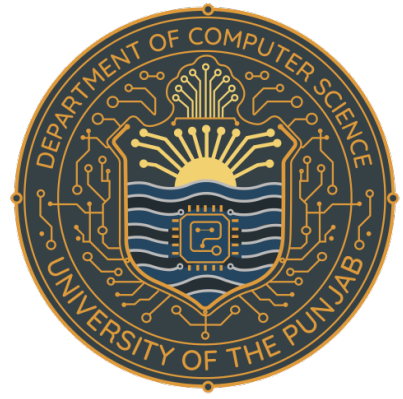
CS-566 Deep Reinforcement Learning

MDP Terminology - I



Nazar Khan
Department of Computer Science
University of the Punjab

MDP Terminology

- RL makes use of some ideas from probability and statistics
  - Random variable
  - Probability distribution
  - Expectation
- It also uses some common terms and notations
  - Traces
  - Return
  - Value functions

Random Variables

- Day of the week is a *variable*. It *varies* between 7 values *deterministically* – Monday after Sunday and so on.
- Day of birth of students in a class is also a variable. It also varies between 7 values but *randomly*. It is a *random variable*.
- Random variables can be *discrete* or *continuous*.
- Discrete random variables
  - Number when you roll a die.
  - Students present in a class.
  - Votes in an election.
- Continuous random variables
  - Amount of liquid in your next cold drink.
  - Height of the person sitting next to you in a bus.
  - Time taken to complete an exam.
  - Time spent waiting in a queue.

Probability

- Since a random variable *must* take some value, the probability of *all* its values equals 1.
- Probability of any individual value will be less than 1.
- For *discrete* random variables, how the probability is *distributed* among the different values is given by the *probability distribution* table.
- It can be distributed evenly

X	1	2	3	4	5	6
$P(X)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

  which is also called a *uniform distribution*.
- It can also be distributed unevenly

X	1	2	3	4	5	6
$P(X)$	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{1}{6}$

Probability

- For *continuous* random variables, the distribution of probability is represented by areas under a non-negative function with total area 1.
- Such functions are called *probability density functions*.

$$\text{probability density} \neq \text{probability}$$

- Probability is given by computing area under a probability density function $p(x)$

$$P(a \leq X \leq b) = \int_a^b p(x)dx$$

- When total area is 1, any sub-area will be between 0 and 1. So rules of probability will be satisfied.
- Most well-known density function for continuous random variables is the *Gaussian*.

Expectation

- The **expected value** of a random variable X is the long-run average outcome.
- Definition:

$$\mathbb{E}[X] = \sum_x p(x) \cdot x \quad \text{(discrete)}$$

$$\mathbb{E}[X] = \int x\, p(x)\, dx \quad \text{(continuous)}$$

- Think of expectation as a **weighted average**, where outcomes are weighted by their probability.

Expectation Example (Unfair Die)

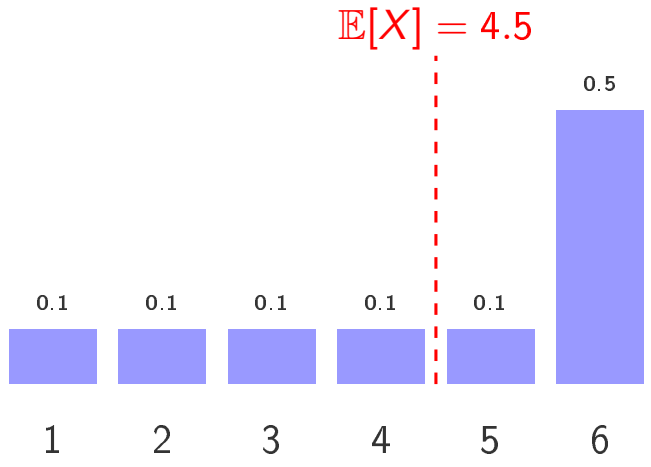
- Suppose a die is biased:

X	1	2	3	4	5	6
$P(X)$	0.1	0.1	0.1	0.1	0.1	0.5
$X \cdot P(X)$	0.1	0.2	0.3	0.4	0.5	3.0

- Expected value:

$$\mathbb{E}[X] = 0.1(1 + 2 + 3 + 4 + 5) + 0.5 \cdot 6 = 0.1 \cdot 15 + 3 = 4.5$$

- Unlike the fair die ($\mathbb{E}[X] = 3.5$), our unfair die's expectation is skewed towards 6.



