MARKOV DECISION PROCESSES

MARKOV DECISION PROCESSES (MDP)

DEFINITION

An MDP is a model of the agent's world and objective

Is defined by the tuple $M = (\mathcal{S}, \mathcal{A}, \mathcal{R}, p, d_0, \gamma)$

MARKOV DECISION PROCESSES (MDP)

DEFINITION

$\mathcal{S}$ set of all states — Information the agent uses to make a decision

S_t the state at time t

A state needs to include all information necessary to predict the future

Examples:

1. Board state in chess
2. Joint position and velocities of simulated robot
3. Position and momentum of all particles of the universe

MARKOV DECISION PROCESSES (MDP)

DEFINITION

$\mathscr{A}$ set of all actions — possible decisions

A_t the state at time t

Examples:

1. Moving a chess piece to a new position, the set is all valid moves in that state
2. The amount of current applied to a motor
3. Driving a car: Steering wheel angle, gas pedal position, brake position, gear setting

MARKOV DECISION PROCESSES (MDP)

DEFINITION

$\mathscr{R}$ set of all rewards

R_t the reward the agent receives when it transitions to the state S_t

Examples:

1. Money won in a gamble, e.g., total amount - bet
2. {0,1} for successfully completing a task or not
3. -1 every step to make the agent get to a goal faster

MARKOV DECISION PROCESSES (MDP)

DEFINITION

$p : \mathcal{S} \times \mathcal{R} \times \mathcal{S} \times \mathcal{A} \to [0,1]$ — a function that returns the probability of observing the next state and reward after taking an action in a particular state

$$p(s', r \mid s, a) \doteq \Pr\left(S_t = s', R_t = r \mid S_{t-1} = s, A_{t-1} = a\right)$$

Example:

1. Physics equations + reward
2. Chess: next state transition is agent's move + opponent's move, reward +1 if win -1 loss

MARKOV DECISION PROCESSES (MDP)

DEFINITION

$d_0 : \mathcal{S} \rightarrow [0,1]$ — function that returns the probability of starting in a given state

$$d_0(s) = \Pr(S_0 = s)$$

Examples:

1. Chess white player: deterministic always starts in the same way
2. Chess black player: stochastic distribution depends on white player's strategy

MARKOV DECISION PROCESSES (MDP)

DEFINITION

$\gamma \in [0,1]$ — discount factor to downweight rewards in the future

MARKOV DECISION PROCESSES (MDP)

USEFUL FUNCTIONS

$$p(s, a, s') = \Pr(S_t = s' \mid S_{t-1} = s, A_{t-1} = a)$$

$$r(s, a, s') \doteq \mathbb{E}[R_t \mid S_{t-1} = s, A_{t-1} = a, S_t = s']$$

$$r(s, a) \doteq \mathbb{E}[R_t \mid S_t = s, A_t = a]$$

MARKOV DECISION PROCESSES (MDP)

TYPES OF TASKS

Episodic Tasks: naturally broken up into independent sequences

- Balancing a pen on your finger. The pen falls, and then you start over with it, balancing

Continuing tasks: go on forever

- Maximize the time you balance a pen on your finger and you have to quickly pick up the pen quickly
- Maintain the temperature of a refrigerator at 1 deg C

Other tasks:

- Life: you only get one long episode

OBJECTIVE

CUMULATIVE REWARD: EPISODIC TASKS

$$\max \sum_{t=1}^{T} R_t$$

Only episodic MDPs must terminate in finite time, i.e., $\Pr(T < \infty) = 1$

Note: This object says it does not matter when big rewards occur

OBJECTIVE

CUMULATIVE REWARD: CONTINUING TASKS

Problem: $\sum_{t=1}^{T} R_t -$ can be infinite, we need bounded objectives

Solution: place lower weight on future rewards

$$\max \sum_{t=1}^{\infty} \gamma^{t-1} R_t \quad \gamma \in [0,1)$$

$\gamma < 1$ Implies that getting higher rewards sooner is better.

OBJECTIVE

CUMULATIVE REWARD: CONTINUING TASKS

$$R_{\max} = \max_{r \in \mathscr{R}} r$$

OBJECTIVE

CUMULATIVE REWARD: CONTINUING TASKS

$$R_{\max} = \max_{r \in \mathscr{R}} r$$

$$\max \sum_{t=1}^{\infty} \gamma^{t-1} R_t \leq$$

OBJECTIVE

CUMULATIVE REWARD: CONTINUING TASKS

$$R_{\max} = \max_{r \in \mathscr{R}} r$$

$$\max \sum_{t=1}^{\infty} \gamma^{t-1} R_t \leq \sum_{t=1}^{\infty} \gamma^{t-1} R_{\max}$$

OBJECTIVE

CUMULATIVE REWARD: CONTINUING TASKS

$$R_{\max} = \max_{r \in \mathscr{R}} r$$

$$\max \sum_{t=1}^{\infty} \gamma^{t-1} R_t \leq \sum_{t=1}^{\infty} \gamma^{t-1} R_{\max}$$

$$= R_{\max} \sum_{t=1}^{\infty} \gamma^{t-1}$$

OBJECTIVE

CUMULATIVE REWARD: CONTINUING TASKS

$$R_{\max} = \max_{r \in \mathscr{R}} r$$

$$\max \sum_{t=1}^{\infty} \gamma^{t-1} R_t \leq \sum_{t=1}^{\infty} \gamma^{t-1} R_{\max}$$

$$= R_{\max} \sum_{t=1}^{\infty} \gamma^{t-1}$$

$$\alpha \in [0,1), \quad \sum_{i=0}^{\infty} \alpha^i = \frac{1}{1-\alpha}$$

OBJECTIVE

CUMULATIVE REWARD: CONTINUING TASKS

$$R_{\max} = \max_{r \in \mathscr{R}} r$$

$$\max \sum_{t=1}^{\infty} \gamma^{t-1} R_t \leq \sum_{t=1}^{\infty} \gamma^{t-1} R_{\max}$$

$$= R_{\max} \sum_{t=1}^{\infty} \gamma^{t-1}$$

$$= R_{\max} \frac{1}{1-\gamma}$$

$$\alpha \in [0,1), \quad \sum_{i=0}^{\infty} \alpha^i = \frac{1}{1-\alpha}$$

OBJECTIVE

CUMULATIVE REWARD: UNIFIED OBJECTIVE

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

$\gamma \in [0,1]$ for episodic tasks,

- recall $R_t = 0$ when $S_t = s_\infty$, So G_t is bounded

$\gamma \in [0,1)$ for continuing tasks

OBJECTIVE

CUMULATIVE REWARD: NOTATION AND NAMING

G_t — has several names:

