Sequential Decision-Making Under Uncertainty:
Markov Decision Processes

Siddharth Srivastava, Ph.D.
Assistant Professor
Arizona State University

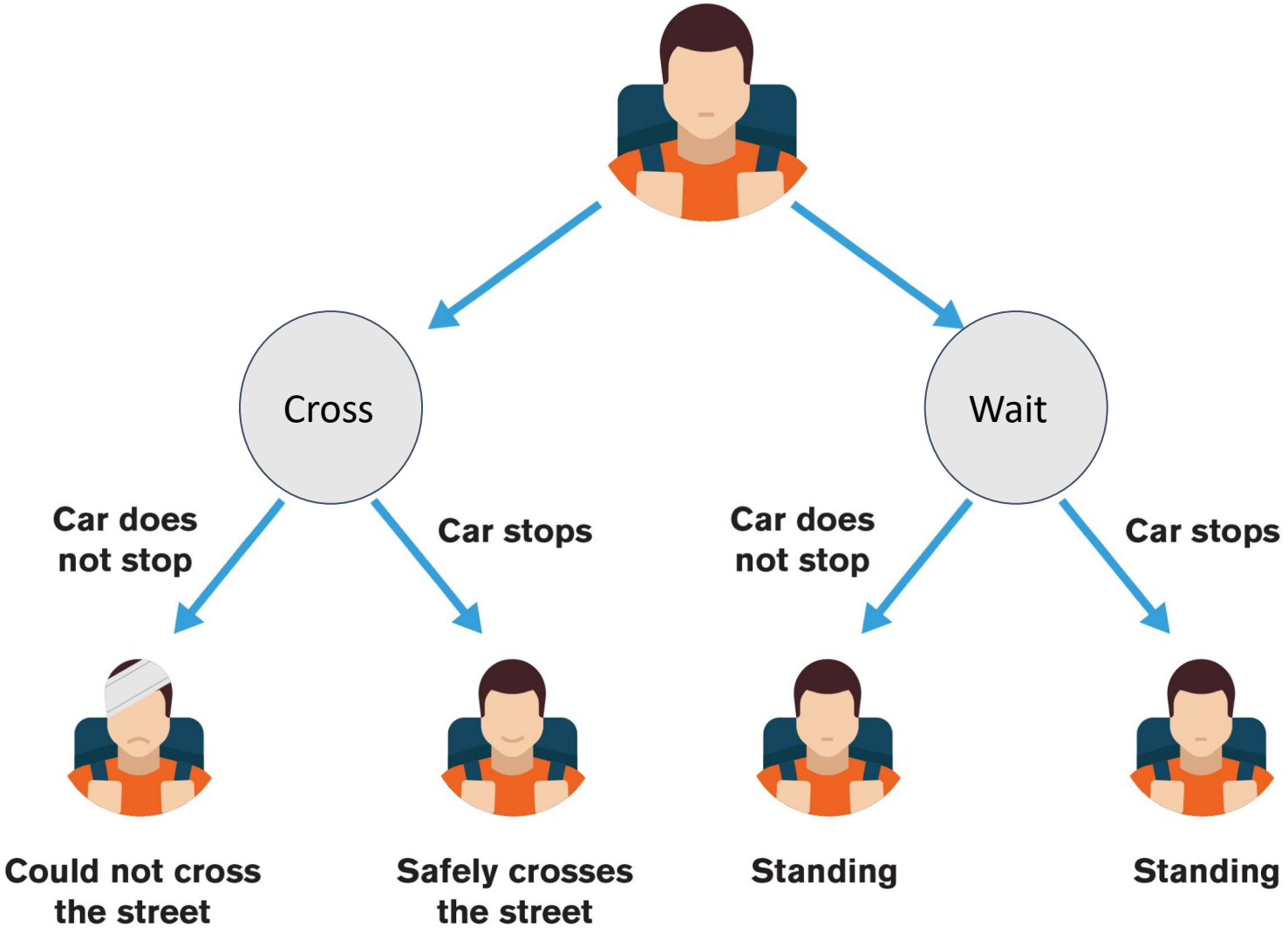
Ira A. Fulton Schools of Engineering
Arizona State University

Dealing with Non-Deterministic Environments

Actions: To cross the street or not to cross the street.

