Markov Decision Processes

Last Time

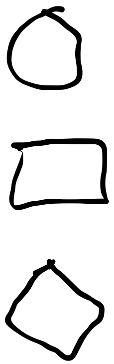
- What does "Markov" mean in "Markov Process"?

Guiding Questions

- What is a **Markov decision process**?
- What is a **policy**?
- How do we **evaluate** policies?

Decision Networks and MDPs

Decision Network

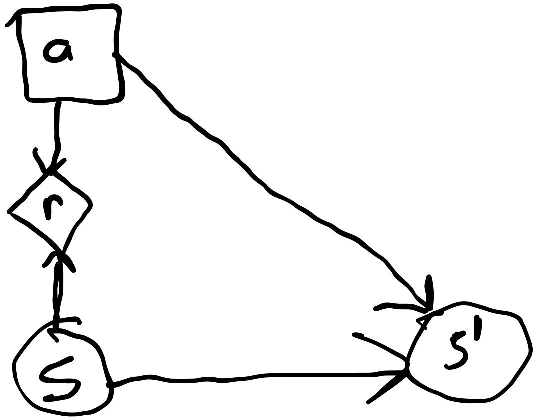


Chance node

Decision node

Utility node

MDP Dynamic Decision Network



MDP Optimization problem

$$\text{maximize} \quad \mathrm{E}\left[\sum_{t=1}^{\infty} r_t\right]$$

Not well formulated!
Infinite

Finite MDP Objectives

1. Finite time

$$\mathbf{E}\left[\sum_{t=0}^{T} r_t\right]$$

2. Average reward

$$\lim_{n\to\infty} \mathbf{E}\left[\frac{1}{n}\sum_{t=0}^{n} r_t\right]$$

3. Discounting

$$\boxed{\mathbf{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right]}$$

discount $\gamma \in [0, 1)$

typically 0.9, 0.95, 0.99

if $\underline{r} \leq r_t \leq \bar{r}$

then

$$\frac{\underline{r}}{1-\gamma} \leq \sum_{t=0}^{\infty} \gamma^t r_t \leq \frac{\bar{r}}{1-\gamma}$$

4. Terminal States

Infinite time, but a terminal state (no reward, no leaving) is always reached with probability 1.

MDP "Tuple Definition"

(S, A, T, R, γ) (and b in some contexts)

- S (state space) - set of all possible states

$\{1, 2, 3\}$ $(x, y) \in \mathbb{R}^2$ $\{0, 1\} \times \mathbb{R}^4$

$\{\text{healthy}, \text{pre-cancer}, \text{cancer}\}$ $(s, i, r) \in \mathbb{N}^3$

- A (action space) - set of all possible actions

$\{1, 2, 3\}$ $\mathbb{R}^2$ $\{0, 1\} \times \mathbb{R}^2$

$\{\text{test}, \text{wait}, \text{treat}\}$

- T (transition distribution) - explicit or implicit ("generative") model of how the state changes

$T(s' \mid s, a)$

- R (reward function) - maps each state and action to a reward

$R(s, a)$ or $R(s, a, s')$

- γ: discount factor

$s', r = G(s, a)$

- b: initial state distribution

MDP Example

Imagine it's a cold day and you're ready to go to work. You have to decide whether to bike or drive.

- If you drive, you will have to pay $15 for parking; biking is free.
- On 1% of cold days, the ground is covered in ice and you will crash if you bike, but you can't discover this until you start riding. After your crash, you limp home with pain equivalent to losing $100.

Policies and Simulation

- A *policy*, denoted with π, as in $a_t = \pi(s_t)$ is a function mapping every state to an action.
- When a policy is combined with a Markov decision process, it becomes a Markov stochastic process with
$$P(s' \mid s) = T(s' \mid s, \pi(s))$$

Break

- Suggest a policy that you think is optimal for the icy day problem

Utility

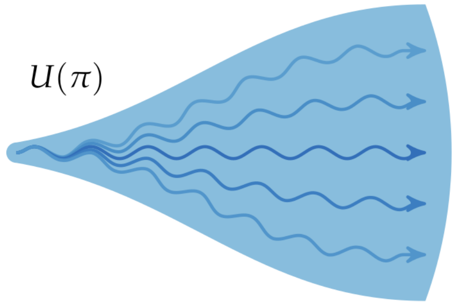
Slide not on Exam

Policy Evaluation

Naive Policy Evaluation not on Exam

Monte Carlo Policy Evaluation

- Running a large number of simulations and averaging the accumulated reward is called *Monte Carlo Evaluation*



Let $\tau = (s_0, a_0, r_0, s_1, \ldots, s_T)$ be a *trajectory* of the MDP

$$U(\pi) \approx \frac{1}{m} \sum_{i=1}^{m} R(\tau^{(i)})$$

also an R.V.

$$U(\pi) \approx \bar{u}_m = \frac{1}{m} \sum_{i=1}^{m} \hat{u}^{(i)}$$

where $\hat{u}^{(i)}$ is generated by a rollout simulation

How can we quantify the accuracy of $\bar{u}_m$?

C.L.T. $\quad \dfrac{\bar{u}_m - U(\pi)}{\sigma_m / \sqrt{m}} \xrightarrow{d} \mathcal{N}(0, 1)$

CLT not on exam

Standard Error of the Mean $\quad \text{s.e.m.} = \dfrac{\text{std}(\hat{u})}{\sqrt{m}}$

Value Function-Based Policy Evaluation

Guiding Questions

- What is a **Markov decision process**?
- What is a **policy**?
- How do we **evaluate** policies?