FAKULTÄT FÜR WIRTSCHAFTSINFORMATIK
UND WIRTSCHAFTSMATHEMATIK

UNIVERSITÄT MANNHEIM

Prof. Dr. Leif Döring
André Ferdinand, Sara Klein

Reinforcement Learning

10. Excercise Sheet

1. SoftMax parameterisation

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 $\Phi(s,a)$

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

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

Show that the constant baseline b in Theorem 5.2.1 can be replaced by any deterministic state-dependent baseline $b : \mathcal{S} \to \mathbb{R}$, i.e.

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

4. (Batch-)Stochastic policy gradient algorithm

Implement algorithm 32 (REINFORCE- Batch Stochastic Policy Gradient Algorithm) for the finite Ice Vendor example of the lecture.