Reinforcement Learning

Model-Free RL

Lecturer: *Nicolò Cesa-Bianchi* version February 11, 2024

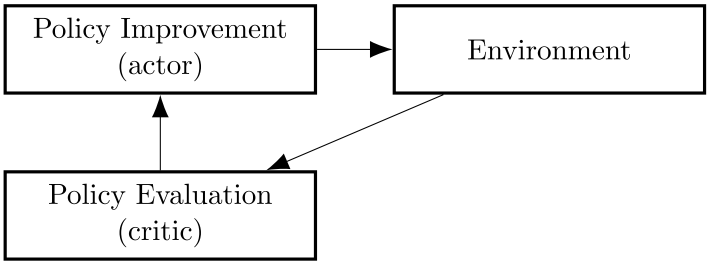
This material is partially based on the book draft "Reinforcement Learning: Foundations" by Shie Mannor, Yishay Mansour, and Aviv Tamar.

Generally speaking, we can distinguish two main tasks in RL:

1. **Prediction** is concerned with computing V^π given π. Hence, we want to measure how good is a policy with respect to a certain evaluation criterion (e.g., finite, infinite or discounted horizon). This is akin to estimating the statistical risk of a predictor in supervised learning.
2. **Control** is concerned with learning the optimal policy π^*. This is akin to learning the Bayes optimal predictor in supervised learning.

In model-free RL, we avoid learning the structure of the MDP. Rather, we directly learn the optimal policy by interacting with the MDP. There are two main approaches: methods based on policy iteration and methods based on value iteration.

- **Policy iteration methods:** Recall that policy iteration methods loop over two phases: policy evaluation, where V^π is computed for the current policy π, and policy improvement, where π is updated. Using a traditional terminology, we call **critic** the block that performs policy evaluation and **actor** the block that performs policy improvement. The main algorithm used to implement the critic block is TD(λ).



- **Value iteration methods:** These methods use online versions of value iteration. They can be off-policy (Q-learning) when they learn the optimal policy by observing the trajectory generated by a different policy, or on-policy (SARSA) when the policy generating the trajectories converges to the optimal policy.

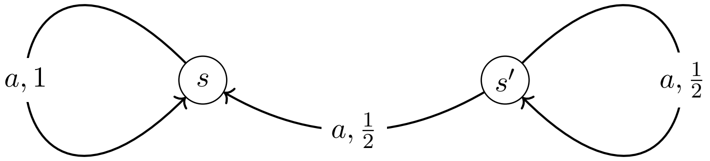
We consider the discounted infinite horizon criterion and focus on MDP with finite state space $\mathcal{S}$, finite action space $\mathcal{A}$ such that $\mathcal{A}(s) = \mathcal{A}$ for all $s \in \mathcal{S}$, transition kernel $\{p(\cdot \mid s, a) : s \in \mathcal{S}, a \in \mathcal{A}\}$, and time-independent reward function $r : \mathcal{S} \times \mathcal{A} \to [-1, 1]$.

A stochastic policy $\boldsymbol{\pi}$ is **fully mixed** if $\pi_t(a \mid s) > 0$ for all $t \geq 0$ and all $(s, a) \in \mathcal{S} \times \mathcal{A}$.

A Markov chain is **irreducible** if for any two states s, s', there is a non-zero probability of going from s to s' in a finite number of steps. As a consequence, every state of an irreducible Markov chain is visited infinitely often with probability 1.

An MDP is **communicating** if the Markov chain induced by any stationary fully mixed policy is irreducible.

The MDP below here, with two states s, s' and a single action a is non-communicating because state s is absorbing: the probability of going from s to s' in any number of steps is zero.



The next result states that if any stationary fully mixed policy ensures that all states are visited infinitely often, then even nonstationary fully mixed policies have the same guarantee.

Theorem 1 *The Markov chain induced on a communicating MDP by any (possibly nonstationary) fully mixed policy is irreducible.*

By definition of fully mixed policy, we also have that every state-action pairs $(s, a) \in \mathcal{S} \times \mathcal{A}$ occurs infinitely often. This will be key to prove the convergence of Q-learning and SARSA.

