CSE 250A. Principles of AI

Probabilistic Reasoning and Decision-Making

Lecture 19 – Value iteration and temporal differences

Lawrence Saul
Department of Computer Science and Engineering
University of California, San Diego

Fall 2021

Outline

1. Review and demos

2. Convergence of value iteration

3. Model-free reinforcement learning
   - Stochastic approximation theory
   - Temporal difference prediction

Planning in MDPs

Given an MDP $= \{\mathcal{S}, \mathcal{A}, P(s'|s,a), R(s)\}$ and discount factor γ, how to compute an optimal policy $\pi^*(s)$ or value function $V^*(s)$?

- **Policy iteration**

  Initialize policy at random.
  Iterate until convergence:

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

- **Value iteration**

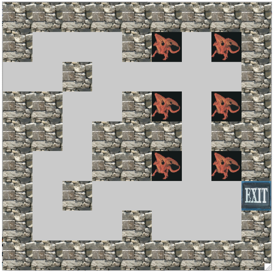
  Initialize $V_0(s) = 0$ for all states $s \in \mathcal{S}$.
  Iterate until convergence:

$$V_{k+1}(s) \;=\; R(s) + \gamma \max_a \sum_{s'} P(s'|s,a)\, V_k(s')$$

Value iteration demos

$$\mathcal{A} = \{\leftarrow, \uparrow, \rightarrow, \downarrow\}$$

$$R(s) = \begin{cases} -1 & \text{if } s \text{ has a dragon} \\ +1 & \text{if } s \text{ is an exit} \\ 0 & \text{otherwise} \end{cases}$$



demo	discount factor	drift
1	0.95	none
2	0.95	low
3	0.95	high
4	0.85	low

Outline

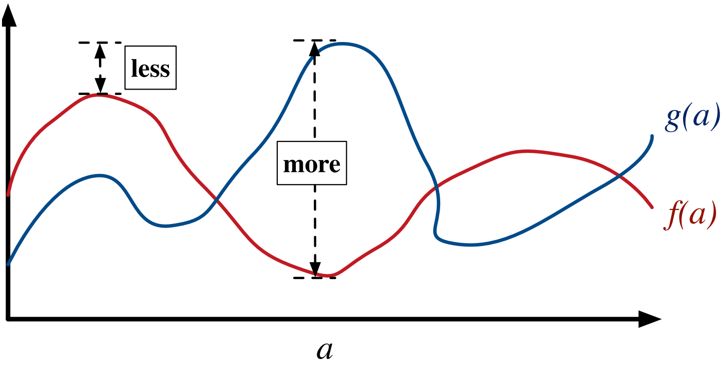
- **Review and demos**

- $\boxed{\textbf{Convergence of value iteration}}$

$$\lim_{k \to \infty} V_k(s) = V^*(s)$$

- **Model-free reinforcement learning**

Useful lemma



Lemma: for all functions f and g, we have:

$$\left|\max_a f(a) - \max_a g(a)\right| \leq \max_a \left|f(a) - g(a)\right|$$

Proof of lemma

1. As shorthand, let $a^* = \text{argmax}_a f(a)$.

2. It follows that:

$$\begin{aligned} \max_a f(a) - \max_a g(a) &= f(a^*) - \max_a g(a) \\ &\leq f(a^*) - g(a^*) \\ &\leq \max_a \left[f(a) - g(a) \right] \\ &\leq \max_a \left| f(a) - g(a) \right|. \end{aligned}$$

3. By symmetry:

$$\max_a g(a) - \max_a f(a) \leq \max_a \left| g(a) - f(a) \right|.$$

4. Together these inequalities establish the lemma:

$$\left| \max_a f(a) - \max_a g(a) \right| \leq \max_a \left| f(a) - g(a) \right|.$$

Value iteration

- **Algorithm**

$$\begin{aligned} \text{Initialize:} \quad V_0(s) &= 0 \text{ for all } s \in \mathcal{S}. \\ \text{Iterate:} \quad V_{k+1}(s) &= R(s) + \gamma \max_a \sum_{s'} P(s'|s,a) V_k(s'). \end{aligned}$$

