Markov Decision Processes

Chen-Yu Wei

Sequence of Actions

To win the game, the learner has to take a sequence of actions $a_1 \to a_2 \to \cdots \to a_H$.

One option: view every sequence as a “meta-action”: $\bar{a} = (a_1, a_2, \cdots, a_H)$

Drawback:

- The number of actions is exponential in horizon
- In stochastic environments, this does not leverage intermediate observations

Solution idea: dynamic programming

Interaction Protocol: Fixed-Horizon Case

For **episode** $t = 1, 2, \ldots, T$:

  For **step** $h = 1, 2, \ldots, H$:

    Learner observes an observation $x_{t,h}$

    Learner chooses an action $a_{t,h}$

    Learner receives instantaneous reward $r_{t,h}$

General case:

$$\mathbb{E}[r_{t,h}] = R(x_{t,1}, a_{t,1}, \ldots, x_{t,h}, a_{t,h}), \quad x_{t,h+1} \sim P(\cdot \mid x_{t,1}, a_{t,1}, \ldots, x_{t,h}, a_{t,h})$$

$\Rightarrow$ Optimal decisions may depend on the entire history $\mathcal{H}_t = (x_{t,1}, a_{t,1}, \ldots, x_{t,h})$

Interaction Protocol: Fixed-Horizon Case

For **episode** $t = 1, 2, \dots, T$:

  For **step** $h = 1, 2, \dots, H$:

    Learner observes an observation $x_{t,h}$

    Learner chooses an action $a_{t,h}$

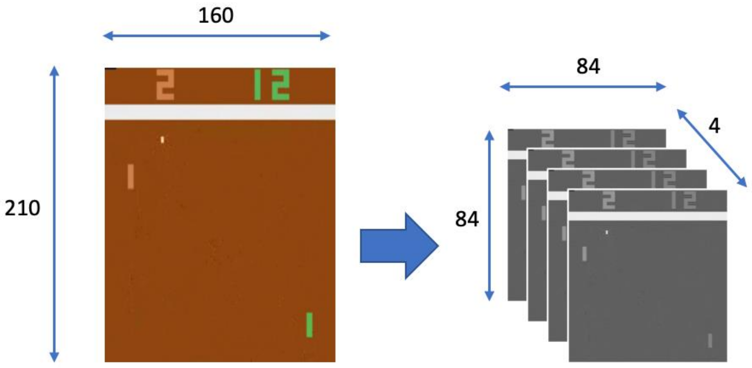
    Learner receives instantaneous reward $r_{t,h}$

We assume that the history $\mathcal{H}_t = (x_{t,1}, a_{t,1}, \dots, x_{t,h})$ can be summarized as a **horizon-length-independent** representation $s_{t,h} = \Phi(x_{t,1}, a_{t,1}, \dots, x_{t,h}) \in \mathcal{S}$ so that

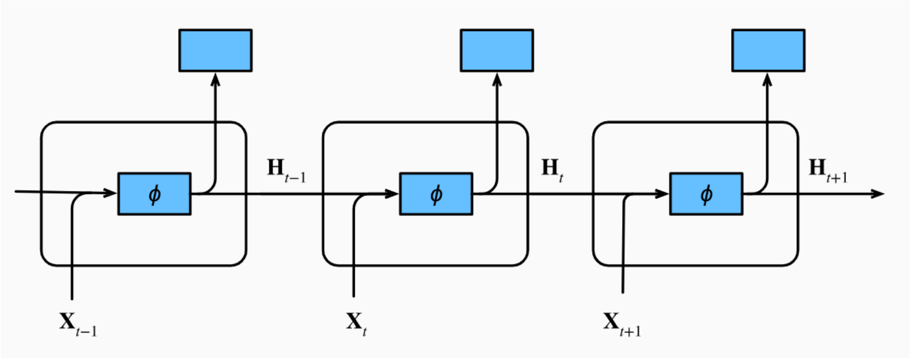
$$\mathbb{E}[r_{t,h}] = R(s_{t,h}, a_{t,h}), \quad x_{t,h+1} \sim P(\cdot \mid s_{t,h}, a_{t,h})$$

$s_{t,h}$ is called the "state" at the step h of episode t.

