arise as consequences of the geometric ergodicity assumption on the underlying Markov chain and of the properties of the solution υ_θ of the Poisson equation [15, 26].

We now proceed as in [9]. Define the sequences

$$\begin{aligned}
\tau_0 &= 0; & \tau_k &= \sum_{t=0}^{k-1} \alpha_t; \\
T_0 &= 0; & T_{n+1} &= \min\{\tau_k \mid \tau_k > T_n + T\},
\end{aligned}$$

and let $\{\theta_t\}$ be a sample trajectory obtained using (A.4) with constant ε. From $\{\theta_t\}$ build the continuous-time process $\bar{\theta}^0(t)$ by taking $\bar{\theta}^0(\tau_k) = \theta_k$ and using linear interpolation in the interval $[\tau_k, \tau_{k+1}]$. In particular (A.5) yields

$$\bar{\theta}^0(\tau_{t+1}) = \bar{\theta}^0(\tau_t) + (\tau_{t+1} - \tau_t) h^\varepsilon(\theta^0(\tau_t)) + \rho(\tau_t), \tag{A.6}$$

with

$$\rho(\tau_t) = \alpha_t(e_{t+1} + \eta_{t+1}).$$

By interpreting (A.6) as a discretized version of the ODE (A.3) with noise $\rho(\tau_t)$, we can once again resort to the Gronwall inequality to bound the distance between $\bar{\theta}^0(t)$ and θ_t^ε in each interval $[T_k, T_{k+1}]$ by a constant that can be made arbitrarily small for large enough k. But this means that, for any $\sigma > 0$, the process $\bar{\theta}^{t_0}(t) = \bar{\theta}^0(t + t_0)$ is a $(T, \frac{\sigma}{2})$-perturbation of (A.3) for t_0 large enough.

By combining this conclusion with the previous conclusion on θ_t^ε, we see that, for any $T > 0$ and $\sigma > 0$, $\bar{\theta}^{t_0}(t)$ is a (T, σ)-perturbation of (A.2) (for ε sufficiently small).

Consider now the ODE

$$\frac{d}{dt}\varepsilon_t = 0. \tag{A.7}$$

and the approximate process

$$\varepsilon_{t+1} = \varepsilon_t - \alpha_t \frac{\beta_t}{\alpha_t}\varepsilon_t. \tag{A.8}$$

Repeating the exact same procedure used for $\bar{\theta}_t^0$, and by noticing that $\beta_t = \mathrm{o}(\alpha_t)$, we build a process $\bar{\varepsilon}_t^{t_0}$ such that, for any $\sigma > 0$, $\bar{\varepsilon}_t^{t_0}$ is a (T, σ)-perturbation of (A.7) for t_0 large enough. Finally, since $\sum_t \hat{\alpha}_t(i) = \infty$, this leads to the conclusion that, given any $T > 0$ and $\sigma > 0$, $(\bar{\theta}_t^{t_0}, \bar{\varepsilon}_t^{t_0})$ is a (T, σ)-perturbation of the system of ODEs

$$\frac{d}{dt}\theta_t(i) = (\wp \mathbf{H} Q(\theta_t))_i - \theta_t(i); \tag{A.9a}$$

$$\frac{d}{dt}\varepsilon_t = 0, \tag{A.9b}$$

for t_0 large enough.

We now notice that, in general, a smaller ε will require a larger t_0 to ensure that $\bar{\theta}^{t_0}(t)$ is a (T, σ)-perturbation of (A.3). On the other hand, ε_t as generated by (3.6b) converges to 0. Therefore, to guarantee that there is t_0 such that $(\bar{\theta}_t^{t_0}, \bar{\varepsilon}_t^{t_0})$ is a (T, σ)-perturbation of (A.9), it is necessary to ensure that ε_t approaches zero sufficiently slowly. By choosing the sequence α_t so that $\sum_t \hat{\alpha}_t(i) = \infty$, we ensure that the trajectories $\{\theta_t\}$ approach the trajectories of the ODE (A.3) faster than g_{ε_t} approaches the Dirac delta and are in position to apply the Hirsh lemma to conclude that, for any $\rho > 0$, the process $(\bar{\theta}_t^{t_0}, \bar{\varepsilon}_t^{t_0})$ converges to a ρ-neighborhood of $(\theta^*, 0)$. This, in turn, implies that $\theta_t \to \theta^*$ as long as the sequence $\{\theta_t\}$ remains bounded, which we establish in the continuation.

