CSE 150A-250A AI: Probabilistic Models

Lecture 15

Fall 2025

Trevor Bonjour
Department of Computer Science and Engineering
University of California, San Diego

Slides adapted from previous versions of the course (Prof. Lawrence, Prof. Alvarado, Prof Berg-Kirkpatrick)

Agenda

Review

Value functions

Planning in MDPs

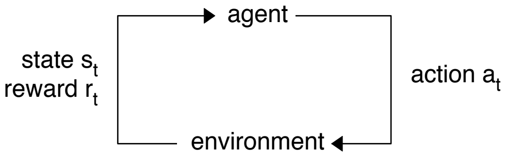
Policy Based

- Policy Evaluation
- Policy Improvement
- Policy Iteration

Review

Reinforcement learning (RL)

- Learning from experience in the world



- Formalization as Markov decision process

$\mathcal{S}$	state space
$\mathcal{A}$	action space
$P(s'\|s,a)$	transition probabilities
$R(s)$	reward function

MDP $\{\mathcal{S}, \mathcal{A}, P(s'|s,a), R(s)\}$

Decision-making in MDPs

- **Definition**

  A **policy** $\pi : \mathcal{S} \to \mathcal{A}$ is a mapping of states to actions.
  In this class we will only consider deterministic policies.

- **Number of policies**

  If there are $|\mathcal{A}|$ possible actions in each of $|\mathcal{S}|$ states, then there are *combinatorially* many policies:

  $$\#\,\text{policies} \;=\; |\mathcal{A}|^{|\mathcal{S}|}$$

- **Experience under policy π**

  $$\begin{array}{rcccccc} \text{state} & s_0 & \xrightarrow{\quad \pi(s_0) \quad} & s_1 & \xrightarrow{\quad \pi(s_1) \quad} & s_2 & \cdots \\ \text{reward} & r_0 & & r_1 & & r_2 & \cdots \end{array}$$

  Transitions occur with probabilities $P(s'|s, \pi(s))$.

Test your understanding

A policy π completely determines the next state s' that an agent will end up in after taking an action from state s.

True (A) or False (B)?

How to measure long-term return?

1. **Finite-horizon return**

$$\text{return} \;=\; \frac{1}{T}(r_0 + r_1 + \cdots + r_{T-1}) \quad \text{for a } T\text{-step horizon}$$

2. **Undiscounted return with infinite horizon**

$$\text{return} \;=\; \lim_{T\to\infty}\left[\frac{1}{T}\sum_{t=0}^{T-1} r_t\right]$$

These are the most obvious ways to accumulate rewards.
But they are **not** the most commonly used in practice …

How to measure long-term return? (con't)

3. **Discounted return with infinite horizon**

   Let $\gamma \in [0, 1)$ denote the so-called **discount factor**.
   Then define

$$\text{return} \;=\; r_0 + \gamma r_1 + \gamma^2 r_2 + \gamma^3 r_3 + \cdots \;=\; \sum_{t=0}^{\infty} \gamma^t r_t$$

What does it mean when the discount factor $\gamma << 1$?

A. Immediate and future rewards are valued equally.

B. Future rewards are heavily discounted compared to immediate.

C. Future rewards are lightly discounted compared to immediate.

D. Only future rewards are considered.

How to measure long-term return? (con't)

3. **Discounted return with infinite horizon**

   Let $\gamma \in [0, 1)$ denote the so-called **discount factor**.
   Then define

$$\text{return} \;=\; r_0 + \gamma r_1 + \gamma^2 r_2 + \gamma^3 r_3 + \cdots \;=\; \sum_{t=0}^{\infty} \gamma^t r_t$$

What does it mean when the discount factor $\gamma \sim 1$?

A. Immediate and future rewards are valued equally.

B. Future rewards are heavily discounted compared to immediate.

C. Future rewards are lightly discounted compared to immediate.

D. Only future rewards are considered.

How to measure long-term return? (con't)

3. **Discounted return with infinite horizon** Let $\gamma \in [0, 1)$

   denote the so-called **discount factor**.
   Then define

   $$\text{return} \;=\; r_0 + \gamma r_1 + \gamma^2 r_2 + \gamma^3 r_3 + \cdots \;=\; \sum_{t=0}^{\infty} \gamma^t r_t$$

   When $\gamma \ll 1$, future rewards are heavily discounted.
   These returns can be optimized by short-sighted agents.

