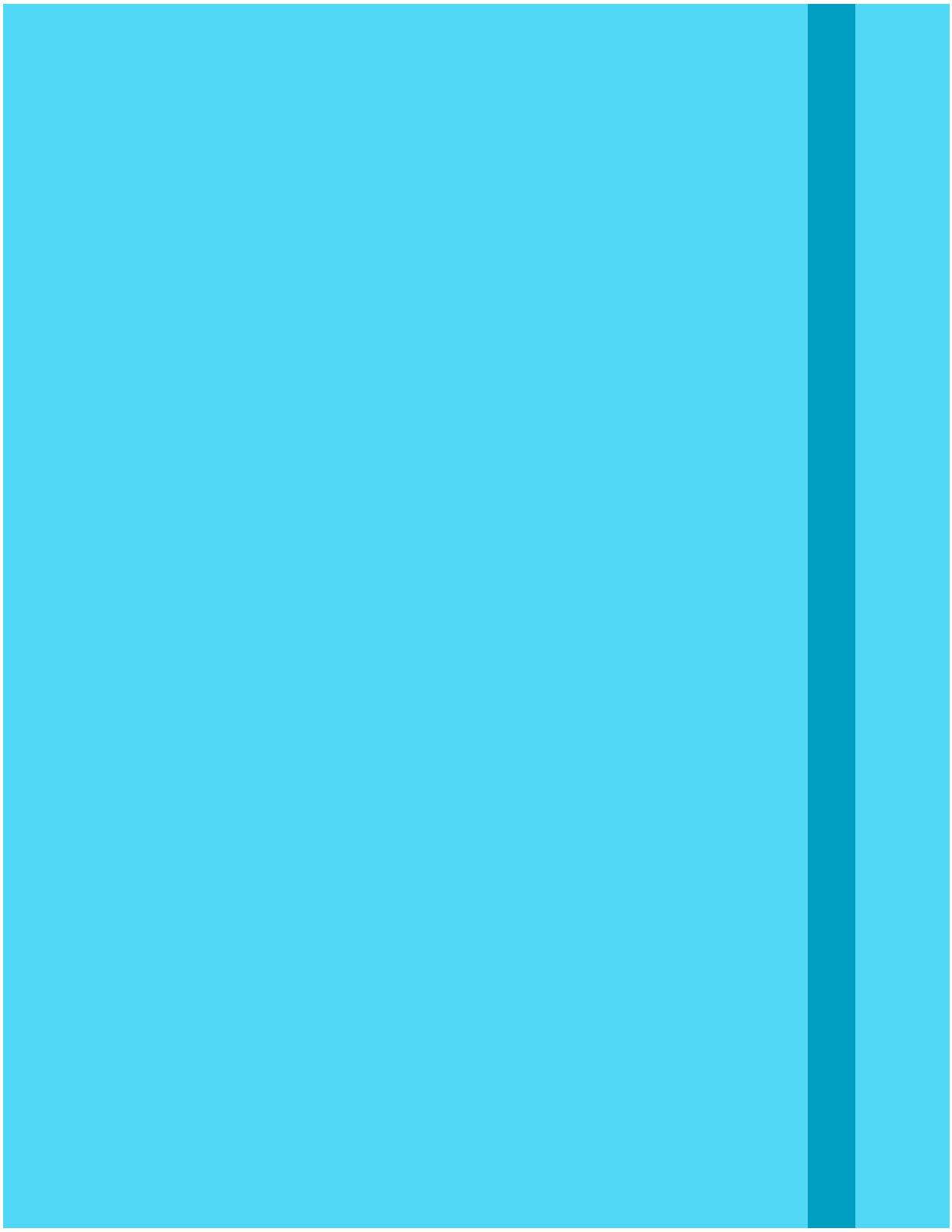
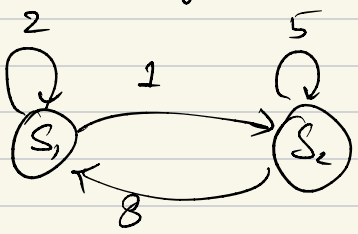


25/1/2024

Reinforcement Learning.