Recall the action-value function $Q^\pi : \mathcal{S} \times \mathcal{A} \to [-1, 1]$ for a stationary Markov policy π,

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

This is the expected return of executing action a in state s and then following policy π.

Similarly to the Bellman system of equations for the optimal state-value function V^*,

$$V^*(s) = \max_{a \in \mathcal{A}} \left(r(s, a) + \gamma \sum_{s' \in \mathcal{S}} p(s' \mid s, a) V^*(s') \right) \qquad s \in \mathcal{S}$$

we can define a corresponding system for the optimal action-value function $Q^* = Q^{\pi^*}$,

$$Q^*(s, a) = r(s, a) + \gamma \sum_{s' \in \mathcal{S}} p(s' \mid s, a) V^*(s') \qquad (s, a) \in \mathcal{S} \times \mathcal{A}$$

This is the expected return of executing action a in state s and then following the optimal policy π^*. Clearly, Q^* gives access to π^* because

$$\pi^*(s) \in \operatorname*{argmax}_{a \in \mathcal{A}} Q^*(s, a) \qquad s \in \mathcal{S}$$

Also, comparing the definitions of V^* and Q^* we get

$$V^*(s) = \max_{a \in \mathcal{A}} Q^*(s, a) \qquad s \in \mathcal{S}$$

and so we obtain

$$Q^*(s,a) = r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s' \mid s,a) \max_{a' \in \mathcal{A}} Q^*(s',a') \qquad (s,a) \in \mathcal{S} \times \mathcal{A}$$

The above is equivalent to

$$Q^*(s,a) - r(s,a) - \gamma \, \mathbb{E}\left[\max_{a' \in \mathcal{A}} Q^*(s',a') \,\middle|\, s,a \right] = 0 \tag{1}$$

where the expectation is with respect to the draw of the next state s' according to the distribution $p(\cdot \mid s,a)$.

Q-learning. We now study the problem of learning Q^* without knowing the transition kernel $p(\cdot \mid s,a)$. Let Q_t be the current guess for Q^*. Given any sequence of state-action pairs (s_t,a_t), we could use the identity (1) and run gradient descent with respect the square loss function

$$\ell_t(x) = \frac{1}{2}\left(x - r(s_t,a_t) - \gamma \, \mathbb{E}\left[\max_{a \in \mathcal{A}} Q_t(s,a) \,\middle|\, s_t,a_t \right] \right)^2$$

The gradient descent step is $x_{t+1} = x_t - \eta_t \frac{d}{dx}\ell_t(x_t)$, which for $x_t = Q_t(s_t,a_t)$ takes the form

$$\begin{aligned} Q_{t+1}(s_t,a_t) &= Q_t(s_t,a_t) - \eta_t \left(Q_t(s_t,a_t) - r(s_t,a_t) - \gamma \, \mathbb{E}\left[\max_{a \in \mathcal{A}} Q_t(s,a) \,\middle|\, s_t,a_t \right] \right) \\ &= (1-\eta_t) Q_t(s_t,a_t) + \eta_t \left(r(s_t,a_t) + \gamma \, \mathbb{E}\left[\max_{a \in \mathcal{A}} Q_t(s,a) \,\middle|\, s_t,a_t \right] \right) \end{aligned}$$

This looks fine, except that we cannot compute the expectation because the transition function is unknown. The solution is to run gradient descent on a **perturbed gradient**, in which the conditional expectation

$$\mathbb{E}\left[\max_{a \in \mathcal{A}} Q_t(s,a) \,\middle|\, s_t,a_t \right]$$

is replaced by $\max_{a \in \mathcal{A}} Q_t(s_{t+1},a)$ where s_{t+1} is drawn from $p(\cdot \mid s_t,a_t)$.