   When γ is close to 1, future rewards are lightly discounted.
   These returns can only be optimized by far-sighted agents.

Motivation for $\gamma \in [0, 1)$

Psychologist: *Why discount rewards from the distant future?*
Economist: *Why favor investments with short-term payoffs?*

1. Intuition

   Many models are only approximations to the real world; we should not attempt to extrapolate them indefinitely.

2. Mathematical convenience

   Discounted returns lead to simple iterative algorithms with strong guarantees of convergence.

What to optimize?

The discounted return $\sum_{t=0}^{\infty} \gamma^t r_t$ is a random variable.
But we can try to optimize its expected value:

$$\mathrm{E}^{\pi}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t) \,\middle|\, s_0 = s\right]$$

the expected value of the discounted infinite-horizon return, starting in state s at time $t = 0$, and following policy π.

Maximizing the expected return is:

– generally wiser than maximizing the best-case return,
– but not as robust as minimizing the worst-case return.

Value functions

State value function

$$V^{\pi}(s) \;=\; \mathrm{E}^{\pi}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t) \,\middle|\, s_0 = s\right]$$

expected return,
starting in state s,
following policy π

- **Values versus rewards:**

  The reward $R(s)$ give immediate feedback to the agent.
  The value $V^{\pi}(s)$ computes the expected long-term return.

- **Types of behaviors:**

  Sacrifice now for long-term gain: $R(s) < 0$, $V^{\pi}(s) > 0$.
  Win now at the expense of later: $R(s) > 0$, $V^{\pi}(s) < 0$.

Properties of the state value function

- Experience under policy π

$$\begin{array}{rccccc} \text{state} & s_0 & \xrightarrow{\pi(s_0)} & s_1 & \xrightarrow{\pi(s_1)} & s_2 \;\cdots \\ \text{reward} & r_0 & & r_1 & & r_2 \;\cdots \end{array}$$

- **Adjacent states**

  States (s, s') can be visited in succession if $P(s'|s, \pi(s)) > 0$.

  The values $V^{\pi}(s)$ and $V^{\pi}(s')$ should be related, but how?

The **Bellman equation** tells us how.

Bellman equation

$$
\begin{aligned}
V^\pi(s) &= \mathrm{E}^\pi\left[R(s_0) + \gamma R(s_1) + \gamma^2 R(s_2) + \cdots \,\middle|\, s_0 = s\right] \\
&= R(s) + \gamma\, \mathrm{E}^\pi\left[R(s_1) + \gamma R(s_2) + \cdots \,\middle|\, s_0 = s\right] \\
&= R(s) + \gamma \sum_{s'} P(s'|s,\pi(s))\, \mathrm{E}^\pi\left[R(s_1) + \gamma R(s_2) + \cdots \,\middle|\, s_1 = s'\right] \\
&= R(s) + \gamma \sum_{s'} P(s'|s,\pi(s))\, V^\pi(s')
\end{aligned}
$$

The Bellman equation is the basis for much that will follow:

$$
\boxed{V^\pi(s) = R(s) + \gamma \sum_{s'} P(s'|s,\pi(s))\, V^\pi(s')}
$$

Action value function

$$Q^\pi(s,a) = \mathrm{E}^\pi\left[\sum_{t=0}^{\infty} \gamma^t R(s_t) \middle| s_0 = s, a_0 = a\right]$$

expected return,
starting from state s,
taking action a,
then following policy π

- **Motivation**

  Useful to imagine how small changes affect expected outcomes.
  What if (just once) the agent acted differently in state s?

- **Analogous to the Bellman equation:**

$$\begin{aligned} Q^\pi(s,a) &= R(s) + \gamma \sum_{s'} P(s'|s,a)\, V^\pi(s') \\ V^\pi(s) &= R(s) + \gamma \sum_{s'} P(s'|s,\pi(s))\, V^\pi(s') \end{aligned}$$

Optimality

- **Goal**

  Find the optimal policy given the environment that the agent is in.

- **Planning**

  If reward function and transition probabilities are known.

- **Reinforcement Learning**
  If reward function and transition probabilities are unknown.

Optimality

- **Theorem**

  There exists at least one policy π^* (and perhaps many) such that $V^{\pi^*}(s) \geq V^{\pi}(s)$ for all policies π and states s of the MDP.

