CONVERGENCE OF SARSA WITH LINEAR FUNCTION APPROXIMATION: THE RANDOM HORIZON CASE

LINA PALMBORG

ABSTRACT. The reinforcement learning algorithm SARSA combined with linear function approximation has been shown to converge for infinite horizon discounted Markov decision problems (MDPs). In this paper, we investigate the convergence of the algorithm for random horizon MDPs, which has not previously been shown. We show, similar to earlier results for infinite horizon discounted MDPs, that if the behaviour policy is ε-soft and Lipschitz continuous with respect to the weight vector of the linear function approximation, with small enough Lipschitz constant, then the algorithm will converge with probability one when considering a random horizon MDP.

1. INTRODUCTION

In reinforcement learning, an agent learns by interacting with an environment, and adjusts its behaviour (or policy) based on rewards received. The goal is to behave in such a way that the expected total rewards over the time horizon considered are maximised. Central to most reinforcement learning algorithms is the estimation of values, in terms of either a state-value function representing the expected total reward of each state, or an action-value function representing the expected total reward of each action and state.

Temporal difference (TD) learning algorithms are among the most popular reinforcement learning algorithms. In the tabular case (when the state and action set are small enough so that the values in each state or each state-action pair can be stored as tables) these type of algorithms have been shown to converge with probability one, both when estimating the value function given a specific policy [3], and in TD control algorithms such as SARSA [10] and Q-learning [13]. However, when the state space becomes large, tabular solution methods are no longer feasible. In this case, the algorithms need to be combined with function approximation when estimating the value (or action-value) function.

However, when combining these methods with function approximation, convergence results have proven more difficult to obtain, even for algorithms using linear function approximation. In fact, there are examples of divergence of off-policy algorithms (such as Q-learning) when combined with linear function approximation in the literature, see e.g. [12]. Semi-gradient TD learning methods, in which the value function for a specific given policy is estimated using linear function approximation, have been shown to converge with

probability one, see e.g. [3, 12]. As for SARSA combined with linear function approximation, some convergence results where obtained in the 2000s, for the case of infinite horizon discounted Markov decision problems (MDPs).

To begin, de Farias & Van Roy [4] considered an infinite horizon discounted MDP and showed that a variant of approximate policy iteration, motivated by TD learning algorithms, using linear function approximation and a softmax behaviour policy, is guaranteed to possess at least one fixed point. The continuity of the behaviour policy with respect to the weight vector of the function approximation was needed for this result. However, convergence results were not obtained. Next, Gordon [5] considered a random horizon MDP, and showed that SARSA combined with linear function approximation will converge to a region, when using an ε-greedy behaviour policy, and that the parameters of the function approximation (and hence the derived policy) might oscillate in that region. Since an ε-greedy policy is discontinuous in the action-values (and thus the weight vector of the approximation), the results of [5] and [8] gave an indication that the continuity of the behaviour policy with respect to the weight vector of the function approximation might be of importance for providing a convergence result. This was further explored by Perkins & Precup [8], who considered an infinite horizon discounted MDP. They showed that a variant of SARSA combined with linear function approximation converges to a unique policy, under the condition that the behaviour policy used is ε-soft and Lipschitz continuous w.r.t. the weight vector of the function approximation with a sufficiently small Lipschitz constant. However, in the variant of SARSA that they used, the behaviour policy was only updated based on the new action-values after the weight vector had converged, i.e. not in an online fashion. Hence, this version of the algorithm is likely to converge slowly in practice. Melo et al. [7], who again considered an infinite horizon discounted MDP, later extended the results from [8] to the case when the policy is updated after each iteration, i.e. the standard online version of SARSA with linear function approximation commonly used in practice.

However, to date, similar results have not been shown for random horizon MDPs. The results in [4, 8, 7] are obtained for infinite horizon MDPs, and all use the assumption that the Markov chain induced by any policy is irreducible and aperiodic [4, 8] or uniformly ergodic [7]. For a random horizon MDP, the Markov chain induced by a policy is (under certain assumptions) absorbing, hence the results of [4, 8, 7] are not directly applicable in this case.

We combine results and ideas from [4, 8, 7], but adjusted and applied to random horizon MDPs, together with results from [2]. The algorithm studied in this paper is SARSA with linear function approximation, but where the weight vector and the policy is updated at the end of each trajectory, after reaching the absorbing state. We show that using this algorithm to solve a random horizon MDP, when the behaviour policy is ε-soft and Lipschitz continuous w.r.t. the weight vector with sufficiently small Lipschitz constant, the weight vector will converge with probability one, hence extending the results of [4, 8, 7] to the random horizon MDP case.

The paper is organised as follows. Section 2 presents the random horizon Markov decision problem and its associated state-value function and action-value function. Section 3 defines

the algorithm studied in this paper. Section 4 presents our main convergence result, the assumptions used, and the proof of the convergence result. We conclude with a discussion of the results in Section 5.

2. Markov decision problem

We consider a random horizon (episodic) Markov decision problem, with a finite state set $\mathcal{S}^+$, and a finite action set $\mathcal{A}$. We let $\mathcal{S}$ denote the set of non-terminal states, hence the set of terminal (absorbing) states is given by $\mathcal{S}^+ \setminus \mathcal{S}$. If the process is in state s at time t and the agent chooses action a, the process will transition to state s' with probability $p(s' \mid s, a)$. After choosing action a, the agent receives the reward $r(s, a)$. For the case when the reward also depends on the state s' at time $t + 1$, we denote the reward by $r(s, a, s')$, and let $r(s, a)$ be the expected value of the reward:

$$r(s, a) = \sum_{s' \in \mathcal{S}^+} p(s' \mid s, a) r(s, a, s').$$

A policy π can be deterministic or stochastic. A deterministic policy determines what action to take in each state, while a stochastic policy assigns a probability distribution over the set of available actions to each state $s \in \mathcal{S}$. The probability of choosing action a in state s is denoted by $\pi(a \mid s)$. The Markov decision problem consists of finding the policy π that maximises the sum of the expected rewards received

$$\operatorname*{maximise}_{\pi} \mathrm{E}_\pi \left[\sum_{t=0}^{T-1} r(S_t, A_t, S_{t+1}) \mid S_0 = s \right],$$

where S_t is the state at time t, A_t is the action taken at time t, T denotes the terminal time, i.e. $T := \min\{t : S_t \in \mathcal{S}^+ \setminus \mathcal{S}\}$, and $\mathrm{E}_\pi[\cdot]$ denotes the expectation given that policy π is used. The state-value function, v_π, and the action-value function, q_π, under policy π are defined as

$$v_\pi(s) = \mathrm{E}_\pi \left[\sum_{t=0}^{T-1} r(S_t, A_t, S_{t+1}) \mid S_0 = s \right],$$

$$q_\pi(s, a) = \mathrm{E}_\pi \left[\sum_{t=0}^{T-1} r(S_t, A_t, S_{t+1}) \mid S_0 = s, A_0 = a \right].$$

Note that a policy π can be seen as a matrix, with $|\mathcal{S}|$ rows and $|\mathcal{A}|$ columns, where each row sums to one. Here we view π as a vectorised version of this matrix, i.e. π is an element of $\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$. Let Δ_ε denote the set of ε-soft policies,

$$\Delta_\varepsilon = \left\{ \pi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} : \sum_a \pi(a \mid s) = 1 \text{ for all } s, \pi(a \mid s) \geq \varepsilon \text{ for all } (s, a) \right\}.$$

Note that Δ_ε can be viewed as a compact subset of $\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, since it is closed and bounded.

A policy π is said to be proper if the Markov chain induced by π reaches the terminal state with probability one, irrespective of starting state, see further [2, Def 2.1].

3. SARSA with linear function approximation

The algorithm we consider is SARSA with linear function approximation. Hence the action-value function is approximated by a parameterised function $\hat{q}(\cdot; \theta)$ which is a linear function of the weight vector $\theta \in \mathbb{R}^d$:

$$\hat{q}(s, a; \theta) = \phi(s, a)^\top \theta,$$

where $\phi(s, a)$ are basis functions. We let $\Phi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times d}$ denote the matrix whose (s, a)th row is $\phi(s, a)^\top$.

Suppose that (S_t) and (A_t) are sampled trajectories of states and actions, according to some behaviour policy π. The update equation for the weight vector using SARSA with linear function approximation is then

$$\theta_{t+1} = \theta_t + \alpha_{t+1} \phi(S_t, A_t)(r(S_t, A_t, S_{t+1}) + \phi(S_{t+1}, A_{t+1})^\top \theta_t - \phi(S_t, A_t)^\top \theta_t),$$

with the convention that $\phi(S_t, \cdot) = 0$ when $S_t \in \mathcal{S}^+ \setminus \mathcal{S}$, where α_t is the step-size parameter. We consider a slightly modified version of SARSA, where the weight vector is only updated at the end of each trajectory, when we have reached the terminal state:

$$(1) \qquad \theta_{t+1} = \theta_t + \alpha_{t+1} \sum_{u=0}^{T^{(t)}-1} \phi_u^{(t+1)}(r_u^{(t+1)} + (\phi_{u+1}^{(t+1)})^\top \theta_t - (\phi_u^{(t+1)})^\top \theta_t),$$

where $\phi_u^{(t)} = \phi(S_u^{(t)}, A_u^{(t)})$, $r_u^{(t)} = r(S_u^{(t)}, A_u^{(t)}, S_{u+1}^{(t)})$, and $(S_u^{(t)})_{u=0}^{T^{(t)}}$ and $(A_u^{(t)})_{u=0}^{T^{(t)}-1}$ are the sampled states and actions during trajectory t, and $T^{(t)}$ is the time the terminal state is reached during trajectory t. Let $X_t = (S_0^{(t)}, A_0^{(t)}, S_1^{(t)}, A_1^{(t)}, \ldots, A_{T^{(t)}-1}^{(t)}, S_{T^{(t)}}^{(t)})$ denote the tth sampled trajectory. Then (1) can be written as

$$\theta_{t+1} = \theta_t + \alpha_{t+1} H(\theta_t, X_{t+1}),$$

where

$$(2) \qquad H(\theta_t, X_{t+1}) = \sum_{u=0}^{T^{(t+1)}-1} \phi_u^{(t+1)}(r_u^{(t+1)} + (\phi_{u+1}^{(t+1)})^\top \theta_t - (\phi_u^{(t+1)})^\top \theta_t).$$

We further assume that the behaviour policy generating actions is updated at the end of each trajectory, and is dependent on the weight vector θ. Hence the policy generating actions during trajectory t will be denoted π_{θ_t}. This algorithm corresponds to Algorithm 1 below.

