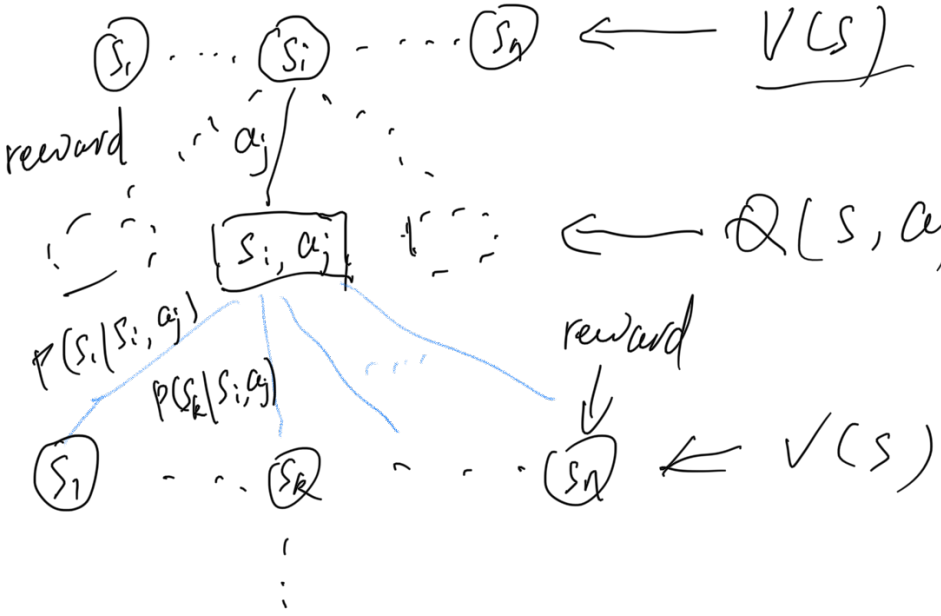
Lecture 12

$$V(s) = \max_{a_i} Q(s, a_i) \quad \text{action}$$

$$Q(s,a) = r(a) + \sum_{s'} P(s'|s,a)\left(r(s') + \lambda V(s')\right) \leftarrow \text{stochastic}$$



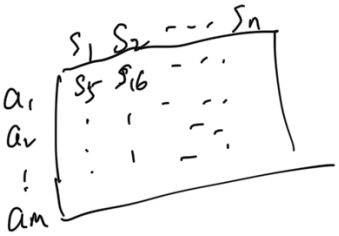
reward from action vs. reward at a state

Deterministic vs. stochastic

$(s,a) \to s'$

$P(s'|s,a) \quad s' \in \{s_1, \cdots, s_n\}$

$P(s,a,s')$ or $T(s,a,s')$



Deterministic:

$$\cdots r(s') + \lambda V(s') \quad \Leftarrow$$

$Q(s,a) = r(a) + \gamma \max_{a'} Q(s',a')$

stochastic:

$Q(s,a) = r(a) + \sum P(s'|s,a)\left(r(s') + \gamma V(s')\right)$ ←

on policy vs. off policy

Q-learning — off policy

Q-table with some initial values.

Go through a number of episode

start an episode with a s

Go through the episode step by step

↑ $Q = Q_{new}$ (Q table)

(assign updated Q_{new} as the Q)

episode — a sequece of action & state.

Goal state is the end of an episode.

or use the length limit to stop an episode.

→ based on Q, choose an action a,

calculate V.

$V = \max_a (Q(s,a))$

decide with a to take

Q-table

So take action a_{ij}.

→ Take action a, get reward, get to a state s'

if deterministic or real-time

to current Q

compute current Q

$Q_{curr} = r + \lambda V(s')$

if stochastic, in simulation collect data

collection (known transition probability $P(s'|s,a)$

$Q_{curr} = r(a) + \sum_{s'} (r(s') + \lambda V(s'))$

→ Update Q table

$Q_{new}(s,a) = Q(s,a) + \alpha (Q_{curr}(s,a) - Q(s,a))$

learning rate

$s \leftarrow s'$

continue until to the end of the episode