Book: *return*

Others: cumulative reward, discounted cumulative reward, (negative) cost-to-go

Common mistake: call G_t the reward

CUMULATIVE REWARD

CONNECTIONS

What does G_t represent?

$$G_t = R_{t+1} + \gamma \underbrace{R_{t+2} + \gamma^2 R_{t+3} + \dots}_{G_{t+1}}$$

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

Exponential moving average: $\mu_{t+1} = (1 - \alpha)\mu_t + \alpha x_t$

μ is an exponentially weighted average of x

CUMULATIVE REWARD

CONNECTIONS

Total weight in G_t is $\frac{1}{1-\gamma}$

To get an average, divide by the total weight, e.g., $\bar{R}_t = (1-\gamma)G_t$

CUMULATIVE REWARD

CONNECTIONS

Total weight in G_t is $\frac{1}{1-\gamma}$

To get an average, divide by the total weight, e.g., $\bar{R}_t = (1-\gamma)G_t$

$$\begin{aligned}\bar{R}_t = (1-\gamma)G_t &= (1-\gamma)R_{t+1} + \gamma(1-\gamma)G_{t+1} \\ &= (1-\gamma)R_{t+1} + \gamma\bar{R}_{t+1}\end{aligned}$$

CUMULATIVE REWARD

CONNECTIONS

Total weight in G_t is $\frac{1}{1-\gamma}$

To get an average, divide by the total weight, e.g., $\bar{R}_t = (1-\gamma)G_t$

$$\begin{aligned}\bar{R}_t = (1-\gamma)G_t = &\ (1-\gamma)R_{t+1} + \gamma(1-\gamma)G_{t+1} \\ = &\ (1-\gamma)R_{t+1} + \gamma\bar{R}_{t+1}\end{aligned}$$

$\bar{R}_t$ is an exponential moving average of reward computed going backward in time with weight $(1-\gamma)$

G_t is an unnormalized exponential moving average

MDP: EXAMPLE

GRIDWORLD

$\mathscr{A} = \{ \leftarrow , \rightarrow , \uparrow , \downarrow \}$ – Agent attempts to go in a direction

With probability, c an orthogonal action will be executed

If the agent would hit a wall, it will stay in the same state

Start 1	2	3	4	5
6	7	8	9	10
11	12	13	14	15
16	17	18	19	20
21	22	23	24	Goal 25

MDP: EXAMPLE

GRIDWORLD

$\mathscr{A} = \{ \leftarrow , \rightarrow , \uparrow , \downarrow \}$ – Agent attempts to go in a direction

With probability, c an orthogonal action will be executed

$$p(13, \uparrow ,8) = 1 - c, p(13, \uparrow ,14) = c/2$$

If the agent would hit a wall, it will stay in the same state

$$p(3, \uparrow ,3) = 1 - c$$

Start 1	2	3	4	5
6	7	8	9	10
11	12	13	14	15
16	17	18	19	20
21	22	23	24	Goal 25

MDP: EXAMPLE

GRIDWORLD

$\mathscr{A} = \{ \leftarrow , \rightarrow , \uparrow , \downarrow \}$ – Agent attempts to go in a direction

With probability, c an orthogonal action will be executed

$$p(13, \uparrow ,8) = 1 - c, p(13, \uparrow ,14) = c/2$$

If the agent would hit a wall, it will stay in the same state

$$p(3, \uparrow ,3) = 1 - c$$

$$\Pr(S_0 = 1) = 1.0$$

$$\forall a,\ p(s_\infty,0 \mid s_{25}, a) = 1.0$$

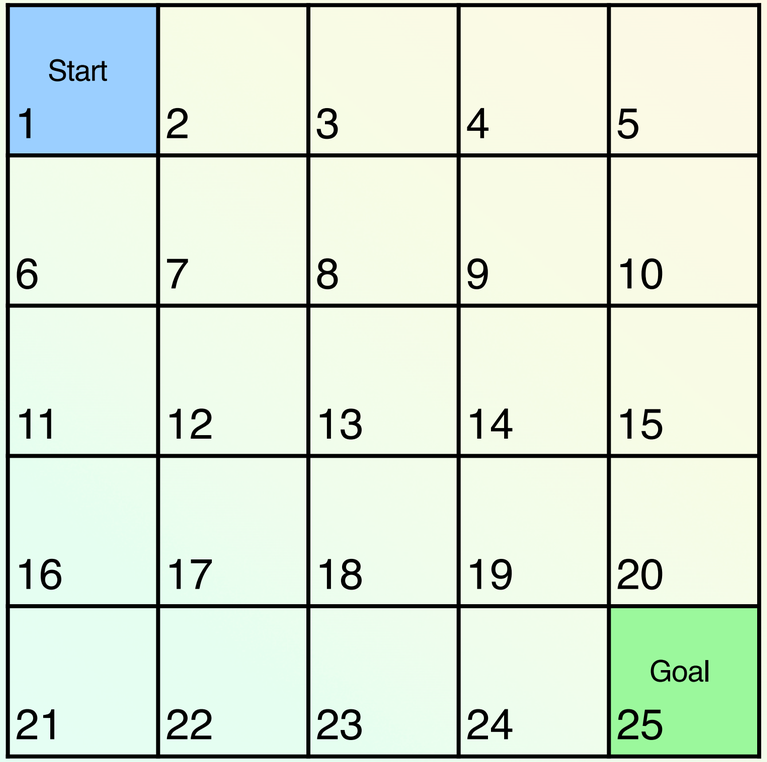
Start 1	2	3	4	5
6	7	8	9	10
11	12	13	14	15
16	17	18	19	20
21	22	23	24	Goal 25

MDP: EXAMPLE

GRIDWORLD — REWARD FUNCTION

Want the agent to get to the goal as fast as possible

What values of the reward function and γ make this happen?



MDP: EXAMPLE

GRIDWORLD — REWARD FUNCTION

Want the agent to get to the goal as fast as possible

What reward function and value of γ lead to this behavior?