Algorithm 1 (Q-learning)

```
Input: Fully mixed policy $\pi : \mathcal{S} \to \Delta_{\mathcal{A}}$, initial state $s_0 \in \mathcal{S}$
Set $Q_0(s,a) = 0$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$
for $t = 0, 1, \dots$ do
    Observe action $a_t \sim \pi(\cdot \mid s_t)$, reward $r_t = r(s_t,a_t)$ and next state $s_{t+1}$ drawn from $p(\cdot \mid s_t,a_t)$
    Update $Q_{t+1}(s_t,a_t) = (1-\eta_t) Q_t(s_t,a_t) + \eta_t \Big( r(s_t,a_t) + \gamma \max_{a \in \mathcal{A}} Q_t(s_{t+1},a) \Big)$
end for
```

Note that we learn Q^* by running an arbitrary fully mixed policy π. For this reason, Q-learning is called an **off-policy** learning algorithm.

The proof of convergence of Q-learning (and SARSA) is based on this result from the field of iterative stochastic approximation.

Lemma 2 *Let $\mathcal{X}$ be a finite set and consider the iterates $F_0, F_1, \ldots$, where $F_t : \mathcal{X} \to \mathbb{R}$ for all $t \geq 0$, $F_0(x) = 0$ for all $x \in \mathcal{X}$, and*

$$F_{t+1}(x) = \big(1 - \eta_t(x)\big)F_t(x) + \eta_t(x)\Big(H(F_t)(x) + \nu_t + \zeta_t\Big)$$

for some operator H on the space of functions $F : \mathcal{X} \to \mathbb{R}$, for some sequence $\eta_0, \eta_1, \ldots$ such that $\eta_t : \mathcal{X} \to [0,1]$ for all $t \geq 0$, and for random variables ν_t and ζ_t for $t \geq 0$. If the following properties hold:

1. ***Stepsize*** *For every $x \in \mathcal{X}$,*
$$\sum_{t \geq 0} \eta_t(x) = \infty \qquad \text{and} \qquad \sum_{t \geq 0} \eta_t(x)^2 < \infty$$
2. ***Noise*** *For all $t \geq 0$, $\mathbb{E}\big[\nu_t \mid \nu_0, \zeta_0, \ldots, \nu_{t-1}, \zeta_{t-1}\big] = 0$ and $|\nu_t| \leq M$*
3. ***Bias*** *$\lim_{t \to \infty} \zeta_t = 0$ with probability 1*
4. ***Contraction*** *There exist F^* and $0 \leq \gamma < 1$ such that for any F we have $\|H(F) - F^*\|_\infty \leq \gamma \|F - F^*\|_\infty$*

Then

$$\lim_{t \to \infty} F_t(x) = F^*(x) \qquad x \in \mathcal{X}$$

with probability 1.

We now prove the convergence of Q-learning when η_t is a function $\eta_t : \mathcal{S} \times \mathcal{A} \to [0,1]$ of the state-action pairs defined by

$$\eta_t(s,a) = \frac{\mathbb{I}\{s = s_t, a = a_t\}}{N_t(s,a)} \qquad \text{where} \qquad N_t(s,a) = \sum_{\tau=0}^{t} \mathbb{I}\{s_\tau = s, a_\tau = a\} \tag{2}$$

where $\eta_t(s,a) = 0$ for $(s,a) \neq (s_t, a_t)$.

Theorem 3 *Assume that Q-learning is run with a fully mixed policy π on a communicating MDP. Then*

$$\lim_{t \to \infty} Q_t(s,a) = Q^*(s,a) \qquad (s,a) \in \mathcal{S} \times \mathcal{A}$$

with probability 1.

