FAKULTÄT FÜR WIRTSCHAFTSINFORMATIK
UND WIRTSCHAFTSMATHEMATIK

UNIVERSITÄT MANNHEIM

Prof. Dr. Leif Döring
Daniel Schmidt, Benedikt Wille

Reinforcement Learning

11. Exercise Sheet

11.05.2026

For this assignment, knowledge from lectures 1 to 24 is assumed.

1. SoftMax parameterisation

a) Show for the tabular softmax parametrisation from Example 5.0.2 that

$$\frac{\partial \log(\pi^\theta(a\,;\,s))}{\partial \theta_{s',a'}} = \mathbf{1}_{\{s=s'\}}(\mathbf{1}_{\{a=a'\}} - \pi^\theta(a'\,;\,s'))$$

and for the linear softmax with features $\Phi(s,a)$

$$\nabla \log(\pi^\theta(a\,;\,s)) = \Phi(s,a) - \sum_{a'} \pi^\theta(a'\,;\,s)\Phi(s,a').$$

b) Show that the tabular and linear softmax parametrisation fulfill Assumption 5.1.11, i.e. that $\log\big(\pi^\theta(a;s)\big)$ is L-smooth and $\nabla \log\big(\pi^\theta(a;s)\big)$ has bounded norm for any $(s,a) \in \mathcal{S} \times \mathcal{A}$.

2. Policy Gradient Theorems

For episodic MDPs (the MDP terminates almost surely under all policies π_θ), we can get rid of the assumption of the existence of $\nabla J_s(\theta)$. Go through the proof of Theorem 5.1.6 and argue why it is enough to assume the existence of $\nabla \pi_\theta(\cdot\,;\,s)$ for all $s \in \mathcal{S}$.

3. Baseline Trick

a) Show that the constant baseline b in Theorem 5.1.16 can be replaced by any deterministic state-dependent baseline $b : \mathcal{S} \to \mathbb{R}$, i.e.

$$\nabla_\theta J(\theta) = \mathbb{E}_s^{\pi_\theta}\Big[\sum_{t=0}^{T-1} \nabla_\theta\big(\log \pi^\theta(A_t\,;\,S_t)\big)\big(Q_t^{\pi^\theta}(S_t,A_t) - b(S_t)\big)\Big].$$

b) Write down and prove the baseline gradient representation with baseline $b : \mathcal{S} \to \mathbb{R}$ for infinite discounted MDPs.

The solution to the following exercises has to be turned in until 25.05.2026. Groups of up to three people are allowed.

4. *Programming task: Actor-Critic Algorithms

The aim of this exercise is to compare various Policy-Gradient-like algorithms on more complex environments than the tabular ones you've used so far.

(a) Conduct an evaluation study across all Gymnasium classic control environments, using as many seeds per environment as feasible. Include your own implementations of REINFORCE and Mini-batch REINFORCE, as well as the following pre-implemented (partially SOTA) algorithms: ARS, A2C, DDPG, PPO, SAC, TD3, TQC, and TRPO.

(b) Compare the evaluation metrics to those used in the tabular setting. What differences do you observe? Why do these differences arise? What suggestions do you derive in order to better compare how good algorithms are?

The solution to the theoretical exercises will be discussed in the exercise class in B2 on May 18, 2026, at Mathelounge in B6 B301.