CSE 150A-250A AI: Probabilistic Models

Lecture 18

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

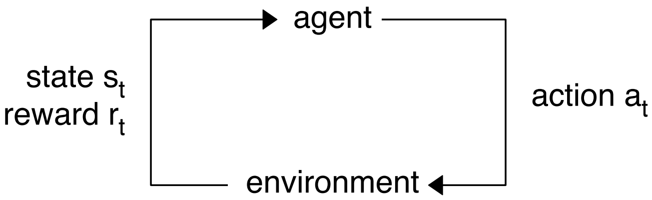
Review

TD Prediction

Q-learning

Review

Review



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

Given a model, we can plan using policy or value iteration.
But what if we aren't given the model?

1. Model-based approach: estimate $\hat{P}(s'|s,a)$ from experience.
2. Model-free approach: more on this today.

Stochastic approximation theory (con't)

How to estimate the mean of a random variable X from IID samples?

$$x_1,\ x_2,\ x_3,\ x_4,\ x_5,\ x_6,\ x_7,\ x_8,\ x_9, \ldots$$

2. **Incremental update**

$$\textbf{Initialize:} \quad \mu_0 = 0$$

$$\textbf{Update:} \quad \mu_t = (1-\alpha_t)\mu_{t-1} + \alpha_t x_t \qquad \text{for} \quad \alpha_t \in (0,1)$$

The update is a convex sum of the old estimate and latest sample.
It can also be written as:

$$\mu_t = \mu_{t-1} + \alpha_t(x_t - \mu_{t-1})$$

The corrective term $\boxed{x_t - \mu_{t-1}}$ is known as a **temporal difference**.
This is the simplest example of a temporal difference (TD) update.

What are the effects of using a higher step size (or learning rate) α when updating μ_t?

A. It gives more weight to recent samples.

B. It helps the estimate adapt more quickly to changes in the data.

C. It reduces sensitivity to noise and outliers.

D. A and B

E. A, B, and C

Temporal differences

- Update rule:

$$\mu_t = \mu_{t-1} + \alpha_t(x_t - \mu_{t-1})$$

Note how the corrective term is small on average when $\mu_{t-1} \approx \mathrm{E}[X]$

For convergence of the stochastic approximation estimate μ_t to the true mean $\mathrm{E}[X]$, what conditions must the step sizes α_t satisfy?

A. $\sum_{t=1}^{\infty} \alpha_t = \infty$ and $\sum_{t=1}^{\infty} \alpha_t^2 < \infty$

B. $\sum_{t=1}^{\infty} \alpha_t < \infty$ and $\sum_{t=1}^{\infty} \alpha_t^2 = \infty$

Temporal differences

- **Update rule:**

$$\mu_t = \mu_{t-1} + \alpha_t(x_t - \mu_{t-1})$$

Note how the corrective term is small on average when $\mu_{t-1} \approx \mathrm{E}[X]$

- **Theorem:** $\mu_t \to \mathrm{E}[X]$ as $t \to \infty$ with probability 1 if

$$\text{(i)} \quad \sum_{t=1}^{\infty} \alpha_t = \infty \quad (\textit{diverges})$$

$$\text{and} \quad \text{(ii)} \quad \sum_{t=1}^{\infty} \alpha_t^2 < \infty \quad (\textit{converges})$$

- **Intuition:**

  (i) α_t decays sufficiently slowly to incorporate many examples
  (ii) α_t decays sufficiently fast to converge in the limit

Temporal differences

- **Update rule:**

$$\mu_{t+1} \;=\; \mu_t \;+\; \alpha_t(x_{t+1} - \mu_t)$$

$$V_{t+1}(s_t) = V_t(s_t) \;+\; \alpha_v(s_t)\Big[x_t - V_t(s_t)\Big]$$

But what is x_t?
TD estimate of the expected future reward.

$$V_{t+1}(s_t) = V_t(s_t) \;+\; \alpha_v(s_t)\Big[R(s_t) + \gamma V_t(s_{t+1}) - V_t(s_t)\Big]$$

TD Prediction

Model-free policy evaluation

How to estimate $V^\pi(s)$ directly from experience w/o knowing $P(s'|s,a)$?