- **Theorem**

Value iteration converges asymptotically: i.e., for all $s \in \mathcal{S}$,

$$\lim_{k \to \infty} V_k(s) \to V^*(s).$$

Also, the error at the k^{th} iteration is bounded by

$$\max_s \left| V_k(s) - V^*(s) \right| \leq \gamma^k \left(\frac{\max_s |R(s)|}{1-\gamma} \right).$$

Convergence of value iteration

- **Proof sketch**

Define the error at the k^{th} iteration by

$$\Delta_k \;=\; \max_s \left| V_k(s) - V^*(s) \right|.$$

Use the Bellman optimality equation (and lemma) to show

$$\Delta_{k+1} \;\leq\; \gamma \Delta_k.$$

Note that $\gamma < 1$, which establishes convergence.

Convergence of value iteration (con't)

- **Proof**

$$
\begin{aligned}
\Delta_{k+1} &= \max_s \left| V_{k+1}(s) - V^*(s) \right| \\
&= \max_s \left| \left[R(s) + \gamma \max_a \sum_{s'} P(s'|s,a) V_k(s') \right] \right. \\
&\qquad \left. - \left[R(s) + \gamma \max_a \sum_{s'} P(s'|s,a) V^*(s') \right] \right| \\
&= \gamma \max_s \left| \max_a \underbrace{\sum_{s'} P(s'|s,a) V_k(s')}_{f(a)} - \max_a \underbrace{\sum_{s'} P(s'|s,a) V^*(s')}_{g(a)} \right|
\end{aligned}
$$

Now apply the lemma ...

Convergence of value iteration (con't)

- **Proof (con't)**

From the lemma:

$$\left|\max_a f(a) - \max_a g(a)\right| \;\leq\; \max_a \left|f(a) - g(a)\right|.$$

It follows that:

$$\begin{aligned}
\Delta_{k+1} &= \gamma \max_s \left| \max_a \sum_{s'} P(s'|s,a) V_k(s') - \max_a \sum_{s'} P(s'|s,a) V^*(s') \right| \\
&\leq \gamma \max_s \max_a \left| \sum_{s'} P(s'|s,a) V_k(s') - \sum_{s'} P(s'|s,a) V^*(s') \right| \\
&= \gamma \max_s \max_a \left| \sum_{s'} P(s'|s,a) \Big[V_k(s') - V^*(s') \Big] \right|
\end{aligned}$$

Convergence of value iteration (con't)

- **Proof (con't)**

$$
\begin{aligned}
\Delta_{k+1} &\leq \gamma \max_s \max_a \left| \sum_{s'} P(s'|s,a) \left[V_k(s') - V^*(s') \right] \right| \\
&\leq \gamma \max_s \max_a \left| \max_{s'} \left[V_k(s') - V^*(s') \right] \right| \\
&\leq \gamma \max_s \max_a \max_{s'} \left| V_k(s') - V^*(s') \right| \\
&= \gamma \Delta_k \qquad \boxed{\text{with } \gamma < 1}
\end{aligned}
$$

Since $\Delta_{k+1} \leq \gamma \Delta_k$, by induction we have $\Delta_k \leq \gamma^k \Delta_0$.

Thus if Δ_0 is bounded, we have $\lim_{k \to \infty} \Delta_k \to 0$.

Convergence of value iteration (con't)

- **Proof (con't)**

  Assume the rewards are bounded. Then:

$$\begin{aligned}\Delta_0 &= \max_s |V_0(s) - V^*(s)|, \\ &= \max_s |V^*(s)|, \\ &\leq \max_s |R(s)| \, (1+\gamma+\gamma^2+\cdots), \\ &= \max_s |R(s)| \left(\frac{1}{1-\gamma}\right).\end{aligned}$$

  Thus we have shown:

$$\Delta_k \leq \left(\frac{\gamma^k}{1-\gamma}\right) \max_s |R(s)|,$$

  suggesting (intuitively) that smaller γ leads to faster convergence.

