Reinforcement Learning 1

Kintan Saha

January 6, 2026

Outline

Why Machine Learning Theory?

- Machine Learning Theory seeks to answer a fundamental question: *When, why, and how do learning algorithms work?*
- It provides **provable guarantees** on performance, generalization, and stability of learning algorithms.
- When algorithms fail in practice, theory helps explain *why failure is unavoidable* under certain assumptions.
- Through this seminar series, we aim to showcase key ideas, techniques, and results across modern ML theory.

Why Not Just Take Courses?

- A full, rigorous treatment of ML theory typically requires multiple advanced courses.
- Completing these courses often takes 1–2 years of sustained effort.
- This seminar series is designed as a **gentle, structured introduction** to several subfields of ML theory.
- The goal is to provide intuition, key results, and clear pointers for deeper independent study.

Format of the Seminar Series

- The series will be organized as a sequence of **mini-courses** on different topics in ML theory.
- Example themes include:
  - Reinforcement Learning and Stochastic Approximation
  - Over-parameterized Models
  - Diffusion Models
- Each mini-course will consist of 3–4 lectures.
- These will be followed by faculty or industry talks showcasing state-of-the-art research.
- Where possible, we will include hands-on components such as paper reproductions or small projects.

Tentative Schedule

- **Jan 6, 8, 10, 12, 14**: Reinforcement Learning + Stochastic Approximation
- **Jan 15–18**: Reinforcement Learning Workshop (CSA)
- **Jan 24/25**: Talks by Prof. Shubhada Agrawal / Prof. Gugan Thoppe
- **Jan 31, Feb 7, 14, 21**: Over-parameterized Models
- **Feb 28**: Talk by Prof. Anant Raj
- **March**: Diffusion Models + Industry Talks (Microsoft Research)

Note

The continuation of the series depends on audience interest and participation. If you want to interact with leading researchers, make sure to attend!

RL Workshop

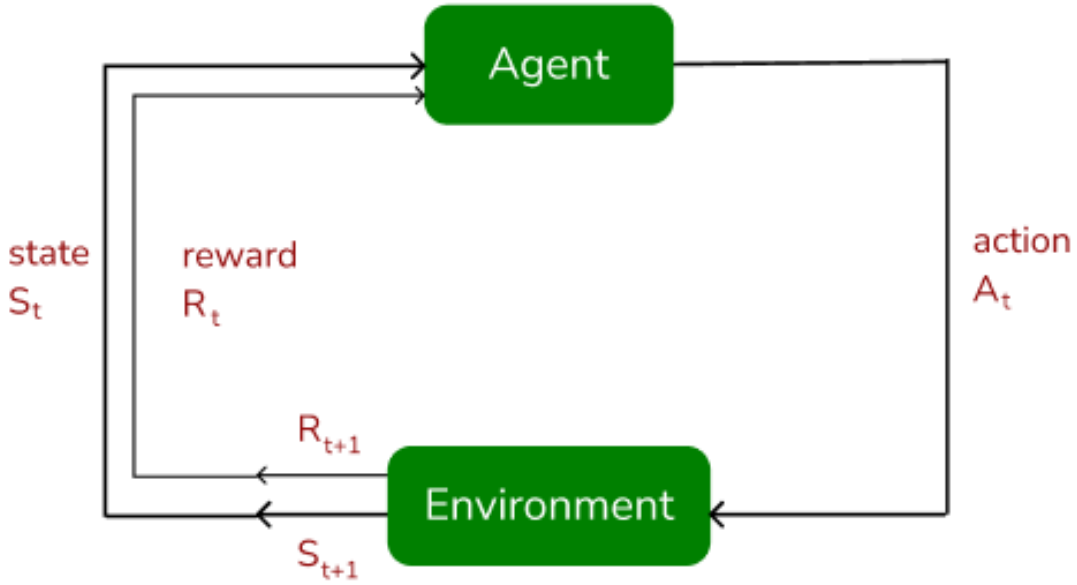
Timings are as per the Indian Standard Time