1. $\forall t,\ R_t = -1$ except for $S_t = s_\infty, \gamma \in (0,1]$

   Penalty for not ending the episode

2. $r(s,a,s') = 0$ except $r(s,a,25) = R_{\text{goal}} > 0, \gamma \in (0,1)$

   Reward for getting to the goal

3. $r(s,a,s') = \text{dist}(s',25) - \text{dist}(s,25), \gamma \in (0,1)$

   The reward for making progress towards the goal

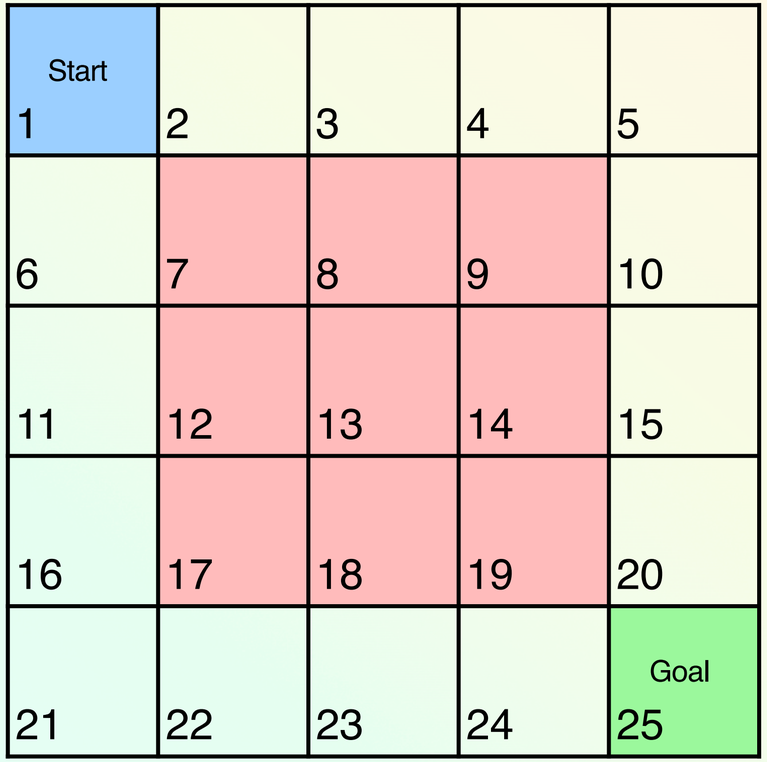
Start 1	2	3	4	5
6	7	8	9	10
11	12	13	14	15
16	17	18	19	20
21	22	23	24	Goal 25

MDP: EXAMPLE

GRIDWORLD — REWARD FUNCTION

Want the agent to get to the goal as fast as possible but avoid the reward squares?

What reward function and value of γ lead to this behavior?



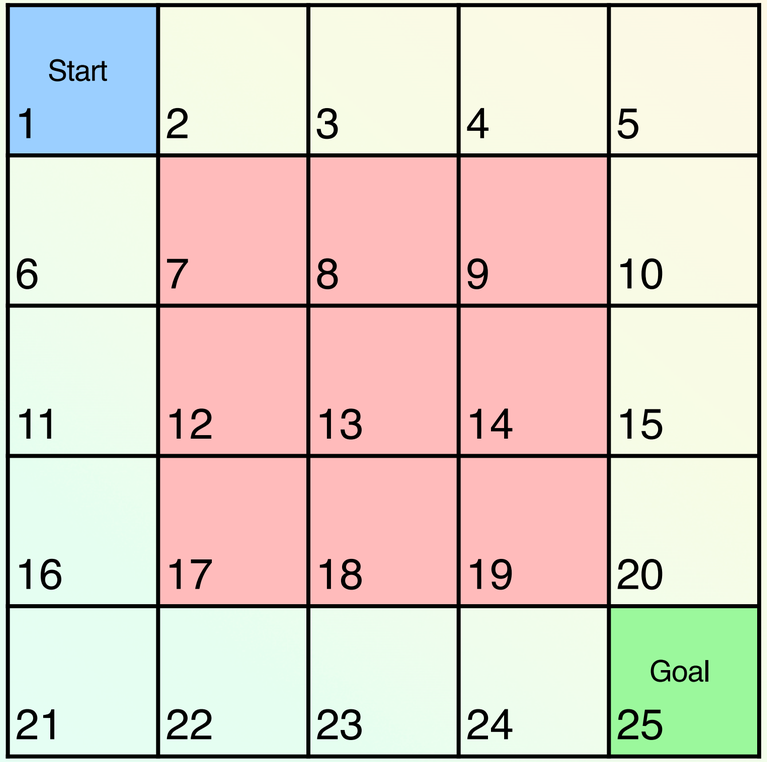
MDP: EXAMPLE

GRIDWORLD — REWARD FUNCTION

Want the agent to get to the goal as fast as possible but avoid the reward squares?

What reward function and value of γ lead to this behavior?

- $\forall t,\ R_t = -1$ in normal squares
- $s' \in \mathcal{S}_{\text{red}}\ \ r(s, a, s') = -R_{\text{red}} < -1$
- What value for R_{red}? Depends on c and γ

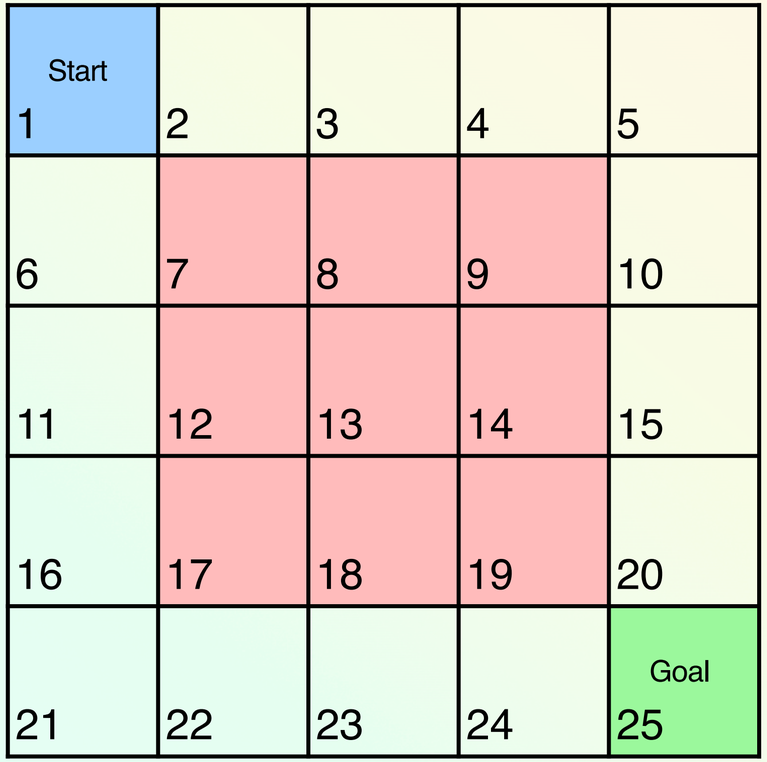


MDP: EXAMPLE

GRIDWORLD — REWARD FUNCTION

$c = 0$ deterministic transitions

- $R_{\text{red}} < -1, \gamma \in (0,1]$



MDP: EXAMPLE

GRIDWORLD — REWARD FUNCTION

$c > 0, \gamma = 0$ best one step choice for state 24

- $$\begin{aligned} r(24, \rightarrow) = \; & p(24, \rightarrow, 25)(-1) \\ & + p(24, \rightarrow, 24)(-1) \\ & + p(24, \rightarrow, 19)R_{\text{red}} \end{aligned}$$