- Explore state space via policy π

$$
\begin{array}{rccccccc}
\text{action} & & \pi(s_0) & & \pi(s_1) & & \\
\text{state} & s_0 & \longrightarrow & s_1 & \longrightarrow & s_2 & \cdots \\
\text{reward} & r_0 & & r_1 & & r_2 & \cdots
\end{array}
$$

- Bellman equation (BE)

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

- Temporal difference prediction

$$
\begin{aligned}
&\textit{Initialize:} \quad & V_0(s) &= 0 \quad \text{for all} \quad s \in \mathcal{S} \\
&\textit{Update:} \quad & V_{t+1}(s_t) &= V_t(s_t) + \alpha_v(s_t)\Big[R(s_t) + \gamma V_t(s_{t+1}) - V_t(s_t)\Big]
\end{aligned}
$$

Model-free policy evaluation

How to estimate $V^\pi(s)$ directly from experience w/o knowing $P(s'|s,a)$?

- Explore state space via policy π

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

- Bellman equation (BE)

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

- Temporal difference prediction

$$\begin{array}{lrcl} \textit{Initialize:} & V_0(s) & = & 0 \quad \text{for all} \quad s \in \mathcal{S} \\ \textit{Update:} & V_{t+1}(s_t) & = & \underbrace{V_t(s_t)}_{\text{previous}} + \underbrace{\alpha_v(s_t)}_{\text{step}} \Big[\underbrace{R(s_t) + \gamma V_t(s_{t+1})}_{\text{sample from right side of BE}} - V_t(s_t) \Big] \end{array}$$

TD prediction

- **Incremental, model-free update**

  The state value function $V^\pi(s)$ is iteratively re-estimated from the most recent experience at each time step:

$$\begin{array}{rccc} \text{action} & & \pi(s_t) & \\ \text{state} & s_t & \longrightarrow & s_{t+1} \\ \text{reward} & r_t & & r_{t+1} \end{array}$$

$$V_{t+1}(s_t) \;=\; V_t(s_t) \;+\; \alpha_v(s_t)\Big[R(s_t) + \gamma V_t(s_{t+1}) - V_t(s_t)\Big]$$

- **Asymptotic convergence**

  Under suitable conditions, the TD update converges in the limit:

$$V_t(s) \to V^\pi(s) \quad \text{as} \quad t \to \infty \quad \text{for all} \quad s \in \mathcal{S}$$

Theorem

Assume that each state $s \in \mathcal{S}$ is visited infinitely often by policy π.

Allow the step size $\alpha_v(s)$ in each state $s \in \mathcal{S}$ to depend on the number of previous visits v to the state.

Assume the step sizes satisfy:

$$\sum_{v=1}^{\infty} \alpha_v(s) = \infty \quad \text{and} \quad \sum_{v=1}^{\infty} \alpha_v^2(s) < \infty.$$

Then the TD update

$$V_{t+1}(s_t) = V_t(s_t) + \alpha_v(s_t)\Big[R(s_t) + \gamma V_t(s_{t+1}) - V_t(s_t)\Big]$$

converges with probability one:

$$V_t(s) \to V^\pi(s) \quad \text{as} \quad t \to \infty.$$

Theory versus practice

- **Theory**

  For rigorous guarantees of convergence, agents should use step sizes that satisfy

$$\sum_{v=1}^{\infty} \alpha_v(s) = \infty \quad \text{and} \quad \sum_{v=1}^{\infty} \alpha_v^2(s) < \infty.$$

- **Practice**

  Many implementations choose small but constant step sizes.

  Remember — the MDP may only be an approximation to a world that is not completely stationary!

  In this situation, small constant step sizes are justified.

Q-learning

Model-free RL

- **Motivation**

  How to optimize policy π^* without model $P(s'|s,a)$?
  How to estimate $Q^*(s,a)$ without model $P(s'|s,a)$?