Trace τ (Trajectory)

- As we start interacting with the MDP, at each timestep t:
  - Observe state s_t
  - Take an action a_t
  - Observe next state $s_{t+1} \sim T_{a_t}(s_t)$
    - $x \sim P$ should be read as *x is a random variable with probabilities distributed according to P*
    - So s_{t+1} is a random variable with probabilities distributed according to $T_{a_t}(s_t)$
  - Receive reward $r_t = R_{a_t}(s_t, s_{t+1})$
- Repeating this process leads to a sequence (trace/trajectory/episode).

$$\tau_t^n = \{s_t, a_t, r_t, s_{t+1}, \ldots, a_{t+n}, r_{t+n}, s_{t+n+1}\}$$

Finite vs Infinite Trace

- $n =$ length of the trace.
- Often assume $n = \infty$, i.e., run until termination.
- In that case, write:

$$\tau_t = \tau_t^{\infty}$$

- Traces are fundamental in RL:
  - A single full rollout of decisions
  - Also called trajectory, episode, or sequence

Trace Visualization

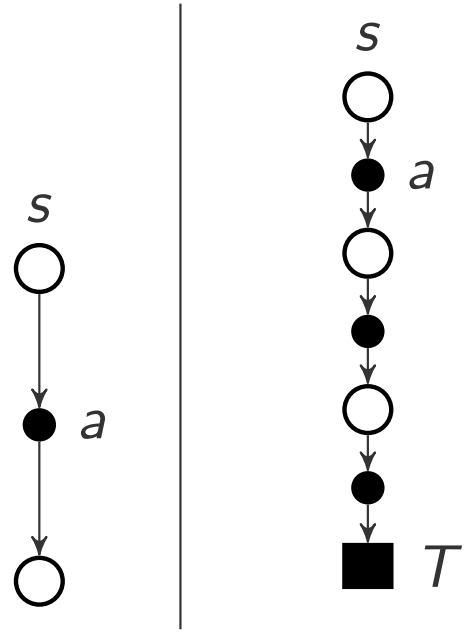


Figure: Single Transition Step vs. Full 3-Step Trace/Episode/Trajectory

Example of a Trace

Example: A short trace with three actions:

$$\begin{aligned}\tau_0^2 = \{&s_0 = 1, a_0 = \text{up}, r_0 = -1,\\ &s_1 = 2, a_1 = \text{up}, r_1 = -1,\\ &s_2 = 3, a_2 = \text{left}, r_2 = 20,\\ &s_3 = 5\}\end{aligned}$$

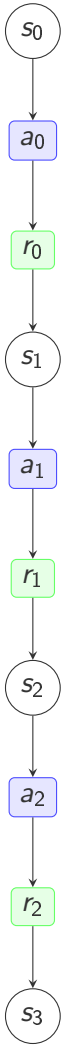
Trace τ: Step-by-Step Expansion

Recall: A trace (trajectory/episode) unfolds step by step:

- At each timestep t, observe s_t,
- Take action a_t,
- Observe reward r_t and next state s_{t+1}.

Trace Example:

$$\tau_0^2 = \{s_0, a_0, r_0, s_1, a_1, r_1, s_2, a_2, r_2, s_3\}$$



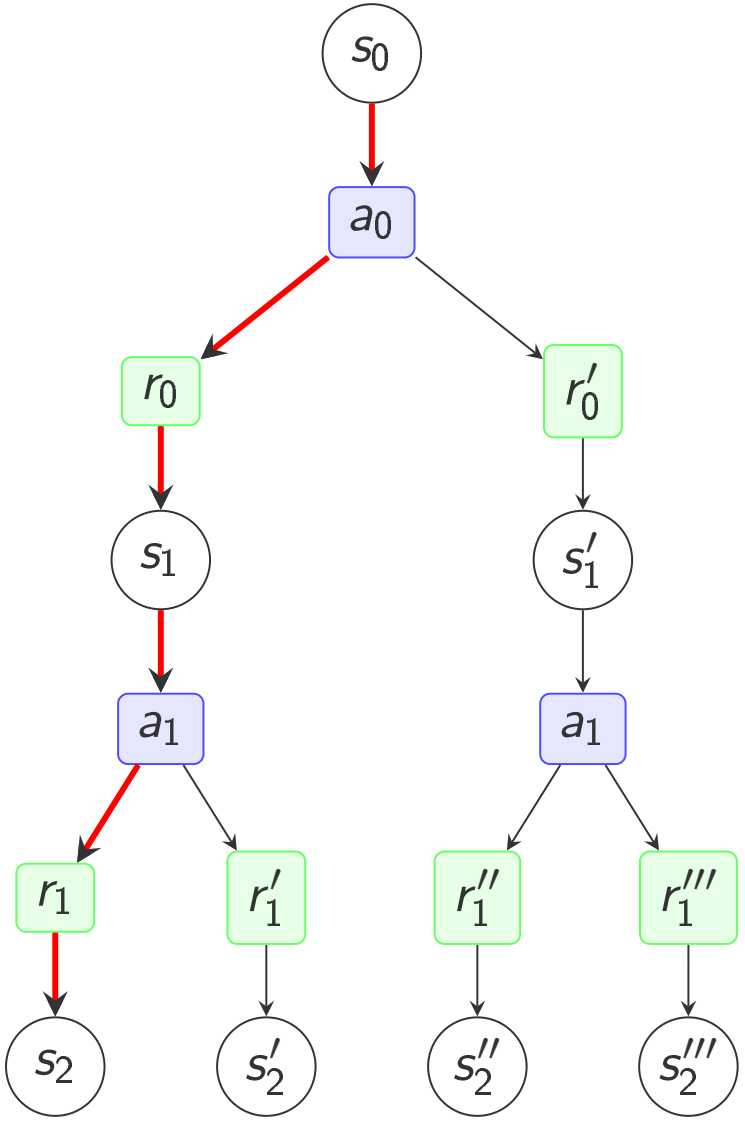
Trace τ: Branching Expansion

Trace as a sequence:

$$\tau_0 = \{s_0, a_0, r_0, s_1, a_1, r_1, s_2, \dots\}$$

But in practice:

- From each state s_t and action a_t,
- The environment may branch into different s_{t+1}.
- So a trace is one *path* through this tree.



