A. What you thought 25%.
B. Better than exp. 25%.
C. Worse than you exp. 48%.

CSE 150A-250A AI: Probabilistic Models

D. Test2 Scores are out? 2%.

Lecture 17

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

Exploration vs Exploitation

Reinforcement Learning

- Stochastic approximation theory
- Temporal difference prediction

A. Milestone 1 / Milestone 2 or spoke TA

B. You haven't.

Review

Policy Improvement

- **Greedy policy:**

$$\pi'(s) \quad = \quad \underset{a}{\text{argmax}}\; Q^{\pi}(s, a)$$

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

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

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'.

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

Value iteration

- Idea in a nutshell

  Replace the equality sign in the Bellman optimality equation by an assignment operation:

$$V^*(s) = \max_a \left[R(s) + \gamma \sum_{s'} P(s'|s,a)\, V^*(s') \right] \quad \boxed{\text{BOE}}$$

$$V_{\text{new}}(s) \longleftarrow \max_a \left[R(s) + \gamma \sum_{s'} P(s'|s,a)\, V_{\text{old}}(s') \right] \quad \boxed{\text{algorithm}}$$

Algorithm for value iteration

1. **Initialize:** $V_0(s) = 0$ for all $s \in \mathcal{S}$.

2. **Iterate until convergence:**

$$V_{k+1}(s) \;=\; \max_a \left[R(s) \;+\; \gamma \sum_{s'} P(s'|s,a)\, V_k(s') \right] \text{ for all } s \in \mathcal{S}.$$

3. **Solve for optimal policy**:

$$\begin{aligned} Q_k(s,a) &= R(s) + \gamma \sum_{s'} P(s'|s,a)\, V_k(s'), \\ \pi^*(s) &= \lim_{k \to \infty} \operatorname{argmax}_a Q_k(s,a). \end{aligned}$$

Value iteration (VI) versus policy iteration (PI)

- **Compare and contrast:**

  PI searches through the combinatorial space of policies.
  VI searches through the continuous space of value functions.

- **Convergence:**

  PI converges in a finite number of steps.
  VI converges asymptotically (in the limit).

Exploration vs Exploitation


GOLD CHAMPION
BAR
Slot Machine!
GOLD CHAMPION
BAR
Slot Machine!
GOLD CHAMPION
BAR
Slot Machine!

Multi-Armed Bandit

- Stateless MDP: N one-armed bandits.

Multi-Armed Bandit

- Stateless MDP: N one-armed bandits.
- Each bandit pays a random reward from an unknown probability distribution. Some bandits are more likely to get a winning payoff than others.

Multi-Armed Bandit

- Stateless MDP: N one-armed bandits.
- Each bandit pays a random reward from an unknown probability distribution. Some bandits are more likely to get a winning payoff than others.
- **Goal**: Maximize the total rewards of a sequence of lever pulls.

Definition: A **multi-armed bandit** is defined by a set of random variables R_{at} where:

- $1 \leq a \leq N$, such that a is the arm of the bandit; and
- t the index of the play of arm a;

Multi-Armed Bandit

Definition: A **multi-armed bandit** is defined by a set of random variables R_{at} where:

- $1 \leq a \leq N$, such that a is the arm of the bandit; and
- t the index of the play of arm a;

Successive plays are assumed to be independently distributed, but we do not know the probability distributions of the random variables.

MAB - Action Value

Action value can be estimated:

$$Q(a) = \frac{1}{N(a)} \sum_{t=1}^{T} R_{at}$$

where t: number of rounds so far,

$N(a)$: number of times a was selected in previous rounds

R_{at} : reward obtained in the round t for playing arm a.

Exploitation vs Exploration dilemma

Goal: Maximize the reward

Exploitation vs Exploration dilemma

Goal: Maximize the reward

- Ideally, keep playing the actions that have given us the **best** reward.

Exploitation vs Exploration dilemma

Goal: Maximize the reward

- Ideally, keep playing the actions that have given us the **best** reward.
- Initially, we do not have enough information to tell us what the best actions are.

Exploitation vs Exploration dilemma

Goal: Maximize the reward

- Ideally, keep playing the actions that have given us the **best** reward.
- Initially, we do not have enough information to tell us what the best actions are.
- We want strategies that **exploit** what we think are the best actions so far, but still **explore** other actions.

Exploitation vs Exploration dilemma

Goal: Maximize the reward

- Ideally, keep playing the actions that have given us the **best** reward.
- Initially, we do not have enough information to tell us what the best actions are.
- We want strategies that **exploit** what we think are the best actions so far, but still **explore** other actions.

But how much should we exploit and how much should we explore? This is known as the exploration vs. exploitation dilemma.