4. Convergence of the algorithm

4.1. **Preliminaries.** We make the following assumptions:

Assumption 4.1. $|r(s, a, s')| \le r_{\max} < \infty$.

Assumption 4.2. *(i) The columns of Φ are linearly independent, (ii) $\|\Phi\|_\infty = \Phi_{\max} < \infty$.*

Assumption 4.3. *The step-size parameters satisfy $\sum_t \alpha_t = \infty$, $\sum_t \alpha_t^2 < \infty$.*

Algorithm 1

Input: θ-dependent policy π_θ
Algorithm parameters: step size parameters (α_t)
Initialise $\theta_0 \in \mathbb{R}^d$ arbitrarily

$\pi_0 = \pi_{\theta_0}$
repeat for $t = 0, 1, 2, \ldots$
    for $u = 0, 1, 2, \ldots$ **do**
        Simulate/observe state S_u
        if $S_u \in \mathcal{S}$ **then**
            Choose action $A_u \sim \pi_t(\cdot | S_u)$
        else
            $T = u$
            break
        end if
    end for
    $\theta_{t+1} = \theta_t + \alpha_{t+1} \sum_{u=0}^{T-1} \phi(S_u, A_u)(r(S_u, A_u, S_{u+1}) + \phi(S_{u+1}, A_{u+1})^\top \theta_t - \phi(S_u, A_u)^\top \theta_t)$
    $\pi_{t+1} = \pi_{\theta_{t+1}}$
until approximate convergence of (θ_t)

Assumption 4.4. *All states in $\mathcal{S}$ are reachable with a positive probability, i.e. $\mathrm{P}_\pi(S_t = s) > 0$ for some t, for all s and π, where $\mathrm{P}_\pi(\cdot)$ denotes the probability given that policy π is used.*

Assumption 4.5. *All policies are proper.*

Remark 4.1. That all policies are proper means that the Markov chain induced by any policy π will reach the terminal state with probability one, irrespective of starting state. Naturally, this might not hold for all random horizon Markov decision problems. However, if one were to consider the discounted version of the problem, i.e.

$$\underset{\pi}{\text{maximise}} \, \mathrm{E}_\pi \bigg[\sum_{t=0}^{T-1} \gamma^t r(S_t, A_t, S_{t+1}) \mid S_0 = s \bigg],$$

where $\gamma < 1$ is the discount factor, then discounting can be seen as a form of termination. This is due to that the problem with discounting can be seen as a problem without discounting where the state space is augmented by an additional (policy independent) absorbing state, and where the probability of reaching this new absorbing state is $1 - \gamma$ from any transient state. See further [9, Ch. 5.3]. Based on this, the transition probability becomes $\widetilde{p}(s' \mid s, a) = \gamma p(s' \mid s, a)$. Hence, by considering the discounted version of the problem, and reformulating this in terms of the equivalent Markov decision problem augmented with an additional absorbing state, one can ensure that all policies are proper, even if the original Markov decision problem does not have this property.

Under Assumption 4.5, the Markov chain induced by any policy π has $|\mathcal{S}|$ transient states, and $|\mathcal{S}^+ \setminus \mathcal{S}|$ absorbing (terminal) states. We now consider the Markov chain over

state-action pairs induced by a policy π. Since no action is taken in an absorbing state, we augment the set of actions with an additional action (or "no-action") a^+, which corresponds to "take no action". Under Assumption 4.5 $\{(s, a^+) : s \in \mathcal{S}^+ \setminus \mathcal{S}\}$ is the set of absorbing states of this Markov chain over state-action pairs, and $\{(s, a) : s \in \mathcal{S}, a \in \mathcal{A}\}$ is the set of transient states. Let P_π denote the $|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}||\mathcal{A}|$ transition matrix of the Markov chain over state-action pairs induced by policy π corresponding to transitions between transient states, i.e. the element in the (s, a)th row and (s', a')th column of P_π is

$$p_\pi(s', a' \mid s, a) = \mathrm{P}_\pi(S_t = s', A_t = a' \mid S_{t-1} = s, A_{t-1} = a) = p(s' \mid s, a)\pi(a' \mid s'),$$

for $s, s' \in \mathcal{S}$ and $a, a' \in \mathcal{A}$. From [9, Prop. A.3] we know that $(I - P_\pi)^{-1}$ exists and satisfies

$$(I - P_\pi)^{-1} = \sum_{k=0}^{\infty} P_\pi^k.$$

Furthermore, let $\lambda(s) = \mathrm{P}(S_0 = s)$, and let η_π denote the length $|\mathcal{S}||\mathcal{A}|$ vector whose (s, a)th element is the expected number of visits to state-action pair (s, a) before absorption. Then

$$\eta_\pi(s, a) := \mathrm{E}_\pi \Big[\sum_{t=0}^{\infty} \mathbf{1}_{\{S_t=s, A_t=a\}} \Big] = \pi(a \mid s)\lambda(s) + \sum_{t=1}^{\infty} \mathrm{P}_\pi(S_t = s, A_t = a)$$

$$= \pi(a \mid s)\lambda(s) + \sum_{t=1}^{\infty} \sum_{s' \in \mathcal{S}, a' \in \mathcal{A}} p_\pi(s, a \mid s', a') \, \mathrm{P}_\pi(S_{t-1} = s', A_{t-1} = a')$$

$$= \pi(a \mid s)\lambda(s) + \sum_{s' \in \mathcal{S}, a' \in \mathcal{A}} p_\pi(s, a \mid s', a') \sum_{t=0}^{\infty} \mathrm{P}_\pi(S_t = s', A_t = a')$$

$$(3) \qquad = \pi(a \mid s)\lambda(s) + \sum_{s' \in \mathcal{S}, a' \in \mathcal{A}} p_\pi(s, a \mid s', a')\eta_\pi(s', a'),$$

or, in matrix form, $\eta_\pi^\top = \lambda_\pi^\top + \eta_\pi^\top P_\pi$, hence $\eta_\pi^\top = \lambda_\pi^\top (I - P_\pi)^{-1}$, where λ_π is the length $|\mathcal{S}||\mathcal{A}|$ vector whose (s, a)th element is $\mathrm{P}_\pi(S_0 = s, A_0 = a) = \pi(a \mid s)\lambda(s)$. Let D_π be the diagonal matrix whose diagonal is η_π, and let r denote the length $|\mathcal{S}||\mathcal{A}|$ vector whose (s, a)th element is $r(s, a)$.

Assumption 4.6. *At least one of the following statements holds: (i) $\lambda_\pi(s, a) > 0$ for all s, a, (ii) $\sum_{s', a'} p_\pi(s', a' \mid s, a) < 1$ for all s, a.*

Remark 4.2. If $\pi \in \Delta_\varepsilon$, then Assumption 4.6(i) holds if $\lambda(s) = P(S_0 = s) > 0$ for all s. Furthermore, since $\sum_{s', a'} p_\pi(s', a' \mid s, a) = \sum_{s'} p(s' \mid s, a) \sum_{a'} \pi(a' \mid s') = \sum_{s'} p(s' \mid s, a)$, Assumption 4.6(ii) holds if $\sum_{s'} p(s' \mid s, a) < 1$ for all s, a, irrespective of which policy is used. Note that, similarly to Remark 4.1, $\sum_{s'} p(s' \mid s, a) < 1$ for all s, a can be achieved by instead considering the discounted version of the problem, since the augmented probabilities are then $\widetilde{p}(s' \mid s, a) = \gamma p(s' \mid s, a)$, hence $\sum_{s'} \widetilde{p}(s' \mid s, a) = \sum_{s'} \gamma p(s' \mid s, a) \leq \gamma < 1$ for all s, a.

We make the following assumptions regarding the behaviour policy π_θ:

Assumption 4.7. *(i) π_θ is Lipschitz continuous with respect to θ, i.e. there exists a constant C such that $\|\pi_\theta - \pi_{\theta'}\| \leq C\|\theta - \theta'\|$, (ii) π_θ is ε-soft, i.e. $\pi_\theta(a \mid s) \geq \varepsilon$ for all s and a, for some $\varepsilon > 0$.*

Note that the set of behaviour policies considered here can be viewed as a subset of Δ_ε. The norm denoted by $\|\cdot\|$ here corresponds to the Euclidean norm if applied to vectors, and to the spectral norm if applied to matrices. For further details on different norms and norm inequalities used throughout the paper, see Appendix A.

4.2. **Convergence theorem.** We begin by stating our main result, Theorem 4.1, and then briefly go through an overview of the proof. The full proof can be found in Section 4.5, using results from Section 4.4.

Theorem 4.1. *Assume that the assumptions listed in the previous section are satisfied. Then, for any $\varepsilon > 0$, there exists $C_0 > 0$ such that, if $C < C_0$, the sequence (θ_t) generated by Algorithm 1 converges with probability one.*

Proof overview: To prove Theorem 4.1, we want to use Theorem 1, Section 5.1 in Benveniste et al. [1] (restated in Theorem 4.2, Section 4.3, for completeness). Hence, we need to show that the Robbins-Monro assumption (4), the square integrability condition (5) and the stability condition (6), are satisfied.

The Robbins-Monro assumption clearly holds for the algorithm considered, see Section 4.5.1 below.