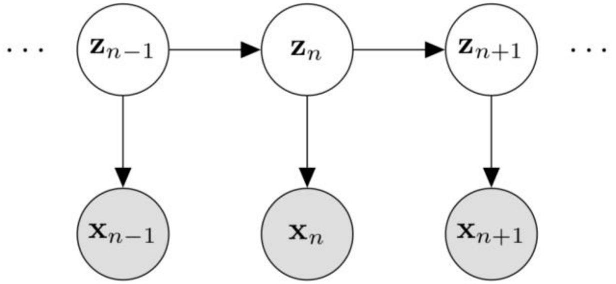
From Observations to States



Stacking recent observations



Recurrent neural network



Hidden Markov model

Interaction Protocol: Fixed-Horizon Case

For **episode** $t = 1, 2, \ldots, T$:

For **step** $h = 1, 2, \ldots, H$:

Environment reveals state $s_{t,h}$

Learner chooses an action $a_{t,h}$

Learner observes instantaneous reward $r_{t,h}$ with $\mathbb{E}[r_{t,h}] = R(s_{t,h}, a_{t,h})$

Next state is generated as $s_{t,h+1} \sim P(\cdot \mid s_{t,h}, a_{t,h})$

This is called the Markov decision process.

MDP as Contextual Bandits?

Viewing states as contexts, and viewing the problem as a contextual bandit problem with TH rounds.

$(x_{t,1}, a_{t,1} \ldots x_{t,h})$

$$\text{Regret (contextual bandit)} = \sum_{t=1}^{T}\sum_{h=1}^{H} \underbrace{\max_a R(s_{t,h}, a)} - \sum_{t=1}^{T} R(s_{t,h}, a_{t,h})$$

$$\text{Regret} = \sum_{t=1}^{T}\left[\sum_{h=1}^{H} R(s^*_{t,h}, a^*_{t,h})\right] - \sum_{t=1}^{T}\sum_{h=1}^{H} R(s_{t,h}, a_{t,h})$$

$$s^*_{t,1} = s_{t,1}$$

$$s^*_{t,h} \neq s_{t,h}$$

Formulations

- Interaction Protocol
  - Fixed-Horizon
  - Variable-Horizon (Goal-Oriented)
  - Infinite-Horizon
- Performance Metric
  - Total Reward
  - Average Reward
  - Discounted Reward
- Policy
  - History-Dependent Policy
  - Markov Policy
  - Stationary Policy

Horizon = Length of an episode

Interaction Protocols (1/3): Fixed-Horizon

Horizon length is a fixed number H

$h \leftarrow 1$

Observe initial state $s_1 \sim \rho$

While $h \leq H$:

Choose action a_h

Observe reward r_h with $\mathbb{E}[r_h] = R(s_h, a_h)$

Observe next state $s_{h+1} \sim P(\cdot \mid s_h, a_h)$

Examples: games with a fixed number of time

Interaction Protocols (2/3): Goal-Oriented

The learner interacts with the environment until reaching **terminal states** $\mathcal{T} \subset \mathcal{S}$

$h \leftarrow 1$

Observe initial state $s_1 \sim \rho$

While $s_h \notin \mathcal{T}$:

  Choose action a_h

  Observe reward r_h with $\mathbb{E}[r_h] = R(s_h, a_h)$

  Observe next state $s_{h+1} \sim P(\cdot \mid s_h, a_h)$

  $h \leftarrow h + 1$

Examples: video games, robotics tasks, personalized recommendations, etc.

