UVA | SCHOOL of DATA SCIENCE

Summary of Solution Concepts

Reinforcement Learning

School of Data Science

University of Virginia

Last updated: September 29, 2025

Agenda

> Backup Diagrams for Understanding MDP Solution Methods

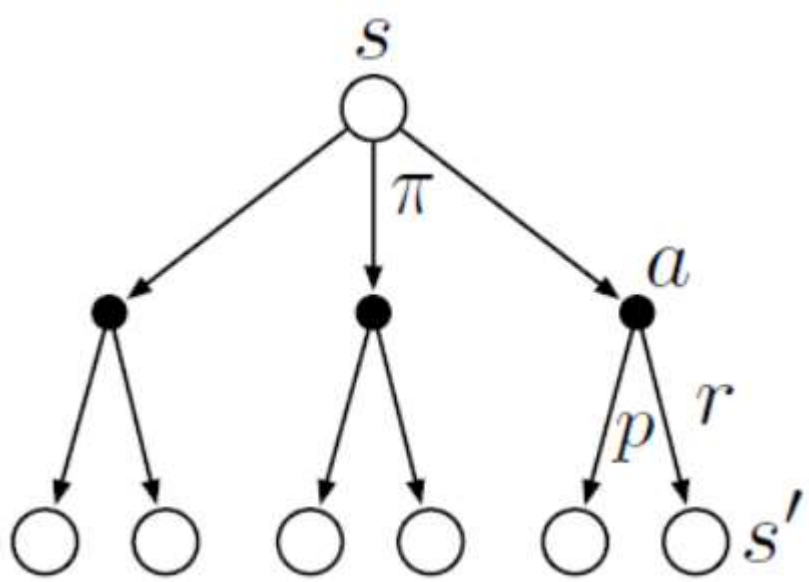
> On-policy vs off-policy learning

Overarching Goal for Solutions to MDP

The overarching goal is to estimate value functions

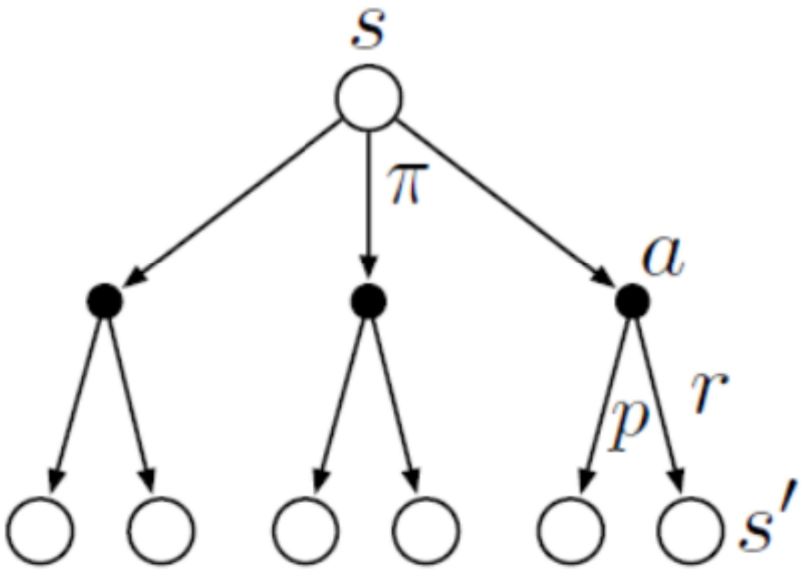
We study three solution methods:
1) Dynamic Programming
2) Monte Carlo Simulation
3) Temporal Difference

Backup diagrams (example on right)
can help understand how they work.
- Black circles are actions
- Hollow circles are states
- Time flows top down



Backup Diagrams

Dynamic Programming

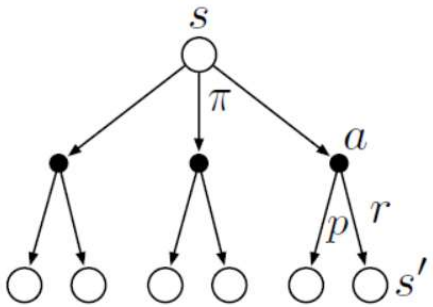


Monte Carlo

Temporal Difference: TD(0)

Backup Diagrams with update equations

Dynamic Programming



transition distribution

$$V(s) \leftarrow \sum_a \pi(a|s) \sum_{s',r} P(s',r|s,a)\,[r + \gamma V(s')]$$

- Update equation computes an expectation
- Need to know transition, reward distributions

Monte Carlo

$$V(s_t) \leftarrow V(s_t) + \alpha\Big[G_t - V(s_t)\Big]$$

- Update equation uses Gain
- Need to wait for episode to end

Temporal Difference: TD(0)

$$V(s_t) \leftarrow V(s_t) + \alpha\Big[R_{t+1} + \gamma V(s_{t+1}) - V(s_t)\Big]$$

- Update equation uses single transition

On-Policy vs. Off-Policy Learning

We study algorithms that use on-policy learning and off-policy learning

It is important to realize that two things are happening:
1. An agent follows a policy to guide behavior
2. Based on historical actions, value estimates are calculated

We have a choice:

Option 1: We can use the same policy for both (on-policy learning)

Option 2: We can use different policies (off-policy learning)

On-Policy vs. Off-Policy Learning

On-policy algorithms learn and improve on the same policy that is being used to collect data in the environment:

1. The agent uses its policy to select actions
2. Data is generated in the form of transitions (s, a, r, s')
3. This data is used to update the value estimates and the policy

---

Off-policy algorithms are more flexible and efficient as they separate behavior from estimates.

1. The agent uses behavior policy *b* to select actions
2. Data is generated in the form of transitions (s, a, r, s')
3. This data is used to update the value estimates and the target policy

Example of Off-Policy Learning

In healthcare, it is not safe or possible to run experimental policies

We may have collected historical data. This can be regarded as behavior policy *b*.

Algorithm can be trained on this data for use (offline reinforcement learning)

These behaviors may have been conducted by current and past doctors

Going forward, doctors may make different choices.