Start 1	2	3	4	5
6	7	8	9	10
11	12	13	14	15
16	17	18	19	20
21	22	23	24	Goal 25

MDP: EXAMPLE

GRIDWORLD — REWARD FUNCTION

$c > 0, \gamma = 0$ best one step choice for state 24

- $$r(24, \rightarrow) = (1-c)(-1) + \frac{c}{2}(-1) + \frac{c}{2}R_{\text{red}}$$

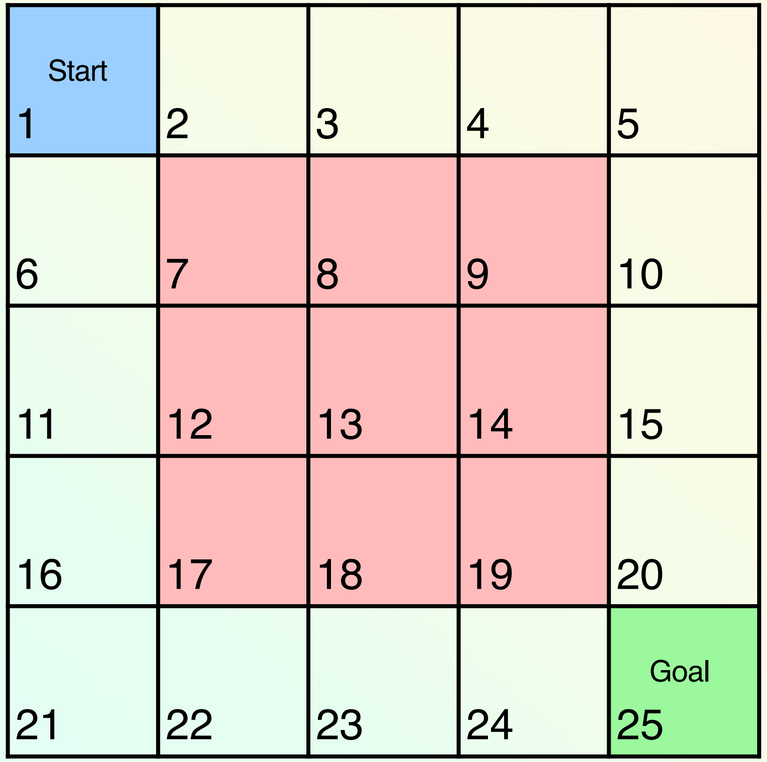
Start 1	2	3	4	5
6	7	8	9	10
11	12	13	14	15
16	17	18	19	20
21	22	23	24	Goal 25

MDP: EXAMPLE

GRIDWORLD — REWARD FUNCTION

$c > 0, \gamma = 0$ best one step choice for state 24

- $r(24, \rightarrow) = -(1-c) - \frac{c}{2} + \frac{c}{2} R_{\text{red}}$

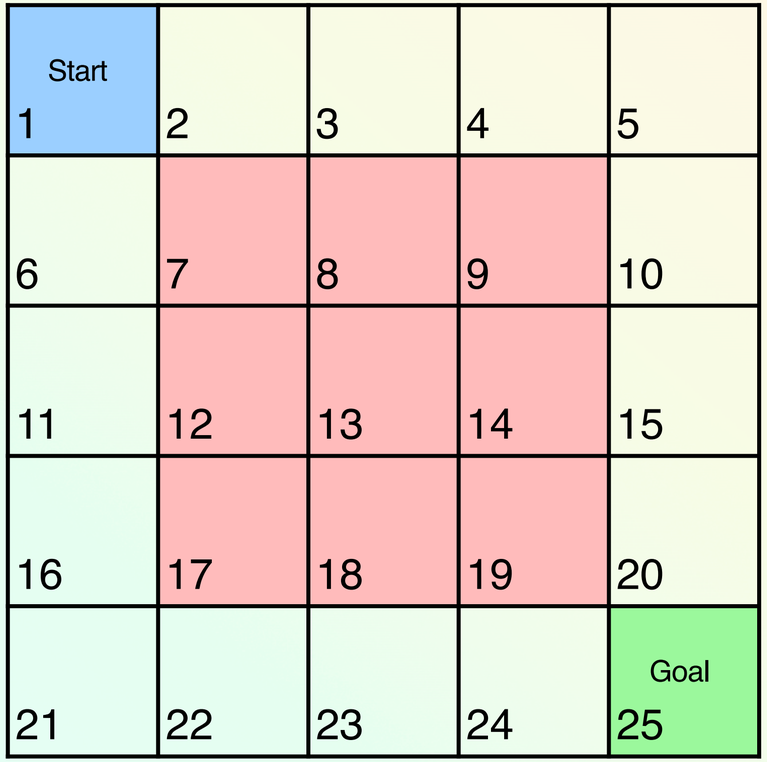


MDP: EXAMPLE

GRIDWORLD — REWARD FUNCTION

$c > 0, \gamma = 0$ best one step choice for state 24

- $r(24, \rightarrow) = -(1 - \frac{c}{2}) + \frac{c}{2} R_{\text{red}}$



MDP: EXAMPLE

GRIDWORLD — REWARD FUNCTION

$c > 0, \gamma = 0$ best one step choice for state 24

- $r(24, \rightarrow) = -(1 - \frac{c}{2}) + \frac{c}{2} R_{\text{red}}$
- $r(24, \downarrow) = p(24, \downarrow, 24)(-1) + p(24, \downarrow, 23)(-1) + p(24, \downarrow, 25)(-1)$