Stochastic Spaces and Distribution over Traces

- A process or sequence with randomness is called a *stochastic process*.
- In RL, both the policy π and transitions T can be stochastic.
- So proceeding from the start state will not always produce the same trace.

> The trace will be a stochastic process.
> Each trace will have some probability of occuring.

- So we get a **probability distribution over traces**:

$$p(\tau_0) = \text{probability of complete trace from start state } s_0$$

- Probability of a trace = product of probabilities of its transitions:

$$p(\tau_0) = p_0(s_0) \cdot \prod_{t=0}^{\infty} \pi(a_t|s_t) \cdot T_{a_t}(s_t, s_{t+1})$$

Distribution over Traces: Breaking Down the Equation

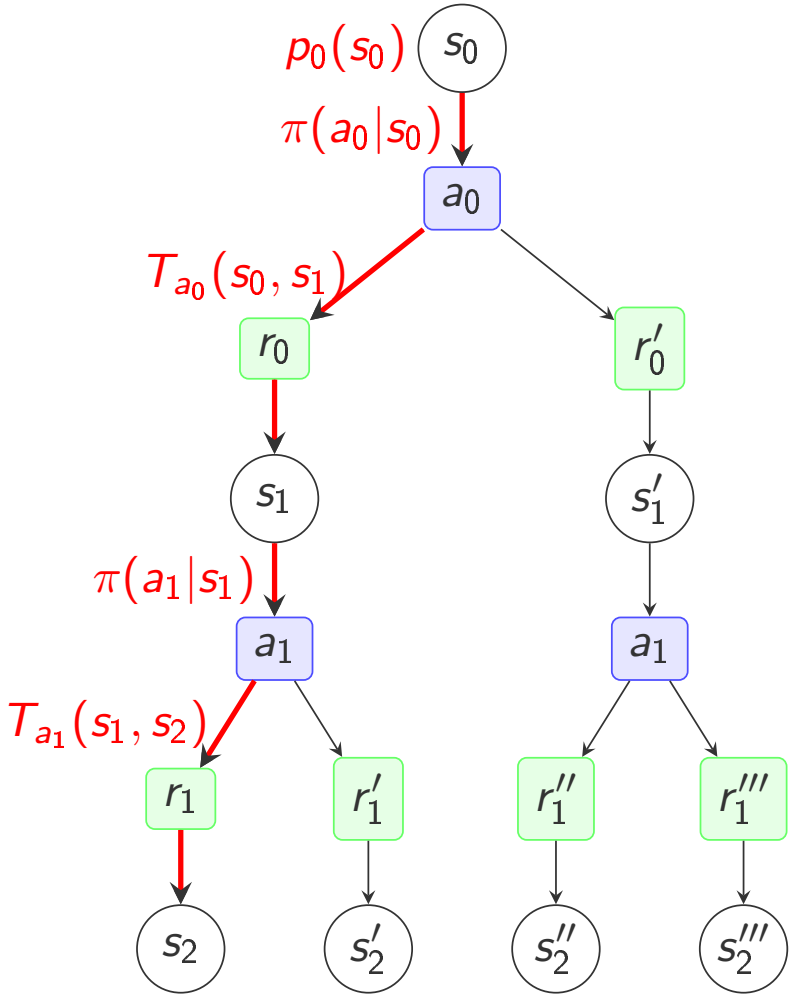
- The trace $\tau_0 = \{s_0, a_0, r_0, s_1, a_1, r_1, \dots\}$ is one possible path.
- Its probability depends on:
  - $p_0(s_0)$ = probability of starting in state s_0,
  - $\pi(a_t|s_t)$ = probability of choosing action a_t in state s_t,
  - $T_{a_t}(s_t, s_{t+1})$ = probability of transitioning to s_{t+1} after action a_t.
- Multiply these step probabilities together for the full trace:

$$p(\tau_0) = p_0(s_0)\, \pi(a_0|s_0)\, T_{a_0}(s_0, s_1)\, \pi(a_1|s_1)\, T_{a_1}(s_1, s_2)\, \cdots$$

- Compact notation:

$$p(\tau_0) = p_0(s_0) \cdot \prod_{t=0}^{\infty} \pi(a_t|s_t)\, T_{a_t}(s_t, s_{t+1})$$

Distribution over Traces: Breaking Down the Equation



A trace is just one path in the MDP tree, and its probability is the product of the branching probabilities along that path.

Traces in RL

- **Policy-based RL**: depends heavily on full traces.
- **Value-based RL**: often uses single transition steps.
- Both approaches build on the idea of traces.