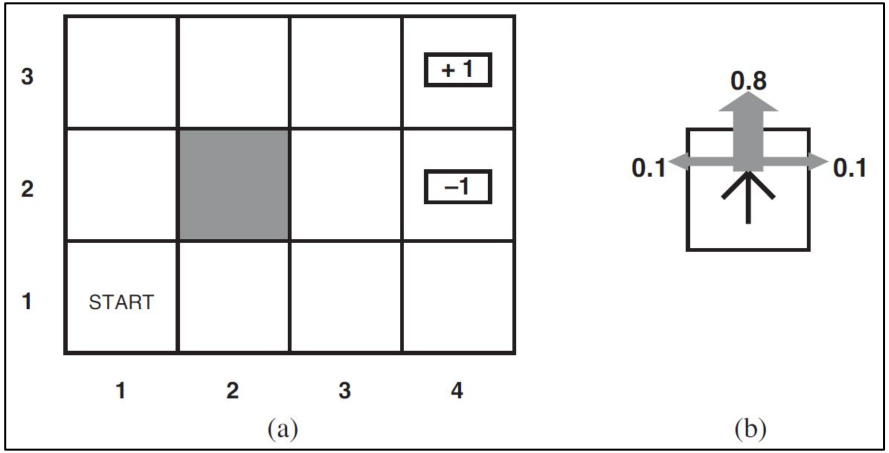
# AIL 7022: Reinforcement Learning

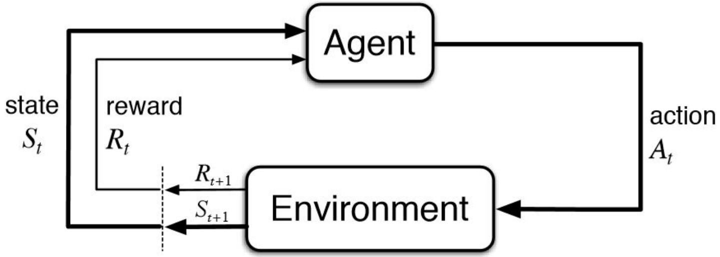
## Lecture 4: MDPs & Value Functions

Instructor: Raunak Bhattacharyya

YARDI SCHOOL OF ARTIFICIAL INTELLIGENCE
INDIAN INSTITUTE OF TECHNOLOGY DELHI

# Recap



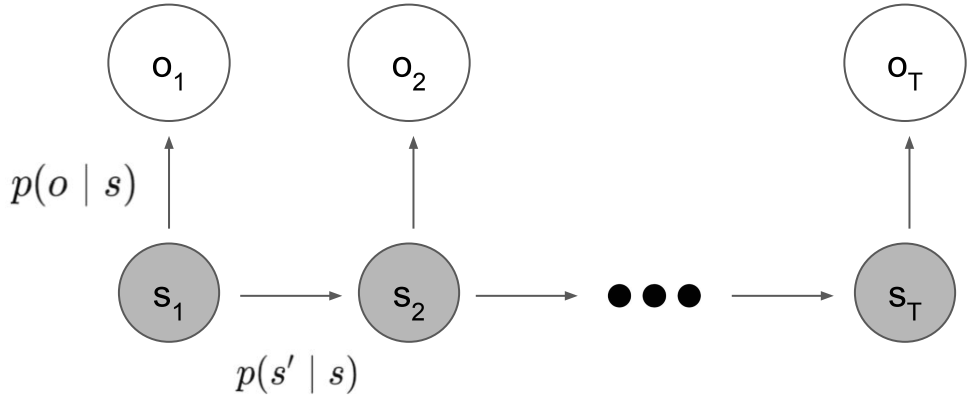



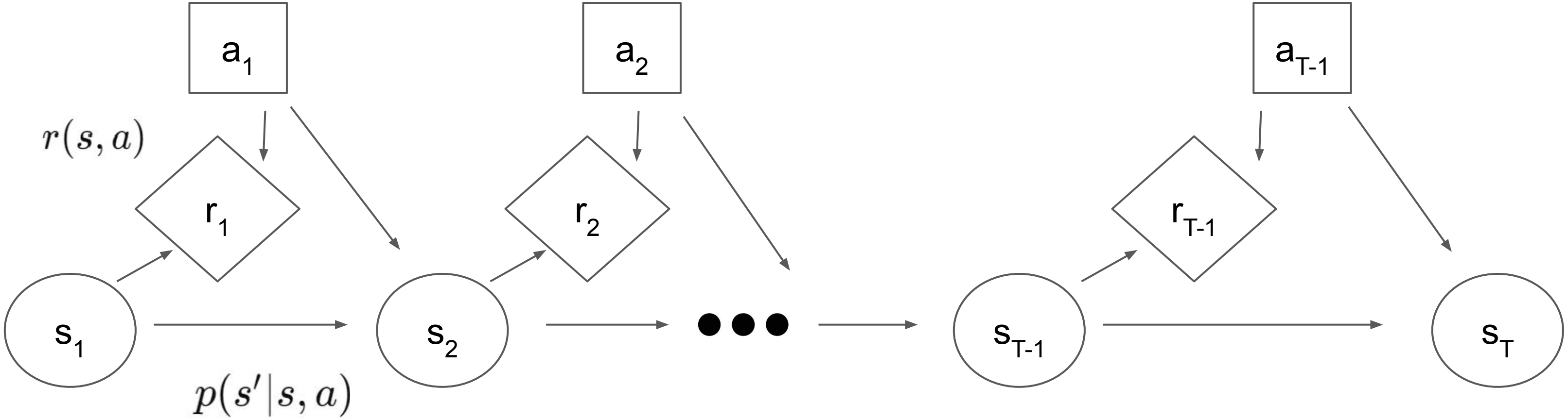

Source: [Sutton & Barto]



Russell & Norvig

# MDP: State Evolution



# Queuing Problem


Source: Dreamstime

- Customers line up in a queue. There is only one line. Line is empty initially
- We can serve one customer at a time. There are two modes of service: fast and slow
- Each timestep, a new customer arrives with probability p. The horizon length is T
- Waiting cost: gamma * queue length

# Queuing Problem: Formulation

$\mathcal{S} = \{0, 1, 2, \ldots\}$ : Length of the queue $x_t$ $\quad x_0 = 0$

$\mathcal{U} = \{\text{Fast (F)}, \text{Slow (S)}\}$ $\quad$ Completion probs: $q(F) > q(S)$

$c(x_t, u_t) = \gamma x_t + d(u_t)$ $\quad$ Service costs: $d(F) > d(S)$

If $x = 0$:

$$p(x' = 1 \mid x = 0, u = F/S) = p$$

$$p(x' = 0 \mid x = 0, u = F/S) = 1 - p$$

If $x > 0$:

$$p(x' = x + 1 \mid x, u) = p \cdot (1 - q(u))$$

$$p(x' = x \mid x, u) = (1 - p) \cdot (1 - q(u)) + p \cdot q(u$$

$$p(x' = x - 1 \mid x, u) = q(u) \cdot (1 - p)$$

# Plan

- Queuing Problem
- Value functions
- Policy Evaluation

# Policy

We need a policy, a rule for action selection that works in any state

Search problem: path (sequence of actions)

MDP:

