Prof. Joschka Boedecker, Gabriel Kalweit, Maria Huegle, Andreas Saelinger

REINFORCEMENT LEARNING
Exercise 10

1 Score Function

Assume a task with two discrete actions 0 and 1. Instead of a Gaussian policy or a softmax, we can define the policy to follow a Bernoulli distribution by the sigmoid function $\sigma(\cdot)$ over a linear combination of state features s and parameters θ, i.e. $\pi(a = 1|s) = \sigma(s^T\theta)$ and $\pi(a = 0|s) = 1 - \sigma(s^T\theta)$.

Derive the score function for this policy.
Hint: the derivative of the sigmoid function is $\frac{d}{dx}\sigma(x) = \sigma(x)(1 - \sigma(x))$.

2 REINFORCE

(a) Implement the missing parts of the `Policy` class and the REINFORCE algorithm in `reinforce.py`. In this first part, we won't use a baseline. We again apply the algorithm on the discrete Mountain Car Environment, so use a softmax output for your policy network. An exemplary parameter setting is given in the script. You can use the neural network implementation provided or choose to replace it by your own.

(b) Introduce and fit the value function as a baseline. Write a short report about your experiments and compare with Q-learning.

3 Experiences

Make a post in thread *Week 10: Policy Gradient Methods* in the forum[1], where you provide a brief summary of your experience with this exercise and the corresponding lecture.

[1] `https://ilias.uni-freiburg.de/goto.php?target=crs_1837295&client_id=unifreiburg`