Proof. We verify the conditions ensuring that we can apply Lemma 2. Let H be the operator $Q \mapsto H(Q)$ acting on the set of functions $Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ defined by

$$H(Q)(s,a) = r(s,a) + \gamma \mathbb{E}\left[\max_{a' \in \mathcal{A}} Q(s', a') \,\middle|\, s, a\right] \tag{3}$$

where the expectation is with respect to the random draw of s' from $p(\cdot \mid s, a)$. Let also

$$\nu_t = r(s_t, a_t) + \gamma \max_{a \in \mathcal{A}} Q_t(s_{t+1}, a) - H(Q_t)(s_t, a_t)$$

Then

$$Q_{t+1}(s_t,a_t) = (1-\eta_t)Q_t(s_t,a_t) + \eta_t\Big(H(Q_t)(s_t,a_t) + \nu_t\Big)$$

With our choice of η_t, and using the fact that, with probability 1, for each $(s,a) \in \mathcal{S}\times\mathcal{A}$ there are infinitely many t for which $s_t = s, a_t = a$, we have

$$\sum_{t\ge 0} \eta_t(s,a) = \sum_{n=1}^{\infty} \frac{1}{n} = \infty \qquad \text{and} \qquad \sum_{t\ge 0} \eta_t(s,a)^2 = \sum_{n=1}^{\infty} \frac{1}{n^2} < \infty$$

As for the noise condition, it is easy to prove by induction on $t = 0,1,\dots$ that $\|Q_t\|_\infty \le \frac{1}{1-\gamma}$. Hence $|\nu_t| \le \frac{2}{1-\gamma}$. Moreover,

$$\mathbb{E}\big[\nu_t \mid s_t,a_t\big] = \gamma\,\mathbb{E}\left[\max_{a\in\mathcal{A}} Q_t(s_{t+1},a) \,\middle|\, s_t,a_t\right] - \gamma\,\mathbb{E}\left[\max_{a\in\mathcal{A}} Q_t(s,a) \,\middle|\, s_t,a_t\right] = 0$$

as s_{t+1} and s are drawn from the same distribution $p(\cdot \mid s_t,a_t)$. Finally, because of (1), $H(Q^*) = Q^*$ and thus

$$\begin{aligned}
\|H(Q) - Q^*\|_\infty &= \|H(Q) - H(Q^*)\|_\infty \\
&= \gamma \max_{s,a}\left|\mathbb{E}\Big[\max_b Q(s',b) - \max_{b'} Q^*(s',b') \,\Big|\, s,a\Big]\right| \\
&\le \gamma \max_{s,a} \mathbb{E}\left[\left|\max_b Q(s',b) - \max_{b'} Q^*(s',b')\right| \,\middle|\, s,a\right]
\end{aligned}$$

where in the last step we used $\big|\mathbb{E}[X]\big| \le \mathbb{E}\big[|X|\big]$ that holds for any random variable X because the absolute value is a convex function. Now consider $\big|\max_b Q(s',b) - \max_{b'} Q^*(s',b')\big|$ and assume $\max_b Q(s',b) - \max_{b'} Q^*(s',b') \ge 0$. Then

$$\begin{aligned}
\big|\max_b Q(s',b) - \max_{b'} Q^*(s',b')\big| &= \max_b Q(s',b) - \max_{b'} Q^*(s',b') \\
&\le \max_b \big(Q(s',b) - Q^*(s',b)\big) \\
&\le \max_b \big|Q(s',b) - Q^*(s',b)\big|
\end{aligned}$$

If $\max_b Q(s',b) - \max_{b'} Q^*(s',b') \le 0$, then we proceed similarly to bound

$$\big|\max_b Q(s',b) - \max_{b'} Q^*(s',b')\big| = \max_{b'} Q^*(s',b') - \max_b Q(s',b) \le \max_{b'} \big|Q^*(s',b') - Q(s',b')\big|$$

Hence, in both cases we have

$$\begin{aligned}
\|H(Q) - Q^*\|_\infty &\le \gamma \max_{s,a} \mathbb{E}\left[\max_b \big|Q(s',b) - Q^*(s',b)\big| \,\middle|\, s,a\right] \\
&\le \gamma \max_{s',b} \big|Q(s',b) - Q^*(s',b)\big| && \text{(because } E\big[f(X)\big] \le \max_x f(x)\text{)} \\
&= \gamma\, \|Q - Q^*\|_\infty
\end{aligned}$$

This concludes the proof. □

SARSA. Q-learning is an off-policy method: we learn Q^* while running an arbitrary policy π satisfying certain minimal properties. As V^π can be much smaller than V^*, the algorithm has no control on the return while learning Q^*. SARSA, instead, is an on-policy method: Q^* is learned by a policy that is being updated. This allows the algorithm to control the return during the learning process.