Day	8:45am to 9:00am	9:00am to 10:00am	10:00am to 10:30am	10.30am to 11.30am	11:30 am to 12:30 pm	12:30 pm to 2:00 pm	2:00 pm to 3:00 pm	3:00 pm to 4:00 pm	4:00 pm to 4:30 pm	4:30 pm to 5:30 pm
Thursday 15th January 2026		**Tutorial: RL Theory Gugan Thoppe (IISc)**	Coffee Break	**Tutorial: RL Theory Gugan Thoppe (IISc)**	**Tutorial: RL Theory Gugan Thoppe (IISc)**	Lunch	**Tutorial: RL Theory Gugan Thoppe (IISc)**	**Tutorial: RL Simulations Gugan Thoppe (IISc)**	Coffee Break	**Tutorial: RL Simulations Gugan Thoppe (IISc)**
Friday 16th January 2026	**Inauguration by CSA Chair**	**Furong Huang** *(UMD)* (Online)	Coffee Break	**Vivek Borkar** *(IIT Bombay)*	**M Vidyasagar** *(IIT Hyderabad)*	Lunch	**Dheeraj Nagraj** *(Google)*	**Raunak B** *(IIT Delhi)*	Coffee Break	**Arunselvan Ramaswamy** *(Karistad University)*
Saturday 17th January 2026		**Serdar Yuksel** *(Queens)* (Online)	Coffee Break	**Sridharan Devarajan** *(CMU)*	**Prashanth L A** *(IIT Madras)*	Lunch	**Paras Chopra** *(LossFunk)*	**Karthik Sundaresan** *(Walmart)*	Coffee Break	**Pranay Sharma** *(IIT Bombay)*
Sunday 18th January 2026		**Aviral Kumar** *(IISc)* (Online)	Coffee Break	**Shubhada Agarwal** *(IISc)*	**Swanand Khare** *(IIT Madras)*	Lunch	**Networking and Casual Discussions**	**Networking and Casual Discussions**	Coffee Break	**Networking and Casual Discussions**

Outline

Setup: Agent–Environment Interaction

Reinforcement Learning (RL) studies *sequential decision making under uncertainty*.



Key idea

An **agent** repeatedly interacts with an **environment**, observing states, choosing actions, and receiving rewards.

What is Sequential and What is Uncertain?

Sequential decision making

At time t:

$$S_t \xrightarrow{A_t} S_{t+1}$$

Actions affect not only immediate reward but also *future states*.

Uncertainty

- The reward function $R(s, a)$ is unknown
- The transition kernel $P(\cdot \mid s, a)$ is unknown

Objective

Maximize expected discounted return:

$$\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_t\right], \qquad \gamma \in (0, 1)$$

Markov Decision Processes (MDPs)

Definition

A Markov Decision Process (MDP) is a tuple

$$(\mathcal{S}, \mathcal{A}, R, P, \gamma)$$

where:

- $\mathcal{S}$: state space
- $\mathcal{A}$: action space
- $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}_+$: reward function
- $P(\cdot \mid s, a)$: transition kernel (probability simplex over $\mathcal{S}$)
- $\gamma \in (0, 1)$: discount factor

Markov property

$$\mathbb{P}(S_{t+1} \mid S_0, A_0, \ldots, S_t, A_t) = P(S_{t+1} \mid S_t, A_t)$$

Policies and Value Functions

Policy

A (stationary) policy π is a mapping

$$\pi : \mathcal{S} \to \Delta(\mathcal{A})$$

where $\pi(a \mid s)$ is the probability of selecting action a in state s.

State-value function

For a policy π,

$$V^\pi(s) := \mathbb{E}^\pi \left[\sum_{t=0}^{\infty} \gamma^t R(S_t, A_t) \,\middle|\, S_0 = s \right]$$

Action-value function

$$Q^\pi(s,a) := \mathbb{E}^\pi \left[\sum_{t=0}^{\infty} \gamma^t R(S_t, A_t) \,\middle|\, S_0 = s, A_0 = a \right]$$

$$V^{\pi}(s) = \mathbb{E}^{\pi}\left[\sum_{t=0}^{\infty} \gamma^t R_t \mid S_0 = s\right]$$

$$= \underbrace{\mathbb{E}^{\pi}\left[R(S_0, A_0) \mid S_0 = s\right]}_{\| \; \sum_{a \in A} \pi(a|s)\, r(s,a)} + \gamma \underbrace{\mathbb{E}^{\pi}\left[R(S_1, A_1) \mid S_0 = s\right]}_{\downarrow \; S_1, A_1}$$

$$S_1 \sim P(\cdot \mid s_0, A_0), \quad A_0 \sim \pi(\cdot \mid s_0)$$

$$\sum_a \pi(a|s) \sum_{s'} P(s' \mid s, a) \sum_{a'} \pi(a'|s')\, r(s', a')$$

$$\mathbb{E}\left[R(S_1, A_1)\right] = \sum_{s' \in S} R(s', A_1) \cdot \underbrace{P(s')}_{\| \; \sum_{a \in A} P(s' \mid s, a)\pi(a|s)}$$