That the stability condition (6) holds is shown using similar ideas to Melo et al. [7]. We begin by showing that there exists θ^* such that $A_{\pi_{\theta^*}}\theta^* + b_{\pi_{\theta^*}} = 0$, where A_{π_θ} and b_{π_θ} are given by (7). To this end, we use results from de Farias & Van Roy [4], but adapted to the current situation, where the MDP has terminal states. See further Lemmas 4.1-4.7, Section 4.4. In the next step, we show that A_{π_θ} and b_{π_θ} are Lipschitz continuous in θ, based on results from Perkins & Precup [8], adapted to an MDP with terminal states. This is done in Lemmas 4.8-4.10, Section 4.4. Furthermore, we note that A_{π_θ} is negative definite (Lemma 4.3), as shown in Bertsekas & Tsitsiklis [2]. Using these three results it is possible to show that the stability condition (6) is satisfied if the Lipschitz constant C in Assumption 4.7(i) is sufficiently small. This is done in Section 4.5.3.

That the square integrability condition (5) holds is shown by rewriting (2) in a way suggested by Sutton [11], as described in Gordon [5]. Then, using Assumptions 4.1 and 4.2, together with general expressions for the expectation and variance of the number of steps before being absorbed in an absorbing Markov chain, it is shown that the square integrability condition holds (see Section 4.5.2).

4.3. **Theorem from Benveniste et al.** The algorithm studied is of the form

$$\theta_{t+1} = \theta_t + \alpha_{t+1}H(\theta_t, X_{t+1}).$$

Robbins-Monro assumption: For any positive Borel function g,

$$(4) \qquad\qquad \mathrm{E}[g(\theta_t, X_{t+1}) \mid \mathcal{F}_t] = \sum_x g(\theta_t, x) p_{\theta_t}(x),$$

where $p_\theta(x) = \mathrm{P}_\theta(X_t = x)$, where $\mathrm{P}_\theta(\cdot)$ denotes the probability given parameter θ, and $\mathcal{F}_t$ is the σ-field generated by $X_t, X_{t-1}, \ldots, X_1, \theta_t, \theta_{t-1}, \ldots, \theta_0$, i.e. the conditional distribution of X_{t+1}, knowing the past, depends only on θ_t.

Square integrability condition:

$$(5) \qquad\qquad \text{For any } \theta, \text{ there exists } K \text{ s.t. } \mathrm{E}_\theta[\|H(\theta, X_t)\|^2] \leq K(1 + \|\theta\|^2)$$

where $\mathrm{E}_\theta[\cdot]$ denotes the expectation given parameter θ.

Stability condition:

$$(6) \qquad \text{For any } \nu > 0, \text{ there exists } \theta^* \text{ s.t. } \sup_{\nu \leq \|\theta - \theta^*\| \leq \frac{1}{\nu}} (\theta - \theta^*)^\top \mathrm{E}_\theta[H(\theta, X_{t+1})] < 0.$$

Theorem 4.2 (Theorem 1, Section 5.1 in Benveniste et al. [1]). *Under the assumptions above, if the sequence of step-size parameters satisfies $\sum_t \alpha_t = \infty$, $\sum_t \alpha_t^2 < \infty$, then the sequence $(\theta_t)_{t \geq 0}$ converges almost surely to θ^* satisfying* (6).

4.4. **Preliminary results.** We begin by showing that there exists θ^* such that $A_{\pi_{\theta^*}} \theta^* + b_{\pi_{\theta^*}} = 0$, where

$$(7) \qquad\qquad A_{\pi_\theta} = \Phi^\top D_{\pi_\theta}(P_{\pi_\theta} - I)\Phi, \quad b_{\pi_\theta} = \Phi^\top D_{\pi_\theta} r.$$

This is done by using results from de Farias & Van Roy [4] and Bertsekas & Tsitsiklis [2]. The proofs of Lemmas 4.5-4.7 follow directly from the proofs of corresponding Lemmas 5.3-5.5 in [4]. These proofs are included in Appendix B for completeness. Lemmas 4.1-4.2 and 4.4 require new proofs to take into account that we consider an MDP with absorbing states. These proofs are included below. Lemma 4.3 is used in the proof of Lemma 4.4, and later on used to show that the stability condition (6) is satisfied. The proof of Lemma 4.3 is identical to the last part of the proof of Lemma 6.10 in Bertsekas & Tsitsiklis [2], and is included in Appendix B for completeness.

In order to show the above, we study the operators H_π and H_{π_θ}, defined by

$$H_\pi \Phi\theta = \operatorname*{arg\,min}_{\bar{q} \in \{\Phi\theta : \theta \in \mathbb{R}^d\}} \|r + P_\pi \Phi\theta - \bar{q}\|_{\eta_\pi}, \quad H_{\pi_\theta} \Phi\theta = \operatorname*{arg\,min}_{\bar{q} \in \{\Phi\theta : \theta \in \mathbb{R}^d\}} \|r + P_{\pi_\theta} \Phi\theta - \bar{q}\|_{\eta_{\pi_\theta}},$$

where

$$\|q\|_{\eta_\pi} = \left(\sum_{s \in \mathcal{S}, a \in \mathcal{A}} \eta_\pi(s, a) q(s, a)^2 \right)^{1/2} = \left(q^\top D_\pi q \right)^{1/2},$$

for any $|\mathcal{S}||\mathcal{A}|$-dimensional vector q. We can also write $H_\pi = \Pi_\pi T_\pi$ and $H_{\pi_\theta} = \Pi_{\pi_\theta} T_{\pi_\theta}$, where

$$T_\pi \Phi\theta = r + P_\pi \Phi\theta, \quad \Pi_\pi q = \operatorname*{arg\,min}_{\bar{q} \in \{\Phi\theta : \theta \in \mathbb{R}^d\}} \|q - \bar{q}\|_{\eta_\pi}.$$

Note that by Assumption 4.4 and that π is ε-soft, all transient state-action pairs have a positive probability of being visited, hence D_π is positive definite. Thus, using Assumption 4.2(i) $\Phi^\top D_\pi \Phi$ is positive definite. Hence, the projection operator Π_π is given by

$$(8) \qquad \Pi_\pi = \Phi(\Phi^\top D_\pi \Phi)^{-1}\Phi^\top D_\pi.$$

The aim is to show that H_{π_θ} has a fixed point, i.e. that there exist $\Phi\theta$ such that $\Phi\theta = H_{\pi_\theta}\Phi\theta$. The reason for this is that any solution $\bar{q} = \Phi\theta^*$ to $\Phi\theta^* = H_{\pi_\theta}\Phi\theta$ must satisfy

$$\Phi^\top D_{\pi_\theta}(r + P_{\pi_\theta}\Phi\theta - \Phi\theta^*) = 0,$$

hence the fixed point of H_{π_θ}, if it exists, must satisfy

$$\Phi^\top D_{\pi_\theta}(r + P_{\pi_\theta}\Phi\theta - \Phi\theta) = 0,$$

which can also be written as $A_{\pi_\theta}\theta + b_{\pi_\theta} = 0$. Hence, if H_{π_θ} has a fixed point, then there exists θ^* such that $A_{\pi_{\theta^*}}\theta^* + b_{\pi_{\theta^*}} = 0$.

Lemma 4.1. *The expected number of visits to each state-action pair before absorption, η_π, is a continuous function of π.*

Proof. By (3) we have $\eta_\pi^\top = \lambda_\pi^\top(I - P_\pi)^{-1}$. λ_π and P_π are continuous functions of π, $I - P_\pi$ is nonsingular (see e.g. [9, Prop. A.3]), and the matrix inverse function of a nonsingular matrix is continuous (see e.g. [9, Prop. C.5]). Hence η_π is a continuous function of π. $\square$

Lemma 4.2. *For each policy π, H_π is a contraction, and there exists a unique vector θ_π such that $\Phi\theta_\pi = H_\pi\Phi\theta_\pi$.*

Proof. We begin by showing that T_π is a contraction, i.e. that there exists $\beta \in [0, 1)$ such that $\|T_\pi q - T_\pi q'\|_{\eta_\pi} \leq \beta\|q - q'\|_{\eta_\pi}$, where $q, q' \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$. This will hold if there exists $\beta \in [0, 1)$ such that $\|P_\pi q\|_{\eta_\pi} \leq \beta\|q\|_{\eta_\pi}$, since $T_\pi q - T_\pi q' = P_\pi(q - q')$. Let p_{sum} be the $|\mathcal{S}||\mathcal{A}|$-dimensional vector corresponding to the row sums of P_π, i.e. $p_{\mathrm{sum}} = P_\pi \mathbf{1}$, and $p_{\mathrm{sum}}(s, a)$ the (s, a)th element of p_{sum}. For $q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$,

$$\|P_\pi q\|_{\eta_\pi}^2 = q^\top P_\pi^\top D_\pi P_\pi q = \sum_{s,a} \eta_\pi(s, a)\left(\sum_{s',a'} p_\pi(s', a' \mid s, a)q(s', a')\right)^2$$

$$= \sum_{(s,a):p_{\mathrm{sum}}(s,a)>0} \eta_\pi(s, a)p_{\mathrm{sum}}(s, a)^2\left(\sum_{s',a'} \frac{p_\pi(s', a' \mid s, a)}{p_{\mathrm{sum}}(s, a)}q(s', a')\right)^2$$

$$\leq \sum_{(s,a):p_{\mathrm{sum}}(s,a)>0} \eta_\pi(s, a)p_{\mathrm{sum}}(s, a)^2 \sum_{s',a'} \frac{p_\pi(s', a' \mid s, a)}{p_{\mathrm{sum}}(s, a)}q(s', a')^2$$

$$= \sum_{s,a} \eta_\pi(s, a)p_{\mathrm{sum}}(s, a) \sum_{s',a'} p_\pi(s', a' \mid s, a)q(s', a')^2.$$