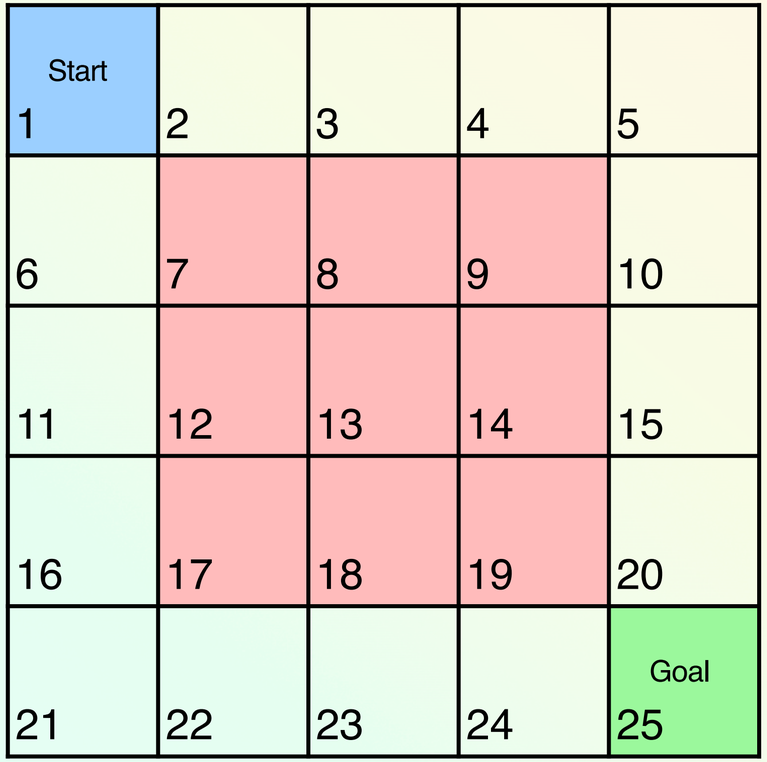
Start 1	2	3	4	5
6	7	8	9	10
11	12	13	14	15
16	17	18	19	20
21	22	23	24	Goal 25

MDP: EXAMPLE

GRIDWORLD — REWARD FUNCTION

$c > 0, \gamma = 0$ best one step choice for state 24

- $r(24, \rightarrow) = -(1 - \frac{c}{2}) + \frac{c}{2} R_{\text{red}}$
- $r(24, \downarrow) = -1$



MDP: EXAMPLE

GRIDWORLD — REWARD FUNCTION

$c > 0, \gamma = 0$ best one step choice for state 24

- $r(24, \rightarrow) = -(1 - \frac{c}{2}) + \frac{c}{2} R_{\text{red}}$
- $r(24, \downarrow) = -1$
- $$r(24, \downarrow) - r(24, \rightarrow) = -1 - \left(-(1 - \frac{c}{2}) + \frac{c}{2} R_{\text{red}} \right)$$
$$= -1 + (1 - \frac{c}{2}) - \frac{c}{2} R_{\text{red}}$$
$$= -\frac{c}{2} - \frac{c}{2} R_{\text{red}} = -\frac{c}{2}(1 - R_{\text{red}})$$
- $R_{\text{red}} < -1$ means $r(24, \downarrow) > r(24, \rightarrow)$

Start 1	2	3	4	5
6	7	8	9	10
11	12	13	14	15
16	17	18	19	20
21	22	23	24	Goal 25

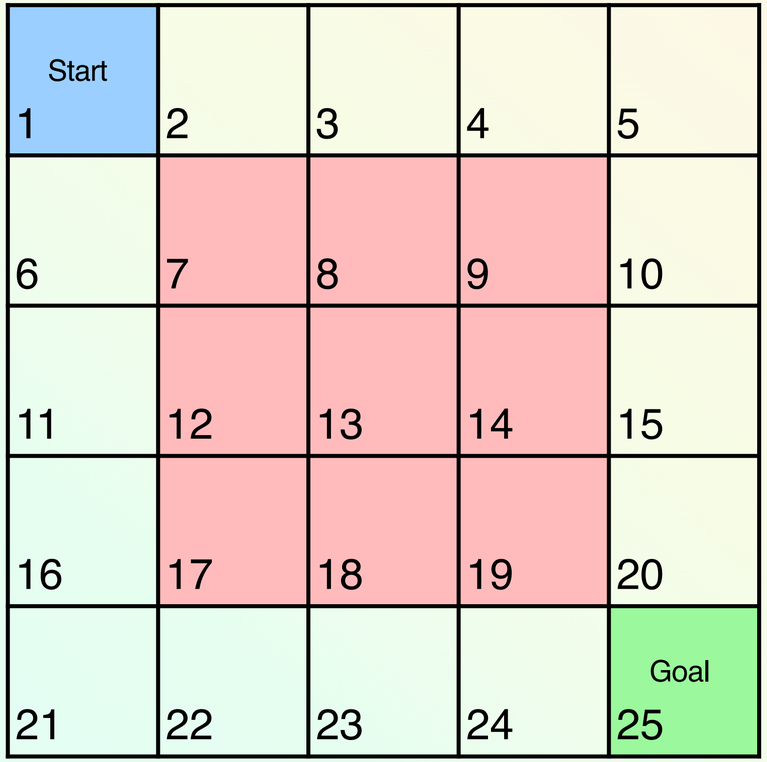
MDP: EXAMPLE

GRIDWORLD — REWARD FUNCTION

$c > 0, \gamma > 0$

Need to reason about entering red states versus taking longer to reach the goal.

Need tools developed in next weeks material



MDP: EXAMPLE

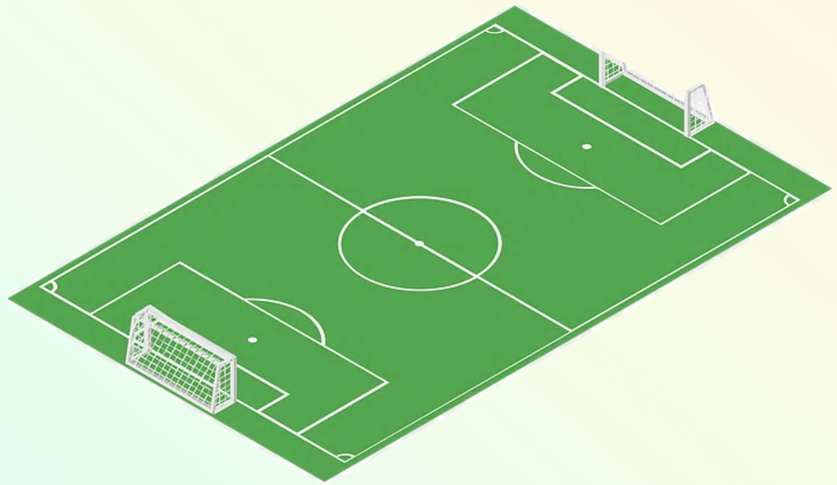
SOCCER AND REWARD FUNCTIONS

$s =$ (player positions, ball position, time remaining, $\ldots$)