$\epsilon-$first strategy

Explore the options uniformly for some time, and then once we are confident we have enough samples (when the changes to the $Q(a)$ of start to stabilize), we **exploit** $argmax_a Q(a)$.

$\epsilon-$first strategy

Explore the options uniformly for some time, and then once we are confident we have enough samples (when the changes to the $Q(a)$ of start to stabilize), we **exploit** $argmax_a Q(a)$.

ϵ determines how many rounds to select random actions before moving to the greedy action.

$\epsilon-$first strategy

Explore the options uniformly for some time, and then once we are confident we have enough samples (when the changes to the $Q(a)$ of start to stabilize), we **exploit** $argmax_a Q(a)$.

ϵ determines how many rounds to select random actions before moving to the greedy action.

Can we do better?

$\epsilon-$first strategy

Explore the options uniformly for some time, and then once we are confident we have enough samples (when the changes to the $Q(a)$ of start to stabilize), we **exploit** $argmax_a Q(a)$.

ϵ determines how many rounds to select random actions before moving to the greedy action.

Can we do better? Time is wasted equally in all actions using the uniform distribution. Instead, we can focus on the most promising actions given the rewards we have received so far.

$\epsilon-$greedy strategy

With some probability, $\epsilon \in [0, 1]$

- Choose a random arm with uniform probability. Update $Q(a)$.

$\epsilon-$greedy strategy

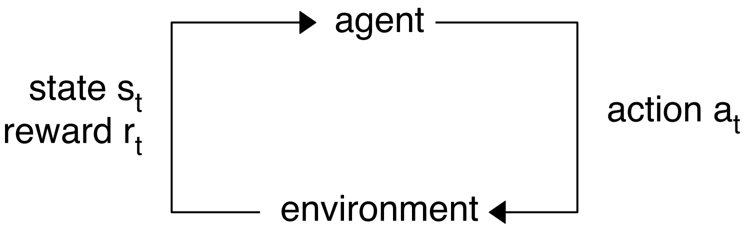
With some probability, $\epsilon \in [0, 1]$

- Choose a random arm with uniform probability. Update $Q(a)$.

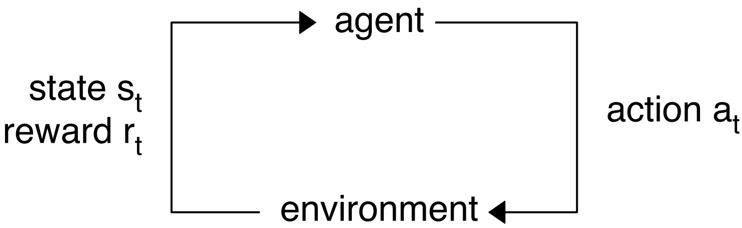
With probability, 1 - ϵ

- Choose arm with maximum action value: $argmax_a Q(a)$

Reinforcement Learning

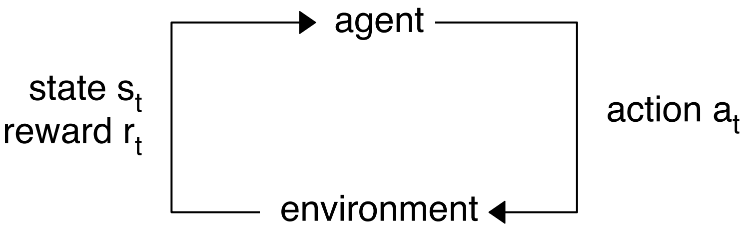

agent
state s_t
reward r_t
action a_t
environment

Reinforcement learning



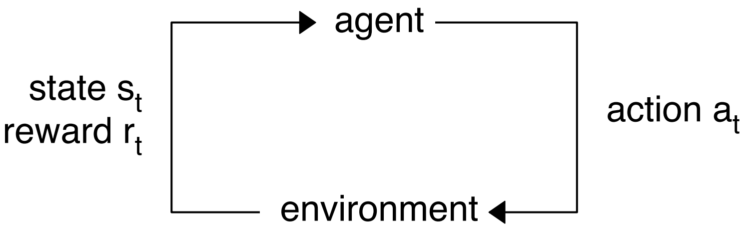
Consider the model $\{\mathcal{S}, \mathcal{A}, P(s'|s, a), R(s)\}$ defined by an MDP.

Reinforcement learning



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

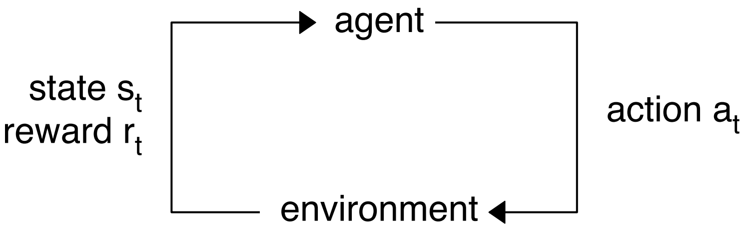
If we know the model, we can plan using policy or value iteration.



