Prof. Joschka Bödecker, Gabriel Kalweit

Reinforcement Learning
Exercise 6

0 Lecture

Watch *Lecture 08: Advanced Policy Gradient Algorithms*[1] before the upcoming session on Friday, December 21.

1 REINFORCE

Algorithm 1 Monte Carlo Policy Gradient

procedure REINFORCE
  initialize parameters θ of policy π arbitrarily
  for each episode $\{s_1, a_1, r_2, \ldots, s_{T-1}, a_{T-1}, r_T\} \sim \pi_\theta$ **do**
    for $t = 1$ to T-1 **do**
      $\theta \leftarrow \theta + \alpha \nabla_\theta \log \pi_\theta(s_t, a_t)(G_t - V(s_t))$
  return θ

(a) Implement the missing parts of the `Policy` class and the REINFORCE algorithm – which is a simplification of the general Vanilla Policy Gradient algorithm from the lecture – 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) Now introduce and fit the value function as a baseline. Write a short report about your experiments and compare with Q-learning.

2 Line Walker

Solve this task using pen and paper. Imagine an agent that walks along a line. The states are described by the position of the agent and the goal it has to reach, i.e. $s_t = \binom{x_t}{g_t}$, where x_t is the position and g_t is the goal. The agent is always starting in state $s_0 = \binom{0}{1}$. The action space is continuous, so the agent follows a Gaussian policy, $a \sim \mathcal{N}(\mu(s), \sigma^2)$, with a parameterized mean $\mu(s) = s^{\mathrm{T}}\theta$, where θ are the

[1] `https://ilias.uni-freiburg.de/goto.php?target=xvid_1121353&client_id=unifreiburg`

parameters of the policy, and a fixed variance of 0.1, i.e. $\sigma^2 = 0.1$. The agent gets in state $s_{t+1} = s_t + a_t$ after applying action a_t in state s_t.

If the agent reaches $[0.95, 1.05]$, it gets a reward of 1 and the episode ends. If the agent is in a state with $x_i > 1.05$, it gets a reward of -1 and the episode ends. The same holds, if the agent is in a state with $x_i < 0$. The agent gets 0 reward otherwise.

Assume the weights are initialized with $\theta_0 = \binom{-0.4}{0.6}$. Given the two trajectories:

$$\text{traj}_1 = \left(\binom{0}{1}, \binom{0.5}{1}, \binom{1.1}{1} \right)$$

and

$$\text{traj}_2 = \left(\binom{0}{1}, \binom{0.5}{1}, \binom{1.0}{1} \right),$$

provide all updates of θ following the REINFORCE algorithm with a learning rate of $\alpha = 0.1$ and no baseline.

3 Experiences

Make a posts in thread *Week 07: Introduction to Policy Gradients* in the forum[2], where you provide a brief summary of your experience with this exercise, the corresponding lecture and the last meeting.

[2] `https://ilias.uni-freiburg.de/goto.php?target=frm_1121060&client_id=unifreiburg`