- **Notation**

$$\begin{aligned} V^*(s) &= V^{\pi^*}(s) \\ Q^*(s,a) &= Q^{\pi^*}(s,a) \end{aligned}$$

  These optimal value functions are **unique**.
  (All optimal policies share the same value functions.)

Relations at optimality

- From the optimal action value function:

$$
\begin{aligned}
V^*(s) &= \max_a \left[Q^*(s,a)\right] \\
\pi^*(s) &= \operatorname*{argmax}_a \left[Q^*(s,a)\right]
\end{aligned}
$$

- From the optimal state value function:

$$
\begin{aligned}
Q^*(s,a) &= R(s) + \gamma \sum_{s'} P(s'|s,a) V^*(s') \\
\pi^*(s) &= \operatorname*{argmax}_a \left[R(s) + \gamma \sum_{s'} P(s'|s,a) V^*(s')\right]
\end{aligned}
$$

- Why are these relations useful?

  Sometimes it can be easier to estimate $Q^*(s,a)$ or $V^*(s)$ (which are continuous) than to learn $\pi^*(s)$ (which is discrete).

Planning in MDPs

Planning in MDPs

Given a complete model of the agent and its environment as a Markov decision process, namely

$$\text{MDP} = \{\mathcal{S}, \mathcal{A}, P(s'|s,a), R(s), \gamma\},$$

how can we *efficiently* compute (i.e., in time *polynomial in the number of states*) any of the following:

1. an optimal policy $\pi^*(s)$?
2. the optimal state value function $V^*(s)$?
3. the optimal action value function $Q^*(s,a)$?

This is the problem of **planning** in MDPs.

Policy Based

Algorithms

1. **Policy evaluation**

   How to compute $V^{\pi}(s)$ for some fixed policy π?

2. **Policy improvement**

   How to compute a policy π' such that $V^{\pi'}(s) \geq V^{\pi}(s)$?

3. **Policy iteration**

   How to compute an optimal policy $\pi^{*}(s)$?

Policy evaluation

- How to compute the state value function?

$$V^{\pi}(s) \;=\; \mathrm{E}^{\pi}\left[\sum_{t=0}^{\infty} \gamma^{t} R(s_t) \,\middle|\, s_0 = s\right]$$

- Bellman equation:

$$V^{\pi}(s) \;=\; R(s) \,+\, \gamma \sum_{s'} P(s'|s, \pi(s)) V^{\pi}(s')$$

- **Solve linear system:** There are n equations for n unknowns (where $s = 1, 2, \ldots, n$).

Solving the linear system

- From the Bellman equation:

$$V^\pi(s) \;=\; R(s) + \gamma \sum_{s'} P(s'|s,\pi(s))\, V^\pi(s').$$

- Rearranging terms:

$$\begin{aligned} R(s) &= V^\pi(s) - \gamma \sum_{s'} P(s'|s,\pi(s)) V^\pi(s') \\ &= \sum_{s'} \Big[\underbrace{I(s,s')}_{\text{identity matrix}} - \gamma P(s'|s,\pi(s)) \Big] V^\pi(s') \end{aligned}$$

- In matrix-vector form:

$$R \;=\; \left[I - \gamma P^\pi \right] V^\pi$$

$$\begin{bmatrix} \text{column vector of} \\ n \text{ known rewards} \end{bmatrix} = \begin{bmatrix} n \times n \text{ matrix} \\ \text{(known)} \end{bmatrix} \begin{bmatrix} \text{column vector of} \\ n \text{ unknown values} \end{bmatrix}$$

Solving the linear system (con't)

- **Solution**

$$R = \Big[I - \gamma P^{\pi}\Big] V^{\pi} \quad \Longrightarrow \quad V^{\pi} = \underbrace{(I - \gamma P^{\pi})^{-1}}_{\text{matrix inverse}} R$$

- **Complexity**

  It takes $O(n^3)$ operations to solve this system of equations.