Under Assumption 4.6(i), since $\eta_\pi(s,a) - \lambda_\pi(s,a) < \eta_\pi(s,a)$ and $p_{\text{sum}}(s,a) \le 1$ for all s,a,

$$
\begin{aligned}
\|P_\pi q\|_{\eta_\pi}^2 &\le \sum_{s,a} \eta_\pi(s,a) \sum_{s',a'} p_\pi(s',a' \mid s,a) q(s',a')^2 \\
&= \sum_{s',a'} q(s',a')^2 \sum_{s,a} \eta_\pi(s,a) p_\pi(s',a' \mid s,a) \\
&= \sum_{s',a'} q(s',a')^2 (\eta_\pi(s',a') - \lambda_\pi(s',a')) \\
&\le \beta \sum_{s',a'} q(s',a')^2 \eta_\pi(s',a') = \beta \|q\|_{\eta_\pi}^2,
\end{aligned}
$$

where

$$
\beta = 1 - \min_{s,a} \frac{\lambda_\pi(s,a)}{\eta_\pi(s,a)}
$$

i.e. $\beta \in [0,1)$ since $\lambda_\pi(s,a) > 0$ and $\eta_\pi(s,a) \ge \lambda_\pi(s,a)$ for all s,a. Under Assumption 4.6(ii), since $p_{\text{sum}}(s,a) < 1$ and $\eta_\pi(s,a) - \lambda_\pi(s,a) \le \eta_\pi(s,a)$ for all s,a,

$$
\begin{aligned}
\|P_\pi q\|_{\eta_\pi}^2 &\le \sum_{s,a} \eta_\pi(s,a) p_{\text{sum}}(s,a) \sum_{s',a'} p_\pi(s',a' \mid s,a) q(s',a')^2 \\
&\le \max_{s,a} p_{\text{sum}}(s,a) \sum_{s',a'} q(s',a')^2 \sum_{s,a} \eta_\pi(s,a) p_\pi(s',a' \mid s,a) \\
&\le \beta \sum_{s',a'} q(s',a')^2 \eta_\pi(s',a') = \beta \|q\|_{\eta_\pi}^2,
\end{aligned}
$$

where $\beta = \max_{s,a} p_{\text{sum}}(s,a) < 1$. Hence, T_π is a contraction under Assumption 4.6.

Furthermore, Π_π is a non-expansion in the sense that $\|\Pi_\pi q\|_{\eta_\pi} \le \|q\|_{\eta_\pi}$ (this is a generic property of projections, and can also easily be shown based on (8), see e.g. the proof of Proposition 6.9 in [2]). Hence, H_π is a contraction. By the contraction mapping theorem, it follows that H_π has a unique fixed point $\Phi\theta_\pi$. By Assumption 4.2(i), we can conclude that θ_π is unique. $\qquad\square$

Lemma 4.3. *For any ε-soft policy π, the matrix $A_\pi = \Phi^\top D_\pi (P_\pi - I)\Phi$ is negative definite.*

Lemma 4.4. *Let θ_π be the unique solution to $\Phi\theta_\pi = H_\pi \Phi\theta_\pi$. For any $\varepsilon > 0$, the function $\Delta_\varepsilon \ni \pi \mapsto \theta_\pi$ is continuous.*

Proof. By Assumption 4.4 and that π is ε-soft, the projection operator Π_π is given by (8), and the solution θ_π to $\Phi\theta_\pi = H_\pi \Phi\theta_\pi$ must satisfy

$$
\Phi\theta_\pi = \Phi(\Phi^\top D_\pi \Phi)^{-1} \Phi^\top D_\pi (r + P_\pi \Phi\theta_\pi).
$$

Hence,

$$
\Phi^\top D_\pi (I - P_\pi)\Phi\theta_\pi = \Phi^\top D_\pi r.
$$

Let $A_\pi = \Phi^\top D_\pi (P_\pi - I)\Phi$ and $b_\pi = \Phi^\top D_\pi r$, so that $A_\pi \theta_\pi + b_\pi = 0$. By Lemma 4.3 A_π is negative definite, and thus $\theta_\pi = -A_\pi^{-1} b_\pi$. P_π is a continuous function of π, by Lemma 4.1, D_π is a continuous function of π, and the matrix inverse function of a nonsingular matrix is continuous (see e.g. [9, Prop. C.5]), hence both A_π^{-1} and b_π are continuous functions of π, for any ε-soft policy π. $\qquad \square$

Similarly to de Farias and Van Roy [4], but for any ε-soft policy π, and for any policy π_θ satisfying Assumption 4.7, we define

$$s_\pi(\theta) = \Phi^\top D_\pi (T_\pi \Phi\theta - \Phi\theta) \quad \text{and} \quad s_{\pi_\theta}(\theta) = \Phi^\top D_{\pi_\theta}(T_{\pi_\theta}\Phi\theta - \Phi\theta),$$

and functions $F_\pi^\alpha : \mathbb{R}^d \to \mathbb{R}^d$ and $F_{\pi_\theta}^\alpha : \mathbb{R}^d \to \mathbb{R}^d$ by

$$F_\pi^\alpha(\theta) = \theta + \alpha s_\pi(\theta) \quad \text{and} \quad F_{\pi_\theta}^\alpha(\theta) = \theta + \alpha s_{\pi_\theta}(\theta).$$

Lemma 4.5. *For any $\alpha > 0$, θ is a fixed point of F_π^α ($F_{\pi_\theta}^\alpha$) if and only if $\Phi\theta$ is a fixed point of H_π (H_{π_θ}).*

The following lemma is used to prove Lemma 4.7.

Lemma 4.6. *There exists $\alpha^* > 0$ such that, for all ε-soft policies π and any $\alpha \in (0, \alpha^*)$, there exists a scalar β_α such that*

$$\|F_\pi^\alpha(\theta) - \theta_\pi\| \leq \beta_\alpha \|\theta - \theta_\pi\|.$$

Lemma 4.7. *For any $\alpha > 0$, the function $F_{\pi_\theta}^\alpha$ possesses a fixed point.*

Hence, using Lemma 4.7 and Lemma 4.5 we can show the desired result, i.e. that H_{π_θ} has a fixed point.

Next, we show that A_{π_θ} and b_{π_θ} are Lipschitz continuous w.r.t. θ, which, since π_θ is Lipschitz continuous w.r.t. θ, will hold if A_π and b_π are Lipschitz continuous w.r.t. π, where

$$A_\pi = \Phi^\top D_\pi (P_\pi - I)\Phi, \quad b_\pi = \Phi^\top D_\pi r.$$

This will be done using results from Perkins & Precup [8]. The proof of Lemma 4.8 follows directly from the proof of the correspoding Lemma 1 in [8]. This proof is included in Appendix B for completeness. Lemma 4.9 requires a new proof to take into account that we consider an MDP with absorbing states, and this proof is included below. The proof of Lemma 4.10 is similar to the proof of Lemma 3 in [8], but some adjustments are required for the case of an absorbing MDP, hence this proof is also included below.

Lemma 4.8. *For any $\varepsilon > 0$, there exists C_P such that $\|P_{\pi_1} - P_{\pi_2}\| \leq C_P \|\pi_1 - \pi_2\|$ for all $\pi_1, \pi_2 \in \Delta_\varepsilon$.*

Lemma 4.9. *For any $\varepsilon > 0$, there exists C_D such that $\|D_{\pi_1} - D_{\pi_2}\| \leq C_D \|\pi_1 - \pi_2\|$ for all $\pi_1, \pi_2 \in \Delta_\varepsilon$.*

Proof. Note that

$$(\eta_{\pi_1}^\top - \eta_{\pi_2}^\top)(I - P_{\pi_1}) = \eta_{\pi_1}^\top - \eta_{\pi_1}^\top P_{\pi_1} - \eta_{\pi_2}^\top + \eta_{\pi_2}^\top P_{\pi_1} = \lambda_{\pi_1}^\top - \eta_{\pi_2}^\top + \eta_{\pi_2}^\top P_{\pi_1}$$
$$= \lambda_{\pi_1}^\top - \lambda_{\pi_2}^\top + \eta_{\pi_2}^\top(P_{\pi_1} - P_{\pi_2}),$$

hence, since $(I - P_\pi)^{-1}$ exists for all proper policies π,

$$\eta_{\pi_1}^\top - \eta_{\pi_2}^\top = \big(\lambda_{\pi_1}^\top - \lambda_{\pi_2}^\top + \eta_{\pi_2}^\top(P_{\pi_1} - P_{\pi_2})\big)(I - P_{\pi_1})^{-1}.$$

Then

$$\|\eta_{\pi_1} - \eta_{\pi_2}\| = \|((I - P_{\pi_1})^{-1})^\top\big(\lambda_{\pi_1} - \lambda_{\pi_2} + (P_{\pi_1}^\top - P_{\pi_2}^\top)\eta_{\pi_2}\big)\|$$
$$\leq \|(I - P_{\pi_1})^{-1}\|(\|\lambda_{\pi_1} - \lambda_{\pi_2}\| + \|(P_{\pi_1}^\top - P_{\pi_2}^\top)\eta_{\pi_2}\|).$$

Now,

$$\|\lambda_{\pi_1} - \lambda_{\pi_2}\| = \sqrt{\sum_{s,a}|\pi_1(a \mid s)\lambda(s) - \pi_2(a \mid s)\lambda(s)|^2} = \sqrt{\sum_{s\in\mathcal{S}}\lambda(s)^2\sum_{a\in\mathcal{A}}|\pi_1(a \mid s) - \pi_2(a \mid s)|^2}$$
$$\leq \sqrt{\sum_{s,a}|\pi_1(a \mid s) - \pi_2(a \mid s)|^2} = \|\pi_1 - \pi_2\|,$$

and

$$\|(P_{\pi_1}^\top - P_{\pi_2}^\top)\eta_{\pi_2}\| \leq \|P_{\pi_1} - P_{\pi_2}\|\|\eta_{\pi_2}\| \leq \|P_{\pi_1} - P_{\pi_2}\|\|(I - P_{\pi_2})^{-1}\|\|\lambda_{\pi_2}\|$$
$$\leq \|P_{\pi_1} - P_{\pi_2}\|\|(I - P_{\pi_2})^{-1}\|\|\lambda_{\pi_2}\|_1 \leq C_P\|(I - P_{\pi_2})^{-1}\|\|\pi_1 - \pi_2\|$$

using $\|\lambda_\pi\|_1 = 1$ for any policy π and Lemma 4.8. Hence

$$\|\eta_{\pi_1} - \eta_{\pi_2}\| \leq \|(I - P_{\pi_1})^{-1}\|\big(1 + \|(I - P_{\pi_2})^{-1}\|C_P\big)\|\pi_1 - \pi_2\|$$
$$\leq \zeta\big(1 + \zeta C_P\big)\|\pi_1 - \pi_2\|,$$

where $\zeta := \max_{\pi\in\Delta_\varepsilon}\|(I - P_\pi)^{-1}\|$. The maximum is attained since P_π is a continuous function of π, $I - P_\pi$ is nonsingular, the matrix inverse function of a nonsingular matrix is continuous (see e.g. [9, Prop. C.5]), any norm is a continuous function, and Δ_ε can be viewed as a compact subset of $\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$. Let $C_D = \zeta\big(1 + \zeta C_P\big)$. Then

$$\|D_{\pi_1} - D_{\pi_2}\| = \max_{s,a}|\eta_{\pi_1}(s,a) - \eta_{\pi_2}(s,a)| = \|\eta_{\pi_1} - \eta_{\pi_2}\|_\infty \leq \|\eta_{\pi_1} - \eta_{\pi_2}\| \leq C_D\|\pi_1 - \pi_2\|.$$

