Problem Solving with AI Techniques
Sequential Decision-Making under Uncertainty

Paul Weng

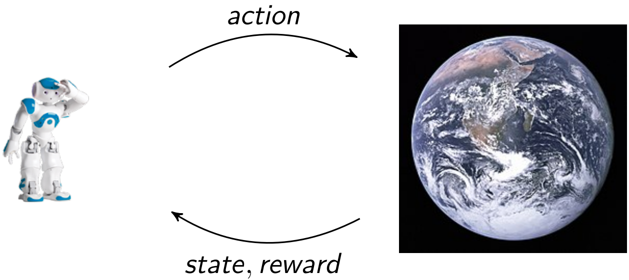
UM-SJTU Joint Institute

VE593, Fall 2018

JOINT INSTITUTE
交大密西根学院

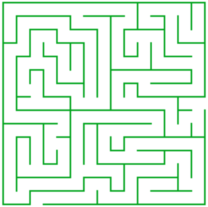
Decision-making

- Goal: design adaptive autonomous systems that can act on our behalf
- Examples:
  - Robotics
  - Traffic light controller
  - Data center control
  - Streaming video player
  - Intelligent tutoring system
- Interaction loop



Example: Autonomous Navigation

- Goal: move autonomously from point A to point B

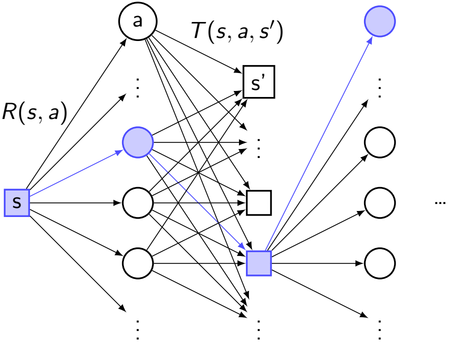


- Repeat: choose an action, observe new state, check if goal state
- Optimize travel time, power consumption...
- Search problem if effects of actions are deterministic
- Stochastic model: exogenous events, imperfect sensors/actuators, model error...

Markov Decision Process

Markov Decision Process (MDP) defined as a quadruplet $(\mathcal{S}, \mathcal{A}, T, R)$

- $\mathcal{S}$ is a set of states
- $\mathcal{A}$ is a set of actions
- $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ is a transition function, $T(s, a, s') = \mathbb{P}(s' \mid s, a)$
- $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is a reward function



Examples of MDP

An MDP is defined by:

- $\mathcal{S}$ set of states
- $\mathcal{A}$ set of actions
- $T(s, a, s')$ transition function
- $R(s, a)$ reward function

Examples of MDP

An MDP is defined by:

- $\mathcal{S}$ set of states
- $\mathcal{A}$ set of actions
- $T(s, a, s')$ transition function
- $R(s, a)$ reward function

Example: Navigation of a robot

- Positions of the robot
- Possible moves
- Probabilistic effects of a move
- Reward encoding a certain task

Examples of MDP

An MDP is defined by:

- $\mathcal{S}$ set of states
- $\mathcal{A}$ set of actions
- $T(s, a, s')$ transition function
- $R(s, a)$ reward function

Example: Inventory control

- Stock levels
- Orders
- Stochastic demand
- Earnings - costs

Examples of MDP

An MDP is defined by:

- $\mathcal{S}$ set of states
- $\mathcal{A}$ set of actions
- $T(s, a, s')$ transition function
- $R(s, a)$ reward function

Example: Video games

- Resources, Units, Structures
- Explore, Exploit, Build...
- Uncertain consequences
- Utility

Examples of MDP

An MDP is defined by:

- $\mathcal{S}$ set of states
- $\mathcal{A}$ set of actions
- $T(s, a, s')$ transition function
- $R(s, a)$ reward function

Generally, T and R are not known $\Rightarrow$ Reinforcement learning
Moreover, if there is only one state with random rewards $\Rightarrow$ Multi-armed bandit

Goal in an MDP

- High-level goal: determine which action to choose in every state at every time step
- Important notions:
  - Trajectory $s_0, s_1, s_2 \ldots$: sequence of states
  - History $s_0, a_0, s_1, a_1, s_2 \ldots$: sequence of state/action
  - Policy π: stationary or not, Markov/history-dependent, deterministic/randomized
  - Horizon: finite or infinite
- Problem: find "best" policy for a given MDP

Discussions

- Assumptions
  - Stationary (time-homogeneous) model
  - Markov property
  - Discrete time
  - States are observable (MDP can be generalized to Partially Observable MDP or POMDP)
- MDP can be seen as an extension of
  - a deterministic state-space where actions have stochastic effects
  - a Markov chain where in every state we can choose the distribution over next states

How to Define "Best"?