Outline

- **Review and demos**

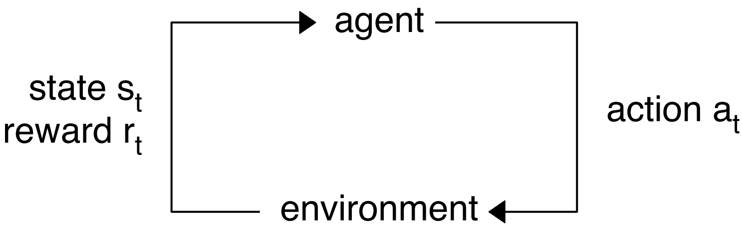
- **Convergence of value iteration**

- **Model-free reinforcement learning**

  Stochastic approximation theory

  Temporal difference prediction

Reinforcement learning



Consider the model $\{\mathcal{S}, \mathcal{A}, P(s'|s,a), R(s)\}$ defined by an MDP.

If we know the model, we can plan using policy or value iteration.

But what if we don't know $P(s'|s,a)$ and $R(s)$?

Can we learn an optimal policy *directly from experience*?

Model-based approach

- **Estimate model from experience**

  Explore world and estimate $\hat{P}(s'|s,a) \approx P(s'|s,a)$ from samples.
  Compute $\hat{\pi}^*(s)$ or $\hat{V}^*(s)$ from $\hat{P}(s'|s,a)$.

- **Benefits**

  A model $P(s'|s,a)$ is useful for *task transfer* — to retain knowledge when $R(s)$ or γ change but $P(s'|s,a)$ stays the same.

- **Costs**

  $P(s'|s,a)$ has $O(n^2)$ elements when $|\mathcal{S}| = n$.
  But $\pi^*(s)$, $V^*(s)$, and $Q^*(s,a)$ have only $O(n)$ elements.

Is it really necessary to estimate a model?

Model-free approach

- **Haiku**

  It is possible
  to optimize policies
  without a model.

- **But for this we need new tools:**

  Stochastic approximation theory
  Temporal difference (TD) learning

Stochastic approximation theory

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

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

1. **Sample average**

$$\mu_T \;=\; \frac{1}{T}\big(x_1 + x_2 + x_3 + \cdots + x_T\big)$$

This estimate converges to the mean by the law of large numbers:

$$\mu_T \rightarrow \mathrm{E}[X] \quad \text{as} \quad T \rightarrow \infty.$$

This is the most obvious estimate, but not the only one ...

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**

$$\begin{aligned} \textbf{Initialize:} \quad \mu_0 &= 0 \\ \textbf{Update:} \quad \mu_t &= (1-\alpha_t)\mu_{t-1} + \alpha_t x_t \qquad \text{for} \quad \alpha_t \in (0,1) \end{aligned}$$

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.

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

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**

$$\begin{array}{rccccc} \text{action} & & \pi(s_0) & & \pi(s_1) & \\ \text{state} & s_0 & \xrightarrow{\hspace{2em}} & s_1 & \xrightarrow{\hspace{2em}} & s_2 \quad \cdots \\ \text{reward} & r_0 & & r_1 & & r_2 \quad \cdots \end{array}$$

- **Bellman equation (BE)**

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

- **Temporal difference prediction**

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

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:

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

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.

Next lecture

1. **Wrap-up of reinforcement learning**

   Q-learning

   Exploration-exploitation tradeoff

   MDPs in large state spaces

   Partially observed MDPs

2. **What you've learned in 250A, and what comes next ...**

Final exam update

Basic information:

- It will be a take-home, open-book exam.
- It is designed to take 3-6 hours.
- Okay to check (but not do) work in Python, R, Matlab, Mathematica, etc.
- No collaboration is allowed.
- Questions will be broken down into simpler parts (like HW).
- Roughly speaking: 50% straightforward, 30% familiar, 20% stimulating.
- Solutions will be collected via Gradescope.
- For MS students in CSE: the final is the comprehensive exam.

Confirmed:

- Sun Dec 5 @ noon to Mon Dec 6 @ noon (PST)
- 10 questions with parts of varying difficulty
- 100 total points