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 $R(s)$?



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

Model-based approach

- Estimate model 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.

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 $\hat{P}(s'|s,a)$.

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

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*

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.

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

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 $|\mathcal{S}| = n$.

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.

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

Model-free approach

- Haiku

Model-free approach

- Haiku

  It is possible

Model-free approach

- Haiku

  It is possible
  to optimize policies

Model-free approach

- Haiku

  It is possible
  to optimize policies
  without a model.

Model-free approach

- Haiku

  It is possible
  to optimize policies
  without a model.

- But for this we need new tools:

Model-free approach

- Haiku

  It is possible
  to optimize policies
  without a model.

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

  Stochastic approximation theory

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

Taking Averages Sample by Sample

Let's say I'm playing a game where I can score between 1 and 10 points. What would you predict my score would be the next time I play it? What if you knew that in the past, I have scored these scores (not necessarily in this order)

8, 8, 2, 5, 7, 2, 5, 7, 1, 3

What score would you predict I will get?

A. 3

B. 5

C. 7

D. 9

E. 10

Stochastic approximation theory

Stochastic approximation theory

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

Stochastic approximation theory

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

$$X_1,\ X_2,\ X_3,\ X_4,\ X_5,\ X_6,\ X_7,\ X_8,\ X_9\ \ldots$$

Stochastic approximation theory

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

$$X_1,\ X_2,\ X_3,\ X_4,\ X_5,\ X_6,\ X_7,\ X_8,\ X_9\ \ \dots$$

1. Sample average

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

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

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**

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

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

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:

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

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**

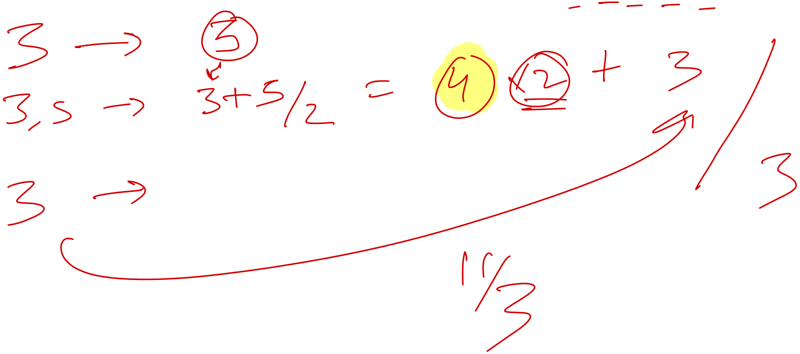
$$\mu_T \;=\; \frac{1}{T}\big(x_1 + x_2 + x_3 + \cdots + x_T\big)$$

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

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

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

Let’s average these numbers, in 5 iterations: 3, 5, 3, 8, 10



Taking averages, sample by sample

Let's average these numbers, in 5 iterations: 3, 5, 3, 8, 10

1. Score 3, Avg: 3

Taking averages, sample by sample

Let's average these numbers, in 5 iterations: 3, 5, 3, 8, 10

1. Score 3, Avg: 3
2. Score 5, Avg: $(1/2)3 + (1/2)5 = 4$

Taking averages, sample by sample

Let's average these numbers, in 5 iterations: 3, 5, 3, 8, 10

1. Score 3, Avg: 3
2. Score 5, Avg: (1/2)3 + (1/2)5 = 4
3. Score 3, Avg: (2/3)4 + (1/3)3 = (1 - 1/3)4 + (1/3)3 = 11/3

$1 - 1/3$

$= 2/3$

Taking averages, sample by sample

Let's average these numbers, in 5 iterations: 3, 5, 3, 8, 10

1. Score 3, Avg: 3
2. Score 5, Avg: (1/2)3 + (1/2)5 = 4
3. Score 3, Avg: (2/3)4 + (1/3)3 = (1 - 1/3)4 + (1/3)3 = 11/3
4. Score 8, Avg: (3/4)(11/3) + (1/4)8 = (1 - 1/4)(11/3) + (1/4)8 = 19/4

Taking averages, sample by sample

Let's average these numbers, in 5 iterations: 3, 5, 3, 8, 10

1. Score 3, Avg: 3
2. Score 5, Avg: (1/2)3 + (1/2)5 = 4
3. Score 3, Avg: (2/3)4 + (1/3)3 = (1 - 1/3)4 + (1/3)3 = 11/3
4. Score 8, Avg: (3/4)(11/3) + (1/4)8 = (1 - 1/4)(11/3) + (1/4)8 = 19/4
5. Score 10, Avg: (4/5) (19/4) + (1/5)10 = (1-1/5)(19/4) + (1/5)10 = 29/5 = 5.8