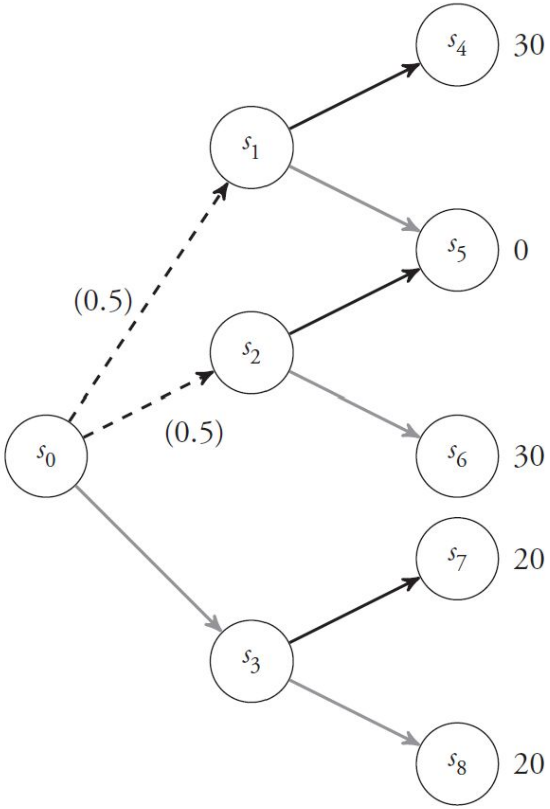
**Definition: policy**

A **policy** $\pi$ is a mapping from each state $s \in \text{States}$ to an action $a \in \text{Actions}(s)$.



Adapted from Percy Liang

# Open Loop Plan

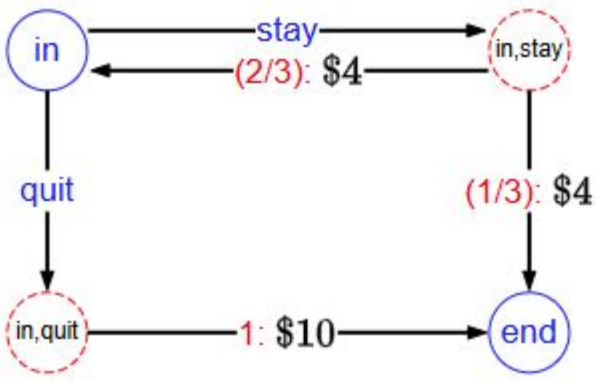



$$U(\text{up}, \text{up}) = 0.5 \times 30 + 0.5 \times 0 = 15$$
$$U(\text{up}, \text{down}) = 0.5 \times 0 + 0.5 \times 30 = 15$$
$$U(\text{down}, \text{up}) = 20$$
$$U(\text{down}, \text{down}) = 20$$

**Open loop plan chooses down action from $s_0$**

From Decision Making under Uncertainty, Mykel Kochenderfer

# Dice Game



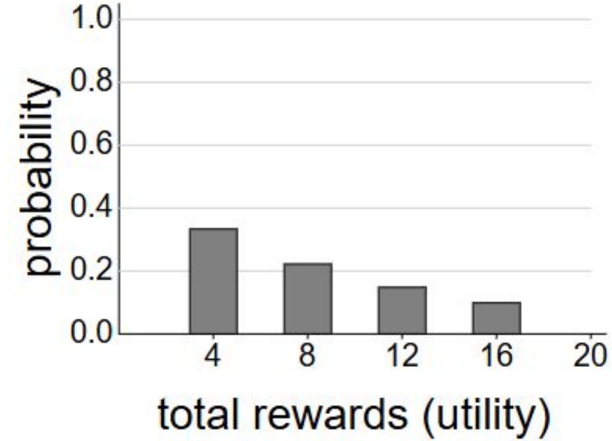

| $s$ | $a$ | $s'$ | $T(s,a,s')$ |
|---|---|---|---|
| in | quit | end | $1$ |
| in | stay | in | $2/3$ |
| in | stay | end | $1/3$ |

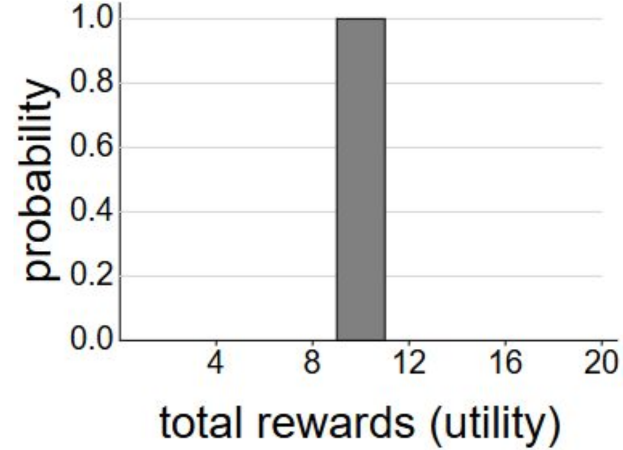
| Path | Utility |
|---|---|
| [in; stay, 4, end] | 4 |
| [in; stay, 4, in; stay, 4, in; stay, 4, end] | 12 |
| [in; stay, 4, in; stay, 4, end] | 8 |
| [in; stay, 4, in; stay, 4, in; stay, 4, in; stay, 4, end] | 16 |