Interaction Protocols (3/3): Infinite-Horizon

The learner continuously interacts with the environment

$h \leftarrow 1$

Observe initial state $s_1 \sim \rho$

Loop forever:

Choose action a_h

Observe reward r_h with $\mathbb{E}[r_h] = R(s_h, a_h)$

Observe next state $s_{h+1} \sim P(\cdot \mid s_h, a_h)$

$h \leftarrow h + 1$

Examples: network management, inventory management

Formulations for Markov Decision Processes

- Interaction Protocol
  - Fixed-Horizon
  - Variable-Horizon (Goal-Oriented)
  - Infinite-Horizon

Episodic setting

- Performance Metric
  - Total Reward
  - Average Reward
  - Discounted Reward
- Policy
  - History-Dependent Policy
  - Markov Policy
  - Stationary Policy

Performance Metric

Total Reward (for episodic settings): $$\sum_{h=1}^{\tau} r_h$$ (τ: the step where the episode ends)

Average Reward (for infinite-horizon setting): $$\lim_{T\to\infty} \frac{1}{T}\sum_{h=1}^{T} r_h$$

If $|r_h| \le 1$,

Discounted Total Reward (for episodic or infinite-horizon): $$\sum_{h=1}^{\tau} \gamma^{h-1} r_h \le \frac{1}{1-\gamma}$$

τ: the step where the episode ends, or ∞ in the infinite-horizon case
$\gamma \in [0,1)$: discount factor

Interaction Protocols vs. Performance Metrics

"natural" objective

Fixed-Horizon → Total Reward

Goal-Oriented → Total Reward   Could be unbonded

Infinite-horizon → Average Reward   Could have constant change for an infinitesimal change in policy

Discounted Total Reward?

Focusing more on the **recent** reward

Our Focus

In most of the following lectures, we focus on the **goal-oriented / infinite-horizon** setting with **discount total reward** as the performance metric.

Policy

A mapping from observations/contexts/states to (distribution over) actions

- Contextual bandits

$$a = \pi(x)$$

or $a \sim \pi(\cdot\,|x)$

- Multi-armed bandits

$$a \sim \pi$$

or $a = a^{\star}$

Policy for MDPs

History-dependent Policy

$$a_h \sim \pi(\cdot \mid s_1, a_1, r_1, s_2, a_2, r_2, \ldots, s_h)$$
$$a_h = \pi(s_1, a_1, r_1, s_2, a_2, r_2, \ldots, s_h)$$

Markov Policy

$$a_h \sim \pi(\cdot \mid s_h, h)$$
$$a_h = \pi(s_h, h)$$

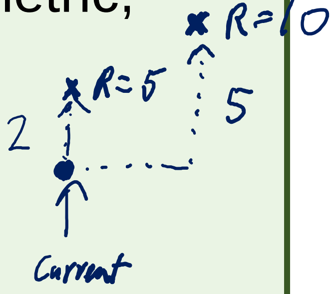
Stationary Policy

$$a_h \sim \pi(\cdot \mid s_h)$$
$$a_h = \pi(s_h)$$

Existence of a Stationary and Deterministic Optimal Policy

Theorem.

For goal-oriented or infinite-horizon setting with discounted total reward metric, there exists an optimal policy that is **stationary** and **deterministic**.



That is, there exists a stationary and deterministic policy $\pi^\star$ such that

$$\mathbb{E}\left[\sum_{h=1}^{\infty}\gamma^{h-1}r_h \;\middle|\; P,R,\rho,\pi^\star\right] \geq \left[\sum_{h=1}^{\infty}\gamma^{h-1}r_h \;\middle|\; P,R,\rho,\pi\right]$$

for any history-dependent, randomized policy π.

Remark. For fixed-horizon setting, we can only guarantee that there is an optimal policy which is **Markov** and **deterministic.** There may not be a stationary optimal policy.

Value Functions and Occupancy Measures

Value Functions

Let π be a stationary policy

$$V^\pi(s) = \mathbb{E}\left[\sum_{i=0}^{\infty} \gamma^i R(s_i, a_i) \;\middle|\; s_0 = s, \quad \forall i \geq 0:\ a_i \sim \pi(\cdot\,|s_i), \quad s_{i+1} \sim P(\cdot\,|s_i, a_i)\right]$$

$$Q^\pi(s,a) = \mathbb{E}\left[\sum_{i=0}^{\infty} \gamma^i R(s_i, a_i) \;\middle|\; (s_0, a_0) = (s, a), \quad \forall i \geq 1:\ a_i \sim \pi(\cdot\,|s_i), \quad \forall i \geq 0:\ s_{i+1} \sim P(\cdot\,|s_i, a_i)\right]$$

$$V^\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|\,s) Q^\pi(s,a)$$

$$Q^\pi(s,a) = R(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a)\, V^\pi(s')$$

