Markov Decision Processes

Terminology

States

Actions

Transition Model

Rewards

Markov property – results only depend on the current state, not on any history

All states are observable – we can always tell where we are

Stationarity - Transition model does not change with time

MDPs

More terminology

Utility or return – captures not just the current reward but also future rewards

Value function – the reward or utility associated with being in that state

Action-Value (or Q) function – the reward or utility assoc with a state and action

Policy – given a state, determine an action

Optimal policy – one that maximizes long term reward

How do we calculate the optimal policy?

TD Learning

Learning from experience

Gain some experience following a policy π

Update estimate of V for the nonterminal states St occurring in that experience.

$$V(S_t) \leftarrow V(S_t) + \alpha\Big[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)\Big].$$

Q - Learning

Q-learning: An off-policy TD control algorithm

Initialize $Q(s,a), \forall s \in \mathcal{S}, a \in \mathcal{A}(s)$, arbitrarily, and $Q(\textit{terminal-state}, \cdot) = 0$
Repeat (for each episode):
  Initialize S
  Repeat (for each step of episode):
    Choose A from S using policy derived from Q (e.g., ϵ-greedy)
    Take action A, observe R, S'
    $Q(S,A) \leftarrow Q(S,A) + \alpha\big[R + \gamma \max_a Q(S',a) - Q(S,A)\big]$
    $S \leftarrow S'$
  until S is terminal