Taking averages, sample by sample

Let's average these numbers, in 5 iterations: 3, 5, 3, 8, 10

1. Score 3, Avg: 3
2. Score 5, Avg: (1/2)3 + (1/2)5 = 4
3. Score 3, Avg: (2/3)4 + (1/3)3 = (1 - 1/3)4 + (1/3)3 = 11/3
4. Score 8, Avg: (3/4)(11/3) + (1/4)8 = (1 - 1/4)(11/3) + (1/4)8 = 19/4
5. Score 10, Avg: (4/5) (19/4) + (1/5)10 = (1-1/5)(19/4) + (1/5)10 = 29/5 = 5.8

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

Taking averages, sample by sample

Let's average these numbers, in 5 iterations: 3, 5, 3, 8, 10

1. Score 3, Avg: 3
2. Score 5, Avg: (1/2)3 + (1/2)5 = 4
3. Score 3, Avg: (2/3)4 + (1/3)3 = (1 - 1/3)4 + (1/3)3 = 11/3
4. Score 8, Avg: (3/4)(11/3) + (1/4)8 = (1 - 1/4)(11/3) + (1/4)8 = 19/4
5. Score 10, Avg: (4/5) (19/4) + (1/5)10 = (1-1/5)(19/4) + (1/5)10 = 29/5 = 5.8

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

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

Taking averages, sample by sample

Let's average these numbers, in 5 iterations: 3, 5, 3, 8, 10

1. Score 3, Avg: 3
2. Score 5, Avg: (1/2)3 + (1/2)5 = 4
3. Score 3, Avg: (2/3)4 + (1/3)3 = (1 - 1/3)4 + (1/3)3 = 11/3
4. Score 8, Avg: (3/4)(11/3) + (1/4)8 = (1 - 1/4)(11/3) + (1/4)8 = 19/4
5. Score 10, Avg: (4/5) (19/4) + (1/5)10 = (1-1/5)(19/4) + (1/5)10 = 29/5 = 5.8

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

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

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

Let's say I'm playing a game where I can score between 1 and 10 points. What would you predict my score would be the next time I play it? What if you knew that in the past, I have scored these scores in **this order**:

1, 2, 3, 2, 5, 7, 5, 7, 8, 8

Is your guess about my next score higher, lower or the same as last time (5)?

A. Higher

B. Lower

C. The same

Stochastic approximation theory (con't)

Stochastic approximation theory (con't)

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

$$X_1,\ X_2,\ X_3,\ X_4,\ X_5,\ X_6,\ X_7,\ X_8,\ X_9, \ldots$$

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

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

    Initialize: $\mu_0 \quad = \quad 0$

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

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

$$\text{Update:} \quad \mu_t = (1-\alpha_t)\mu_{t-1} + \alpha_t X_t$$

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

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

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

Stochastic approximation theory (con't)

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

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

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.

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:

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

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**.

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:

Temporal differences

- Update rule:

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

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 $\mu_{t-1} \approx \mathrm{E}[X]$

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:

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

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

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

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

(i) $$\sum_{t=1}^{\infty} \alpha_t = \infty$$ (*diverges*)

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

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

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

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

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

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

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:**

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

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

Temporal differences

- Update rule:

Temporal differences

- Update rule:

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

Temporal differences

- Update rule:

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

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

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?

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.

- **Update rule:**

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

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

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

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

Model-free policy evaluation

Model-free policy evaluation

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

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

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

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

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

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

- Bellman equation (BE)

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)

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

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

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

Initialize: $V_0(s) = 0 \quad \text{for all} \quad s \in \mathcal{S}$

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

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

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

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

- Bellman equation (BE)

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

- Temporal difference prediction

$$\textit{Initialize:} \quad V_0(s) = 0 \quad \text{for all} \quad s \in \mathcal{S}$$

$$\textit{Update:} \quad V_{t+1}(s_t) = \underbrace{V_t(s_t)}_{\text{previous}} + \underbrace{\alpha_v(s_t)}_{\text{step}} \Big[\underbrace{R(s_t) + \gamma V_t(s_{t+1})}_{\text{sample from right side of BE}} - V_t(s_t) \Big]$$

TD prediction

- Incremental, model-free update

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:

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:

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

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:

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

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

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**

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:

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

Theorem

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

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.

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:

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:

$$\sum_{v=1}^{\infty} \alpha_v(s) \; = \; \infty$$

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:

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

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

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

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

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:

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 versus practice

- Theory

- Theory

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

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

- **Theory**

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

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

- **Practice**

Theory versus practice

- **Theory**

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

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

- **Practice**

  Many implementations choose small but constant step sizes.

Theory versus practice

- **Theory**

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

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

- **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!

Theory versus practice

- **Theory**

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

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

- **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.

That's all folks!