$\lambda = 0.8$



$D_1 = \max\{2 + 0.8 D_1 , \quad 1 + 0.8 D_2\}$

$D_2 = \max(5 + 0.8 D_2 , \quad 8 + 0.8 D_1 \}.$

Suppose $D_1 = 21$ & $D_2 = 25$.

$$21 = \max\{2 + 0.8 \times 21, \quad 1 + 0.8 \times 25\}$$
$$= \max\{2 + 16.8, \quad 1 + 20.0\}$$
$$= \max\{18.8, \quad 21\} = 21 //.$$

$$D_2 = \max\{5 + 0.8 \times 25, \quad 8 + 0.8 \times 21\}$$
$$= \max\{5 + 20, \quad 8 + 16.8\}$$
$$= \max\{25, \quad 24.8\}$$
$$= 25 //.$$

We get optimal discounted state for all states.

D is function : state → [D] → Number ($\mathbb{R}$).

Graph G. you also have OPT(G). → (The two D_1 & D_2 equations).

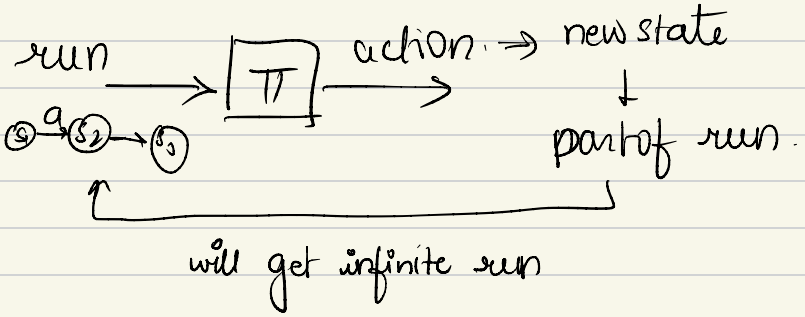
if you find D_1 & D_2 then. $D(s)$ = optimal discounted reward from s.

The policy will be memoryless and deterministic.

for every graph, there always will be a unique solution of OPT(G).

Solution in poly time.

policy:



General policies: probability distribution over available actions.
(but not necessary for deterministic / MDPs).

Discounted Rewards: $r : s_0, a_1, s_1, a_2, s_2 \ldots$

$$D(r) = \sum_{i=0}^{\infty} \lambda^i \cdot r(s_i, a_{i+1}).$$

$$D(s_1, \overset{b}{1}, s_2, \overset{a}{5}, s_2, \overset{b}{8}, s_1, \overset{a}{2}, s_1, \overset{b}{1} \ldots)$$

$$= 1 + \lambda 5 + 8\lambda^2 + 2\lambda^3 + \lambda^4 + \ldots$$

* Can be infinite policies.

for a state s, value or optimal discounted reward is:

$$D_*(s) = \sup_{u \in \Sigma} \left[D(s, u) \right].$$

A policy is optimum if $u_* \in \Sigma$ $\quad D(s, u) = D_*(s)$ for every $s \in S$.

Proof for Correctness of Bellman Optimality Equations:

If $D \underset{\text{Satisfies}}{\models} \text{Opt}(G)$ then $D(s) = D_*(s)$ & u^D is an optimal (memoryless) policy.

Proof: to show: $D(s) = \sup_{u \in \Sigma} D(s, u)$

① $D(s) \geq \text{Disc}(s, u)$

② $D(s) = \text{Disc}(s, u^D)$.

$\because$ you are using the optimal policy u^D.

$D(s_0) =$ ~~

$D(s_1) =$ ~~

$\therefore D(s) = \text{Disc}(s, u^D)$.

$s_0 \xrightarrow{a_1} s_1 \xrightarrow{a_2} s_2 \ldots$ $\quad \text{Run}(s, u)$

$D \models \text{OPT}(G)$.

$D(s_0) \geq r(s_0, a_1) + \lambda D(s_1)$

$D(s_1) \geq r(s_1, a_2) + \lambda D(s_2)$

$\vdots$

$$D(s_0) \geq \lim_{n \to \infty} \left(\sum_{i=0}^{n} \lambda^i \cdot r(s_i, a_{i+1}) + \lambda^{n+1} D(s_{n+1}) \right)$$

$\geq \text{Disc}(s, u)$.