Recall the Q-learning update step:

$$Q_{t+1}(s_t, a_t) = (1 - \eta_t) Q_t(s_t, a_t) + \eta_t \big(r(s_t, a_t) + \gamma \max_{a \in \mathcal{A}} Q_t(s_{t+1}, a) \big)$$

where $s_{t+1} \sim p(\cdot \mid s_t, a_t)$ and $a_t \sim \pi(\cdot \mid s_t)$. SARSA replaces $\max_{a \in \mathcal{A}} Q_t(s_{t+1}, a)$ with $Q_t(s_{t+1}, a_{t+1})$, where a_{t+1} is selected by a policy π_t based on the current approximation Q_t of the action-value function.

Algorithm 2 (SARSA)

```
Input: Initial state $s_0 \in \mathcal{S}$
Set $Q_0(s,a) = 0$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$
Draw a random initial action $a_0$
for $t = 0, 1, \dots$ do
    Observe reward $r_t = r(s_t, a_t)$ and next state $s_{t+1}$ drawn from $p(\cdot \mid s_t, a_t)$
    Draw action $a_{t+1} \sim \pi_t\big(\cdot \mid s_{t+1}, Q_t\big)$
    Update $Q_{t+1}(s_t, a_t) = (1 - \eta_t) Q_t(s_t, a_t) + \eta_t \big( r(s_t, a_t) + \gamma Q_t(s_{t+1}, a_{t+1}) \big)$
end for
```

Concretely, we consider the two following approaches for the policy $\pi_t(\cdot \mid s, Q)$:

- Greedy: play any action in $\operatorname*{argmax}_{a \in \mathcal{A}} Q(s,a)$
- ε-greedy: If in state s at time t, then play Greedy with probability $1 - \varepsilon_t(s)$; otherwise play a random action in $\mathcal{A}$.

Theorem 4 *Assume that the SARSA (Algorithm 2) is run on a communicating MDP with a ε_t-greedy policy such that $\varepsilon_t = \varepsilon_t(s) = 1/N_t(s)$, where $N_t(s)$ is the number of visits of state s in the time steps from $0, \dots, t$. If the learning rate η_t is chosen according to* (2)*, then*

$$\lim_{t \to \infty} Q_t(s,a) = Q^*(s,a) \qquad (s,a) \in \mathcal{S} \times \mathcal{A}$$

Proof. The proof applies Lemma 2 using the same operator H and the same noise parameter ν_t as in the proof of Q-learning. However, this time the bias term ζ_t is not equal to zero. We have

$$Q_{t+1}(s_t, a_t) = \big(1 - \eta_t(s_t, a_t)\big) Q_t(s_t, a_t) + \eta_t(s_t, a_t) \Big(r(s_t, a_t) + \gamma\, Q_t(s_{t+1}, a_{t+1}) \Big)$$

where $\eta_t(s,a) = 0$ for $(s,a) \neq (s_t, a_t)$. Now,

$$r(s_t, a_t) + \gamma\, Q_t(s_{t+1}, a_{t+1}) = r(s_t, a_t) + \gamma \max_b Q_t(s_{t+1}, b) + \underbrace{\gamma \Big(Q_t(s_{t+1}, a_{t+1}) - \max_b Q_t(s_{t+1}, b) \Big)}_{\zeta_t}$$

Let $\nu_t = r(s_t, a_t) + \gamma \max_b Q_t(s_{t+1}, b) - H(Q_t)(s_t, a_t)$ where the operator H is defined in (3). Therefore

$$Q_{t+1}(s_t, a_t) = \big(1 - \eta_t(s_t, a_t)\big) Q_t(s_t, a_t) + \eta_t(s_t, a_t)\Big(H(Q_t)(s_t, a_t) + \nu_t + \zeta_t\Big)$$