A.2 Boundedness of the iterates

To establish the boundedness of the iterates, we replicate the procedure by Borkar and Meyn [10].

Let $\{\theta_t\}$ be a trajectory generated by (3.6). We build a scaled sequence $\{\hat{\theta}_t\}$ by setting

$$\hat{\theta}_t = \frac{\theta_t}{\lambda_n}$$

where $\lambda_n = \max\{\|\theta_{T_n}\|, 1\}$, for every t in $[T_n, T_{n+1})$ and T_i are as defined in Subsection A.1. If the sequence θ_t is unbounded, this means that $\limsup \lambda_n = \infty$, so we analyze the behavior of $\hat{\theta}_t$ as $\lambda_n \to \infty$.

In Subsection A.1, we established the trajectories $\{\theta_t\}$ to closely follow those of the ODE

$$\frac{d}{dt}\theta_t = h(\theta_t),$$

where $h(\theta) = \wp \mathbf{H} Q(\theta) - \theta$. For the scaled sequence $\hat{\theta}_t$, we now consider the function $h_\lambda(\theta)$, given by

$$h_\lambda(\theta) = \frac{h(\lambda\theta)}{\lambda},$$

with $\lambda > 0$. Notice that, as $\lambda \to \infty$, h_λ approaches the function h_∞ given by

$$h_\infty(\theta)_i = \gamma \max_{b \in \mathcal{A}} \int \xi^\top(y, b)\theta \mathsf{P}_{a_i}(x_i, dy) - \theta_i.$$

Define the operator $\mathbf{F} : \mathbb{R}^M \longrightarrow \mathbb{R}^M$ with ith component given by

$$\mathbf{F}(\theta)_i = \gamma \max_{b \in \mathcal{A}} \int \xi^\top(y, b)\theta \mathsf{P}_{a_i}(x_i, dy).$$

This operator is a contraction in the sup-norm (due to the fact that $\sum_i |\xi_i(x,a)| \leq 1$) and, hence has a single fixed point. Since the origin is a fixed point of $\mathbf{F}$, the ODE associated with h_∞ has a single equilibrium point at the origin and this equilibrium point is globally asymptotically (exponentially) stable.

By repeating the procedure used in Subsection A.1, we can build by interpolation a continuous time process from the scaled sequence $\hat{\theta}_t$. In each interval $[T_n, T_{n+1}]$, this continuous-time process is a (T,σ)-perturbation of the ODE,

$$\frac{d}{dt}\theta_t = h_{\lambda_n}(\theta_t), \tag{A.10}$$

for any $T > 0$ and any $\sigma > 0$ (by eventually considering a time-shifted version of the continuous time process, as in Subsection A.1). This implies the boundedness of θ_t as a consequence.

In fact, suppose that $\{\theta_t\}$ is not bounded. This implies that $\lambda_n \to \infty$ eventually along a subsequence. Since the solutions of (A.10) converge exponentially fast to an arbitrarily small neighborhood of the origin (depending on λ_n), by taking n large enough we can ensure that $\left\|\hat{\theta}_{T_{n+1}}\right\| \leq C$ for any $C < 1$. But this implies that

$$\frac{\left\|\theta_{T_{n+1}}\right\|}{\|\theta_{T_n}\|} \leq C$$

or, equivalently, $\left\|\theta_{T_{n+1}}\right\| \leq C \|\theta_{T_n}\|$. Therefore, whenever θ_t leaves, say, the unit ball in $\mathbb{R}^M$, it returns exponentially fast toward it, and θ_t remains bounded.

We refer to [8, 10], where a similar process is apllied to establish boundedness of an iterative process.