- There is a chance of being run over 😟

Dealing with Non-Deterministic Environments



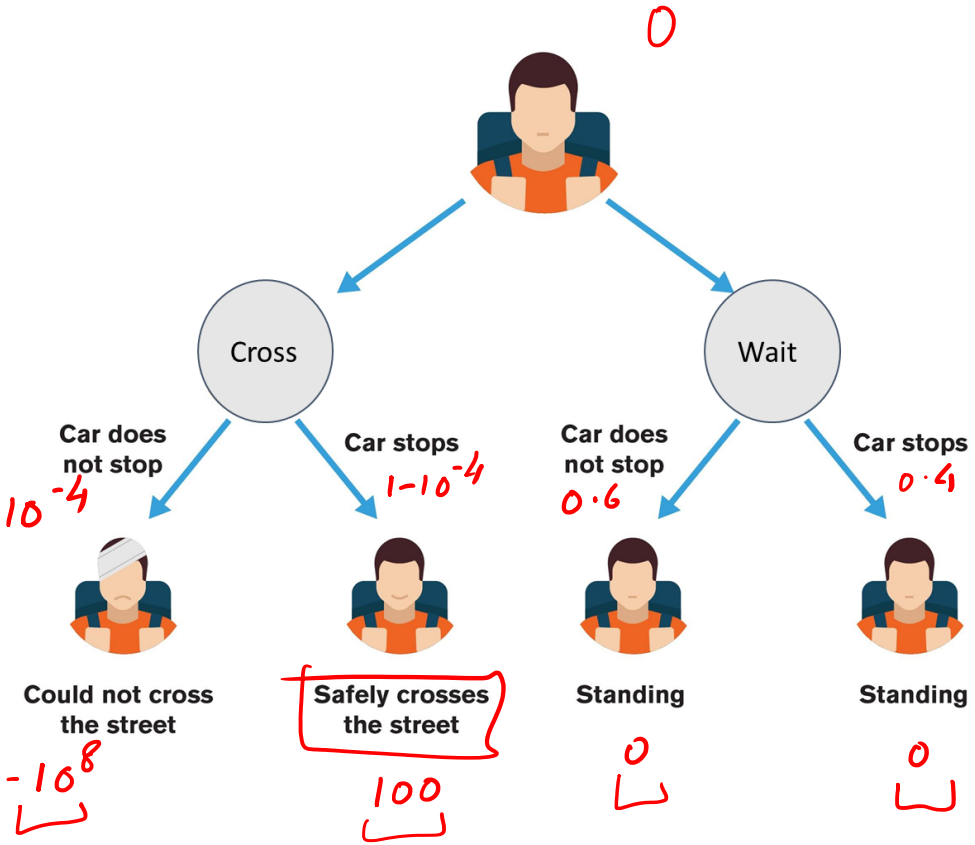
Dealing with Non-Deterministic Environments

Problem with this form of modeling:

- Cannot make trade-offs based on relative likelihoods

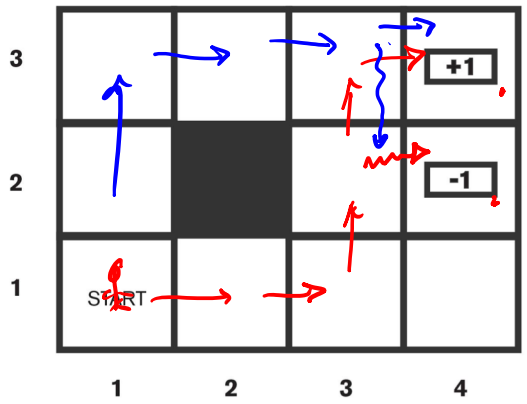
Can we do better?

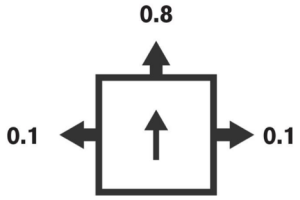
We could assign probabilities to each action outcome



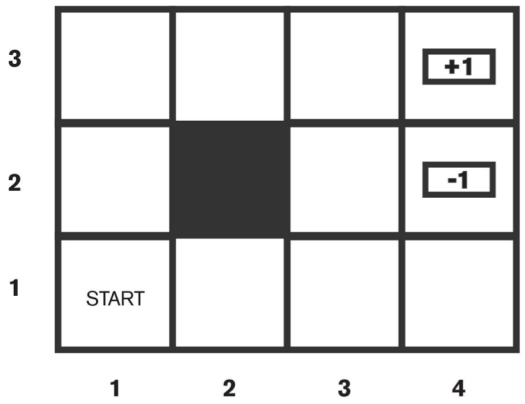
Markov Decision Processes (MDPs)

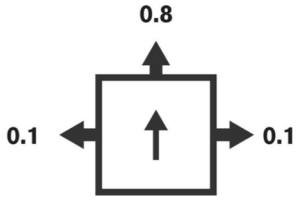
- S set of states
  - At(1,1)
- A set of actions
- P transition model
  - $P(s'|s,a)$
- $R: S \rightarrow \mathbb{R}$ reward, or utility function
- Agent can "drift", end up in unintended states
- Solutions take the form of policies: $\pi: S \rightarrow A$





