Prof. Dr. Leif Döring

André Ferdinand, Sara Klein

Reinforcement Learning

8. Solution Sheet

1. Second version of Theorem 4.2.9 for SARSA

Show that the statement of Theorem 4.2.9 also holds if $\mathbb{E}[\varepsilon_n \,|\, \mathcal{F}_n] \neq 0$ but instead satisfies

$$\sum_{n=1}^{\infty} \alpha_i(n)\big|\mathbb{E}[\varepsilon_i(n) \,|\, \mathcal{F}_n]\big| < \infty \tag{1}$$

almost surely. It is enough to prove an improved version of Lemma 4.2.5 where the condition $\mathbb{E}[\varepsilon(t) \,|\, \mathcal{F}_t] = 0$ is replaced with

$$\sum_{n=1}^{\infty} \alpha(t)\big|\mathbb{E}[\varepsilon(t) \,|\, \mathcal{F}_t]\big| < \infty. \tag{2}$$

Apply the Robbins-Siegmund theorem to W^2 and use that $W \leq 1 + W^2$.
Solution:

$$\begin{aligned}
\mathbb{E}\big[W(t+1)^2 \,\big|\, \mathcal{F}_t\big] &= \mathbb{E}\big[(1-\alpha(t))^2 W^2(t) + \alpha^2(t)\varepsilon^2(t) + 2\alpha(t)(1-\alpha(t))W(t)\varepsilon(t) \,\big|\, \mathcal{F}_t\big] \\
&\leq (1 - 2\alpha(t) + \alpha^2(t))W^2(t) + \alpha^2(t)C + 2\alpha(t)(1-\alpha(t))W(t)\mathbb{E}\big[\varepsilon(t) \,\big|\, \mathcal{F}_t\big] \\
&\leq (1 - 2\alpha(t) + \alpha^2(t))W^2(t) + \alpha^2(t)C + 2\alpha(t)(1-\alpha(t))(1+W^2(t))\big|\mathbb{E}\big[\varepsilon(t) \,\big|\, \mathcal{F}_t\big]\big| \\
&\leq (1 - 2\alpha(t) + \alpha^2(t) + 2\alpha(t)\big|\mathbb{E}\big[\varepsilon(t) \,\big|\, \mathcal{F}_t\big]\big| - \underbrace{2\alpha(t)^2\big|\mathbb{E}\big[\varepsilon(t) \,\big|\, \mathcal{F}_t\big]\big|}_{\geq 0})W^2(t) \\
&\quad + \alpha^2(t)C + 2\alpha(t)\big|\mathbb{E}\big[\varepsilon(t) \,\big|\, \mathcal{F}_t\big]\big| - \underbrace{2\alpha(t)^2\big|\mathbb{E}\big[\varepsilon(t) \,\big|\, \mathcal{F}_t\big]\big|}_{\geq 0} \\
&\leq (1 - a_t + b_t)W^2(t) + c_t,
\end{aligned}$$

with $a_t = -2\alpha(t)$, $b_t = \alpha^2(t) + 2\alpha(t)\big|\mathbb{E}\big[\varepsilon(t) \,\big|\, \mathcal{F}_t\big]\big|$, *and* $c_t = \alpha^2(t)C + 2\alpha(t)\big|\mathbb{E}\big[\varepsilon(t) \,\big|\, \mathcal{F}_t\big]\big|$. *Now the claim follows from Robbins-Siegmund.*

2. n-step TD

a) Write pseudocode for n-step TD algorithms for evaluation of V^π and Q^π in the non-terminating case and prove the convergence by checking that using the n-step Bellman expectation equations

$$T_1^\pi V(s) = \mathbb{E}_s^\pi\Big[R(s, A_0) + \sum_{t=1}^{n-1} \gamma^t R(S_t, A_t) + \gamma^n V(S_n)\Big]$$

and

$$T_2^\pi Q(s,a) = \mathbb{E}_s^{\pi^a}\Big[R(s,a) + \sum_{t=1}^{n-1}\gamma^t R(S_t,A_t) + \gamma^n Q(S_n,A_n)\Big]$$

and the corresonding error terms fulfill the conditions of Theorem 4.2.9. Note that the algorithm only starts to update after the MDP ran for n steps. Can you also write down a version in the terminating case?