How to get optimal discounted solns (D)?

Banach's Theorem

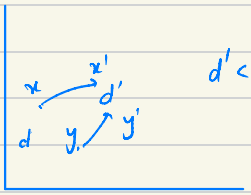
for every graph G, exists a unique soln of DPT(G).

$$D(s) = \max_{a \in A} \{ x(s,a) + \lambda \cdot D(T(s,a)) \} \quad \forall s \in S.$$

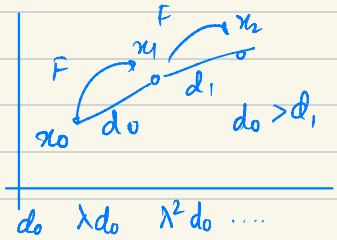
Start from (0,0) → (2,8) → (7.4, 11.4) → ---- → (21, 25)

any (51, 3) → → → ⟶

Contraction: closeness of points. [fns that bring points closer]



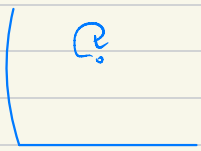
$d' < d$.



F is contraction if there exists $\lambda \in (0,1)$ s.t.

$$d(F(x), F(y)) \leq \lambda \cdot d(x,y)$$

Fixed point: applying the fn gives you the same point.



We can say that F is a contraction ⇒ fixed point is unique

* Amazing: $x_* = \lim_{n \to \infty} x_n$

$$F(x_*) = F\left(\lim_{n \to \infty} x_n\right) = \lim_{n \to \infty} F(x_n) = \lim_{n \to \infty} x_n = x_*$$

Banach's fixed point Theorem:

Every contraction mapping in a non-empty complete metric space admits a unique fixed point

complete: all points in the same space

metric space: some space with a sense of distance b|w points.

Cauchy Sequence:

Contraction

Operator T : old value $\longrightarrow$ New value

$$V : S \to R$$

$$T(V(s)) = \max_{a \in A(s)} \{ x(s,a) + \lambda \cdot V(s') \}$$

$$s' = T(s,a).$$

To prove:

$V_1, V_2 \quad d(T(V_1), T(V_2)) \leq \lambda \, d(V_1, V_2).$

def: $d(V_1, V_2) = \max_i |V_1(i), V_2(i)|$ max - norm.

1. T is mon

(proof given in Lec).

$$T(V+c) = T(V) + \lambda \cdot c.$$

Theorem: T is a contraction. (To prove).

$V_1, V_2 \quad d(T(V_1), T(V_2)) \leq \lambda \cdot d(V_1, V_2).$

assume $d(V_1, V_2) = c$. & $V_2 - c \leq V_1 \leq V_2 + c.$

$$T(V_2 - c) \leq T(V_1) \leq T(V_2 + c).$$

$$\Rightarrow T(V_2) - \lambda c \leq T(V_1) \leq T(V_2) + \lambda \cdot c$$

$$\Rightarrow |T(V_1), T(V_2)| \leq \lambda \, \|V_1, V_2\|$$

$\therefore$ T is a contraction.

As a Linear Programming outline:

$$d_1 = \max(2 + 0.8 d_1 + 1 + 0.8 d_2)$$
$$d_2 = \max(5 + 0.8 d_2 + 8 + 0.8 d_1)$$

Obj: Min $d_1 + d_2$.

s.t
$$d_1 \geq 2 + 0.8 d_1$$
$$d_1 \geq 1 + 0.8 d_2$$
$$d_2 \geq 5 + 0.8 d_2$$
$$d_2 \geq 8 + 0.8 d_1$$

$D \geq T(D)$

$\Rightarrow V \geq T(V)$

$T(V) \geq T(T(V))$

$\Rightarrow V \geq T(V) \geq T^2(V) \geq \dots V_*$.

Markov Decision Process :-

Probability : finite & infinite tosses of coin :-

Random variable: object that depends on random process.

X : Sample space $\longrightarrow$ measurable E.

Expected value : probabilistic weighted average.

$$E(X) = \sum_{x \in E} x \cdot P(X = x).$$

$E[X_n] \not\to E[X]$ even if $X_n \to X$.

$$X_n = n \cdot 1\{U \in (0, \tfrac{1}{n})\}. \quad \text{(uniform dist)}$$

$X_n \to 0$ but $\lim_{n \to \infty} E[X_n] = \lim_{n \to \infty} 1 = 1$.