A.3 Limit of convergence and error bounds

We have established that the sequence $\{\theta_t\}$ generated by (3.6) converges w.p.1 to a point θ^*. The limit point θ^* is the globally asymptotically stable equilibrium of the ODE (A.2), verifying the following recursive relation:

$$\theta^* = \wp\mathbf{H}Q(\theta^*).$$

This provides an interpretation for the limit point of $\{\theta_t\}$ as the fixed point of the combined operator $\wp\mathbf{H}Q(\cdot)$, where Q is now understood as a mapping from $\mathbb{R}^M$ to $\mathcal{B}$.

To conclude the proof of Theorem 3.1, it remains to establish statement 3, thus providing the error bounds for the approximation. To this, we perform some explicit computations, yielding

$$\begin{aligned}
\|Q(\theta^*) - Q^*\|_\infty &= \|Q(\theta^*) - Q(\wp Q^*) + Q(\wp Q^*) - Q^*\|_\infty \leq \\
&\leq \|Q(\theta^*) - Q(\wp Q^*)\|_\infty + \|Q(\wp Q^*) - Q^*\|_\infty = \\
&= \|Q(\theta^* - \wp Q^*)\|_\infty + \|Q(\wp Q^*) - Q^*\|_\infty = \\
&= \|Q(\wp\mathbf{H}Q(\theta^*) - \wp\mathbf{H}Q^*)\|_\infty + \|Q(\wp Q^*) - Q^*\|_\infty
\end{aligned}$$

Using the fact that $\|\xi_i\| = 1$, we get

$$\begin{aligned}
&\|Q(\theta^*) - Q^*\|_\infty \leq \\
&\quad = \|\wp\mathbf{H}Q(\theta^*) - \wp\mathbf{H}Q^*\|_\infty + \|Q(\wp Q^*) - Q^*\|_\infty \leq \\
&\quad = \gamma \|Q(\theta^*) - Q^*\|_\infty + \|Q(\wp Q^*) - Q^*\|_\infty ,
\end{aligned}$$

and this finally leads to

$$\|Q(\theta^*) - Q^*\|_\infty \leq \frac{1}{1-\gamma} \|Q(\wp Q^*) - Q^*\|_\infty .$$

This concludes the proof of Theorem 3.1.

B Proof of Theorem 4.1

Before embarking in the proof of Theorem 4.1, we briefly survey some fundamental concepts on Markov chains. The proofs of all results stated can be found in [26].

B.1 Markov chains

We start by reviewing the concept *stability* in the context of Markov chains. We refer to the book by Meyn and Tweedie [26] for a more detailed treatment and formal proofs of the statements presented here.

A *homogeneous Markov chain* is a discrete-time stochastic process $\{X_t\}$ defined by a pair $(\mathcal{X}, \mathsf{P})$, where $\mathcal{X}$ is the state-space and P is a transition probability kernel defining the transition probabilities

$$\mathsf{P}(x, U) = \mathbb{P}\left[X_t \in U \mid X_{t-1} = x\right],$$

which are independent of the particular time instant t considered. The kernels P^m and K_a, defined as

$$\mathsf{P}^m(x, U) = \int_{\mathcal{X}} \mathsf{P}(y, U)\mathsf{P}^{m-1}(x, dy);^{9}$$
$$\mathsf{K}_a(x, U) = \sum_{k=0}^{\infty} a(k)\mathsf{P}^k(x, U),$$

are the *m-step transition* and *sampled-chain* kernels, respectively; a denotes a discrete probability measure on $\mathbb{N}$, known as a *sampling distribution*. Note that, if $a(k) = 1$ for $k = m$ and 0 otherwise, $\mathsf{K}_a = \mathsf{P}^m$.

A set $C \subset \mathcal{X}$ is ν_m*-small* if there exists some $m > 0$ and a non-trivial measure ν_m such that

$$\mathsf{P}^m(x, U) \geq \nu_m(U),$$

for all $x \in C$ and all measurable sets $U \subset \mathcal{X}$. It is ν_a*-petite* if there exists a non-trivial measure ν_a such that

$$\mathsf{K}_a(x, U) \geq \nu_a(U),$$

where a is some sampling distribution.