- **Example**

  Let $\mathcal{S} = \{1, 2\}$ and $P(s'|s, \pi(s)) = 0.5$ for all (s, s').

$$\begin{bmatrix} V^{\pi}(1) \\ V^{\pi}(2) \end{bmatrix} = \left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \gamma \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix} \right)^{-1} \begin{bmatrix} R(1) \\ R(2) \end{bmatrix}.$$

Policy improvement

- **Problem statement**

  Given a policy π and its state value function $V^{\pi}(s)$, how to compute a policy π' such that

  $$V^{\pi'}(s) \geq V^{\pi}(s) \quad \text{for all states } s?$$

- **Definition**

  Given the action value function $Q^{\pi}(s, a)$ for policy π, we define the **greedy policy** π' by

  $$\pi'(s) = \operatorname*{argmax}_a \left[Q^{\pi}(s, a) \right].$$

  Why *greedy*? Because we change the action in state s to whatever appears to improve the expected return.

Greedy policies

- In terms of the state value function:

$$\begin{aligned}\pi'(s) &= \operatorname*{argmax}_a \left[Q^\pi(s,a) \right] \\ &= \operatorname*{argmax}_a \left[R(s) + \gamma \textstyle\sum_{s'} P(s'|s,a)\, V^\pi(s') \right] \\ &= \operatorname*{argmax}_a \left[\textstyle\sum_{s'} P(s'|s,a)\, V^\pi(s') \right]\end{aligned}$$

- Test your understanding:

$\pi'(s) = \pi(s)$ for some $s \in \mathcal{S}$? not necessarily

$\pi'(s) \neq \pi(s)$ for some $s \in \mathcal{S}$? not necessarily

$Q^\pi(s, \pi'(s)) \geq Q^\pi(s, \pi(s))$ for all $s \in \mathcal{S}$? TRUE

Policy improvement

- **Greedy policy:**

$$\pi'(s) \quad = \quad \operatorname*{argmax}_a Q^\pi(s,a)$$

- **Theorem:**
  The greedy policy $\pi'(s) = \arg\max_a Q^\pi(s,a)$ improves everywhere on the policy π from which it was derived:

$$V^{\pi'}(s) \;\geq V^\pi(s) \quad \text{for all states } s \in \mathcal{S}$$

- **Intuition:**
  If it's better to choose action a in state s before following π, then it's always better to make this choice.

- **Proof idea:**
  We'll prove a key inequality for *one-step deviations* from π, then we'll extend this inequality by an iterative argument.

Proof — 1. Deriving the inequality

- Comparing value functions:

$$\begin{aligned} V^\pi(s) &= Q^\pi(s, \pi(s)) \\ &\leq \max_a Q^\pi(s, a) \\ &= Q^\pi(s, \pi'(s)) \\ &= R(s) + \gamma \sum_{s'} P(s'|s, \pi'(s)) V^\pi(s') \end{aligned}$$

- Combining these steps:

$$V^\pi(s) \leq R(s) + \gamma \sum_{s'} P(s'|s, \pi'(s)) V^\pi(s')$$

- Intuition:

  It is better to take one step under π', then revert to π, than to always follow π.

Proof — 2. Leveraging the inequality

- **One-step inequality:**

$$V^{\pi}(s) \;\leq\; R(s) + \gamma \sum_{s'} P(s'|s, \pi'(s)) V^{\pi}(s')$$

  What happens if we plug this inequality into itself?
  Then we obtain ...

- **Two-step inequality:**

$$V^{\pi}(s) \;\leq\; R(s) + \gamma \sum_{s'} P(s'|s, \pi'(s)) \left[R(s') \,+\, \gamma \sum_{s''} P(s''|s', \pi'(s')) V^{\pi}(s'') \right]$$

- **Intuition:**

  It is better to take **two** steps under π', then revert to π, than to always follow π.

Proof — 3. Taking the limit

- **Two-step inequality:**

$$V^\pi(s) \;\leq\; R(s) + \gamma \sum_{s'} P(s'|s,\pi'(s)) \left[R(s') \;+\; \gamma \sum_{s''} P(s''|s',\pi'(s'))V^\pi(s'') \right]$$

- **Apply the inequality t times:**

  It is better to take t steps under π', then revert to π, than to always follow π. Last term is of order $O(\gamma^t)$.

- **Take the limit $t \to \infty$:**

  It is better to follow π' (always) than to follow π (always). Conclude that $V^\pi(s) \leq V^{\pi'}(s)$ for all states $s \in \mathcal{S}$.

Policy iteration

How to compute π^*?