From Percy Liang

# Dice Game



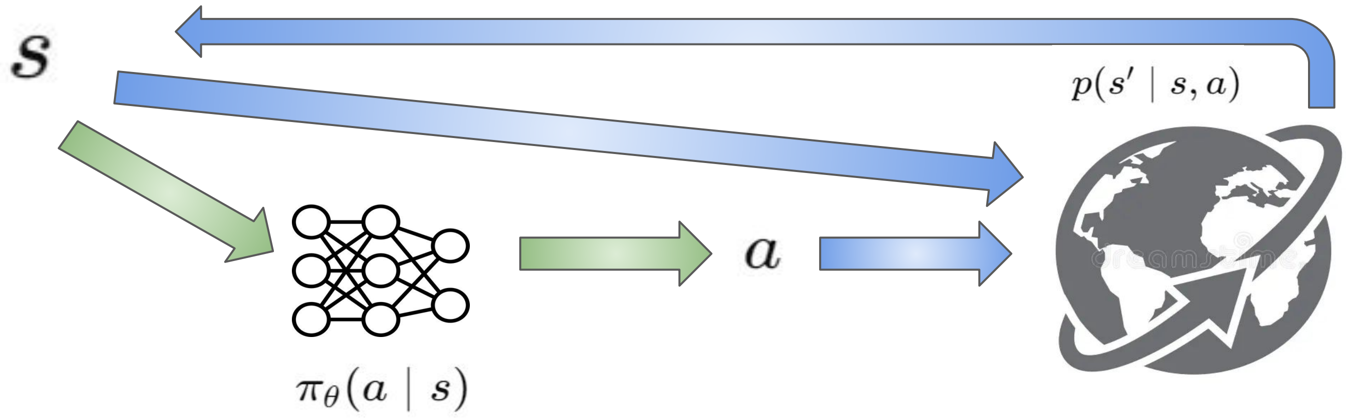






From Percy Liang

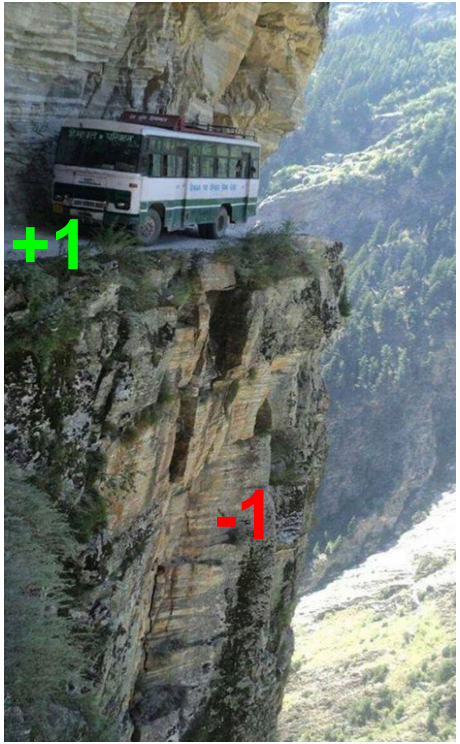
# Value Functions

# Objective




$$\theta^* = \arg\max_\theta \mathbb{E}_{\tau \sim p_\theta(\tau)} \left[ \sum_{t=1}^{T} r(s_t, a_t) \right]$$

# Expectations




Source: Pinterest

$$\theta^* = \arg\max_{\theta} \mathbb{E}_{\tau \sim p_\theta(\tau)} \left[ \sum_{t=1}^{T} r(s_t, a_t) \right]$$