$p -$ physics (assumes fixed strategy for opponent)

Objective win the game:

$$R_T = \begin{cases} +1 & \text{win} \\ -1 & \text{loss} \\ 0 & \text{draw} \end{cases} \quad \text{if win, } -1 \text{ if lost}$$



Sparse reward: The agent needs to learn strategy only from wins and losses

MDP: EXAMPLE

SOCCER AND REWARD FUNCTIONS

Give feedback more often to overcome sparsity

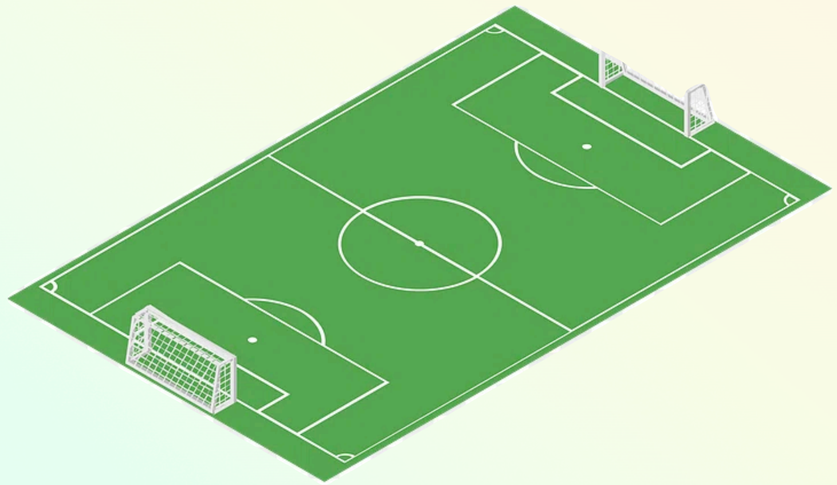
$R_t = R_{\text{scored}}$ if scored a goal

$R_t = 0.0001$ for every minute of possession

Induces unwanted behavior

Sore goals but don’t play defense

Keeping the ball but never scoring



REWARD FUNCTIONS

LESSONS

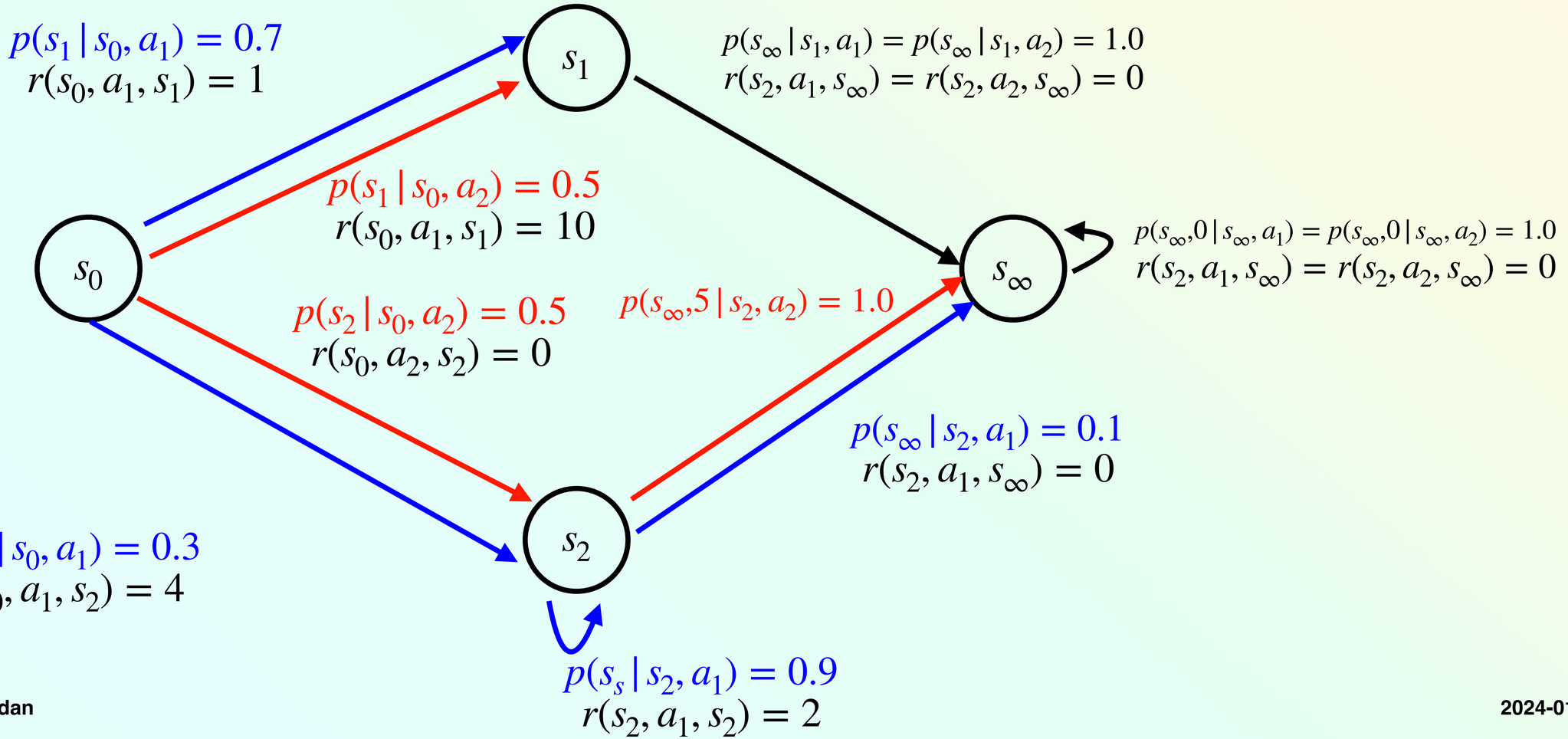
- It is difficult to predict what the optimal behavior is from the reward function
- Might require trial and error to get the behavior correct

Best advice:

- Only reward the agent for achieving the desired behavior

OPTIMAL ACTION?

