

Model Free Control

Generalized Policy Iteration

Marius Lindauer



Winter Term 2021

Recall Policy Iteration

- ▶ Initialize policy π
- ▶ Repeat:
 - ▶ Policy evaluation: compute V^π
 - ▶ Policy improvement: update π

$$\pi'(s) \in \arg \max_{a \in A} R(s, a) + \gamma \sum_{s' \in S} P(s' \mid s, a) V^\pi(s') = \arg \max_{a \in A} Q^\pi(s, a)$$

- ▶ Now want to do the above two steps **without** access to the true dynamics and reward models
- ▶ Before we introduced methods for model-free policy evaluation

Model Free Policy Iteration

- ▶ Initialize policy π
- ▶ Repeat:
 - ▶ Policy evaluation: compute Q^π
 - ▶ Policy improvement: update π

MC for On-Policy Q-Evaluation

- ▶ Initialize $\forall s \in S, a \in A$:
 - ▶ $N(s, a) = 0$
 - ▶ $G(s, a) = 0$
 - ▶ $Q^\pi(s, a) = 0$
- ▶ Loop
 - ▶ Using policy π sample episode $i = s_{i,1}, a_{i,1}, r_{i,1}, s_{i,2}, a_{i,2}, r_{i,2}, \dots, s_{i,T_i}$
 - ▶ $G_{i,t} = r_{i,t} + \gamma r_{i,t+1} + \gamma^2 r_{i,t+2} + \dots + \gamma^{T_i-t} r_{i,T_i}$
 - ▶ For each pair (s, a) visited in episode i
 - ▶ For first (or every) time t that (s, a) is visited in episode i :
 - ▶ $N(s, a) = N(s, a) + 1$
 - ▶ $G(s, a) = G(s, a) + G_{i,t}$
 - ▶ Update estimate $Q^\pi(s, a) = G(s, a) / N(s, a)$

Model-free Generalized Policy Improvement

- ▶ Given an estimate $Q^{\pi_i}(s, a) \forall s \in S, a \in A$
- ▶ Update new policy

$$\pi_{i+1}(s) \in \arg \max_{a \in A} Q^{\pi_i}(s, a)$$