$\square$

Lemma 4.10. *For any $\varepsilon > 0$, there exists C_b and C_A such that $\|b_{\pi_1} - b_{\pi_2}\| \leq C_b\|\pi_1 - \pi_2\|$ and $\|A_{\pi_1} - A_{\pi_2}\| \leq C_A\|\pi_1 - \pi_2\|$ for all $\pi_1, \pi_2 \in \Delta_\varepsilon$.*

Proof. First, note that

$$\|r\| = \sqrt{\sum_{s,a} \Big(\sum_{s'\in\mathcal{S}^+} p(s' \mid s, a) r(s, a, s') \Big)^2}$$

$$\leq \sqrt{\sum_{s,a} \sum_{s'\in\mathcal{S}^+} p(s' \mid s, a) |r(s, a, s')|^2} \leq \sqrt{|\mathcal{S}||\mathcal{A}|} r_{\max},$$

using Assumption 4.1, and

$$\|\Phi^\top\| = \|\Phi\| \leq \sqrt{|\mathcal{S}||\mathcal{A}|} \|\Phi\|_\infty = \sqrt{|\mathcal{S}||\mathcal{A}|} \Phi_{\max},$$

using Assumption 4.2(ii). For the first claim, using Lemma 4.9,

$$\|b_{\pi_1} - b_{\pi_2}\| = \|\Phi^\top (D_{\pi_1} - D_{\pi_2}) r\| \leq \|\Phi^\top\| \|D_{\pi_1} - D_{\pi_2}\| \|r\| \leq C_D |\mathcal{S}||\mathcal{A}| \Phi_{\max} r_{\max} \|\pi_1 - \pi_2\|,$$

i.e. $C_b = C_D |\mathcal{S}||\mathcal{A}| \Phi_{\max} r_{\max}$.

For the second claim, we use that $\|D_\pi\| \leq \max_{\pi\in\Delta_\varepsilon} \|D_\pi\| := \xi$ for any $\pi \in \Delta_\varepsilon$ (By Lemma 4.1 D_π is a continuous function of π, any norm is a continuous function, and Δ_ε is compact, hence the maximum is attained), and $\|P_\pi\| \leq \sqrt{|\mathcal{S}||\mathcal{A}|} \|P_\pi\|_\infty \leq \sqrt{|\mathcal{S}||\mathcal{A}|}$. Hence, using Lemmas 4.8 and 4.9,

$$\|A_{\pi_1} - A_{\pi_2}\| = \|\Phi^\top \big(D_{\pi_1}(P_{\pi_1} - I) - D_{\pi_2}(P_{\pi_2} - I) \big) \Phi\|$$

$$\leq \|\Phi^\top\| \|D_{\pi_2} - D_{\pi_1} + D_{\pi_1}P_{\pi_1} - D_{\pi_2}P_{\pi_2}\| \|\Phi\|$$

$$\leq |\mathcal{S}||\mathcal{A}| \Phi_{\max}^2 \|D_{\pi_2} - D_{\pi_1} + D_{\pi_1}(P_{\pi_1} - P_{\pi_2}) + (D_{\pi_1} - D_{\pi_2})P_{\pi_2}\|$$

$$\leq |\mathcal{S}||\mathcal{A}| \Phi_{\max}^2 \big(\|D_{\pi_1} - D_{\pi_2}\| + \|D_{\pi_1}\| \|P_{\pi_1} - P_{\pi_2}\| + \|D_{\pi_1} - D_{\pi_2}\| \|P_{\pi_2}\| \big)$$

$$\leq |\mathcal{S}||\mathcal{A}| \Phi_{\max}^2 \big((1 + \sqrt{|\mathcal{S}||\mathcal{A}|}) \|D_{\pi_1} - D_{\pi_2}\| + \xi \|P_{\pi_1} - P_{\pi_2}\| \big)$$

$$\leq |\mathcal{S}||\mathcal{A}| \Phi_{\max}^2 \big((1 + \sqrt{|\mathcal{S}||\mathcal{A}|}) C_D + \xi C_P \big) \|\pi_1 - \pi_2\|,$$

i.e. $C_A = |\mathcal{S}||\mathcal{A}| \Phi_{\max}^2 ((1 + \sqrt{|\mathcal{S}||\mathcal{A}|}) C_D + \xi C_P)$. $\qquad\square$

By Lemma 4.10, it is clear that under Assumption 4.7, A_{π_θ} and b_{π_θ} are Lipschitz continuous with respect to θ with Lipschitz constants $C_1 = C_A C$ and $C_2 = C_b C$ respectively.

4.5. **Proof of Theorem 4.1.**

4.5.1. *Robbins-Monro assumption.* The Robbins-Monro assumption (4) holds in our case by the definition of $X_{t+1} = (S_0^{(t+1)}, A_0^{(t+1)}, S_1^{(t+1)}, A_1^{(t+1)}, \ldots, A_{T^{(t+1)}-1}^{(t+1)}, S_{T^{(t+1)}}^{(t+1)})$, which is sampled according to

$$P_{\pi_{\theta_t}}(A_u^{(t+1)} = a \mid S_u^{(t+1)} = s) = \pi_{\theta_t}(a \mid s), \quad \text{for } u = 0, \ldots, T^{(t+1)} - 1,$$

$$P(S_{u+1}^{(t+1)} = s' \mid S_u^{(t+1)} = s, A_u^{(t+1)} = a) = p(s' \mid s, a), \quad \text{for } u = 0, \ldots, T^{(t+1)} - 1,$$

$$P(S_0^{(t+1)} = s) = \lambda(s).$$

Hence each trajectory/episode is independent of the previous episodes given θ_t (and independent of $(\theta_{t-1}, \theta_{t-2}, \ldots)$ given θ_t, since only θ_t affects the behaviour policy used during trajectory $t + 1$).

4.5.2. *Square integrability condition.* Similarly to Gordon [5], we use the equations on p. 25 in Sutton [11] to write $H(\theta, X_{t+1})$ as

$$H(\theta, X_{t+1}) = \sum_{s,a} \sum_{s',a'} \gamma_\theta(s', a' \mid s, a) \phi(s, a) \big(r(s, a, s') + \phi(s', a')^\top \theta - \phi(s, a)^\top \theta \big),$$

where $\gamma_\theta(s', a' \mid s, a)$ denotes the number of times the transition $(s, a) \to (s', a')$ occurs in the sequence $X_{t+1} = (S_0^{(t+1)}, A_0^{(t+1)}, S_1^{(t+1)}, A_1^{(t+1)}, \ldots, A_{T^{(t+1)}-1}^{(t+1)}, S_{T^{(t+1)}}^{(t+1)})$ (for $s' \in \mathcal{S}^+ \setminus \mathcal{S}$ all but one of the $\gamma_\theta(s', a' \mid s, a)$ are 0). Let $\delta_\theta(s, a, s', a')$ be defined by

$$\delta_\theta(s, a, s', a') = r(s, a, s') + \phi(s', a')^\top \theta + \phi(s, a)^\top \theta,$$

so that

$$H(\theta, X_{t+1}) = \sum_{s,a} \sum_{s',a'} \gamma_\theta(s', a' \mid s, a) \phi(s, a) \delta_\theta(s, a, s', a').$$

Then

$$\|H(\theta, X_{t+1})\| \leq \sum_{s,a} \sum_{s',a'} \gamma_\theta(s', a' \mid s, a) |\delta_\theta(s, a, s', a')| \|\phi(s, a)\|$$

$$\leq \Phi_{\max} \sum_{s,a} \sum_{s',a'} \gamma_\theta(s', a' \mid s, a) |\delta_\theta(s, a, s', a')|,$$

where we have used Assumption 4.2(ii). Furthermore, using Assumptions 4.2(ii) and 4.1,

$$|\delta_\theta(s, a, s', a')| \leq |r(s, a, s')| + (\|\phi(s', a')\| + \|\phi(s, a)\|) \|\theta\| \leq r_{\max} + 2\Phi_{\max}\|\theta\|,$$

hence

$$\|H(\theta, X_{t+1})\| \leq \Phi_{\max}(r_{\max} + 2\Phi_{\max}\|\theta\|) \sum_{s,a} \sum_{s',a'} \gamma_\theta(s', a' \mid s, a).$$

By using $(a + b)^2 \leq 2(a^2 + b^2)$ we obtain

$$\|H(\theta, X_{t+1})\|^2 \leq 2\Phi_{\max}^2(r_{\max}^2 + 4\Phi_{\max}^2\|\theta\|^2) \Big(\sum_{s,a} \sum_{s',a'} \gamma_\theta(s', a' \mid s, a) \Big)^2.$$