Given an arbitrary measurable set $U \subset \mathcal{X}$, the *first return time to* U, τ_U, is defined as

$$\tau_U = \min_{t \in \mathcal{T}} \left\{X_t \in U, \quad t \geq 1\right\}.$$

Given a measure φ, a Markov chain is φ*-irreducible* if

$$\varphi(U) > 0 \Rightarrow \mathbb{P}\left[\tau_U < \infty \mid X_0 = x\right] > 0,$$

for any $x \in \mathcal{X}$ and any measurable set $U \subset \mathcal{X}$. If a Markov chain $(\mathcal{X}, \mathsf{P})$ is φ-irreducible, then there is a maximal irreducibility measure ψ for which $(\mathcal{X}, \mathsf{P})$ is ψ-irreducible. All maximal irreducibility measures are equivalent and hence a chain can be classified as being ψ*-irreducible* without specifically identifying the the maximal irreducibility measure ψ.[10]

Given a ψ-irreducible Markov chain $(\mathcal{X}, \mathsf{P})$, it is always possible to determine a maximal family of disjoint sets $\mathcal{D} = \{D_1, \ldots, D_d\}$ such that

- For every $x \in D_i$, $\mathsf{P}(x, D_{i+1}) = 1$, with $i = 1, \ldots, d \pmod d$;
- $\psi\big(\mathcal{X} - \bigcup_{i=1}^{d} D_i\big) = 0$.

[9] We take $\mathsf{P}^1(x, U) = \mathsf{P}(x, U)$.

[10] Equivalent in this context means that any maximal irreducibility measures are absolutely continuous with respect to one another.

Such family $\mathcal{D}$ is called a *d-cycle*. The largest d for which there is a d-cycle for the Markov chain is called the *period* of $(\mathcal{X}, \mathsf{P})$. If $d = 1$, the chain is said to be *aperiodic* and *periodic* otherwise.

If $\mathcal{X}$ is a locally compact, separable and metrizable topological space, a Markov chain $(\mathcal{X}, \mathsf{P})$ verifies the *Feller property* (i.e., is a *weak Feller chain*) if and only if P maps the set $\mathcal{C}(\mathcal{X})$ of all bounded, continuous functions defined on $\mathcal{X}$ into itself. The following result will prove of use later in our work.

Theorem B.1. *Let $(\mathcal{X}, \mathsf{P})$ be a ψ-irreducible Feller chain. If there is an open petite set C such that $\psi(C) > 0$, then all compact subsets of $\mathcal{X}$ are petite.*

Given a Markov chain $(\mathcal{X}, \mathsf{P})$, a probability measure μ is called *invariant* if

$$\int_{\mathcal{X}} \mu(dx)\mathsf{P}(x, U) = \mu(U). \tag{B.1}$$

Stability of Markov chains is deeply related with the convergence of the trajectories of the chain towards stationarity and motivates the following definitions.

Definition B.1. *A Markov chain $(\mathcal{X}, \mathsf{P})$ is* ergodic *if*

$$\left|P^t(x, U) - \mu(U)\right| \to 0$$

for any $x \in \mathcal{X}$ and any $U \subset \mathcal{X}$. It is geometrically ergodic *if, given any initial measure μ_0,*

$$\sum_{t=0}^{\infty} r^t \left\|(\mu_0 \mathsf{P}^t) - \mu\right\| < \infty$$

for some constant $r > 1$, where $\|\cdot\|$ is the total variation norm. If

$$\sup_{x \in \mathcal{X}} \left\|\mathsf{P}^t(x, \cdot) - \mu\right\| \to 0$$

as $t \to \infty$, the chain $(\mathcal{X}, \mathsf{P})$ is uniformly ergodic.

The following theorem identifies sufficient conditions to guarantee uniform ergodicity and binds all concepts of ergodicity presented above.