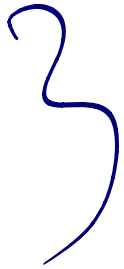
Optimality and Bellman Optimality Equation

Optimal value function

$$V^*(s) := \sup_{\pi} V^{\pi}(s)$$

Bellman optimality equation

$$\boxed{V^*(s) = \max_{a \in \mathcal{A}} \left[R(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V^*(s') \right]}$$

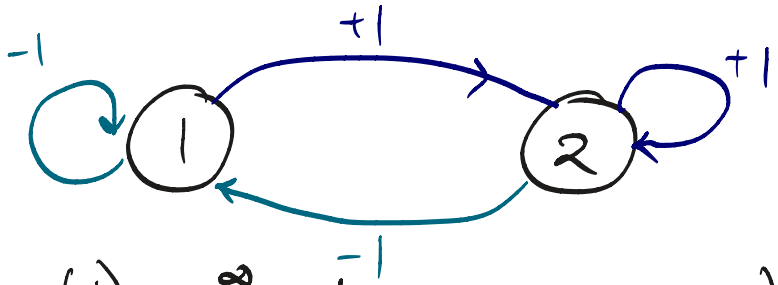


Bellman optimality operator

$$(TV)(s) := \max_{a} \left[R(s,a) + \gamma \mathbb{E} V(s') \right]$$

MDP Example

$\gamma = 0.9$



2 actions: right left

Policy π_1: "always right" $\longrightarrow$ $V_{\pi_1}(1) = \sum_{t=0}^{\infty} \gamma^t \cdot 1 = \frac{1}{1-\gamma} = 10$ $\quad V_{\pi_1}(2) = \frac{1}{1-\gamma} = 10$

π_2: "always left" $\longrightarrow$ $V_{\pi_2}(1) = \sum_{t=0}^{\infty} \gamma^t \cdot (-1) = -10$ $\quad V_{\pi_2}(2) = -10$

Optimal policy: $V^*(2) = \max\{ +1 + \gamma V^*(2), -1 + \gamma V^*(1) \}$

$V^*(1) = \max\{ \underbrace{+1 + \gamma V^*(2)}_{\text{right}}, \underbrace{-1 + \gamma V^*(1)}_{\text{left}} \}$

If right is optimal in both states then $V^*(2) = V^*(1) = 10$

CHECK: at state 1, $1 + \gamma V^*(2) = 10 > -1 + \gamma V^*(1) = 8$

at state 2, $1 + \gamma V^*(2) = 10 > -1 + \gamma V^*(1) = 8$

$\Rightarrow$ right is optimal in both states

Derivation of the Bellman Optimality Equation (I)

Definition of the optimal value function

For any state s,

$$V^*(s) := \sup_\pi \mathbb{E}^\pi \left[\sum_{t=0}^{\infty} \gamma^t R(S_t, A_t) \,\middle|\, S_0 = s \right]$$

Unroll the return (first step)

Separate the immediate reward:

$$\begin{aligned} V^*(s) &= \sup_\pi \mathbb{E}^\pi \left[R(S_0, A_0) + \sum_{t=1}^{\infty} \gamma^t R(S_t, A_t) \,\middle|\, S_0 = s \right] \\ &= \sup_\pi \mathbb{E}^\pi \left[R(s, A_0) + \gamma \sum_{t=0}^{\infty} \gamma^t R(S_{t+1}, A_{t+1}) \,\middle|\, S_0 = s \right] \end{aligned}$$

Derivation of the Bellman Optimality Equation (II)

Apply the tower property

Condition on the next state S_1:

$$\mathbb{E}^{\pi}\left[\sum_{t=0}^{\infty}\gamma^t R(S_{t+1}, A_{t+1}) \,\middle|\, S_0 = s\right] = \mathbb{E}^{\pi}\left[\mathbb{E}^{\pi}\left[\sum_{t=0}^{\infty}\gamma^t R(S_{t+1}, A_{t+1}) \,\middle|\, S_1\right] \,\middle|\, S_0 = s\right]$$