Monotone Convergence Theorem:

$\{X_n : n \geq 0\}$ seq of random variables with $0 \leq X_n \leq X_{n+1}$ & $X_n \to X$ pointwise.

then $\lim_{n\to\infty} E[X_n] = E[\lim_{n\to\infty} X_n] = E[X]$ //

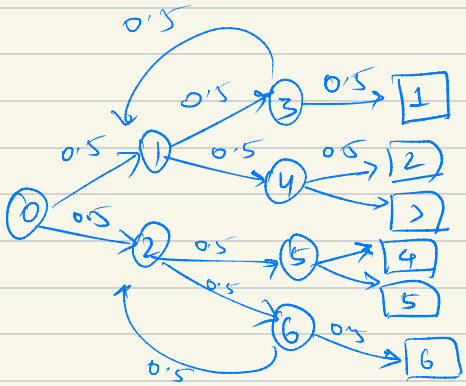
Dominated Convergence Theorem :-

$\{X_n : n \geq 0\}$ $X_n \to X$ pointwise $|X_n| \leq Y < \infty$ and $E[Y] < \infty$

then $\lim_{n\to\infty} E[X_n] = E[X]$.

Markov Chains :

fair coin to simulate a dice :



future depends only on "current."

$M(S, s_0, P, r)$

S is a finite set of states

$s_0 \in S$ initial state

$P : S \times S \to [0,1]$ transition prob matrix.

$r : S \to \mathbb{R}$ (reward func)

Cylinder Sets :- (set of measurable events) [Cyl(ω)]

let finite path $\omega = \langle s_0, s_1, s_2 \ldots s_k \rangle$ is set of all infinite paths with prefix ω

$P(Cyl(\omega)) = \prod_{0 \leq i < k} P(s_i, s_{i+1})$

$$P(\text{getting a } 4) = P(\langle 0,2,5,d4^{w}\rangle) + P(\langle 0,2,6,2,5,d4^{w}\rangle) + \cdots$$

$$= (0.5)^3 + (0.5)^5 + (0.5)^7 + \cdots$$

$$= \frac{(0.5)^3}{1-(0.5)^2} = \frac{1}{6} //$$

what does w at the end mean? (it ends not infinite) rt?

Getting the prob of a event / state reach :

$$x_s = \begin{cases} 1 & \text{if } s \in T \\ \sum_{t \in S} P(s,t)\, x_t & \text{otherwise.} \end{cases}$$

Consider it like the fixed point theorem for discounted values.

$F : [0,1]^S \to [0,1]^S$

$$F(y)(s) = \begin{cases} 1 & \text{if } s \in T \\ \sum_{t \in S} P(s,t)\, y(t) & \text{otherwise.} \end{cases}$$

Markov Chain : $M(S, s_0, P, r)$

discounted reward as V_λ $\quad \lambda \in [0,1)$.

$$V_\lambda(s) = \lim_{N \to \infty} E\left[\sum_{t=1}^{N} \lambda^{t-1} r(X_t)\right] = E\left[\sum_{t=1}^{\infty} \lambda^{t-1} r(X_t)\right]$$

Expected discounted reward :

Policy evaluation :-

2/4/24
Sunday

Deep RL

Learn:

1) value-based Methods
2) differences b/w Monte Carlo and Temporal Difference Learning
3) RL algo: Q-Learning

RL objective: Max. Rewards ← optimum policy (decision making)

2 ways: 1) Policy-based Methods :- train agent directly to learn which action to take in a state
2) Value based Methods :- train agent to learn which state is more valuable and take action that leads to it

Value-based Methods:

Learn a function state ⟶ expected value of being in that state

① $V_\pi(s) = E_\pi\left[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots \mid S_t = s \right]$

for each state,
the state value function outputs
the expected return
if the agent starts in that state.
and then follows the policy forever after

② Action-value function:

$Q_\pi(s,a) = E_\pi[G_t \mid S_t = s, A_t = a]$

the action-value function outputs the expected return
if the agent starts in that state & takes the action
and then follows the policy forever after.

Bellman Equation: