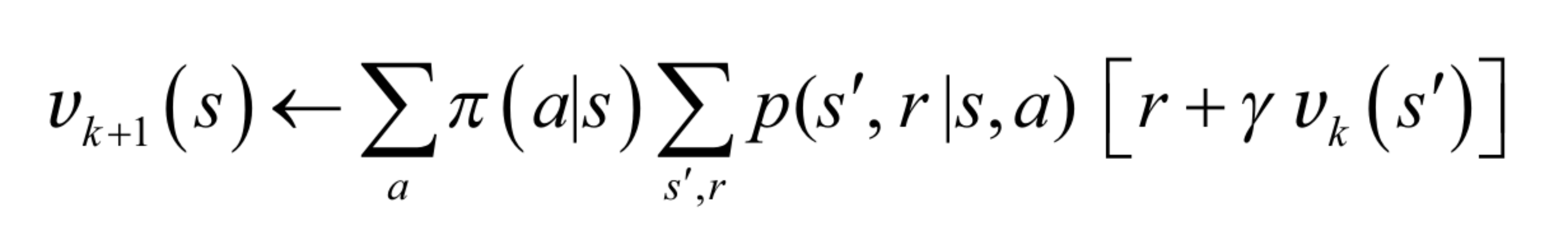
Reinforcement Learning Training 2025

Model-Free Approach

Motivation

Recall in policy iteration



$$v_{k+1}(s) \leftarrow \sum_{a} \pi(a|s) \sum_{s',r} p(s', r|s, a) \left[r + \gamma\, v_k(s') \right]$$

- To make this work, we need to know the model dynamics or $p(s', r|s, a)$.
- However, we do now know p.
- Instead, we will resort to *sampling*.
    - Collecting experience by following some policy in the real world or running the agent through a policy in simulation.

Model-Free Learning

- Monte Carlo (MC) methods
- Temporal difference (TD) methods

Monte Carlo

- We use the law of large numbers (LLN) from statistics.
    - Average of samples is a good estimate for the actual unknown quantity.
    - This estimate becomes better and better as the number of trials of the experiment (samples) increases.