Solution: The algorithms in the non terminating case are 1 and 2. The Algorithm in the

Algorithm 1: n-step TD for evaluation of V^π

```
Data: Policy π ∈ Π_S
Result: Approximation V ≈ V^π
Initialize V ≡ 0
Initialise s arbitrarily
while not converged do
    Set s* = s
    Initialise R = 0
    for i = 0, ... n − 1 do
        a ∼ π(·; s)
        Sample reward R(s, a_i)
        Set R = R + γ^i R(s, a)
        Sample s' ∼ p(· ; s, a)
        s = s'
    end
    Determine stepsize α
    Update V(s*) = V(s*) + α(R + γ^n V(s) − V(s*))
end
```

terminating case would be as stated in algorithm 3. We added a break in the for loop as we cannot continue in a terminating state. As we only which to update after n steps, we will not update V after the break. So it can happen that we never update the value function, if we never run n steps. Next we come to the prove of convergence. Therefore we have to check that the operators T_1 and T_2 are contractions and that the error terms

$$\varepsilon_s(n) := R(s,A_0) + \sum_{t=1}^{n-1}\gamma^t R(S_t,A_t) + \gamma^n V(S_n) - \mathbb{E}_s^\pi\Big[R(s,A_0) + \sum_{t=1}^{n-1}\gamma^t R(S_t,A_t) + \gamma^n V(S_n)\Big]$$

$$\varepsilon_{s,a}(n) := R(s,a) + \sum_{t=1}^{n-1}\gamma^t R(S_t,A_t) + \gamma^n Q(S_n,A_n) - \mathbb{E}_s^{\pi^a}\Big[R(s,a) + \sum_{t=1}^{n-1}\gamma^t R(S_t,A_t) + \gamma^n Q(S_n,A_n)\Big]$$

fulfill the conditions of Theorem 4.2.9. The condition on the error terms is as always given

Algorithm 2: n-step TD for evaluation of Q^π

Data: Policy $\pi \in \Pi_\mathcal{S}$

Result: Approximation $Q \approx Q^\pi$

Initialize $Q \equiv 0$

Initialise s, a arbitrarily

while *not converged* **do**

  Set $s^* = s$ and $a^* = a$

  Initialise $R = 0$

  for $i = 0, \dots n-1$ **do**

    Sample reward $R(s, a)$

    Set $R = R + \gamma^i R(s, a)$

    Sample $s' \sim p(\cdot\,;\, s, a)$

    Sample $a' \sim \pi(\cdot|s')$

    $s = s'$, $a = a^*$

  end

  Determine stepsize α

  Update $Q(s^*, a^*) = Q(s^*, a^*) + \alpha(R + \gamma^n Q(s, a) - Q(s^*, a^*))$

end

by definition and bounded rewards. For the contractions we see that

$$
\begin{aligned}
&\|T_1(V_1) - T_1(V_2)\|_\infty \\
&= \max_{s\in\mathcal{S}} |T_1(V_1)(s) - T_1(V_2)(s)| \\
&= \max_{s\in\mathcal{S}} |\mathbb{E}_s^\pi\Big[R(s, A_0) + \sum_{t=1}^{n-1} \gamma^t R(S_t, A_t) + \gamma^n V_1(S_n)\Big] \\
&\quad - \mathbb{E}_s^\pi\Big[R(s, A_0) + \sum_{t=1}^{n-1} \gamma^t R(S_t, A_t) + \gamma^n V_2(S_n)\Big]| \\
&\leq \max_{s\in\mathcal{S}} \mathbb{E}_s^\pi\Big[\gamma^n |V_1(S_n) - V_2(S_n)|\Big] \\
&\leq \gamma^n \|V_1 - V_2\|_\infty
\end{aligned}
$$

Algorithm 3: n-step TD for evaluation of V^π for terminating MDPs

Data: Policy $\pi \in \Pi_{\mathcal{S}}$
Result: Approximation $V \approx V^\pi$
Initialize $V \equiv 0$
while *not converged* **do**
  Initialise s arbitrarily
  while s *not terminal* **do**
    Set $s^* = s$
    Initialise $R = 0$
    for $i = 0, \dots n-1$ **do**
      if s *terminal* **then**
        Break and beginn with a new while-loop
      end
      $a \sim \pi(\cdot; s)$
      Sample reward $R(s, a_i)$
      Set $R = R + \gamma^i R(s, a)$
      Sample $s' \sim p(\cdot\,; s, a)$
      $s = s'$
    end
    Determine stepsize α
    Update $V(s^*) = V(s^*) + \alpha(R + \gamma^n V(s) - V(s^*))$
  end
end