Theorem B.2. *For any Markov chain $(\mathcal{X}, \mathsf{P})$ the following are equivalent for a ψ-irreducible, aperiodic chain:*

1. *The chain is uniformly ergodic;*
2. *There are constants $r < 1$ and $R < \infty$ such that, for all $x \in \mathcal{X}$,*
$$\left\|\mathsf{P}^t(x, \cdot) - \mu\right\| \leq Rr^t;$$
3. *The state space $\mathcal{X}$ is petite.*

B.2 Three fundamental lemmas

To prove Theorem 4.1, we provide three intermediate results, identifying the conditions under which

- $(\mathbb{S}^n, \hat{\mathsf{P}})$ is ψ-irreducible;
- $(\mathbb{S}^n, \hat{\mathsf{P}})$ is aperiodic;
- $\mathbb{S}^n$ is petite.

Once these facts are properly established, we make use of Theorem B.2 to establish the assertions of the Theorem 4.1.

Lemma B.3. *Let $(\mathcal{X}, \mathcal{Z}, \mathsf{P}, \mathsf{O})$ be a partially observable Markov chain. Then, if $(\mathcal{X}, \mathsf{P})$ is irreducible and there is an observation $z \in \mathcal{Z}$ and a state $x^* \in \mathcal{X}$ such that, for all $y \in \mathcal{X}$, $\mathsf{O}(y, z) = \delta(x^*, y)$,*[11] *the Markov chain $(\mathbb{S}^n, \hat{\mathsf{P}})$ is ψ-irreducible.*

Proof Since $(\mathcal{X}, \mathsf{P})$ is irreducible, for all $y \in \mathcal{X}$ there is $M \in \mathbb{N}$ such that

$$\mathsf{P}^M(y, x^*) > 0.$$

On the other hand, since x^* is admittedly "observable", $X_t = x^* \Rightarrow B_t(y) = \delta(x^*, y)$, where δ denotes the Kronecker delta-function. Let $b^* = \delta(x^*, y)$. Then,

$$\varphi(U) = \begin{cases} 1 & \text{if } b^* \in U; \\ 0 & \text{otherwise} \end{cases}$$

is an irreducibility measure for $(\mathbb{S}^n, \hat{\mathsf{P}})$ since

$$\varphi(U) > 0 \Rightarrow \mathsf{P}_b[\tau_U < \infty] > 0,$$

for any $b \in \mathbb{S}^n$. This implies that there is a maximal irreducibility measure ψ and $(\mathbb{S}^n, \hat{\mathsf{P}})$ is ψ-irreducible. □ □

Lemma B.4. *Let $(\mathcal{X}, \mathcal{Z}, \mathsf{P}, \mathsf{O})$ be a partially observable Markov chain verifying the conditions of Lemma B.3. Then, if $(\mathcal{X}, \mathsf{P})$ is aperiodic, so is $(\mathbb{S}^n, \hat{\mathsf{P}})$.*

Proof Suppose that the conditions of the lemma are met. It is clear that if a d-cycle exists for $(\mathbb{S}^n, \hat{\mathsf{P}})$, with $d > 1$, then there is a set D_k in the d-cycle such that $b^* \in D_k$, where b^* is as defined in the proof of Lemma B.3. But then $\mathsf{P}^m(x^*, x^*) > 0$ only if $m = nd$, for some $n \in \mathbb{N}$ which implies that x^* has period d and, since $(\mathcal{X}, \mathsf{P})$ is irreducible, so have all the states in $\mathcal{X}$, and the chain $(\mathcal{X}, \mathsf{P})$ is periodic with period d. But this is false, by assumption. Then, $(\mathbb{S}^n, \hat{\mathsf{P}})$ must be aperiodic and the proof is complete. □ □

Lemma B.5. *Let $(\mathcal{X}, \mathcal{Z}, \mathsf{P}, \mathsf{O})$ be a partially observable Markov chain. Then, if the conditions of Lemmas B.3 and B.4 are met, the state space $\mathbb{S}^n$ of the chain $\{B_t\}$ is petite.*

Proof Since $(\mathcal{X}, \mathsf{P})$ is irreducible and aperiodic and $\mathcal{X}$ is finite, $(\mathcal{X}, \mathsf{P})$ is ergodic and, consequently, geometrically ergodic.[12] Let b^* be as defined in the proof of Lemma B.3. Take an arbitrary element $z \in \mathcal{Z}$ such that $\sum_y \mathsf{P}(x^*, y)\mathsf{O}(y, z) > 0$ and let $b^*_x = B(b^*, z)$ be the belief state succeeding b^* when observation z occurs. Since $\sum_x b^*_z(x) = 1$, this means that there is $x \in \mathcal{X}$ such that $b^*_z(x) > 0$. Let

$$J = \min_x \{b^*_z(x) \mid b^*_z(x) > 0\}$$
$$I = \arg\min_x \{b^*_z(x) \mid b^*_z(x) > 0\}.$$