The contraction condition and the noise condition in Lemma 2 are both satisfied (see the proof of Theorem 3). Since the MDP is communicating and ε-greedy is fully mixed because $\varepsilon_t(s) > 0$ for all t and s, each pair $(s, a) \in \mathcal{S} \times \mathcal{A}$ is visited infinitely often w.p. 1 and the stepsize condition is satisfied. Since for all s, $\varepsilon_t(s) \to 0$ for $t \to \infty$, $\lim_{t\to\infty} \zeta_t = 0$ with w.p. 1 and the bias condition is satisfied. This concludes the proof. □

We now state and prove two auxiliary results which give us some insights on the discounted return. The first result bounds the variation in discounted return for a greedy policy based on some Q different from Q^*.

Lemma 5 *For any $Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, let $\pi(s) = \operatorname{argmax}_a Q(s, a)$ for all $s \in \mathcal{S}$. Then*

$$\|V^\pi - V^*\|_\infty \leq \frac{2}{1-\gamma} \|Q - Q^*\|_\infty$$

Proof. For any $s \in \mathcal{S}$, let $\pi^*(s) = \operatorname{argmax}_a Q^*(s, a)$ and $\|Q - Q^*\|_\infty = \Delta$. Then

$$\begin{aligned}
V^*(s) - Q^*\big(s, \pi(s)\big) &= Q^*\big(s, \pi^*(s)\big) - Q^*\big(s, \pi(s)\big) \\
&= Q^*\big(s, \pi^*(s)\big) - Q\big(s, \pi^*(s)\big) + Q\big(s, \pi^*(s)\big) + Q\big(s, \pi(s)\big) - Q^*\big(s, \pi(s)\big) - Q\big(s, \pi(s)\big) \\
&\leq 2\Delta + Q\big(s, \pi^*(s)\big) - Q\big(s, \pi(s)\big) \\
&\leq 2\Delta + Q\big(s, \pi^*(s)\big) - Q\big(s, \pi^*(s)\big) && \text{(by definition of } \pi)
\end{aligned}$$

Hence, choosing an initial state s_0,

$$\begin{aligned}
V^*(s_0) &\leq Q^*\big(s_0, \pi(s_0)\big) + 2\Delta \\
&= r\big(s_0, \pi(s_0)\big) + \gamma \mathbb{E}\big[V^*(s_1) \mid s_0\big] + 2\Delta && \text{(where } s_1 \sim p\big(\cdot \mid s_0, \pi(s_0)\big)) \\
&\leq \mathbb{E}\left[\sum_{\tau=0}^{t-1} \gamma^t r\big(s_\tau, \pi(s_\tau)\big)\right] + \gamma^t \mathbb{E}\big[V^*(s_t) \mid s_0\big] + 2\Delta \sum_{\tau=0}^{t-1} \gamma^\tau && \text{(where } s_\tau \sim p\big(\cdot \mid s_{\tau-1}, \pi(s_{\tau-1})\big)) \\
&\leq V^\pi(s_0) + \frac{2\Delta}{1-\gamma} && \text{(for } t \to \infty)
\end{aligned}$$

This concludes the proof. □

The second result shows how the discounted return of a stochastic policy π is affected by perturbations of π.

Lemma 6 *For any two stochastic policies π, ρ let $\pi' = (1 - \varepsilon)\pi + \varepsilon\rho$. Then*

$$\big\|V^\pi - V^{\pi'}\big\|_\infty \leq \frac{2\varepsilon}{(1-\gamma)^2}$$

PROOF. Since we are interested in bounding the difference $\left|V^{\pi}(s) - V^{\pi'}(s)\right|$ for any initial state s, we can map rewards $r(s,a) \in [-1,1]$ to new rewards $1 + r(s,a) \in [0,2]$ without affecting the difference of the state-value functions. Hence, without loss of generality, we may assume rewards are bounded in $[0,2]$. For any $s \in \mathcal{S}$, with probability $1-\varepsilon$ we have $\pi'(\cdot \mid s) \equiv \pi(\cdot \mid s)$. Let T be the stochastic horizon and T_ε be the first time that π and π' choose their action from two different distributions. Hence, $\mathbb{P}(T_\varepsilon > t \mid T \geq t) = (1-\varepsilon)^t$. Let r_t^π be the reward of π at time t starting from $s_0 = s$. Then