1. **Choose an initial policy $\pi : \mathcal{S} \to \mathcal{A}$.**

2. **Repeat until convergence:**

   Compute the action value function $Q^\pi(s, a)$.
   Compute the greedy policy $\pi'(s) = \text{argmax}_a Q^\pi(s, a)$.
   Replace π by π'.

$$\pi_0 \xrightarrow{\text{evaluate}} \begin{matrix} V^{\pi_0}(s) \\ Q^{\pi_0}(s, a) \end{matrix} \xrightarrow{\text{improve}} \pi_1 \xrightarrow{\text{evaluate}} \begin{matrix} V^{\pi_1}(s) \\ Q^{\pi_1}(s, a) \end{matrix} \xrightarrow{\text{improve}} \cdots$$

Policy iteration is guaranteed to terminate.

True (A) or False (B)?

Policy iteration

- **How to compute π^*?**

$$\pi_0 \xrightarrow{\text{evaluate}} \begin{matrix} V^{\pi_0}(s) \\ Q^{\pi_0}(s,a) \end{matrix} \xrightarrow{\text{improve}} \pi_1 \xrightarrow{\text{evaluate}} \cdots$$

  This process is guaranteed to terminate.
  But does it converge to an optimal policy?

- **Theorem**

  If $\pi'(s) = \arg\max_a Q^\pi(s,a)$ and $V^{\pi'}(s) = V^\pi(s)$ for all $s \in \mathcal{S}$, then $V^\pi(s) = V^*(s)$ for all $s \in \mathcal{S}$.

- **Proof idea**

  Prove a key equality/inequality for terminal/non-terminal policies; iterate t times, then compare the limits as $t \to \infty$.

Proof — 1. Bellman optimality equation

- Suppose policy iteration converges to π'.

$$V^{\pi'}(s) = R(s) + \gamma \sum_{s'} P(s'|s,\pi'(s)) V^{\pi'}(s') \qquad \boxed{\text{Bellman equation}}$$

$$V^{\pi}(s) = R(s) + \gamma \sum_{s'} P(s'|s,\pi'(s)) V^{\pi}(s') \qquad \boxed{\text{at convergence}}$$

Now exploit that π' is greedy with respect to π ...

- Bellman optimality equation

$$\boxed{V^{\pi}(s) = R(s) + \gamma \max_a \sum_{s'} P(s'|s,a) V^{\pi}(s')}$$

These equations are nonlinear due to the max operation.
There are n equations for n unknowns (where $s = 1, 2, \ldots, n$).

Proof — 2. Inequality

- Let $\tilde{\pi}$ be any policy of the MDP:

$$
\begin{aligned}
V^{\tilde{\pi}}(s) &= R(s) + \gamma \sum_{s'} P(s'|s, \tilde{\pi}(s)) V^{\tilde{\pi}}(s') && \boxed{\text{Bellman equation}} \\
V^{\tilde{\pi}}(s) &\leq R(s) + \gamma \max_a \sum_{s'} P(s'|s, a) V^{\tilde{\pi}}(s') && \boxed{\text{greedy}}
\end{aligned}
$$

- Compare to Bellman optimality equation (BOE):

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

- Understanding the difference:

  The inequality holds for any policy $\tilde{\pi}$ of the MDP.
  The **BOE** only holds for a solution π from policy iteration.

Proof — 3. Taking the limit

- **Iterating the inequality:**

$$
\begin{aligned}
V^{\tilde{\pi}}(s) &\leq R(s) + \gamma \max_a \sum_{s'} P(s'|s,a) V^{\tilde{\pi}}(s') \\
&\leq R(s) + \gamma \max_a \sum_{s'} P(s'|s,a) \left[R(s') + \gamma \max_{a'} \sum_{s''} P(s''|s',a') V^{\tilde{\pi}}(s'') \right]
\end{aligned}
$$

- **Iterating the BOE:**

$$
\begin{aligned}
V^{\pi}(s) &= R(s) + \gamma \max_a \sum_{s'} P(s'|s,a) V^{\pi}(s') \\
&= R(s) + \gamma \max_a \sum_{s'} P(s'|s,a) \left[R(s') + \gamma \max_{a'} \sum_{s''} P(s''|s',a') V^{\pi}(s'') \right]
\end{aligned}
$$

- **Iterating t times:**

Both right sides agree up to term of order γ^t.
Taking the limit $t \to \infty$, we find $V^{\tilde{\pi}}(s) \leq V^{\pi}(s)$ for all $s \in \mathcal{S}$.

Since $\tilde{\pi}$ is arbitrary, we conclude that π is optimal.

That's all folks!