Prof. Joschka Bödecker, Gabriel Kalweit

Reinforcement Learning
**Sample Solution 6**

# 1 Line Walker

**Algorithm 1** Monte Carlo Policy Gradient

```
procedure REINFORCE
    initialize parameters θ of policy π arbitrarily
    for each episode {s_1, a_1, r_2, ..., s_{T-1}, a_{T-1}, r_T} ~ π_θ do
        for t = 1 to T-1 do
            θ ← θ + α ∇_θ log π_θ(s_t, a_t) G_t
    return θ
```

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 $\theta$ are the 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 $\theta$ following the REINFORCE algorithm with a learning rate of $\alpha = 0.1$ and no baseline.

**Solution** $\pi_\theta(a|s) \sim \mathcal{N}(s^{\mathrm{T}}\theta, \sigma^2)$, i.e. $\pi_\theta(a|s) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2}\frac{(s^{\mathrm{T}}\theta - a)^2}{\sigma^2})$.
The log yields

$$\log \pi(a|s) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(s^{\mathrm{T}}\theta - a)^2$$

and the gradient

$$\nabla_\theta \log \pi(a|s) = -\frac{1}{2\sigma^2}(s^{\mathrm{T}}\theta - a)2s = \frac{(a - s^{\mathrm{T}}\theta)s}{\sigma^2}.$$

$\text{traj}_1$:

- $\theta_0 = \binom{-0.4}{0.6}$, $s_0 = \binom{0}{1}$, $a_0 = 0.5$

  $\rightarrow$ $\nabla_{\theta_0} \log \pi(a_{00}|s_{00}) = \frac{(0.5-(0,1)\binom{-0.4}{0.6})\binom{0}{1}}{0.1} = \binom{0}{-1}$

  $\rightarrow$ return is -1, learning rate is 0.1

  $\rightarrow$ $\theta_1 = \theta_0 + 0.1\binom{0}{-1}(-1) = \binom{-0.4}{0.7}$

- $\theta_1 \rightarrow \theta_2$ analogously

  $\rightarrow$ $\nabla_{\theta_1} \log \pi(a_{01}|s_{01}) = \binom{0.5}{1}$

  $\rightarrow$ return is -1, learning rate is 0.1

  $\rightarrow$ $\theta_2 = \binom{-0.45}{0.6}$

$\text{traj}_2$:

- $\theta_2 = \binom{-0.45}{0.6}$, $s_{10} = \binom{0}{1}$, $a_0 = 0.5$

  $\rightarrow$ $\nabla_{\theta_0} \log \pi(a_{10}|s_{10}) = \frac{(0.5-(0,1)\binom{-0.45}{0.6})\binom{0}{1}}{0.1} = \binom{0}{-1}$

  $\rightarrow$ return is 1, learning rate is 0.1

  $\rightarrow$ $\theta_3 = \theta_2 + 0.1\binom{0}{-1}1 = \binom{-0.4}{0.5}$

- $\theta_3 \rightarrow \theta_4$ analogously

  $\rightarrow$ $\nabla_{\theta_3} \log \pi(a_1|s_1) = \binom{1.125}{2.25}$

  $\rightarrow$ return is 1, learning rate is 0.1

  $\rightarrow$ $\theta_4 = \binom{-0.3375}{0.725}$