RL: Policy Search

Score Function

Marius Lindauer

Leibniz Universität Hannover

Winter Term 2021

Score Function

Define score function as:

$$\nabla_\theta \log \pi_\theta(s, a)$$

Likelihood Ratio + Score Function Policy Gradient

- Putting this together
- Our goal is to find the policy parameters θ^*

$$\theta^* \in \arg\max_\theta V(\theta) = \arg\max_\theta \sum_\tau P(\tau;\theta)R(\tau)$$

- Approximate with empirical estimate for m sample trajectories under policy π_θ:

$$\begin{aligned}\nabla_\theta V(\theta) &\approx \frac{1}{m}\sum_{i=1}^{m} R(\tau^{(i)})\nabla_\theta \log P(\tau^{(i)};\theta) \\ &= \frac{1}{m}\sum_{i=1}^{m} R(\tau^{(i)}) \sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t^{(i)} \mid s_t^{(i)})\end{aligned} \tag{1}$$

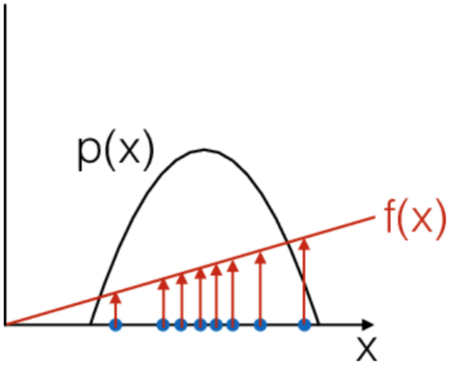
$\leadsto$ Do not need to know dynamics model!

Score Function Gradient Estimator: Intuition

- Consider generic form of $R(\tau^{(i)})\nabla_\theta \log P(\tau^{(i)};\theta)$:
  $\hat{g}_i = f(x_i)\nabla_\theta \log p(x_i \mid \theta)$
- $f(x)$ measures how good the sample x is
- Moving in the direction $\hat{g}_i$ pushes up to the logprob of the sample, in proportion of how good it is
- Valid even if $f(x)$ is discontinuous; and unknown; or sample space (containing x) is a discrete set

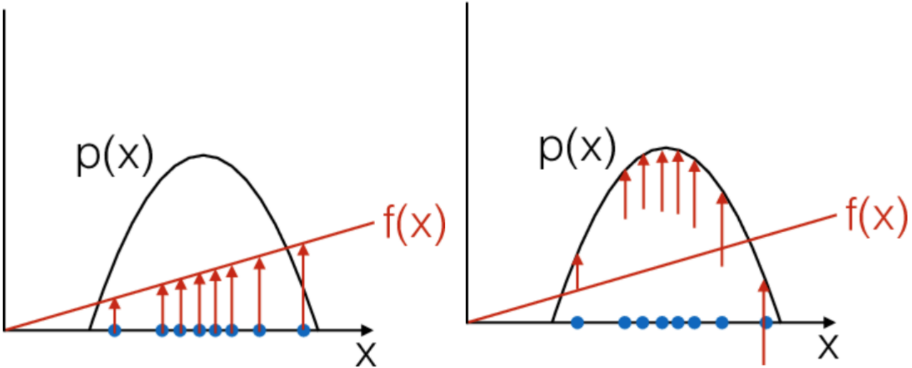
Score Function Gradient Estimator: Intuition

$$\hat{g}_i = f(x_i)\nabla_\theta \log p(x_i \mid \theta)$$



Score Function Gradient Estimator: Intuition

$$\hat{g}_i = f(x_i)\nabla_\theta \log p(x_i \mid \theta)$$



Policy Gradient Theorem

The policy gradient theorem generalizes the likelihood ratio approach:

Theorem

For any differentiable policy π_θ, the policy gradient is

$$\nabla_\theta = \mathbb{E}_{\pi_\theta}[\nabla_\theta \log \pi_\theta(s,a) Q^{\pi_\theta}(s,a)]$$