Optimal substructure (principle of optimality)

For any s',

$$\sup_{\pi}\ \mathbb{E}^{\pi}\left[\sum_{t=0}^{\infty}\gamma^t R(S_{t+1}, A_{t+1}) \,\middle|\, S_1 = s'\right] = \sup_{\pi}\ \mathbb{E}^{\pi}\left[\sum_{t=0}^{\infty}\gamma^t R(S_t, A_t) \,\middle|\, S_0 = s'\right] = V^*(s')$$

Derivation of the Bellman Optimality Equation (III)

Substitute the tail by V^*

Combining the previous steps gives

$$V^*(s) = \sup_{\pi} \mathbb{E}^{\pi}[R(s, A_0) + \gamma V^*(S_1) \mid S_0 = s]$$

Reduce $\sup_{\pi}$ to $\max_{a}$ (first decision)

At time 0 the policy selects a distribution over actions; the supremum is attained by a deterministic choice:

$$V^*(s) = \max_{a \in \mathcal{A}} \left[R(s, a) + \gamma \, \mathbb{E}_{s' \sim P(\cdot|s,a)} V^*(s') \right]$$

Bellman optimality equation and operator form

$$\boxed{V^*(s) = (TV^*)(s) := \max_{a} \left[R(s, a) + \gamma \mathbb{E} V^*(s') \right]}$$

Bellman Evaluation Operator

Bellman expectation equation

For a fixed policy π:

$$V^{\pi}(s) = \sum_{a} \pi(a|s) \Big[R(s,a) + \gamma \mathbb{E} V^{\pi}(s') \Big]$$

Bellman evaluation operator

$$(T^{\pi} V)(s) := \sum_{a} \pi(a|s) \Big[R(s,a) + \gamma \mathbb{E} V(s') \Big]$$

Fixed point

$$V^{\pi} = T^{\pi} V^{\pi}$$

Derivation of the Bellman Expectation Equation (I)

Definition of the value function

For a fixed (stationary) policy π and state s,

$$V^{\pi}(s) := \mathbb{E}^{\pi}\left[\sum_{t=0}^{\infty} \gamma^t R(S_t, A_t) \,\middle|\, S_0 = s\right]$$

Unroll the return

Separate the immediate reward:

$$\begin{aligned} V^{\pi}(s) &= \mathbb{E}^{\pi}\left[R(S_0, A_0) + \sum_{t=1}^{\infty} \gamma^t R(S_t, A_t) \,\middle|\, S_0 = s\right] \\ &= \mathbb{E}^{\pi}\left[R(s, A_0) + \gamma \sum_{t=0}^{\infty} \gamma^t R(S_{t+1}, A_{t+1}) \,\middle|\, S_0 = s\right] \end{aligned}$$

Derivation of the Bellman Expectation Equation (II)

$\mathbb{E}X = \mathbb{E}[\mathbb{E}[X|Y]]$

Apply the tower property

Condition on the next state S_1:

$$\mathbb{E}^{\pi}\left[\sum_{t=0}^{\infty}\gamma^t R(S_{t+1}, A_{t+1}) \,\middle|\, S_0 = s\right] = \mathbb{E}^{\pi}\left[\mathbb{E}^{\pi}\left[\sum_{t=0}^{\infty}\gamma^t R(S_{t+1}, A_{t+1}) \,\middle|\, S_1\right] \,\middle|\, S_0 = s\right]$$

Identify the tail return

By the Markov property and stationarity of π, for any s',

$$\mathbb{E}^{\pi}\left[\sum_{t=0}^{\infty}\gamma^t R(S_{t+1}, A_{t+1}) \,\middle|\, S_1 = s'\right] = V^{\pi}(s')$$

Derivation of the Bellman Expectation Equation (III)

Substitute back

$$V^\pi(s) = \mathbb{E}^\pi[R(s, A_0) + \gamma V^\pi(S_1) \mid S_0 = s]$$

Expand the expectation

Using $A_0 \sim \pi(\cdot \mid s)$ and $S_1 \sim P(\cdot \mid s, A_0)$,

$$V^\pi(s) = \sum_a \pi(a \mid s) \left[R(s, a) + \gamma \sum_{s'} P(s' \mid s, a) V^\pi(s') \right]$$

Bellman evaluation fixed-point equation

$$\boxed{V^\pi = T^\pi V^\pi} \qquad (T^\pi V)(s) := \sum_a \pi(a \mid s) \Big[R(s, a) + \gamma \mathbb{E}_{s'} V(s') \Big]$$

What Do We Want to Solve?

