Prof. Dr. Leif Döring
Sara Klein, Benedikt Wille

9. Solution Sheet

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. Programming task: One-step policy evaluation on grid world

We want to use the grid world example to illustrate how to perform policy evaluation:

a) Implement the grid world example from the lecture notes with target in the lower right corner and trap diagonally above or modify the code from the lecture's webpage.

b) Implement the Algorithms 17 and 18, the one-step policy evaluation schemes for V^π and Q^π respectively, for the grid world example.

c) Think about what you intuitively think the best policy π^+ and the worst policy π^- are for grid world and let additionally π be the policy that chooses the next action uniformly for all available options. Calculate $n = 1000$ steps of each policy evaluation scheme for π^+, π^-, and π.

d) Compare Algorithm 17 to Algorithm 7, the iterative policy evaluation. Which algorithm do you think performs better? Can we always apply both algorithms?

Solution:
See discussion in class and code.