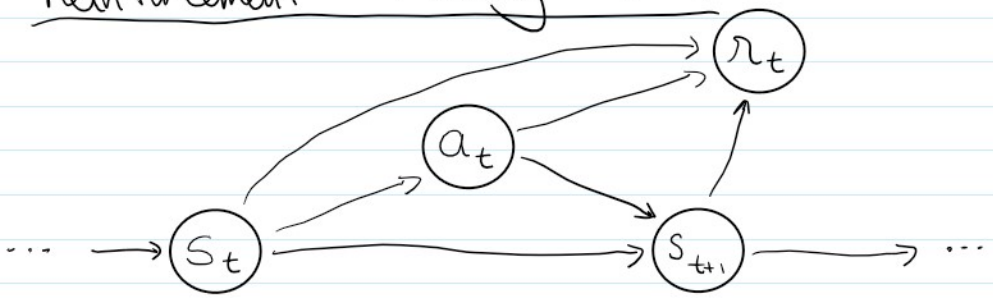
Tutorial #12

December 2, 2021 1:00 PM

Reinforcement learning (RL)



- An agent:
  1) begins in state s_t
  2) takes an action a_t
     $a_t \sim$ policy π
  3) moves to the next state s_{t+1}
     transition probs. $\mathbb{P}[s_{t+1} \mid s_t, a_t]$
  4) receives a reward/cost r_t

- An episode of length T:

$$\tau := (s_0, a_0, r_0, s_1, \dots, a_{T-1}, r_{T-1}, s_T).$$

- Goal of RL: find an optimal policy π that maximizes the value (or Q) function.

$$V^a(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^k r_t\right].$$

- Suppose we have a randomized policy π^θ.

$$\mathbb{P}(\tau \mid \text{randomized } \pi^\theta)$$

$$= \underbrace{\mathbb{P}(s_0)}_{\text{initial state}} \cdot \prod_{t=0}^{T-1} \underbrace{\pi^\theta(a_t \mid s_t)}_{\text{policy for } a_t} \cdot \underbrace{\mathbb{P}[s_{t+1} \mid s_t, a_t]}_{\text{prob. for } s_{t+1}}$$

Then $\nabla_\theta \log \mathbb{P}(\tau \mid \pi^\theta)$

$$= \nabla_\theta \left[\log \mathbb{P}(s_0) + \sum_{t=0}^{T-1} \log \pi^\theta(a_t \mid s_t) + \log \left(\mathbb{P}[s_{t+1} \mid s_t, a_t] \right) \right]$$

$$= \sum_{t=0}^{T-1} \nabla_\theta \log \pi^\theta(a_t \mid s_t).$$

- Likelihood-ratio trick:

$$\mathbb{P}[x;\theta] \nabla_\theta \log \mathbb{P}[x;\theta] = \nabla_\theta \mathbb{P}[x;\theta].$$

- Bellman Principle:

$$V(s) = \max_{a \in \mathcal{A}} \mathbb{E}\left[r(a,s,s') + \gamma V(s') \right]$$

- Q-function:

$$Q(s,a) = \mathbb{E}\left[r(a,s,s') + \gamma V(s') \right]$$

$$= \mathbb{E}\left[r(a,s,s') + \gamma \max_{a' \in \mathcal{A}} Q(s',a') \right]$$