Core RL questions

1. **Policy evaluation**: Given π, compute V^π
2. **Policy improvement**: Improve π using V^π

Goal

Construct a sequence of policies converging to an optimal policy π^*.

Solving Bellman Equations Directly

Observation

Bellman equations are systems of linear equations:

$$V^\pi = R^\pi + \gamma P^\pi V^\pi$$

Why not solve directly?

- State space can be huge
- Matrix inversion is expensive

A Detour: Fixed Points

Operator form

$$V^\pi = T^\pi V^\pi, \qquad V^* = TV^*$$

Key question

Can we compute fixed points using repeated application of operators?

Banach Fixed Point Theorem

Contraction mapping

An operator T is a contraction on $(\mathcal{X}, \|\cdot\|)$ if

$$\|Tx - Ty\| \leq c\|x - y\|, \quad c < 1$$

Banach Fixed Point Theorem

If T is a contraction on a complete normed space, then:

- T has a unique fixed point
- Iteration $x_{k+1} = Tx_k$ converges to it

Choosing a Norm

Common norms

$$\|V\|_\infty = \max_s |V(s)|, \quad \|V\|_2, \quad \|V\|_1$$

Equivalence of norms

In finite-dimensional spaces, all norms are equivalent:

$$c_1\|x\|_a \le \|x\|_b \le c_2\|x\|_a$$

Why ℓ_∞?

It interacts cleanly with max and expectations.

Contraction of Bellman Operators

Theorem

Both T^π and T are γ-contractions in $\|\cdot\|_\infty$:

$$\|TV - TW\|_\infty \leq \gamma \|V - W\|_\infty$$

$$\|T^\pi V - T^\pi W\|_\infty \leq \gamma \|V - W\|_\infty$$

Contraction of the Bellman Evaluation Operator T^π

Recall

$$(T^\pi V)(s) := \sum_a \pi(a \mid s)\Big[R(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V(s')\Big]$$

Proof

For any state s,

$$|(T^\pi V)(s) - (T^\pi W)(s)| = \Big|\sum_a \pi(a|s)\,\gamma \mathbb{E}\big[V(s') - W(s')\big]\Big|$$

$$\leq \gamma \sum_a \pi(a|s)\, \mathbb{E}\big[|V(s') - W(s')|\big]$$

$$\leq \gamma \|V - W\|_\infty$$

Conclusion Taking supremum over s,

$$\boxed{\|T^\pi V - T^\pi W\|_\infty \leq \gamma \|V - W\|_\infty}$$

Contraction of the Bellman Optimality Operator T (I)

Recall

$$(TV)(s) := \max_{a \in \mathcal{A}} \left[R(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V(s') \right]$$

Goal

Show that T is a γ-contraction in the ℓ_∞ norm:

$$\| TV - TW \|_\infty \leq \gamma \| V - W \|_\infty$$

Key inequality

For any real vectors (x_a) and (y_a),

$$|\max_a x_a - \max_a y_a| \leq \max_a |x_a - y_a|$$

Contraction of the Bellman Optimality Operator T (II)

Proof

Fix a state s. Define, for each action a,

$$x_a := R(s,a) + \gamma \mathbb{E}_{s'} V(s'), \qquad y_a := R(s,a) + \gamma \mathbb{E}_{s'} W(s')$$

Then

$$\begin{aligned} |(TV)(s) - (TW)(s)| &= |\max_a x_a - \max_a y_a| \\ &\leq \max_a |x_a - y_a| \\ &= \max_a \gamma \left|\mathbb{E}[V(s') - W(s')]\right| \\ &\leq \gamma \|V - W\|_\infty \end{aligned}$$

