FAKULTÄT FÜR WIRTSCHAFTSINFORMATIK
UND WIRTSCHAFTSMATHEMATIK

UNIVERSITÄT MANNHEIM

Prof. Dr. Leif Döring
Benedikt Wille

Reinforcement Learning

8. Exercise Sheet - Solutions

29.04.2025

1. Sample based policy iteration without bounded rewards

Let the second moments of the rewards given a policy $\pi \in \Pi_S$ exist, i.e.

$$\mathbb{E}_s^\pi[R_0^2] < \infty \ \ \forall s \in \mathcal{S}.$$

Show that the Theorems 4.5.1 and 4.5.2 still apply to this policy, so that the one-step policy evaluation schemes from the lecture converge.

Solution:

The only place where we needed to assume bounded rewards in the proofs of the theorems was when showing $\sup_s \mathbb{E}[\varepsilon_s^2(n) \,|\, \mathcal{F}_n] \leq A + B\|V(n)\|_\infty^2$ *and* $\sup_{s,a} \mathbb{E}[\varepsilon_{s,a}^2(n) \,|\, \mathcal{F}_n] \leq A + B\|Q(n)\|_\infty^2$ *respectively. With the assumption of existing second moments and defining*

$$C := \sup_{s\in\mathcal{S}} \mathbb{E}_s^\pi[R_0^2] < \infty$$

we can proceed as in the lecture notes:

$$\begin{aligned}
&\mathbb{E}[\varepsilon_s^2(n)|\mathcal{F}_n] \\
&= \mathbb{E}[(r_n + \gamma V_{s'_n}(n))^2 \,|\, \mathcal{F}_n] - 2\mathbb{E}_s^\pi[R_0 + \gamma V_{S_1}(n)]\mathbb{E}_s^\pi[r_n + \gamma V_{s'_n}(n) \,|\, \mathcal{F}_n] + (\mathbb{E}[R_0 + \gamma V_{S_1}(n)])^2 \\
&= \mathbb{E}_s^\pi[(R_0 + \gamma V_{S_1}(n))^2] - 2(\mathbb{E}_s^\pi[R_0 + \gamma V_{S_1}(n)])^2 + (\mathbb{E}_s^\pi[R_0 + \gamma V_{S_1}(n)])^2 \\
&\leq \mathbb{E}_s^\pi[(R_0 + \gamma V_{S_1}(n))^2] = \mathbb{E}_s^\pi[R_0^2] + 2\gamma\mathbb{E}_s^\pi[R_0 V_{S_1}(n))^2] + \gamma^2\mathbb{E}_s^\pi[V_{S_1}(n)^2] \\
&\leq C^2 + 2\gamma C\|V(n)\|_\infty + \gamma^2\|V(n)\|_\infty^2 \leq C^2 + 2\gamma C(1 + \|V(n)\|_\infty^2) + \gamma^2\|V(n)\|_\infty^2 \\
&= (C^2 + 2\gamma C) + (2\gamma C + \gamma^2)\|V(n)\|_\infty^2
\end{aligned}$$

The case for $Q(n)$ goes analogously save for the definition of C as the supremum over additionally all $a \in \mathcal{A}$ and the usage of the tower property with given $A_1 = a$ inside the conditional expectation.

2. Convergence theorem 4.3.8 under weaker assumptions

Show that the statement of Theorem 4.3.8 also holds if $\mathbb{E}[\varepsilon_i(n) \,|\, \mathcal{F}_n] \neq 0$ but instead satisfies

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

almost surely for all coordinates $i = 1, \ldots, d$. It is enough to prove an improved version of Lemma 4.4.4 where the condition $\mathbb{E}[\varepsilon_n \,|\, \mathcal{F}_n] = 0$ is replaced with

$$\sum_{n=1}^\infty \alpha_n\big|\mathbb{E}[\varepsilon_n \,|\, \mathcal{F}_n]\big| < \infty. \tag{1}$$