Markov Decision Processes (MDPs)





Markov Assumption

- **Next state depends only on a finite portion of the history**
  - Typically, only the previous state and action
- **Can you think of situations where more than the preceding state and action are needed?**

MDPs: Policies

Policy: $S \rightarrow A$

Every state is mapped to an action

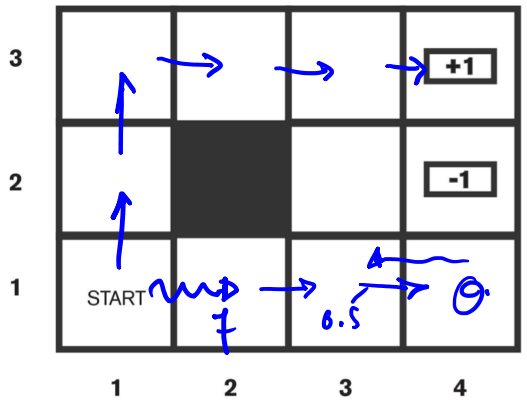
How should we evaluate a policy?

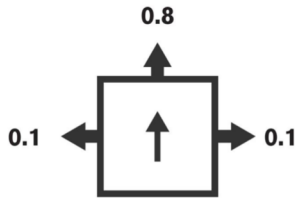
- Evaluate the executions it may produce

Need a function that determines the utility of an execution history

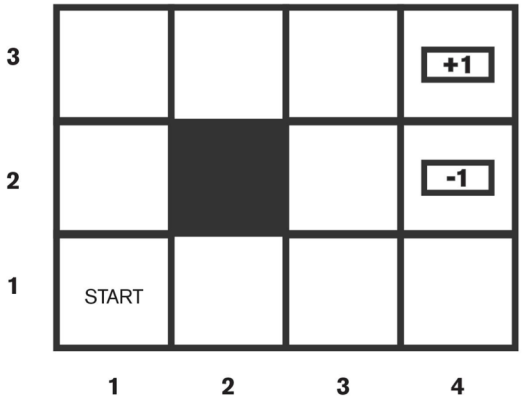
Candidates:

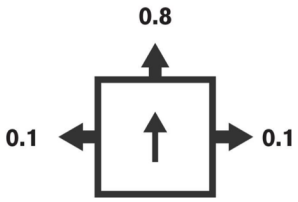
- Sum of utilities of states
- Max state-utility in the execution
- Min state-utility in the execution





Optimal is The “Best” Policy

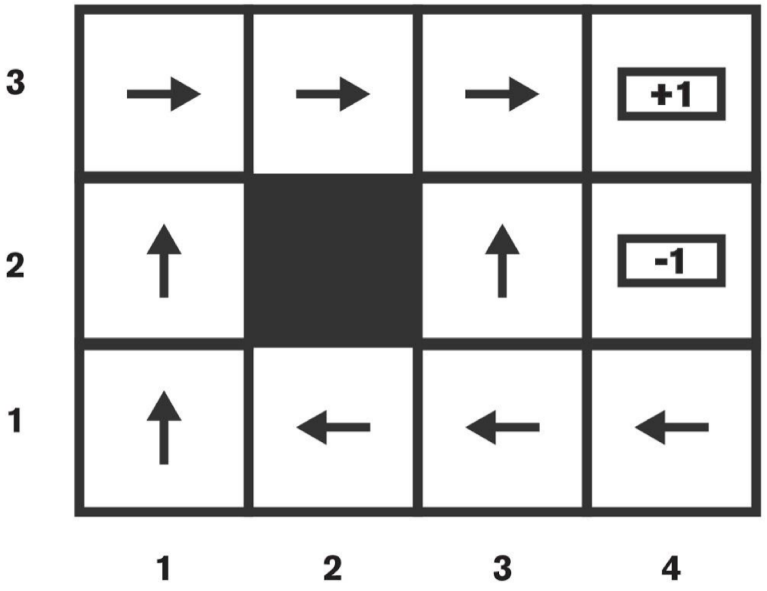


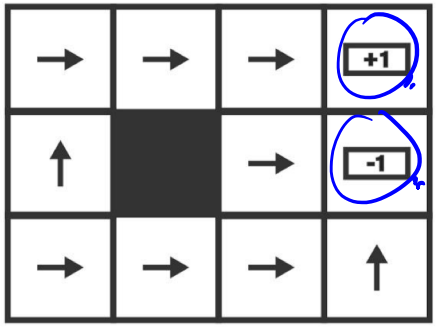


Optimal policy

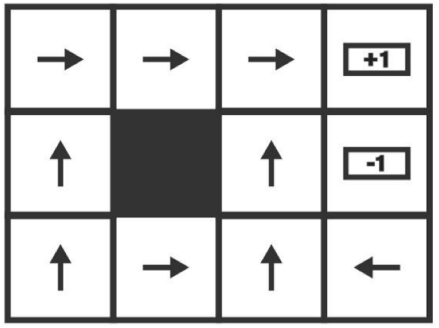
One that yields the highest **expected utility** over all **possible executions**