Monte Carlo

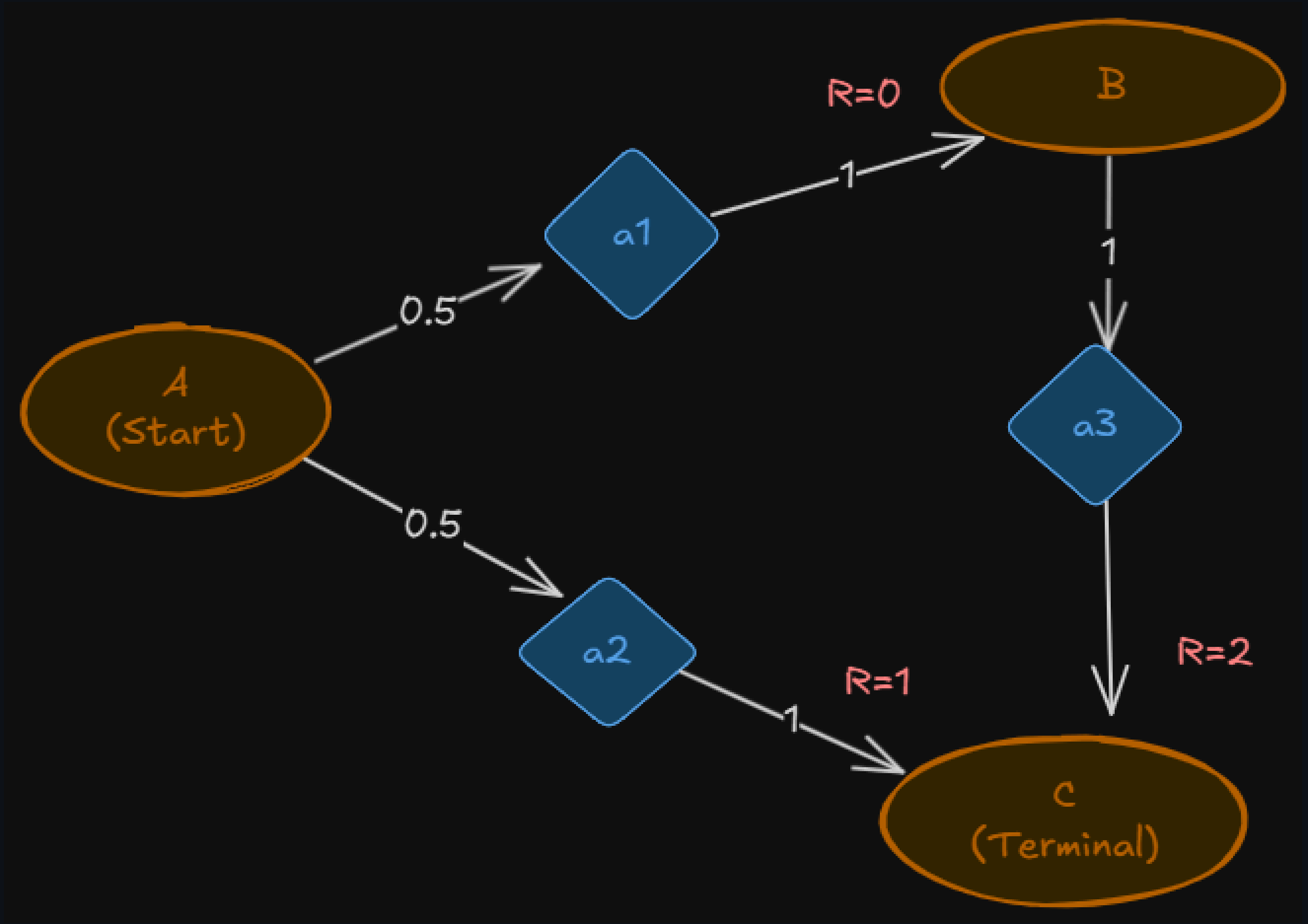
- Re call that We want to calculate

$$v_\pi(s) = \mathbf{E}_\pi[G_t | S_t = s]$$

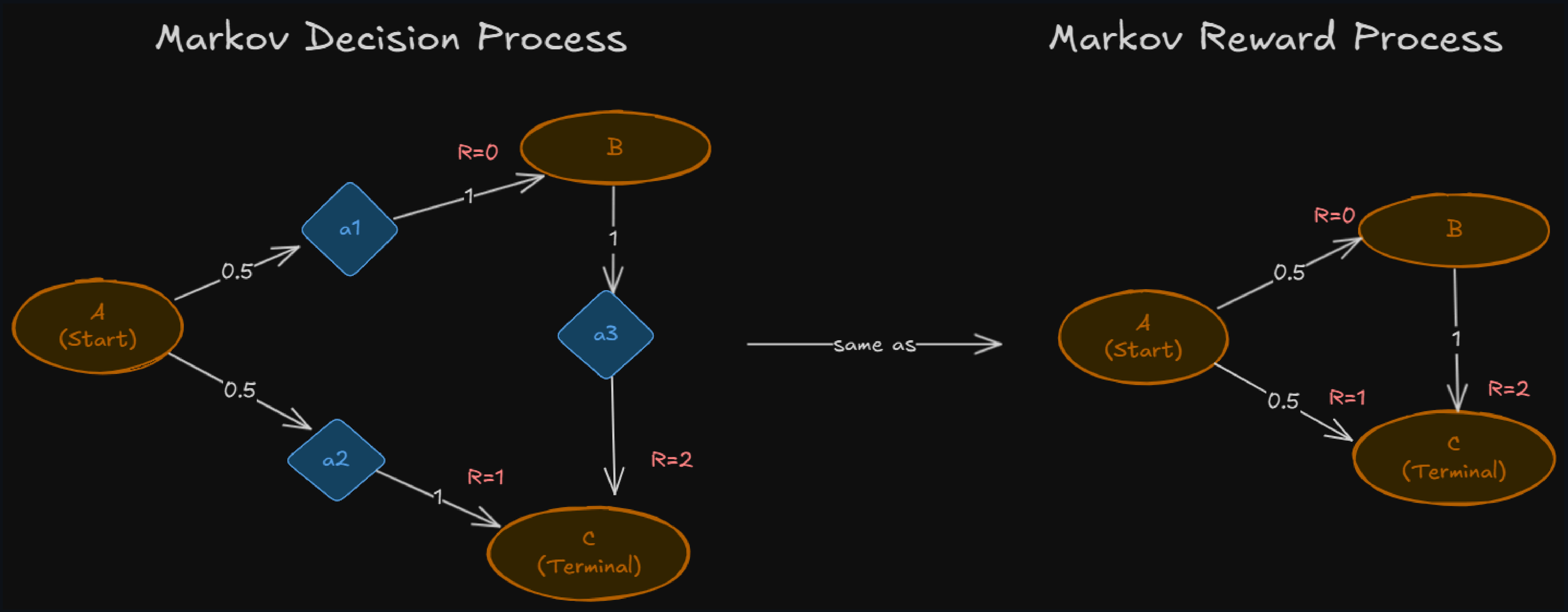
- We let the agent start from this state $S_t = s$, follow the policy π to take actions, and keep doing so until termination.
  - We call one round of actions an **episode**.
- We record the total sum of rewards for each episode.
- We average the rewards to get an estimate of $v_\pi(s)$ for the policy π.

> MC methods replaces expected returns with the average of sample returns.

Worked Example



Note


Markov Decision Process
Markov Reward Process
A (Start)
0.5
0.5
a1
a2
1
1
R=0
R=1
B
1
a3
R=2
C (Terminal)
same as
A (Start)
0.5
R=0
B
1
0.5
R=1
R=2
C (Terminal)

Solution

Sampling

- We simulate 4 episodes.

Episode	**Path**	**Reward from A**
1	A → C	G_1 = 1
2	A → B → C	G_2 = 0 + 2 = 2
3	A → B → C	G_3 = 0 + 2 = 2
4	A → C	G_4 = 1

Results

Monte Carlo estimates the value function $v(A)$ as the average return observed after visiting A.

$$v(A) = \frac{G_1 + G_2 + G_3 + G_4}{4} = \frac{1+2+2+1}{4} = \frac{6}{4} = 1.5$$

Episode	Path	Actions at A	Reward from Action at A
1	A → C	a_2	$G_1 = 1$
2	A → B → C	a_1	$G_2 = 0 + 2$
3	A → B → C	a_1	$G_3 = 0 + 2$
4	A → C	a_2	$G_4 = 1$