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

$$\begin{aligned}
\mathbb{E}\big[W_{n+1}^2 \,\big|\, \mathcal{F}_n\big] &= \mathbb{E}\big[(1-\alpha_n)^2 W_n^2 + \alpha_n^2 \varepsilon_n^2 + 2\alpha_n(1-\alpha_n) W_n \varepsilon_n \,\big|\, \mathcal{F}_n\big] \\
&\leq (1 - 2\alpha_n + \alpha_n^2) W_n^2 + \alpha_n^2 D_n + 2\alpha_n(1-\alpha_n) W_n \mathbb{E}\big[\varepsilon_n \,\big|\, \mathcal{F}_n\big] \\
&\leq (1 - 2\alpha_n + \alpha_n^2) W_n^2 + \alpha_n^2 D_n + 2\alpha_n(1-\alpha_n)(1+W_n^2)\big|\mathbb{E}\big[\varepsilon_n \,\big|\, \mathcal{F}_n\big]\big| \\
&\leq (1 - 2\alpha_n + \alpha_n^2 + 2\alpha_n\big|\mathbb{E}\big[\varepsilon_n \,\big|\, \mathcal{F}_n\big]\big| - \underbrace{2\alpha_n^2\big|\mathbb{E}\big[\varepsilon_n \,\big|\, \mathcal{F}_n\big]\big|}_{\geq 0}) W_n^2 \\
&\quad + \alpha_n^2 D_n + 2\alpha_n\big|\mathbb{E}\big[\varepsilon_n \,\big|\, \mathcal{F}_n\big]\big| - \underbrace{2\alpha_n^2\big|\mathbb{E}\big[\varepsilon_n \,\big|\, \mathcal{F}_n\big]\big|}_{\geq 0} \\
&\leq (1 - a_n + b_n) W_n^2 + c_n,
\end{aligned}$$

with $a_n = 2\alpha_n$, $b_n = \alpha_n^2 + 2\alpha_n\big|\mathbb{E}\big[\varepsilon_n \,\big|\, \mathcal{F}_n\big]\big|$, *and* $c_n = \alpha_n^2 D_n + 2\alpha_n\big|\mathbb{E}\big[\varepsilon_n \,\big|\, \mathcal{F}_n\big]\big|$. *Now the claim follows from the Robbins-Siegmund Corollary 4.4.3.*

3. n-step TD

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

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

and

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

and the conditions of Theorem 4.3.8 on the error term. Note that the algorithm only starts to update after the MDP ran for n steps.
Solution: The algorithms are written down in Algorithm 1 and Algorithm 2.
Regarding convergence we have to check that the operators T_1 *and* T_2 *are contractions and that the error terms*

$$\begin{aligned}
\varepsilon_s(n) :&= R(s, A_0) + \sum_{t=1}^{n-1} \gamma^t R(S_t, A_t) + \gamma^n V(S_n) \\
&\quad - \mathbb{E}_s^\pi\Big[R(s, A_0) + \sum_{t=1}^{n-1} \gamma^t R(S_t, A_t) + \gamma^n V(S_n)\Big] \text{ and} \\
\varepsilon_{s,a}(n) :&= R(s,a) + \sum_{t=1}^{n-1} \gamma^t R(S_t, A_t) + \gamma^n Q(S_n, A_n) \\
&\quad - \mathbb{E}_s^{\pi^a}\Big[R(s,a) + \sum_{t=1}^{n-1} \gamma^t R(S_t, A_t) + \gamma^n Q(S_n, A_n)\Big]
\end{aligned}$$

Data: Policy $\pi \in \Pi_{\mathcal{S}}$
Result: Approximation $V \approx V^\pi$
Initialize $V \equiv 0$
Initialise s arbitrarily
while *not converged* **do**
  Set $s^* = s$
  Initialise $R = 0$
  for $i = 0, \dots n-1$ **do**
    $a \sim \pi(\cdot; s)$
    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

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

fulfill the conditions of Theorem 4.3.8. The conditions on the error terms are checked as always. First, all appearing random variables are $\mathcal{F}_{n+1}$-measurable and so the errors are adapted. The conditional expectation is computed by using that given $\mathcal{F}_n$, the expectation gets taken, which yields 0. Finally, the second moments of the errors are estimated by the second moments of the first terms and using boundedness of the rewards we deduce

$$\mathbb{E}[\varepsilon_s(n)^2|\mathcal{F}_n] \leq A_1 + B_1\|V(n)\|_\infty^2 \text{ and}$$
$$\mathbb{E}[\varepsilon_{s,a}(n)^2|\mathcal{F}_n] \leq A_2 + B_2\|Q(n)\|_\infty^2$$

For the contractions we see that

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

and analogously for T_2.

b) Write pseudocode for an n-step SARSA control algorithm.

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

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

Solution:
Algorithm 3 constitutes an n-step SARSA control algorithm.

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

Algorithm 3: n-step SARSA