- Value of a history = sum of rewards (possibly discounted by $\gamma \in (0, 1)$)
- A policy induces a probability distribution over histories
- Value $v^\pi(s)$ of a policy in a state = expectation of the value of histories it can generate
- This defines the value function v^π of a policy π
- Optimal value function $v^*(s) = \max_\pi \mathbb{E}_\pi[\sum_t \gamma^t R(S_t, A_t) \mid S_0 = s]$

- Theorem: For **finite** horizon problems, finite MDP admits an optimal policy that is Markov and deterministic
- Theorem: For **infinite** horizon problems, finite MDP admits an optimal policy that is stationary, Markov, and deterministic

Induction for Value Function of Fixed Policy

- Induction for π:

$$
\begin{aligned}
v_0^\pi(s) &= 0 \\
v_t^\pi(s) &= R(s,\pi(s)) + \gamma \sum_{s'} T(s,\pi(s),s') v_{t-1}^\pi(s')
\end{aligned}
$$

- v_T^π provides the value function of π for finite horizon T
- v_t^π converges to value function v^π of π for infinite horizon
- value function of π for infinite horizon satisfies:

$$
v^\pi(s) = R(s,\pi(s)) + \gamma \sum_{s'} T(s,\pi(s),s') v^\pi(s')
$$

Backward Induction and Value Iteration

- Induction:

$$
\begin{aligned}
v_0^*(s) &= 0 \\
v_t^*(s) &= \max_a \; R(s,a) + \gamma \sum_{s'} T(s,a,s') v_{t-1}^*(s')
\end{aligned}
$$

- v_T^* provides optimal value function for finite horizon T
- v_t^* converges to optimal value function v^* for infinite horizon
- Optimal value function for infinite horizon satisfies:

$$
v^*(s) = \max_a R(s,a) + \gamma \sum_{s'} T(s,a,s') v^*(s')
$$

Value Iteration: Algorithm

ValueIteration(MDP, ε)
$\forall s, v^*(s) = 0$
repeat
  for $s \in \mathcal{S}$ **do**
    for $a \in \mathcal{A}$ **do**
      $Q^*(s,a) = R(s,a) + \gamma \sum_{s'} T(s,a,s')v^*(s')$
    $nv(s) \leftarrow \max_a Q^*(s,a)$
  $v^* \leftarrow nv$
until $||v^* - nv|| \leq \varepsilon$;
return Q^*

- Computational complexity for one step: $O(|\mathcal{S}|^2|\mathcal{A}|)$
- Computational complexity: $O(\text{poly}(|\mathcal{S}|, |\mathcal{A}|, \frac{1}{1-\gamma})$

Policy Iteration: Principle

- Given current policy π, improve it with one-step look-ahead
- Alternate between
  - Policy evaluation of current policy
    $v^\pi_t(s) = R(s, \pi(s)) + \gamma \sum_{s'} T(s, \pi(s), s') v^\pi_{t-1}(s')$
  - Policy improvement of current policy:
    $\pi'(s) = \arg\max_a R(s, a) + \gamma \sum_{s'} T(s, a, s') v^\pi(s')$

Policy Iteration

PolicyIteration(MDP)
initialize π
repeat
    $\pi' \leftarrow \pi$
    compute v^π (possibly approximately)
    for $s \in \mathcal{S}$ **do**
        for $a \in \mathcal{A}$ **do**
            $Q^\pi(s, a) = R(s, a) + \gamma \sum_{s'} T(s, a, s') v^\pi(s')$
        $\pi(s) \leftarrow \arg\max_a Q^\pi(s, a)$
until $\pi \neq \pi'$;
return π

- Computational complexity: $O(\text{poly}(|\mathcal{S}|, |\mathcal{A}|, \frac{1}{1-\gamma}))$