Chapter 2: Markov reward processes (MRP)

- State: S_t $\quad t = 0, 1, \ldots$

- Reward: $R_t \in \mathbb{R}$ $\quad t = 0, 1, \ldots$
  - Cost / utility at time t
  - RV correlated with transitions $S_t \to S_{t+1}$
  - Sometimes: fct of s, s' $\quad r(s, s')$

- Transition probability (dynamics):

$$p(s', r \mid s) = P(S_{t+1} = s', R_{t+1} = r \mid S_t = s)$$

$$S_0 \xrightarrow{R_1} S_1 \xrightarrow{R_2} S_2 \longrightarrow \cdots$$

  - Reward depends on s, s'
  - Assumed homogeneous (time-independent, stationary)
  - No actions yet: dynamics + rewards

- Reward probability: $p(r \mid s) = \sum_{s'} p(s', r \mid s)$

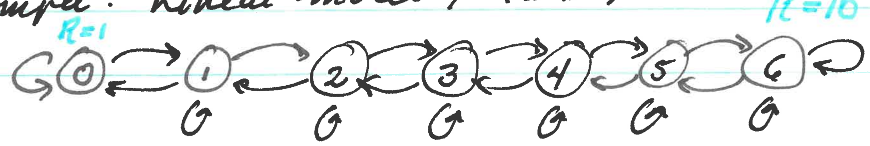
- Transition probability: $p(s' \mid s) = \sum_{r} p(s', r \mid s)$ pure dynamics

- Expected state reward:

$$\rho(s) = E[R_{t+1} \mid S_t = s] = \sum_r r\, p(r \mid s)$$

  - One-time reward from state s

- Example: linear model / Mars rover



R deterministic w/r S_t

$\rho(0) = 1 \quad \rho(6) = 10$

- Return: $G_t = R_{t+1} + R_{t+2} + \cdots + R_T$

$\underset{\text{present}}{G_t}$ ... future

$$\begin{array}{ccccccc} & & R_{t+1} & & R_{t+2} & & R_T \\ S_t & \rightarrow & S_{t+1} & \rightarrow & S_{t+2} & \rightarrow \cdots & S_T \\ \uparrow & & & & & & \uparrow \\ G_t & & & & & & \text{final time} \end{array}$$

- Total reward / cost to go from time t

- Episodic process: T finite
  Process stops/terminates

Calculate backward

$\Rightarrow \quad G_T = 0 \qquad G_{T-1} = R_T \qquad G_{T-2} = R_{T-1} + R_T \qquad \cdots$

- Continuing process: $T = \infty$
  Process goes on and on

- Example:

$$\underset{0}{\overset{R=1}{\text{start}}} \longrightarrow \underset{1}{\overset{R=2}{\text{on}}} \longrightarrow \underset{2}{\overset{R=0}{\text{stop}}} \circlearrowleft$$

$0 \to 1 \to 2 \qquad 1 \to 2 \qquad 2 \qquad$ Episodic

$0 \to 1 \to 2 \to 2 \to 2 \to \cdots \qquad$ Continuing

- Discounted return: $G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 G_{t+3} + \cdots$

$$= \sum_{k=0}^{\infty} \gamma^k R_{t+1+k}$$

- Discount rate $\gamma \in [0,1]$ — Balances present/future rewards
- $G_t < \infty$ for $\gamma \in [0,1)$
- Myopic: $\gamma \approx 0$ — $G_t = R_{t+1}$ for $\gamma = 0$ immediate reward
- Far-sight: $\gamma \approx 1$
- Iteration:

$$G_t = R_{t+1} + \gamma G_{t+1}$$

· Value function: $v(s) = E[G_t \mid S_t = s]$

$$\underset{\overset{\parallel}{s}}{S_t} \to \overbrace{\begin{matrix} R_{t+1} & \gamma R_{t+2} & \gamma^2 R_{t+3} & \\ S_{t+1} \to & S_{t+2} \to & S_{t+3} \to & \cdots \end{matrix}}^{G_t}$$

· State value fct: Expected return / cost to go from state s

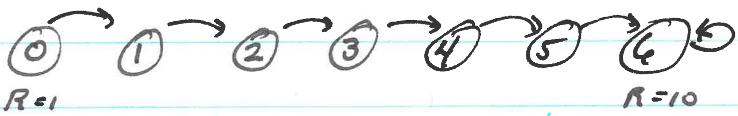
· Doesn't depend on time t

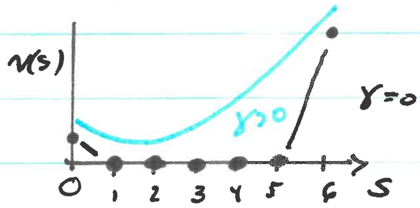
  · Stationary MP

  · Continuing process / infinite return

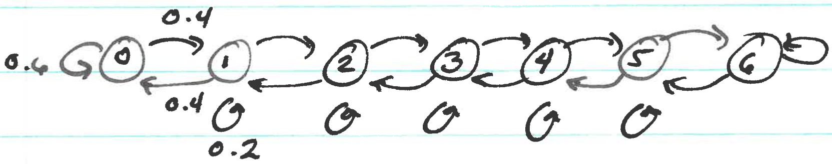
· $\gamma = 0$: $v(s) = \rho(s)$ cost when leaving s

· Example: linear model





deterministic moves / transitions



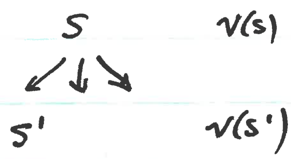
$v(s)$? See CW

· Bellman equation:

$$v(s) = E[R_{t+1} + \gamma G_{t+1} \mid S_t = s]$$

$$= E[R_{t+1} + \gamma v(S_{t+1}) \mid S_t = s]$$

$$\underset{\overset{\parallel}{s}}{S_t} \to \overbrace{\begin{matrix} R_{t+1} & \\ S_{t+1} \to & \overbrace{\begin{matrix} R_{t+2} & \\ S_{t+2} \to & \cdots \end{matrix}}^{G_{t+1}} \end{matrix}}^{G_t}$$



$v(s)$ $\quad v(S_{t+1})$

· Explicit expectation:

$$v(s) = \sum_{r} r\, p(r|s) + \gamma \sum_{s'} v(s')\, p(s'|s)$$

$$= \sum_{r, s'} r\, p(s', r|s) + \gamma \sum_{s', r} v(s')\, p(s', r|s)$$

$$= \rho(s) + \gamma\, (Pv)(s)$$

$P_{ij} = p(j|i)$

· Matrix notation: $v = \rho + \gamma P v$ $\quad v_i = v(i)$

- $v_i = v(i)$ column vector $|S| \times 1$
- $\rho_i = \rho(i)$ " "
- γ scalar
- $P_{ij} = p(j|i)$ $|S| \times |S|$ matrix

· Solution: $v = (I - \gamma P)^{-1} \rho$

- Always exists for $\gamma \in [0, 1)$
- Unique solution

· Bellman operator: $T(v) = \rho + \gamma P v$

$v = T(v)$ Value function = fixed pt of T

· Example: Linear model.

- Write $\underset{7\times1}{\rho}$, $\underset{7\times7}{P}$
- Solve numerically.

· Two ways to define MRPs :

①	②
$p(s', r \mid s)$	
$p(r \mid s)$	
$\rho(s) = E[R_{t+1} \mid S_t = s]$	$\rho(s)$
$p(s' \mid s)$	
P	P

$$v(s) = E[R_{t+1} + \gamma v(S_{t+1}) \mid S_t = s]$$

$$v = \rho + \gamma P v$$