Let $\Gamma_\theta = \sum_{s,a} \sum_{s',a'} \gamma_\theta(s', a' \mid s, a)$ denote the number of steps until absorption. Then

$$\mathrm{E}_\theta \left[\|H(\theta, X_{t+1})\|^2 \right] \leq 2\Phi_{\max}^2(r_{\max}^2 + 4\Phi_{\max}^2\|\theta\|^2) \, \mathrm{E}_\theta \left[\Gamma_\theta^2 \right],$$

where $\mathrm{E}_\theta[\cdot]$ denotes the expectation given parameter θ (and thus also given that policy π_θ is used). The expected number of steps before being absorbed, when starting in state (s, a), is given by the (s, a)th element of t_θ, denoted by $t_\theta(s, a)$, where $t_\theta = N_\theta \mathbf{1}$, and $N_\theta = (I - P_{\pi_\theta})^{-1}$, and the variance of the number of steps before being absorbed, when starting in state (s, a), is given by the (s, a)th element of $(2N_\theta - I)t_\theta - t_\theta \odot t_\theta$, where $\odot$

denotes the Hadamard product, see e.g. [6, Thm 3.3.5]. Let $((2N_\theta - I)t_\theta)_{(s,a)}$ denote the (s,a)th element of $(2N_\theta - I)t_\theta$. Then,

$$
\begin{aligned}
\mathrm{E}_\theta\left[\Gamma_\theta^2\right] &= \mathrm{E}_\theta\left[\mathrm{E}_\theta\left[\Gamma_\theta^2 \mid S_0^{(t+1)}, A_0^{(t+1)}\right]\right] \\
&= \mathrm{E}_\theta\left[\mathrm{Var}_\theta\left(\Gamma_\theta \mid S_0^{(t+1)}, A_0^{(t+1)}\right)\right] + \mathrm{E}_\theta\left[\mathrm{E}_\theta\left[\Gamma_\theta \mid S_0^{(t+1)}, A_0^{(t+1)}\right]^2\right] \\
&= \mathrm{E}_\theta\left[((2N_\theta - I)t_\theta)_{(S_0^{(t+1)}, A_0^{(t+1)})} - t_\theta(S_0^{(t+1)}, A_0^{(t+1)})^2\right] + \mathrm{E}_\theta\left[t_\theta(S_0^{(t+1)}, A_0^{(t+1)})^2\right] \\
&= \lambda_{\pi_\theta}^\top (2N_\theta - I)t_\theta \leq \|\lambda_{\pi_\theta}\|\|(2N_\theta - I)t_\theta\| \leq \|\lambda_{\pi_\theta}\|_1\|2N_\theta - I\|\|t_\theta\| \\
&\leq (2\|N_\theta\| + \|I\|)\|N_\theta\|\|\mathbf{1}\| \leq (2\|(I - P_{\pi_\theta})^{-1}\| + 1)\|(I - P_{\pi_\theta})^{-1}\|\sqrt{|\mathcal{S}||\mathcal{A}|} \\
&\leq \sqrt{|\mathcal{S}||\mathcal{A}|}(2\zeta + 1)\zeta.
\end{aligned}
$$

To conclude,

$$
\mathrm{E}_\theta\left[\|H(\theta, X_{t+1})\|^2\right] \leq 2\Phi_{\max}^2\sqrt{|\mathcal{S}||\mathcal{A}|}(2\zeta + 1)\zeta(r_{\max}^2 + 4\Phi_{\max}^2\|\theta\|^2),
$$

i.e. there exists a constant K such that (5) is satisfied.

4.5.3. *Stability condition.* First, note that

$$
h(\theta) := \mathrm{E}_\theta[H(\theta, X_{t+1})] = \mathrm{E}_\theta\left[\sum_{u=0}^{T^{(t+1)}-1} \phi_u^{(t+1)}(r_u^{(t+1)} + (\phi_{u+1}^{(t+1)})^\top\theta - (\phi_u^{(t+1)})^\top\theta)\right]
$$
$$
= \Phi^\top D_{\pi_\theta} r + \Phi^\top D_{\pi_\theta}(P_{\pi_\theta} - I)\Phi\theta = b_{\pi_\theta} + A_{\pi_\theta}\theta.
$$

This can be shown similarly to the proof of Proposition 6.6 in Bertsekas & Tsitsiklis [2]. The proof is included in Appendix C for completeness.

By Lemmas 4.5 and 4.7, there exists θ^* such that $A_{\pi_{\theta^*}}\theta^* + b_{\pi_{\theta^*}} = 0$, hence

$$
\begin{aligned}
(\theta - \theta^*)^\top h(\theta) &= (\theta - \theta^*)^\top(A_{\pi_\theta}\theta + b_{\pi_\theta}) = (\theta - \theta^*)^\top(A_{\pi_\theta}\theta - A_{\pi_{\theta^*}}\theta^* + b_{\pi_\theta} - b_{\pi_{\theta^*}}) \\
&= (\theta - \theta^*)^\top A_{\pi_\theta}(\theta - \theta^*) + (\theta - \theta^*)^\top(A_{\pi_\theta} - A_{\pi_{\theta^*}})\theta^* + (\theta - \theta^*)^\top(b_{\pi_\theta} - b_{\pi_{\theta^*}}) \\
&\leq (\theta - \theta^*)^\top A_{\pi_\theta}(\theta - \theta^*) + \|\theta - \theta^*\|\|A_{\pi_\theta} - A_{\pi_{\theta^*}}\|\|\theta^*\| + \|\theta - \theta^*\|\|b_{\pi_\theta} - b_{\pi_{\theta^*}}\|.
\end{aligned}
$$

By Lemma 4.10 and Assumption 4.7 we obtain

$$
\begin{aligned}
(\theta - \theta^*)^\top h(\theta) &\leq (\theta - \theta^*)^\top A_{\pi_\theta}(\theta - \theta^*) + C_1\|\theta - \theta^*\|^2\|\theta^*\| + C_2\|\theta - \theta^*\|^2 \\
&= (\theta - \theta^*)^\top(A_{\pi_\theta} + (C_1\|\theta^*\| + C_2)I)(\theta - \theta^*).
\end{aligned}
$$

By Lemma 4.3 A_{π_θ} is negative definite. Hence, for C_1 and C_2 sufficiently small, $A_{\pi_\theta} + (C_1\|\theta^*\| + C_2)I$ is negative definite, i.e. the stability condition (6) is satisfied.

Hence, since the Robbins-Monro assumption (4), the square integrability condition (5), and the stability condition (6) are satisfied, Theorem 4.1 follows.

5. Discussion

We have shown that if the behaviour policy is ε-soft and Lipschitz continuous w.r.t. the weight vector, with small enough Lipschitz constant, then SARSA with linear function approximation will converge with probability one when considering a random horizon MDP. This is in line with earlier convergence results for infinite horizon discounted MDPs in [8, 7].

For the variant of SARSA considered here, the weight vector and the behaviour policy are only updated at the end of each trajectory, not after each iteration. This variant of the algorithm should work well if the trajectories are not too long, but could cause slow convergence in practice in the case of very long trajectories. However, for a random horizon MDP with very long trajectories, it is possible that earlier results for infinite horizon MDPs will hold if replacing the stationary distribution of the Markov chain induced by a policy π with a quasi-stationary distribution, at least for the discounted version of the problem. Obtaining convergence results for the online version of the algorithm, where the weight vector and policy are updated after each iteration, could still be of interest for problems with trajectories of medium length, i.e. too short for the existence of a quasi-stationary distribution, but long enough to cause slow convergence in practice. It is possible that Theorem 17, p. 239, in Benveniste et al. [1] could be used to prove convergence in this case, similarly to what is done in [7], but it would be more complex to ascertain if the various assumptions in this theorem are satisfied, since the Markov chain induced by a policy is not ergodic in the random horizon case.

Furthermore, the theorem in this paper suffers from the same limitations as the theorems obtained in the infinite horizon discounted case, discussed in Perkins & Precup [8] and Melo et al. [7]. As described in [8], the value of C_0 in Theorem 1 is not specified in the theorem, and depends on the properties of the MDP, which might be unknown (e.g. transition probabilities). Moreover, there is no guarantee on how close the approximation is to the true optimal action-value function and the true optimal policy. As discussed in [7], to approximate the true optimal action-value function and policy well, the behaviour policy over time needs to approach the greedy policy, by e.g. having a decaying exploration rate. This would, however, lead to an increased Lipschitz constant (since the greedy policy is discontinuous), hence the condition that the Lipschitz constant is sufficiently small might no longer hold.

References

[1] Albert Benveniste, Michel Métivier, and Pierre Priouret. *Adaptive algorithms and stochastic approximations.* Springer, 1990.

[2] Dimitri P Bertsekas and John N Tsitsiklis. *Neuro-dynamic programming.* Athena Scientific, 1996.

[3] Peter Dayan. The convergence of TD(λ) for general λ. *Machine learning*, 8(3-4):341–362, 1992.

[4] Daniela Pucci De Farias and Benjamin Van Roy. On the existence of fixed points for approximate value iteration and temporal-difference learning. *Journal of Optimization theory and Applications*, 105(3):589–608, 2000.

[5] Geoffrey J Gordon. Reinforcement learning with function approximation converges to a region. In *Advances in neural information processing systems*, pages 1040–1046, 2001.

[6] John G Kemeny and James Laurie Snell. *Finite Markov chains: with a new appendix "Generalization of a fundamental matrix"*. Springer, 1976.

[7] Francisco S Melo, Sean P Meyn, and M Isabel Ribeiro. An analysis of reinforcement learning with function approximation. In *Proceedings of the 25th international conference on Machine learning*, pages 664–671, 2008.

[8] Theodore J Perkins and Doina Precup. A convergent form of approximate policy iteration. In *Advances in neural information processing systems*, pages 1627–1634, 2003.

[9] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2005.

[10] Satinder Singh, Tommi Jaakkola, Michael L Littman, and Csaba Szepesvári. Convergence results for single-step on-policy reinforcement-learning algorithms. *Machine learning*, 38(3):287–308, 2000.

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

[12] 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, 1997.

[13] Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.