- **Bellman equation for optimal policy:**

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

  Equivalently, if we sample many transitions $s \xrightarrow{a} s'$,
  we must find that

$$
Q^*(s,a) = \mathbf{E}_{s'}\left[R(s) + \gamma \max_{a'} \left[Q^*(s',a') \right] \right]
$$

One-step Q-learning

- Explore state space at random:

$$\begin{array}{rccccc} \text{action} & & a_0 & & a_1 & \\ \text{state} & s_0 & \longrightarrow & s_1 & \longrightarrow & s_2 \quad \cdots \\ \text{reward} & r_0 & & r_1 & & r_2 \quad \cdots \end{array}$$

- Incremental update

  Initialize $Q_0(s,a) = 0$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$.
  Then update as follows:

$$Q_{t+1}(s_t, a_t) = \underbrace{Q_t(s_t, a_t)}_{\text{previous estimate}} + \alpha \Big[\underbrace{r_t + \gamma \max_{a'} Q_t(s_{t+1}, a')}_{\text{TD target}} - Q_t(s_t, a_t) \Big]$$

  This update is easy to implement, experience-based, and model-free.

- Q-learning is **off-policy** i.e. independent of current behavior.

Convergence of one-step Q-learning

- **Theorem** (*sketch*)

  If each state-action pair is visited infinitely many times, and each pair's step size $\alpha(s, a)$ is appropriately decayed, then these estimates converge (asymptotically):

$$\lim_{t\to\infty} Q_t(s, a) \to Q^*(s, a) \quad \text{with probability 1}$$

- **Practice**

  It is common to use a small but constant step size.
  An optimal policy π^* can be incrementally estimated by

$$\pi_t(s) = \mathrm{argmax}_a \Big[Q_t(s, a) \Big].$$

Exploration/Exploitation Tradeoff

- Experience

$$\begin{array}{rccccc} \text{action} & & a_0 & & a_1 & \\ \text{state} & s_0 & \longrightarrow & s_1 & \longrightarrow & s_2 \quad \cdots \\ \text{reward} & r_0 & & r_1 & & r_2 \quad \cdots \end{array}$$

- Update

$$Q_{t+1}(s_t, a_t) = \underbrace{Q_t(s_t, a_t)}_{\text{previous estimate}} + \alpha \Big[\underbrace{r_t + \gamma \max_{a'} Q_t(s_{t+1}, a')}_{\text{TD target}} - Q_t(s_t, a_t) \Big]$$

- Fundamental tradeoff

  The agent must explore the full state-action space to converge.
  But it also must exploit high-reward behaviors to converge quickly.
  How to balance?

Exploration strategies

1. **Random exploration**

   Choose action a_t at random for each state s_t.
   Q-learning will converge—but slowly—with this choice.

2. **Greedy exploration**

   Choose action $a_t = \arg\max_a Q_t(s_t, a)$.
   Q-learning is not guaranteed to converge.

3. **ϵ-greedy exploration**

   A compromise: explore greedily with probability $1 - \epsilon$
   and randomly with probability ϵ; this suffices to converge.

Algorithm

Algorithm 4 (Q-learning)

Input : MDP $M = \langle S, s_0, A, P_a(s' \mid s), r(s, a, s') \rangle$
Output : Q-function Q

Initialise Q arbitrarily; e.g., $Q(s, a) \leftarrow 0$ for all s and a

repeat
    $s \leftarrow$ the first state in episode e
    repeat (for each step in episode e)
        Select action a to apply in s;
            e.g. using Q and a multi-armed bandit algorithm such as ϵ-greedy
        Execute action a in state s
        Observe reward r and new state s'
        $\delta \leftarrow r + \gamma \cdot \max_{a'} Q(s', a') - Q(s, a)$
        $Q(s, a) \leftarrow Q(s, a) + \alpha \cdot \delta$
        $s \leftarrow s'$
    until s is the last state of episode e (a terminal state)
until Q converges

Course Evaluations

▸ Link

That's all folks!