Optimal Policies (Utility = Sum of Rewards)

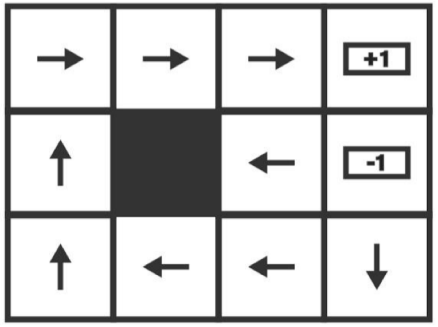




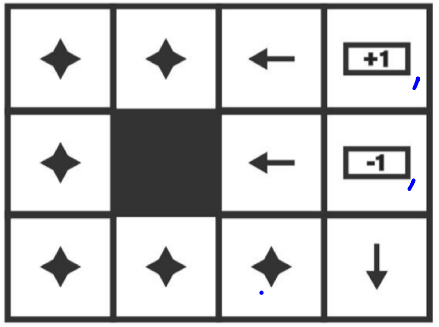
$R(s) < -1.6284$



$-0.4278\ R(s) < -0.0850$



$-0.0221\ R(s) < 0$



$R(s) > 0$

Horizon, Utilities and Policies

- **Finite horizon: agent has a fixed time N after which the execution "ends"**
  - $U([s_1, \ldots s_{N+k}]) = U([s_1, \ldots, s_N])$
- **In such cases, optimal action in a state can depend on time**
  - Optimal policy is *non-stationary*
  - E.g., optimal actions in a timed exam
- **Infinite horizon: optimal action at a state is independent of time**
  - Policies for infinite horizon problems are stationary (simpler!)

Stationary Preferences

We expect the agent's utility function to be stationary:

- Preference between state sequences that have the same prefix depend only on the non-common parts
  - $[s_0, s_1, \ldots, s_n]$ vs $[s_0, s'_1, \ldots, s'_n]$
  - $U([s_0, s_1, \ldots, s_n]) > U([s_0, s'_1, \ldots, s'_n])$ iff $U([s_1, \ldots, s_n]) > U([s'_1, \ldots, s'_n])$

The desirability of a situation is independent of how far into the future it occurs

Note: it can depend on the states!

Stationary Utility Functions

Stationary preferences lead to only two possible ways to assign utilities to sequences

- Additive rewards
  - $U([s_0, \ldots, s_n]) = R(s_0) + \cdots + R(s_n)$
  - Problem: infinite horizon may lead to infinite reward
  - How would we compare solutions?

- Discounted rewards $(0 \leq \gamma \leq 1)$ $\gamma < 1$
  - $U([s_0, \ldots, s_n]) = R(s_0) + \gamma R(s_1) \ldots + \gamma^n R(s_n)$
  - Utility of an infinite sequence (bounded rewards) is guaranteed to converge!

Solving MDPs

Expected utility of using a policy π starting in s

- Let $S_1, \ldots S_k, \ldots$ be the random variables denoting states generated at that timestep
- $U^{\pi}(s) = E[\sum_t \gamma^t R(S_t)]$

We want the optimal policy

- $\pi^*(s) = argmax_{\pi} U^{\pi}(s)$

"Give me the policy that does the best, starting from s"

Should the optimal action depend on the timestep?

Solving MDPs: Infinite Horizons

For infinite horizon problems, π^* is independent of the starting state

- i.e., the optimal action at a state doesn't depend on what you did before reaching it!

Value function, $V(s)$: utility or value of a state under the optimal policy

$$V(s) = U^{\pi^*}(s) = E[\sum_t \gamma^t R(S_t)]$$

where $S_1, S_2, \ldots$ are random variables denoting states generated while following the optimal policy, π^*

Solving MDPs: Useful Equations

Value of a state s under a policy π

- $V^{\pi}(s) =$ expected utility starting in s and following π
- $Exp_{\pi}[\sum_t R(t)]$

Q-function: value of (s, a) pair

- $Q^{\pi}(s, a) =$ expected utility when starting in s, executing a, then following π

Optimal policy: π^*

- $\pi^*(s) = \text{argmax}_{\pi} V^{\pi}(s)$

Convenient:

- Infinite horizon + discounting → π^* independent of starting state!