Appendix A. Norms and norm inequalities

The following norm definitions are used:

- For any vector $x \in \mathbb{R}^n$

$$\|x\| = \sqrt{\sum_{i=1}^{n} x_i^2} \quad \text{(Euclidean norm)},$$

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|,$$

$$\|x\|_\infty = \max_i |x_i| \quad \text{(infinity norm)}.$$

- For any matrix $A \in \mathbb{R}^{m \times n}$

$$\|A\| = \sqrt{\lambda_{\max}(A^\top A)} \quad \text{(spectral norm)},$$

$$\|A\|_\infty = \max_i \sum_{j=1}^n |A_{i,j}| \quad \text{(maximum absolute row sum norm)},$$

$$\|A\|_{\max} = \max_{i,j} |A_{i,j}| \quad \text{(max norm)}.$$

We also use the following well known norm equivalences:

- For any vector $x \in \mathbb{R}^n$

$$\|x\|_\infty \le \|x\| \le \|x\|_1.$$

- For any matrix $A \in \mathbb{R}^{m \times n}$

$$\|A\| \le \sqrt{mn}\|A\|_{\max},$$
$$\|A\| \le \sqrt{m}\|A\|_\infty.$$

Appendix B. Proofs of Lemmas 4.3, 4.5, 4.6-4.7, and 4.8

B.1. **Proof of Lemma 4.3.** The proof of Lemma 4.3 is identical to the last part of the proof of Lemma 6.10 in Bertsekas & Tsitsiklis [2], but here considering a Markov chain over state-action pairs, with P_π (P in [2]) and D_π (Q in [2]) defined accordingly, and the policy π being ε-soft.

Proof. For any ε-soft policy π, similarly to what is shown in the proof of Lemma 4.2, but without requiring Assumption 4.6,

$$\|P_\pi q\|_{\eta_\pi}^2 \le \sum_{s',a'} q(s',a')^2(\eta_\pi(s',a') - \lambda_\pi(s',a')) \le \sum_{s',a'} q(s',a')^2 \eta_\pi(s',a') = \|q\|_{\eta_\pi}^2,$$

for all $q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$. Hence

$$q^\top D_\pi P_\pi q \le \|q\|_{\eta_\pi}\|P_\pi q\|_{\eta_\pi} \le \|q\|_{\eta_\pi}^2 = q^\top D_\pi q.$$

For this inequality to be an equality, we need q and $P_\pi q$ to be colinear, and $\|P_\pi q\|_{\eta_\pi} = \|q\|_{\eta_\pi}$. Thus, the inequality is strict unless $P_\pi q = q$ or $P_\pi q = -q$, which means that $P_\pi^{2m} q = q$ for all $m \ge 0$. Since the policy is proper, P_π^{2m} converges to zero, hence we must have $q = 0$, i.e. the inequality is strict for all $q \ne 0$. Hence, the matrix $D_\pi(P_\pi - I)$ is negative definite. Using Assumption 4.2(i) (linear independence of columns of Φ) A_π is negative definite. $\square$

B.2. **Proof of Lemma 4.5.** The proof of Lemma 4.5 is identical to the proof of Lemma 5.3 in de Farias & Van Roy [4], but using our definitions of s_{π_θ}, $F_{\pi_\theta}^\alpha$, T_{π_θ} and D_{π_θ} (s_δ, F_δ^γ, T_δ, and $D_{\mu_r^\delta}$ in [4]), and that the policy π is ε-soft. We also use that D_π is positive definite for any ε-soft policy π under Assumption 4.4.

Proof. Let θ be a fixed point of F_π^α. Then $s_\pi(\theta) = 0$, hence

$$\Phi^\top D_\pi \Phi \theta = \Phi^\top D_\pi (r + P_\pi \Phi \theta),$$

$$\Phi(\Phi^\top D_\pi \Phi)^{-1} \Phi^\top D_\pi \Phi \theta = \Phi(\Phi^\top D_\pi \Phi)^{-1} \Phi^\top D_\pi (r + P_\pi \Phi \theta),$$

$$\Phi \theta = \Pi_\pi T_\pi \Phi \theta,$$

and $\Phi \theta$ is a fixed point of H_π. The reverse can be shown by reversing the steps, and an entirely analogous argument can be used to show that θ is a fixed point of $F_{\pi_\theta}^\alpha$ if and only if $\Phi \theta$ is a fixed point of H_{π_θ}. $\qquad\square$

B.3. **Proof of Lemma 4.6.** The proof of Lemma 4.6 is essentially identical to the proof of Lemma 5.4 in de Farias & Van Roy [4], but using our definition of H_π and using a ε-soft policy π.

Proof. We have

$$\|F_\pi^\alpha(\theta) - \theta_\pi\|^2 = \|\theta + \alpha s_\pi(\theta) - \theta_\pi\|^2 = \|\theta - \theta_\pi\|^2 + 2\alpha(\theta - \theta_\pi)^\top s_\pi(\theta) + \alpha^2 \|s_\pi(\theta)\|^2.$$

For the second term, note that (using Lemma 4.2 and that H_π is a contraction) for all ε-soft π, there exists $\beta \in [0, 1)$ such that

$$(9) \qquad \|H_\pi \Phi \theta - \Phi \theta_\pi\|_{\eta_\pi} = \|H_\pi \Phi \theta - H_\pi \Phi \theta_\pi\|_{\eta_\pi} \le \beta \|\Phi \theta - \Phi \theta_\pi\|_{\eta_\pi},$$

and furthermore that

$$\begin{aligned}
(\theta - \theta_\pi)^\top s_\pi(\theta) &= (\theta - \theta_\pi)^\top \Phi^\top D_\pi (T_\pi \Phi \theta - \Phi \theta) \\
&= (\theta - \theta_\pi)^\top \Phi^\top D_\pi \Phi (\Phi^\top D_\pi \Phi)^{-1} \Phi^\top D_\pi (T_\pi \Phi \theta - \Phi \theta) \\
&= (\Phi \theta - \Phi \theta_\pi)^\top D_\pi (\Phi(\Phi^\top D_\pi \Phi)^{-1} \Phi^\top D_\pi T_\pi \Phi \theta - \Phi \theta) \\
&= (\Phi \theta - \Phi \theta_\pi)^\top D_\pi (H_\pi \Phi \theta - \Phi \theta) = \langle \Phi \theta - \Phi \theta_\pi, H_\pi \Phi \theta - \Phi \theta \rangle_{\eta_\pi},
\end{aligned}$$

where $\langle \cdot, \cdot \rangle_{\eta_\pi}$ denotes the weighted inner product, i.e. $\langle x, y \rangle_{\eta_\pi} = x^\top D_\pi y$. Now, using (9),

$$\begin{aligned}
\langle \Phi \theta - \Phi \theta_\pi, H_\pi \Phi \theta - \Phi \theta \rangle_{\eta_\pi} &= \langle \Phi \theta - \Phi \theta_\pi, (H_\pi \Phi \theta - \Phi \theta_\pi) - (\Phi \theta_\pi - \Phi \theta) \rangle_{\eta_\pi} \\
&= \langle \Phi \theta - \Phi \theta_\pi, H_\pi \Phi \theta - \Phi \theta_\pi \rangle_{\eta_\pi} - \|\Phi \theta - \Phi \theta_\pi\|_{\eta_\pi}^2 \\
&\le \|\Phi \theta - \Phi \theta_\pi\|_{\eta_\pi} \|H_\pi \Phi \theta - \Phi \theta_\pi\|_{\eta_\pi} - \|\Phi \theta - \Phi \theta_\pi\|_{\eta_\pi}^2 \\
&\le (\beta - 1) \|\Phi \theta - \Phi \theta_\pi\|_{\eta_\pi}^2 = (\beta - 1)(\theta - \theta_\pi)^\top \Phi^\top D_\pi \Phi (\theta - \theta_\pi).
\end{aligned}$$

Hence,

$$(\theta - \theta_\pi)^\top s_\pi(\theta) \le (\beta - 1) \|\theta - \theta_\pi\| \|\Phi^\top D_\pi \Phi (\theta - \theta_\pi)\| \le (\beta - 1) \|\Phi^\top D_\pi \Phi\| \|\theta - \theta_\pi\|^2.$$

Since D_π is a positive definite matrix for all ε-soft policies π, $\Phi^\top D_\pi \Phi$ is positive definite and symmetric, hence $\|\Phi^\top D_\pi \Phi\| > 0$. It follows that there exists a constant $C_1 > 0$ such that

$$(\theta - \theta_\pi)^\top s_\pi(\theta) \le -C_1 \|\theta - \theta_\pi\|^2,$$

namely $C_1 = (1 - \beta) \max_{\pi \in \Delta_\varepsilon} \|\Phi^\top D_\pi \Phi\|$, where the maximum is attained since (by Lemma 4.1) D_π is a continuous function of π, and the set of all ε-soft policies is compact.

Note that $\Phi^\top D_\pi \Pi_\pi = \Phi^\top D_\pi \Phi(\Phi^\top D_\pi \Phi)^{-1}\Phi^\top D_\pi = \Phi^\top D_\pi$. Furthermore, let ϕ_i be the ith column of Φ. Then

$$\|s_\pi(\theta)\|^2 = \|\Phi^\top D_\pi(T_\pi\Phi\theta - \Phi\theta)\|^2 = \sum_{i=1}^d \left(\phi_i^\top D_\pi(T_\pi\Phi\theta - \Phi\theta)\right)^2$$

$$= \sum_{i=1}^d \left(\phi_i^\top D_\pi(\Pi_\pi T_\pi\Phi\theta - \Phi\theta)\right)^2 \leq \sum_{i=1}^d \|\phi_i\|_{\eta_\pi}^2 \|\Pi_\pi T_\pi\Phi\theta - \Phi\theta\|_{\eta_\pi}^2$$

$$\leq \sum_{i=1}^d \|\phi_i\|_{\eta_\pi}^2 (\|\Pi_\pi T_\pi\Phi\theta - \Phi\theta_\pi\|_{\eta_\pi} + \|\Phi\theta_\pi - \Phi\theta\|_{\eta_\pi})^2$$

$$\leq \sum_{i=1}^d \|\phi_i\|_{\eta_\pi}^2 (\beta\|\Phi\theta - \Phi\theta_\pi\|_{\eta_\pi} + \|\Phi\theta_\pi - \Phi\theta\|_{\eta_\pi})^2$$

$$= (\beta+1)^2 \sum_{i=1}^d \|\phi_i\|_{\eta_\pi}^2 \|\Phi\theta_\pi - \Phi\theta\|_{\eta_\pi}^2,$$

and (similarly to above) it follows that there exists a constant $C_2 > 0$ (independent of π) such that $\|s_\pi(\theta)\|^2 \leq C_2\|\theta - \theta_\pi\|^2$. Hence

$$\|F_\pi^\alpha(\theta) - \theta_\pi\|^2 \leq \|\theta - \theta_\pi\|^2 + 2\alpha(\theta - \theta_\pi)^\top s_\pi(\theta) + \alpha^2\|s_\pi(\theta)\|^2$$

$$\leq (1 - 2\alpha C_1 + \alpha^2 C_2)\|\theta - \theta_\pi\|^2.$$

Thus, with $\alpha^* = 2C_1/C_2$ (independent of π) and $\alpha \in (0, \alpha^*)$, we see that

$$1 - 2\alpha C_1 + \alpha^2 C_2 = 1 + C_2\alpha(\alpha - \alpha^*) < 1,$$

i.e. there exists $\beta_\alpha \in (0, 1)$ such that

$$\|F_\pi^\alpha(\theta) - \theta_\pi\|^2 \leq \beta_\alpha\|\theta - \theta_\pi\|^2.$$