Taking the supremum over all states s,

$$\boxed{\|TV - TW\|_\infty \leq \gamma \|V - W\|_\infty}$$

Value Iteration Algorithm

Algorithm

Initialize V_0 arbitrarily. For $k \geq 0$:

$$V_{k+1} = TV_k$$

Policy extraction

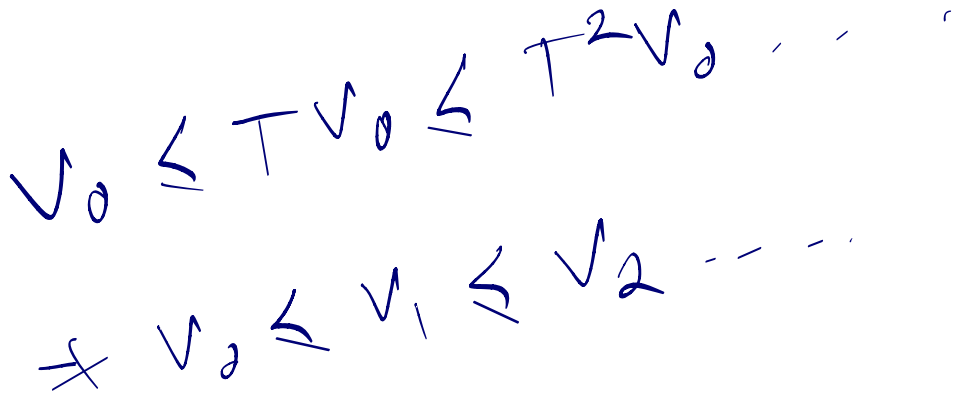
$$\pi_k(s) \in \arg\max_a \left[R(s,a) + \gamma \mathbb{E} V_k(s') \right]$$

When should the algorithm stop btw?

Value Iteration: Guarantees

Main guarantees

- Monotonic value improvement
- Convergence to V^*
- Geometric rate γ^k



Monotonic Value Improvement

Assumption

Rewards are nonnegative and $V_0 \equiv 0$.

Claim

$$V_0 \leq V_1 \leq V_2 \leq \cdots \leq V^*$$

Reason

Bellman operator is monotone:

$$V \leq W \Rightarrow TV \leq TW$$

Monotonicity of the Bellman Operator (I)

Claim (Monotonicity)

If $V(s) \leq W(s)$ for all $s \in \mathcal{S}$, then

$$(TV)(s) \leq (TW)(s) \quad \text{for all } s.$$

Proof

Fix a state s and any action a. Since $V \leq W$ pointwise,

$$\mathbb{E}_{s'}[V(s')] \leq \mathbb{E}_{s'}[W(s')].$$

Therefore,

$$R(s,a) + \gamma \mathbb{E} V(s') \;\leq\; R(s,a) + \gamma \mathbb{E} W(s').$$

Monotonicity of the Bellman Operator (II)

Preservation under maximization

Since the above inequality holds for every action a,

$$\max_a \Big[R(s,a) + \gamma \mathbb{E} V(s')\Big] \;\leq\; \max_a \Big[R(s,a) + \gamma \mathbb{E} W(s')\Big].$$

That is,

$$(TV)(s) \leq (TW)(s).$$

Conclusion

$$\boxed{V \leq W \;\Rightarrow\; TV \leq TW}$$

Hence, the Bellman optimality operator is *order-preserving* (monotone).

Convergence to the Optimal Value

$$\|V_{k+1} - V^*\|_\infty$$
$$= \|TV_k - TV^*\|_\infty$$
$$\leq \gamma \|V_k - V^*\|_\infty$$

Contraction argument

$$\|V_k - V^*\|_\infty \leq \gamma^k \|V_0 - V^*\|_\infty$$

Conclusion

Value Iteration converges geometrically fast to V^*.

Outline

Remarks and Outlook

$$T^\pi V = \sum \pi(a|s) \left[r(s,a) + \gamma \sum p(s'|s,a) V(s') \right]$$

- Recall the MDP setup and the agent–environment interaction framework.
- In practice, the reward function R and transition kernel P are *unknown*.
- Consequently, the Bellman operators T^π and T are not available explicitly.
- As a result, classical value iteration cannot be implemented directly.
- This motivates the study of **model-free** algorithms, which learn from data without explicit knowledge of R and P.
- In the next session, we will study such algorithms and introduce **stochastic approximation** as a key tool for their analysis.

Thank You

Thank you for your attention!

Questions, comments, or clarifications?