Bellman Equation

$$V^{\pi}(s) = \mathbb{E}\left[\sum_{i=0}^{\infty} \gamma^i R(s_i, a_i) \,\middle|\, s_0 = s,\ a_i \sim \pi(\cdot|s_i),\ \forall i \geq 0\right]$$

$$= \mathbb{E}\left[R(s, a_0) + \sum_{i=1}^{\infty} \gamma^i R(s_i, a_i) \,\middle|\, s_0 = s,\ a_i \sim \pi(\cdot|s_i),\ \forall i \geq 0\right]$$

$$= \sum_a \pi(a|s) R(s,a) + \mathbb{E}\left[\sum_{i=1}^{\infty} \gamma^i R(s_i, a_i) \,\middle|\, s_0 = s,\ a_i \sim \pi(\cdot|s_i),\ \forall i \geq 1\right]$$

$$= \sum_a \pi(a|s) \left(R(s,a) + \mathbb{E}\left[\sum_{i=1}^{\infty} \gamma^i R(s_i, a_i) \,\middle|\, s_0 = s,\ a_i \sim \pi(\cdot|s_i),\ \forall i \geq 1\right]\right)$$

$$= \sum_a \pi(a|s)\ \mathbb{E}\left[\sum_{i=0}^{\infty} \gamma^i R(s_i, a_i) \,\middle|\, \underline{s_0 = s,\ a_0 = a},\ a_i \sim \pi(\cdot|s_i)\ \forall i \geq 1\right]$$

$$= \sum_a \pi(a|s)\ Q^{\pi}(s,a)$$

$$Q^{\pi}(s,a) = R(s,a) + \mathbb{E}\left[\sum_{i=1}^{\infty} \gamma^i R(s_i, a_i) \,\middle|\, s_0 = s,\ a_0 = a,\ a_i \sim \pi(\cdot|s_i)\ \forall i \geq 1\right]$$

$$= R(s,a) + \mathbb{E}\left[\sum_{i=1}^{\infty} \gamma^i R(s_i, a_i) \,\middle|\, s_1 \sim P(\cdot|s,a),\ a_i \sim \pi(\cdot|s_i)\ \forall i \geq 1\right]$$

$$= R(s,a) + \gamma\, \mathbb{E}\left[\sum_{i=1}^{\infty} \gamma^{i-1} R(s_i, a_i) \,\middle|\, s_1 \sim P(\cdot|s,a),\ a_i \sim \pi(\cdot|s_i)\ \forall i \geq 1\right] = R(s,a) + \gamma\, \mathbb{E}_{s' \sim P(\cdot|s,a)}\left[V^{\pi}(s')\right]$$

Dynamic Programming Viewpoint

Occupancy Measures

Let π be a stationary policy

$$d_\rho^\pi(s) = (1-\gamma)\mathbb{E}\left[\sum_{i=0}^{\infty} \gamma^i \mathbb{I}\{s_i = s\} \;\middle|\; s_0 \sim \rho, \qquad \forall i \geq 0\colon\ a_i \sim \pi(\cdot\,|s_i), \qquad s_{i+1} \sim P(\cdot\,|s_i, a_i)\right]$$

$$d_\rho^\pi(s,a) = (1-\gamma)\mathbb{E}\left[\sum_{i=0}^{\infty} \gamma^i \mathbb{I}\{s_i = s, a_i = a\} \;\middle|\; s_0 \sim \rho, \qquad \forall i \geq 0\colon\ a_i \sim \pi(\cdot\,|s_i), \qquad s_{i+1} \sim P(\cdot\,|s_i, a_i)\right]$$

$$d_\rho^\pi(s) = (1-\gamma)\rho(s) + \gamma \sum_{s',a'} d_\rho^\pi(s',a') P(s|s',a')$$

$$d_\rho^\pi(s,a) = d_\rho^\pi(s)\pi(a|s)$$