**RL is really about optimising expectations**

$r(s_t, a_t)$ : not smooth

Suppose policy $\pi_\theta(a_t = \text{fall}) = \theta$

$\mathbb{E}_{p_\theta(\tau)} \left[ \sum_{t=1}^{T} r(s_t, a_t) \right]$ : smooth in $\theta$

**Why RL can use smooth optimisation techniques even though rewards are highly discontinuous**

Adapted from Sergey Levine

# Expectations in the Objective

$$J(\theta) = \mathbb{E}_{\tau \sim p_\theta(\tau)} \left[ \sum_{t=1}^{T} r(s_t, a_t) \right]$$

Expanding it out for clarity

$$J(\theta) = \mathbb{E}_{(s_1, a_1, s_2, a_2, \ldots, s_T, a_T) \sim p_\theta(s_1, a_1, \ldots, s_T, a_T)} \left[ \sum_{t=1}^{T} r(s_t, a_t) \right]$$

# Factorising the Trajectory Distribution

$$p_\theta(s_1, a_1, \ldots, s_T, a_T) = p(s_1) \prod_{t=1}^{T} \pi_\theta(a_t \mid s_t)\, p(s_{t+1} \mid s_t, a_t)$$

$$\begin{aligned}
p(s_1, a_1, s_2, a_2, s_3) &= p(s_1) \cdot p(a_1, s_2, a_2, s_3 \mid s_1) \\
&= p(s_1) \cdot p(a_1 \mid s_1) \cdot p(s_2, a_2, s_3 \mid s_1, a_1) \\
&= p(s_1) \cdot p(a_1 \mid s_1) \cdot p(s_2 \mid s_1, a_1) \cdot p(a_2, s_3 \mid s_1, a_1, s_2) \\
&= p(s_1) \cdot p(a_1 \mid s_1) \cdot p(s_2 \mid s_1, a_1) \cdot p(a_2 \mid s_2) \cdot p(s_3 \mid s_2, a_2)
\end{aligned}$$

**Can we use this factorization in the objective function?**

# Conditional Expectations

$$J(\theta) = \mathbb{E}_{(s1,a1,s2,a2,\ldots,sT,aT)\sim p_\theta(s_1,a_1,\ldots,s_T,a_T)}\left[\sum_{t=1}^{T} r(s_t, a_t)\right]$$

$$J(\theta) = \mathbb{E}_{s_1\sim p(s_1)}\Bigg[\mathbb{E}_{a_1\sim\pi_\theta(a_1|s_1)}\bigg[r(s_1, a_1)+\mathbb{E}_{s_2\sim p(s_2|s_1,a_1)}\Big[\mathbb{E}_{a_2\sim\pi_\theta(a_2|s_2)}$$

$$\Big[r(s_2, a_2)+\cdots \mid s_2\Big] \mid s_1, a_1\Big] \mid s_1\bigg]\Bigg]$$

# Introducing the Q-function

$$J(\theta) = E_{s_1 \sim p(s_1)}\left[E_{a_1 \sim \pi_\theta(a_1|s_1)}\left[r(s_1, a_1) + E_{s_2 \sim p(s_2|s_1,a_1)}\left[E_{a_2 \sim \pi_\theta(a_2|s_2)}\left[r(s_2, a_2) + \cdots \mid s_2\right] \mid s_1, a_1\right] \mid s_1\right]\right]$$

Suppose we knew this part

# Definition: Q-function

$$Q^\pi(s_t, a_t) = \mathbb{E}\left[\sum_{t'=t}^{T} r(s_{t'}, a_{t'}) \middle| s_t, a_t\right]$$

**Expected cumulative reward obtained by taking $a_t$ in $s_t$ and then following the policy**

**What is the expectation over?**

**What is the objective in terms of Q?**

$$J(\theta) = \mathbb{E}_{s_1 \sim p(s_1)}\left[\mathbb{E}_{a_1 \sim \pi_\theta(a_1|s_1)}\left[Q(s_1, a_1) \,\middle|\, s_1\right]\right]$$