EXAMPLE GRAPH



OPTIMAL ACTION

ESTIMATING CUMULATIVE REWARD

$$\arg\max_{a} \mathbb{E}[G_t \mid S_t = s, A_t = a]$$

Note: this is not how to find the optimal action, but is a quantity that will help

OPTIMAL ACTION

ESTIMATING CUMULATIVE REWARD

$$\arg\max_a \mathbb{E}[G_t \,|\, S_t = s, A_t = a]$$

Note: this is not how to find the optimal action, but is a quantity that will help

$$\mathbb{E}\left[G_t \,|\, S_t = s, A_t = a\right] = \sum_{s_{t+1}, r_{t+1}, a_{t+1}, \ldots} \Pr\left(S_{t+1} = s_{t+1}, R_{t+1} = r_{t+1}, A_{t+1} = a_{t+1}, \ldots \,|\, S_t = s, A_t = a\right) \sum_{k=0}^{\infty} \gamma^k R_{t+1+k}$$

MDP REASONING

PROBABILITIES

$$\Pr(S_0 = s) = d_0(s)$$

$$\Pr(S_1 = s_1, R_1 = r_1 \,|\, S_0 = s_0, A_0 = a_0) = p(s_1, r_1 \,|\, s_0, a_0)$$

$$\Pr(S_0 = s_0, A_0 = a_0, R_1 = r_1, S_1 = s_1) = ?$$

MDP REASONING

PROBABILITIES

$$\Pr(S_0 = s) = d_0(s)$$

$$\Pr(S_1 = s_1, R_1 = r_1 \,|\, S_0 = s_0, A_0 = a_0) = p(s_1, r_1 \,|\, s_0, a_0)$$

$$\Pr(S_0 = s_0, A_0 = a_0, R_1 = r_1, S_1 = s_1) = \Pr(R_1 = r_1, S_1 = s_1 \,|\, S_0 = s_0, A_0 = a_0) \Pr(S_0 = s_0, A_0 = a_0)$$

MDP REASONING

PROBABILITIES

$$\Pr(S_0 = s) = d_0(s)$$

$$\Pr(S_1 = s_1, R_1 = r_1 \,|\, S_0 = s_0, A_0 = a_0) = p(s_1, r_1 \,|\, s_0, a_0)$$

$$\begin{aligned}\Pr(S_0 = s_0, A_0 = a_0, R_1 = r_1, S_1 = s_1) &= \Pr(R_1 = r_1, S_1 = s_1 \,|\, S_0 = s_0, A_0 = a_0) \Pr(S_0 = s_0, A_0 = a_0) \\ &= \Pr(R_1 = r_1, S_1 = s_1 \,|\, S_0 = s_0, A_0 = a_0) \Pr(A_0 = a_0 \,|\, S_0 = s_0) \Pr(S_0 = s_0)\end{aligned}$$

MDP REASONING

PROBABILITIES

$$\Pr(S_0 = s) = d_0(s)$$

$$\Pr(S_1 = s_1, R_1 = r_1 \,|\, S_0 = s_0, A_0 = a_0) = p(s_1, r_1 \,|\, s_0, a_0)$$

$$\begin{aligned}
\Pr(S_0 = s_0, A_0 = a_0, R_1 = r_1, S_1 = s_1) &= \Pr(R_1 = r_1, S_1 = s_1 \,|\, S_0 = s_0, A_0 = a_0) \Pr(S_0 = s_0, A_0 = a_0) \\
&= \Pr(R_1 = r_1, S_1 = s_1 \,|\, S_0 = s_0, A_0 = a_0) \Pr(A_0 = a_0 \,|\, S_0 = s_0) \Pr(S_0 = s_0) \\
&= p(s_1, r_1 \,|\, s_0, a_0) \Pr(A_0 = a_0 \,|\, S_0 = s_0) d_0(s_0)
\end{aligned}$$

MDP REASONING

PROBABILITIES

$$\Pr(S_0 = s) = d_0(s)$$

$$\Pr(S_1 = s_1, R_1 = r_1 \,|\, S_0 = s_0, A_0 = a_0) = p(s_1, r_1 \,|\, s_0, a_0)$$

$$\begin{aligned}\Pr(S_0 = s_0, A_0 = a_0, R_1 = r_1, S_1 = s_1) &= \Pr(R_1 = r_1, S_1 = s_1 \,|\, S_0 = s_0, A_0 = a_0) \Pr(S_0 = s_0, A_0 = a_0) \\ &= \Pr(R_1 = r_1, S_1 = s_1 \,|\, S_0 = s_0, A_0 = a_0) \Pr(A_0 = a_0 \,|\, S_0 = s_0) \Pr(S_0 = s_0) \\ &= p(s_1, r_1 \,|\, s_0, a_0) \Pr(A_0 = a_0 \,|\, S_0 = s_0) d_0(s_0)\end{aligned}$$

$\Pr(A_t = a \,|\, S_t = s) = ?$ — Defined by the agent's *policy.* We will discuss this in the future.

MDP REASONING

PROBABILITIES

$$\Pr(S_0 = s_0, A_0 = a_0, R_1 = r_1, S_1 = s_1, \ldots, R_t = r_t, S_t = s_t) = ?$$

MDP REASONING

PROBABILITIES

$$\Pr(S_0 = s_0, A_0 = a_0, R_1 = r_1, S_1 = s_1, \ldots, R_t = r_t, S_t = s_t)$$

$$= \Pr(R_t = r_t, S_t = s_t \mid S_{t-1} = s_{t-1}, A_{t-1} = a_{t-1}, \ldots, S_0 = s_0) \Pr(S_{t-1} = s_{t-1}, A_{t-1} = a_{t-1}, \ldots, S_0 = s_0)$$

MDP REASONING

PROBABILITIES

$$\Pr(S_0 = s_0, A_0 = a_0, R_1 = r_1, S_1 = s_1, \ldots, R_t = r_t, S_t = s_t)$$

$$= \Pr(R_t = r_t, S_t = s_t \mid S_{t-1} = s_{t-1}, A_{t-1} = a_{t-1}, \ldots, S_0 = s_0) \Pr(S_{t-1} = s_{t-1}, A_{t-1} = a_{t-1}, \ldots, S_0 = s_0)$$

$$= \Pr(R_t = r_t, S_t = s_t \mid S_{t-1} = s_{t-1}, A_{t-1} = a_{t-1}) \Pr(S_{t-1} = s_{t-1}, A_{t-1} = a_{t-1}, \ldots, S_0 = s_0)$$