$$V^{\pi}(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \mathbb{I}\{T \geq t\} r_t^\pi\right] \quad \text{and} \quad V^{\pi'}(s) \geq \mathbb{E}\left[\sum_{t=0}^{\infty} \mathbb{I}\{T \geq t\}\mathbb{I}\{T_\varepsilon > t\} r_t^\pi\right]$$

where the inequality holds because the reward accumulated by π' for all $t \geq T_\varepsilon$ is nonnegative due to the nonnegativity assumption for the rewards. Therefore

$$\begin{aligned}
V^{\pi}(s) - V^{\pi'}(s) &\leq \mathbb{E}\left[\sum_{t=0}^{\infty} \mathbb{I}\{T \geq t\}\mathbb{I}\{T_\varepsilon \leq t\} r_t^\pi\right] \\
&\leq 2\mathbb{E}\left[\sum_{t=0}^{\infty} \mathbb{I}\{T \geq t\}\mathbb{I}\{T_\varepsilon \leq t\}\right] && \text{(because } r_t^\pi \leq 2\text{)} \\
&\leq 2\sum_{t=0}^{\infty} \mathbb{P}(T \geq t)\mathbb{P}(T_\varepsilon \leq t \mid T \geq t) \\
&\leq 2\sum_{t=0}^{\infty} \gamma^t\bigl(1 - (1-\varepsilon)^t\bigr) \\
&= \frac{2}{1-\gamma} - \frac{2}{1-\gamma(1-\varepsilon)} \\
&\leq \frac{2\varepsilon}{(1-\gamma)^2}
\end{aligned}$$

concluding the proof. □

Using these results, we can prove that, for any $\lambda > 0$, there is a time step t_λ after which the policies used by SARSA are λ-optimal.

Theorem 7 *Assume that the SARSA (Algorithm 2) is run on a communicating MDP with a ε_t-greedy policy such that $\varepsilon_t(s) = 1/N_t(s)$, where $N_t(s)$ is the number of visits of state s in the time steps from $0,\ldots,t$. Then, for any $\lambda > 0$, there is a time t_λ such that for all $t \geq t_\lambda$, $\|V^* - V^{\pi_t}\|_\infty \leq \lambda$.*

PROOF. Since each state is sampled infinitely often, there is a time t_1 such that $n_{t_1}(s) \geq \frac{4}{\lambda(1-\gamma)^2}$ times for all $s \in \mathcal{S}$, implying $\varepsilon_t(s) \leq \lambda(1-\gamma)^2/4$ for all $s \in \mathcal{S}$ and $t \geq t_1$. Since $Q_t \to Q^*$, there is a time t_2 such that $\|Q_t - Q^*\|_\infty \leq \lambda(1-\gamma)/4$ for all $t \geq t_2$. Recall that π_t is the ε_t-greedy policy used by SARSA at time t and let g_t be the greedy policy (based on Q_t). Then for all $t \geq \max\{t_1, t_2\}$,

$$\begin{aligned}
\|V^* - V^{\pi_t}\|_\infty &\leq \|V^* - V^{g_t}\|_\infty + \|V^{g_t} - V^{\pi_t}\|_\infty && \text{(by the triangle inequality)} \\
&\leq \frac{2}{1-\gamma}\,\|Q_t - Q^*\|_\infty + \frac{2\max_s \varepsilon_t(s)}{(1-\gamma)^2} && \text{(by Lemma 5 and 6)} \\
&\leq \frac{\lambda}{2} + \frac{\lambda}{2} = \lambda
\end{aligned}$$

concluding the proof. □