Take any $0 < \varepsilon_1 < J$ and consider the following set

$$C = \{b \in \mathbb{S}^n \mid b(I) > \varepsilon_1\}.$$

Clearly, the set C is open in $\mathbb{R}^n$ and, consequently, it is open in the subspace topology on $\mathbb{S}^n$. On the other hand, $b^*_z \in C$. Since the chain $(\mathcal{X}, \mathsf{P})$ is irreducible by assumption, there is $M > 0$ such that

$$\mathsf{P}^{M-1}(I, x^*) > 0.$$

[11] Recall that $\delta(x^*, y) = 1$ if $x^* = y$ and 0 otherwise.
[12] The two notions coincide in finite state-space chains.

Let $\varepsilon_2 = \mathsf{P}^{M-1}(I, x^*)$. Using the Chapman-Kolmogorov inequality, for any $b \in C$ and any measurable set $U \subset \mathbb{S}^n$,

$$\begin{aligned}\hat{\mathsf{P}}^M(b,U) &\geq \hat{\mathsf{P}}^{M-1}(b,\{b^*\})\hat{\mathsf{P}}(b^*,U) = \\ &= \sum_{y\in\mathcal{X}} b(y)\mathsf{P}^{M-1}(y,x^*)\hat{\mathsf{P}}(b^*,U) > \\ &> \varepsilon_1\varepsilon_2\hat{\mathsf{P}}(b^*,U).\end{aligned}$$

Then, $\nu_M = \varepsilon_1\varepsilon_2\hat{\mathsf{P}}(b^*,U)$ is a non-trivial measure and C is ν_M-small and, consequently, ν_{δ_M}-petite. On the other hand, it is immediate to see that, since $b^*_z \in C$, $\psi(C) > 0$.

Let now f be a bounded continuous function defined on $\mathbb{S}^n$. Then, since $\mathbb{S}^n$ is compact and $\hat{\mathsf{P}}(b,\cdot)$ is a probability measure, $(\hat{\mathsf{P}}f)$ is bounded. On the other hand, notice that, for any $z \in \mathcal{Z}$, the n-dimensional function $b \to B(b,z)$ given coordinate-wise by

$$B(b,z)_y = \frac{\sum_{x\in\mathcal{X}} b(x)\mathsf{P}(x,y)\mathsf{O}(y,z)}{\sum_{x,w\in\mathcal{X}} b(x)\mathsf{P}(x,w)\mathsf{O}(w,x)}, \tag{B.2}$$

is clearly continuous on b. Then, $f_z(b)$ defined for each b and each $z \in \mathcal{Z}$ as

$$f_z(b) = f(B(b,z)),$$

is the composition of two continuous functions and, hence, continuous. Finally,

$$(\hat{\mathsf{P}}f)(b) = \sum_z \sum_{x,y} b(x)\mathsf{P}(x,y)\mathsf{O}(y,z)f_z(b)$$

is clearly also a continuous function of b and $\hat{\mathsf{P}}$ maps $\mathcal{C}(\mathbb{S}^n)$ into $\mathcal{C}(\mathbb{S}^n)$ and $(\mathbb{S}^n, \hat{\mathsf{P}})$ is a weak Feller chain.

We have established $(\mathbb{S}^n, \hat{\mathsf{P}})$ to be a ψ-irreducible Feller chain and have found a petite set C such that $\psi(C) > 0$. Theorem B.1 now ensures that all compact subsets of $\mathbb{S}^n$ are petite and, since $\mathbb{S}^n$ is a compact subset of itself, $\mathbb{S}^n$ is petite and the proof is complete. □ □

Now Lemmas B.3 through B.5 together with Theorem B.2 trivially establish conclusion of Theorem 4.1.