IDA LAB

INTELLIGENT DATA ANALYTICS SALZBURG

PARIS
LODRON
UNIVERSITÄT
SALZBURG

Simon Hirländer

Tutorial RL4AA

Probabilistic trajectories

- Objective if episodic: $J(\theta) = V^{\pi_\theta}(s_0) := V(\theta)$

- ➡ Stochastic search: pure random search, Simplex, Bayesian optimization

- Using the gradient:

➡ $$V(\theta) = \sum_{\tau} P(\tau; \theta) R(\tau)$$

➡ $$\nabla_\theta V(\theta) = \sum_\tau P(\tau;\theta) R(\tau) \nabla_\theta \log P(\tau;\theta) = \mathbb{E}[R(\tau) \nabla_\theta \log P(\tau;\theta)]$$

➡ Sampling of $A_t \sim p(\cdot \mid \tau_t; \theta)$

- Handle probabilistic policies (example)

- High dimensional and continuous action spaces

- Reinforce algorithm considers temporal structure

➡ Finite difference approximation $\hat{=}$ Reinforce algorithm

72

Trajectory probability

Trajectory reward

Log likelihood trick

$$V^{\pi}(s_0) = \mathbb{E}_{\pi}[\sum_{t} \gamma^t R_{t+1} \mid S_t = s_0]$$

Stochastic gradient

Probabilistic trajectories

- Objective if episodic: $J(\theta) = V^{\pi_\theta}(s_0) := V(\theta)$
  - ➡ Stochastic search: pure random search, Simplex, Bayesian optimization
- Using the gradient:

$$V^{\pi}(s_0) = \mathbb{E}_{\pi}[\sum_t \gamma^t R_{t+1} \,|\, S_t = s_0]$$

Trajectory probability

➡ $$V(\theta) = \sum_\tau \boxed{P(\tau;\theta)}\boxed{R(\tau)}$$

Trajectory reward

➡ $$\nabla_\theta V(\theta) = \sum_\tau P(\tau;\theta)R(\tau)\boxed{\nabla_\theta \log P(\tau;\theta)} = \mathbb{E}[R(\tau)\nabla_\theta \log P(\tau;\theta)]$$

Log likelihood trick Stochastic gradient

➡ Sampling of $A_t \sim p(\cdot \,|\, \tau_t;\theta)$

- Handle probabilistic policies (example)
- High dimensional and continuous action spaces
- Reinforce algorithm considers temporal structure

➡ Finite difference approximation $\hat{=}$ Reinforce algorithm

Why optimisation is so popular?

Optimisation and RL address different objectives:

- Optimization objective: searching an optimum a function by varying the parameters of this function
  - ➡ Optimization adapts to changes since it is usually ran from scratch
- RL maximises the cumulative reward on an MDP:
  - ➡ Runs fast if the MDP is not modified to strongly
  - ➡