and similar for T_2

$$\begin{aligned}
&\|T_2(Q_1) - T_2(Q_2)\|_\infty \\
&= \max_{s\in\mathcal{S}, a\in\mathcal{A}_s} |T_2(Q_1)(s,a) - T_2(Q_2)(s,a)| \\
&= \max_{s\in\mathcal{S}, a\in\mathcal{A}_s} |\mathbb{E}_s^{\pi^a}\Big[R(s,a) + \sum_{t=1}^{n-1}\gamma^t R(S_t, A_t) + \gamma^n Q_1(S_n, A_n)\Big] \\
&\quad - \mathbb{E}_s^{\pi^a}\Big[R(s,a) + \sum_{t=1}^{n-1}\gamma^t R(S_t, A_t) + \gamma^n Q_2(S_n, A_n)\Big]| \\
&\le \max_{s\in\mathcal{S}, a\in\mathcal{A}_s} \mathbb{E}_s^{\pi^a}\Big[\gamma^n |Q_1(S_n, A_n) - Q_2(S_n, A_n)|\Big] \\
&\le \gamma^n \|Q_1 - Q_2\|_\infty.
\end{aligned}$$

b) Write pseudocode for an n-step SARSA control algorithm in the non-terminating case. Try to prove convergence in the same way we did for 1-step SARSA in Theorem 4.3.6.

Solution:

Convergence of n-step SARSA in the non-terminating case. Assume that Q_0 has bounded

---

Algorithm 4: n-step SARSA

Result: Approximations $Q \approx Q^*$, $\pi = \text{greedy}(Q) \approx \pi^*$

Initialize Q, e.g. $Q \equiv 0$

Initialise s, a arbitrarily, e.g. uniform.

while *not converged* **do**

  Set $s^* = s$ and $a^* = a$

  Initialise $R = 0$

  Chose new policy π from Q (e.g. ϵ-greedy)

  for $i = 0, \dots n-1$ **do**

    Sample reward $R(s, a)$

    Set $R = R + \gamma^i R(s, a)$

    Sample $s' \sim p(\cdot\,; s, a)$

    Sample $a' \sim \pi(\cdot|s')$

    $s = s'$, $a = a'$

  end

  Determine stepsize α

  Update $Q(s^*, a^*) = Q(s^*, a^*) + \alpha(R + \gamma^n Q(s, a) - Q(s^*, a^*))$

end

---

entries and and the step-sizes satisfy the Robbins-Monro conditions. If furthermore the probabilities $p_n(s, a)$ the the policy π_{n+1} is greedy satisfies are such that

$$\sum_{n=1}^{\infty} \alpha_n(s, a) p_n(s, a) < \infty \qquad a.s.$$

for all (s, a). Then n-step SARSA algorithm converges to Q^* almost surely.

Proof: *We denote by $(\tilde{S}_k, \tilde{A}_k)_{k=0}^{\infty}$ the sequence of state-action pairs obtained from the algorithm. We denote with $I = \{0, n, 2n, 3n, \dots\}$ the set of indices where the Q-function is updated, for $i \in I$ we have the update*

$$Q_{i+n}(\tilde{S}_i, \tilde{A}_i) = Q_i(\tilde{S}_i, \tilde{A}_i) + \alpha_i(\tilde{S}_i, \tilde{A}_i)(T_n^* Q_i(\tilde{S}_i, \tilde{A}_i) - Q_i(\tilde{S}_i, \tilde{A}_i) + \epsilon_i(\tilde{S}_i, \tilde{A}_i)),$$

where

$$T_n^* Q(s, a) = \mathbb{E}_s^{\pi^a\ (\pi\ greedy\ Q^*)} \Big[R(s, a) + \sum_{t=1}^{n-1} \gamma^t R(S_t, A_t) + \gamma^n \max_{\hat{a} \in \mathcal{A}} Q(S_n, \hat{a}) \Big]$$

and

$$\epsilon_i(\tilde{S}_i, \tilde{A}_i) = \sum_{k=0}^{n-1} \gamma^k R(\tilde{S}_{i+k}, \tilde{A}_{i+k}) + \gamma^n Q_i(\tilde{S}_{i+n}, \tilde{A}_{i+n}) - T_n^* Q_i(s, a).$$

For convergence we have to prove that:

(i) *the operator T_n^* is a contraction,*

(ii) *Q^* is a fixpoint of T_n^*,*

(iii) *The error term fulfills the assumptions of the generalised stochastic approximation theorem from Exercise 1 above.*