MDP REASONING

PROBABILITIES

$$\Pr(S_0 = s_0, A_0 = a_0, R_1 = r_1, S_1 = s_1, \ldots, R_t = r_t, S_t = s_t)$$

$$= \Pr(R_t = r_t, S_t = s_t \mid S_{t-1} = s_{t-1}, A_{t-1} = a_{t-1}, \ldots, S_0 = s_0) \Pr(S_{t-1} = s_{t-1}, A_{t-1} = a_{t-1}, \ldots, S_0 = s_0)$$

$$= \Pr(R_t = r_t, S_t = s_t \mid S_{t-1} = s_{t-1}, A_{t-1} = a_{t-1}) \Pr(S_{t-1} = s_{t-1}, A_{t-1} = a_{t-1}, \ldots, S_0 = s_0)$$

$$= p(s_t, r_t \mid s_{t-1}, a_{t-1}) \Pr(S_{t-1} = s_{t-1}, A_{t-1} = a_{t-1}, \ldots, S_0 = s_0)$$

- Applied Markov property

MDP REASONING

PROBABILITIES

$$\Pr(S_0 = s_0, A_0 = a_0, R_1 = r_1, S_1 = s_1, \ldots, R_t = r_t, S_t = s_t)$$

$$= \Pr(R_t = r_t, S_t = s_t \mid S_{t-1} = s_{t-1}, A_{t-1} = a_{t-1}, \ldots, S_0 = s_0) \Pr(S_{t-1} = s_{t-1}, A_{t-1} = a_{t-1}, \ldots, S_0 = s_0)$$

$$= \Pr(R_t = r_t, S_t = s_t \mid S_{t-1} = s_{t-1}, A_{t-1} = a_{t-1}) \Pr(S_{t-1} = s_{t-1}, A_{t-1} = a_{t-1}, \ldots, S_0 = s_0)$$

$$= p(s_t, r_t \mid s_{t-1}, a_{t-1}) \Pr(S_{t-1} = s_{t-1}, A_{t-1} = a_{t-1}, \ldots, S_0 = s_0)$$

- Applied Markov property

$$= p(s_t, r_t \mid s_{t-1}, a_{t-1}) \Pr(A_{t-1} = a_{t-1} \mid S_{t-1} = s_{t-1}) \Pr(S_{t-1} = s_{t-1}, R_{t-1} = r_{t-1}, s_{t-2}, a_{t-2}, \ldots, S_0 = s_0)$$

MDP REASONING

PROBABILITIES

$$\Pr(S_0 = s_0, A_0 = a_0, R_1 = r_1, S_1 = s_1, \ldots, R_t = r_t, S_t = s_t)$$

$$= \Pr(R_t = r_t, S_t = s_t \mid S_{t-1} = s_{t-1}, A_{t-1} = a_{t-1}, \ldots, S_0 = s_0) \Pr(S_{t-1} = s_{t-1}, A_{t-1} = a_{t-1}, \ldots, S_0 = s_0)$$

$$= \Pr(R_t = r_t, S_t = s_t \mid S_{t-1} = s_{t-1}, A_{t-1} = a_{t-1}) \Pr(S_{t-1} = s_{t-1}, A_{t-1} = a_{t-1}, \ldots, S_0 = s_0)$$

$$= p(s_t, r_t \mid s_{t-1}, a_{t-1}) \Pr(S_{t-1} = s_{t-1}, A_{t-1} = a_{t-1}, \ldots, S_0 = s_0)$$

- Applied Markov property

$$= p(s_t, r_t \mid s_{t-1}, a_{t-1}) \Pr(A_{t-1} = a_{t-1} \mid S_{t-1} = s_{t-1}) \Pr(S_{t-1} = s_{t-1}, R_{t-1} = r_{t-1}, s_{t-2}, a_{t-2}, \ldots, S_0 = s_0)$$

$$= p(s_t, r_t \mid s_{t-1}, a_{t-1}) \Pr(A_{t-1} = a_{t-1} \mid S_{t-1} = s_{t-1}) p(s_{t-1}, r_{t-1} \mid s_{t-2}, a_{t-2}) \Pr(S_{t-2} = s_{t-2}, A_{t-2} = a_{t-2}, \ldots, S_0 = s_0)$$

MDP REASONING

PROBABILITIES

$$\Pr(S_0 = s_0, A_0 = a_0, R_1 = r_1, S_1 = s_1, \ldots, R_t = r_t, S_t = s_t)$$

$$= \Pr(R_t = r_t, S_t = s_t \mid S_{t-1} = s_{t-1}, A_{t-1} = a_{t-1}, \ldots, S_0 = s_0) \Pr(S_{t-1} = s_{t-1}, A_{t-1} = a_{t-1}, \ldots, S_0 = s_0)$$

$$= \Pr(R_t = r_t, S_t = s_t \mid S_{t-1} = s_{t-1}, A_{t-1} = a_{t-1}) \Pr(S_{t-1} = s_{t-1}, A_{t-1} = a_{t-1}, \ldots, S_0 = s_0)$$

$$= p(s_t, r_t \mid s_{t-1}, a_{t-1}) \Pr(S_{t-1} = s_{t-1}, A_{t-1} = a_{t-1}, \ldots, S_0 = s_0)$$

- Applied Markov property

$$= p(s_t, r_t \mid s_{t-1}, a_{t-1}) \Pr(A_{t-1} = a_{t-1} \mid S_{t-1} = s_{t-1}) \Pr(S_{t-1} = s_{t-1}, R_{t-1} = r_{t-1}, s_{t-2}, a_{t-2}, \ldots, S_0 = s_0)$$

$$= p(s_t, r_t \mid s_{t-1}, a_{t-1}) \Pr(A_{t-1} = a_{t-1} \mid S_{t-1} = s_{t-1}) p(s_{t-1}, r_{t-1} \mid s_{t-2}, a_{t-2}) \Pr(S_{t-2} = s_{t-2}, A_{t-2} = a_{t-2}, \ldots, S_0 = s_0)$$

$$= d_0(s_0) \prod_{k=1}^{t} p(s_k, r_k \mid s_{k-1}, a_{k-1}) \Pr(A_{k-1} = a_{k-1} \mid S_{k-1} = s_{k-1})$$

NEXT CLASS

WHAT YOU SHOULD DO

1. Read the book (lots of good information and intuition)

Wednesday: In class exercises on MDPs and reward functions