$\square$

B.4. **Proof of Lemma 4.7.** The proof of Lemma 4.7 is essentially identical to the proof of Lemma 5.5 in de Farias & Van Roy [4], but using our definitions of $F_{\pi_\theta}^\alpha$ (F_δ^γ in [4]). Furthermore, [4] use that the set of all stochastic policies is compact, we instead use that the set of ε-soft policies is compact. Moreover, in our case it is the Lipschitz continuity of π_θ that implies that $F_{\pi_\theta}^\alpha$ is continuous in θ, rather than the specific choice of behaviour policy (softmax policy) considered in [4].

Proof. By Lemma 4.4, θ_π is a continuous function of π. Since the set of ε-soft policies, Δ_ε, is compact, the set $\Theta = \{\theta_\pi : \pi \in \Delta_\varepsilon\}$ is also compact. Let $\Theta_{\max} = \max\{\|\theta\| : \theta \in \Theta\}$.

Note that if we establish that a fixed point exists for some $\alpha > 0$, then by Lemma 4.5 this fixed point is also a fixed point for all other $\alpha > 0$. Using Lemma 4.6, we can choose $\alpha > 0$ such that there is a $\beta \in (0, 1)$ with

$$\|F_\pi^\alpha(\theta) - \theta_\pi\| \leq \beta\|\theta - \theta_\pi\|,$$

for all ε-soft π. Then

$$\|F_{\pi_\theta}^\alpha(\theta)\| \le \|F_{\pi_\theta}^\alpha(\theta) - \theta_{\pi_\theta}\| + \|\theta_{\pi_\theta}\| \le \beta\|\theta - \theta_{\pi_\theta}\| + \Theta_{\max} \le \beta\|\theta\| + (\beta+1)\Theta_{\max}.$$

Hence, the set $\bar\Theta = \{\theta : \|\theta\| \le (1+\beta)\Theta_{\max}/(1-\beta)\}$ is closed under $F_{\pi_\theta}^\alpha$, since, if $\theta \in \bar\Theta$, then by the above

$$\|F_{\pi_\theta}^\alpha(\theta)\| \le \beta\|\theta\| + (\beta+1)\Theta_{\max} \le \beta\frac{1+\beta}{1-\beta}\Theta_{\max} + (1+\beta)\Theta_{\max} = \frac{1+\beta}{1-\beta}\Theta_{\max},$$

i.e. $F_{\pi_\theta}^\alpha(\theta) \in \bar\Theta$. Using this, and that $F_{\pi_\theta}^\alpha$ is a continuous function of θ (since π_θ is Lipschitz continuous w.r.t. θ, P_π is a continuous function of π, D_π is a continuous function of π by Lemma 4.1, hence T_{π_θ} is a continuous function of θ, which implies that $F_{\pi_\theta}^\alpha$ is a continuous function of θ), by the Brouwer fixed point theorem $F_{\pi_\theta}^\alpha$ possesses a fixed point. $\qquad\square$

B.5. **Proof of Lemma 4.8.** The proof of Lemma 4.8 is essentially identical to the proof of Lemma 1 in Perkins & Precup [8], but using our definition of P_π, and correcting what appears to be a minor error in the proof, namely that $\|A\| \le n\|A\|_{\max}$ for $A \in \mathbb{R}^{n\times n}$, rather than $\|A\| \le \sqrt{n}\|A\|_{\max}$.

Proof. Let π_1 and π_2 be fixed, and let i and j correspond to the (s,a)th row and (s',a')th column of P_π, respectively. Then

$$\|P_{\pi_1} - P_{\pi_2}\| \le \sqrt{(|\mathcal{S}||\mathcal{A}|)^2}\|P_{\pi_1} - P_{\pi_2}\|_{\max} = |\mathcal{S}||\mathcal{A}|\max_{i,j}|(P_{\pi_1})_{i,j} - (P_{\pi_2})_{i,j}|$$

$$= |\mathcal{S}||\mathcal{A}|\max_{s,a,s',a'}|p(s'|s,a)(\pi_1(a'|s') - \pi_2(a'|s'))| \le |\mathcal{S}||\mathcal{A}|\max_{s',a'}|\pi_1(a'|s') - \pi_2(a'|s')|$$

$$= |\mathcal{S}||\mathcal{A}|\|\pi_1 - \pi_2\|_\infty \le |\mathcal{S}||\mathcal{A}|\|\pi_1 - \pi_2\|,$$

i.e. $C_P = |\mathcal{S}||\mathcal{A}|$. $\qquad\square$

APPENDIX C. PROOF THAT $h(\theta) = b_{\pi_\theta} + A_{\pi_\theta}\theta$

Using the convention $\phi_u^{(t)} = 0$ for $u \ge T^{(t)}$,

$$\mathrm{E}_\theta\left[\sum_{u=0}^{T^{(t+1)}-1}\phi_u^{(t+1)}(\phi_u^{(t+1)})^\top\right] = \mathrm{E}_\theta\left[\sum_{u=0}^\infty \phi(S_u^{(t+1)}, A_u^{(t+1)})\phi(S_u^{(t+1)}, A_u^{(t+1)})^\top\right]$$

$$= \sum_{u=0}^\infty \mathrm{E}_\theta[\phi(S_u^{(t+1)}, A_u^{(t+1)})\phi(S_u^{(t+1)}, A_u^{(t+1)})^\top]$$

$$= \sum_{u=0}^\infty \sum_{s\in\mathcal{S}, a\in\mathcal{A}} P_{\pi_\theta}(S_u^{(t+1)} = s, A_u^{(t+1)} = a)\phi(s,a)\phi(s,a)^\top$$

$$= \sum_{s\in\mathcal{S}, a\in\mathcal{A}} \eta_{\pi_\theta}(s,a)\phi(s,a)\phi(s,a)^\top = \Phi^\top D_{\pi_\theta}\Phi,$$

$$\mathrm{E}_\theta\left[\sum_{u=0}^{T^{(t+1)}-1}\phi_u^{(t+1)}(\phi_{u+1}^{(t+1)})^\top\right] = \mathrm{E}_\theta\left[\sum_{u=0}^{\infty}\phi(S_u^{(t+1)},A_u^{(t+1)})\phi(S_{u+1}^{(t+1)},A_{u+1}^{(t+1)})^\top\right]$$

$$= \sum_{u=0}^{\infty}\mathrm{E}_\theta[\phi(S_u^{(t+1)},A_u^{(t+1)})\phi(S_{u+1}^{(t+1)},A_{u+1}^{(t+1)})^\top]$$

$$= \sum_{u=0}^{\infty}\sum_{s,s'\in\mathcal{S},a,a'\in\mathcal{A}}P_{\pi_\theta}(S_u^{(t+1)}=s,A_u^{(t+1)}=a)p_{\pi_\theta}(s',a'\mid s,a)\phi(s,a)\phi(s',a')^\top$$

$$= \sum_{s,s'\in\mathcal{S},a,a'\in\mathcal{A}}\eta_{\pi_\theta}(s,a)p_{\pi_\theta}(s',a'\mid s,a)\phi(s,a)\phi(s',a')^\top = \Phi^\top D_{\pi_\theta}P_{\pi_\theta}\Phi,$$

and

$$\mathrm{E}_\theta\left[\sum_{u=0}^{T^{(t+1)}-1}\phi_u^{(t+1)}r_u^{(t+1)}\right] = \mathrm{E}_\theta\left[\sum_{u=0}^{\infty}\phi(S_u^{(t+1)},A_u^{(t+1)})r(S_u^{(t+1)},A_u^{(t+1)},S_{u+1}^{(t+1)})^\top\right]$$

$$= \sum_{u=0}^{\infty}\sum_{s\in\mathcal{S},a\in\mathcal{A},s'\in\mathcal{S}^+}P_{\pi_\theta}(S_u^{(t+1)}=s,A_u^{(t+1)}=a)p(s'\mid s,a)\phi(s,a)r(s,a,s')$$

$$= \sum_{s\in\mathcal{S},a\in\mathcal{A}}\sum_{u=0}^{\infty}P_{\pi_\theta}(S_u^{(t+1)}=s,A_u^{(t+1)}=a)\phi(s,a)\sum_{s'\in\mathcal{S}^+}p(s'\mid s,a)r(s,a,s')$$

$$= \sum_{s\in\mathcal{S},a\in\mathcal{A}}\eta_{\pi_\theta}(s,a)\phi(s,a)r(s,a) = \Phi^\top D_{\pi_\theta}r.$$

Thus

$$h(\theta) = \Phi^\top D_{\pi_\theta}r + \Phi^\top D_{\pi_\theta}(P_{\pi_\theta}-I)\Phi\theta = b_{\pi_\theta} + A_{\pi_\theta}\theta.$$

Email address: lina.palmborg@math.su.se