For (i)we have similar to T_2^π of part a), that

$$
\begin{aligned}
&\|T_n^*(Q_1) - T_n^*(Q_2)\|_\infty \\
&= \max_{s\in\mathcal{S}, a\in\mathcal{A}_s} |T_n^*(Q_1)(s,a) - T_n^*(Q_2)(s,a)| \\
&= \max_{s\in\mathcal{S}, a\in\mathcal{A}_s} |\mathbb{E}_s^{\pi^a\,(\pi \text{ greedy } Q^*)}\Big[R(s,a) + \sum_{t=1}^{n-1}\gamma^t R(S_t,A_t) + \gamma^n \max_{\hat{a}\in\mathcal{A}} Q_1(S_n,\hat{a})\Big] \\
&\quad - \mathbb{E}_s^{\pi^a\,(\pi \text{ greedy } Q^*)}\Big[R(s,a) + \sum_{t=1}^{n-1}\gamma^t R(S_t,A_t) + \gamma^n \max_{\hat{a}\in\mathcal{A}} Q_2(S_n,\hat{a})\Big]| \\
&= \gamma^n \max_{s\in\mathcal{S}, a\in\mathcal{A}_s} |\mathbb{E}_s^{\pi^a\,(\pi \text{ greedy } Q^*)}\Big[\max_{\hat{a}\in\mathcal{A}} Q_1(S_n,\hat{a}) - \max_{\hat{a}\in\mathcal{A}} Q_2(S_n,\hat{a})\Big]| \\
&\leq \gamma^n \max_{s\in\mathcal{S}, a\in\mathcal{A}_s} \mathbb{E}_s^{\pi^a\,(\pi \text{ greedy } Q^*)}\Big[\max_{\hat{a}\in\mathcal{A}} |Q_1(S_n,\hat{a}) - Q_2(S_n,\hat{a})|\Big] \\
&\leq \gamma^n \|Q_1 - Q_2\|_\infty.
\end{aligned}
$$

We show (ii) by induction over n. We have that for $n = 1$ that T_1^* is the normal Bellman opator, i.e. $Q^* = T_1^* Q^*$. Assume $Q^* = T_n^* Q^*$, then for $n+1$ we conclude

$$
\begin{aligned}
Q^*(s,a) &= \mathbb{E}_s^{\pi^a\,(\pi \text{ greedy } Q^*)}\Big[R(s,a) + \sum_{t=1}^{n-1}\gamma^t R(S_t,A_t) + \gamma^n \max_{\hat{a}\in\mathcal{A}} Q^*(S_n,\hat{a})\Big] \\
&= \mathbb{E}_s^{\pi^a\,(\pi \text{ greedy } Q^*)}\Big[R(s,a) + \sum_{t=1}^{n-1}\gamma^t R(S_t,A_t) + \gamma^n \sum_{a\in\mathcal{A}} \pi^*(a;S_n) Q^*(S_n,a)\Big] \\
&= \mathbb{E}_s^{\pi^a\,(\pi \text{ greedy } Q^*)}\Big[R(s,a) + \sum_{t=1}^{n-1}\gamma^t R(S_t,A_t) + \gamma^n Q^*(S_n,A_n)\Big] \\
&= \mathbb{E}_s^{\pi^a\,(\pi \text{ greedy } Q^*)}\Big[R(s,a) + \sum_{t=1}^{n-1}\gamma^t R(S_t,A_t) + \gamma^n \Big(R(S_n,A_n) \\
&\quad + \gamma \sum_{s'\in\mathcal{S}} p(s';S_n,A_n) \max_{\hat{a}\in\mathcal{A}} Q^*(s',\hat{a})\Big)\Big] \\
&= \mathbb{E}_s^{\pi^a\,(\pi \text{ greedy } Q^*)}\Big[R(s,a) + \sum_{t=1}^{n}\gamma^t R(S_t,A_t) + \gamma^{n+1} \sum_{s'\in\mathcal{S}} p(s';S_n,A_n) \max_{\hat{a}\in\mathcal{A}} Q^*(s',\hat{a})\Big] \\
&= \mathbb{E}_s^{\pi^a\,(\pi \text{ greedy } Q^*)}\Big[R(s,a) + \sum_{t=1}^{n}\gamma^t R(S_t,A_t) + \gamma^{n+1} \max_{\hat{a}\in\mathcal{A}} Q^*(S_{n+1},\hat{a})\Big] \\
&= T_{n+1}^* Q^*(s,a).
\end{aligned}
$$

For the last claim we first note, that $\epsilon(s,a) = 0$ if $(s,a) \neq (\tilde{S}_i, \tilde{A}_i)$, $i \in I$. We enumerate the elements in I by the index j, i.e $j = i/n$ for $i \in I$ and we have that the next element in I is $j + 1 = \frac{i}{n} + 1 = \frac{i+n}{n}$. Further we denote by $\tilde{\mathcal{F}}_k$ the σ-algebra generated by the process

$(\tilde{S}_k, \tilde{A}_k)$. Then the errors ϵ_i are $\mathcal{F}_{j+1} = \tilde{\mathcal{F}}_{i+2n-1}$ mesurable for every $i \in I$. We define for ever $j \geq 0$ the filtration $\mathcal{F}_j = \tilde{\mathcal{F}}_{j(n+1)-1}$. Then $\epsilon_{j\cdot n}$ is $\mathcal{F}_{j+1}$ measurable and we follow

$$\begin{aligned}
&\mathbb{E}[\epsilon_{j\cdot n}(\tilde{S}_{j\cdot n}, \tilde{A}_{j\cdot n})|\mathcal{F}_j] \\
&= \mathbb{E}[\epsilon_i(\tilde{S}_i, \tilde{A}_i)|\mathcal{F}_{i+n-1}] \\
&= \mathbb{E}[\mathbf{1}_{\{\pi_{i+n}(\cdot;\tilde{S}_{i+n}) \text{ is greedy}\}}\Big(\sum_{k=0}^{n-1}\gamma^k R(\tilde{S}_{i+k}, \tilde{A}_{i+k}) + \gamma^n Q_i(\tilde{S}_{i+n}, \tilde{A}_{i+n}) - T_n^* Q_i(s,a)\Big)|\mathcal{F}_{i+n-1}] \\
&\quad + \mathbb{E}[\mathbf{1}_{\{\pi_{i+n}(\cdot;\tilde{S}_{i+n}) \text{ is non-greedy}\}}\Big(\sum_{k=0}^{n-1}\gamma^k R(\tilde{S}_{i+k}, \tilde{A}_{i+k}) + \gamma^n Q_i(\tilde{S}_{i+n}, \tilde{A}_{i+n}) - T_n^* Q_i(s,a)\Big)|\mathcal{F}_{i+n-1}]
\end{aligned}$$

Do to the choice

$$T_n^* Q(s,a) = \mathbb{E}_s^{\pi^a\,(\pi \text{ greedy } Q^*)}\Big[R(s,a) + \sum_{t=1}^{n-1}\gamma^t R(S_t, A_t) + \gamma^n \max_{\hat{a}\in\mathcal{A}} Q(S_n, \hat{a})\Big]$$

we do not get that the error is 0 for the greedy policy choice. We would need

$$T_n^* Q(s,a) = \mathbb{E}_s^{\pi^a\,(\pi \text{ greedy } Q)}\Big[R(s,a) + \sum_{t=1}^{n-1}\gamma^t R(S_t, A_t) + \gamma^n \max_{\hat{a}\in\mathcal{A}} Q(S_n, \hat{a})\Big]$$

as bellman operator. Unfortunetly then we can no longer show that T_n^* is a contraction:

$$\begin{aligned}
&\|T_n^*(Q_1) - T_n^*(Q_2)\|_\infty \\
&= \max_{s\in\mathcal{S}, a\in\mathcal{A}_s} |T_n^*(Q_1)(s,a) - T_n^*(Q_2)(s,a)| \\
&= \max_{s\in\mathcal{S}, a\in\mathcal{A}_s} |\mathbb{E}_s^{\pi^a\,(\pi \text{ greedy } Q_1)}\Big[R(s,a) + \sum_{t=1}^{n-1}\gamma^t R(S_t, A_t) + \gamma^n \max_{\hat{a}\in\mathcal{A}} Q_1(S_n, \hat{a})\Big] \\
&\quad - \mathbb{E}_s^{\pi^a\,(\pi \text{ greedy } Q_1)}\Big[R(s,a) + \sum_{t=1}^{n-1}\gamma^t R(S_t, A_t) + \gamma^n \max_{\hat{a}\in\mathcal{A}} Q_2(S_n, \hat{a})\Big]|,
\end{aligned}$$

cannot be written in one expectation due to different measures.