- Homework 3 available
- Watch for solutions
- Midterm Exam 1 week from today

RL Model

- States are individuated by perception (i.e., intrinsic features)

State = set of intrinsic features / Markov

- World is a finite exclusive and exhaustive set of states

"Finite" is now required

- World changes are state transitions

Same

- Actions may have probabilistic effects

New action ontology

- Rewards (+ or -) occur probabilistically (for us, on state arrival)

New goal ontology

- Learn how to act so as to maximize rewards

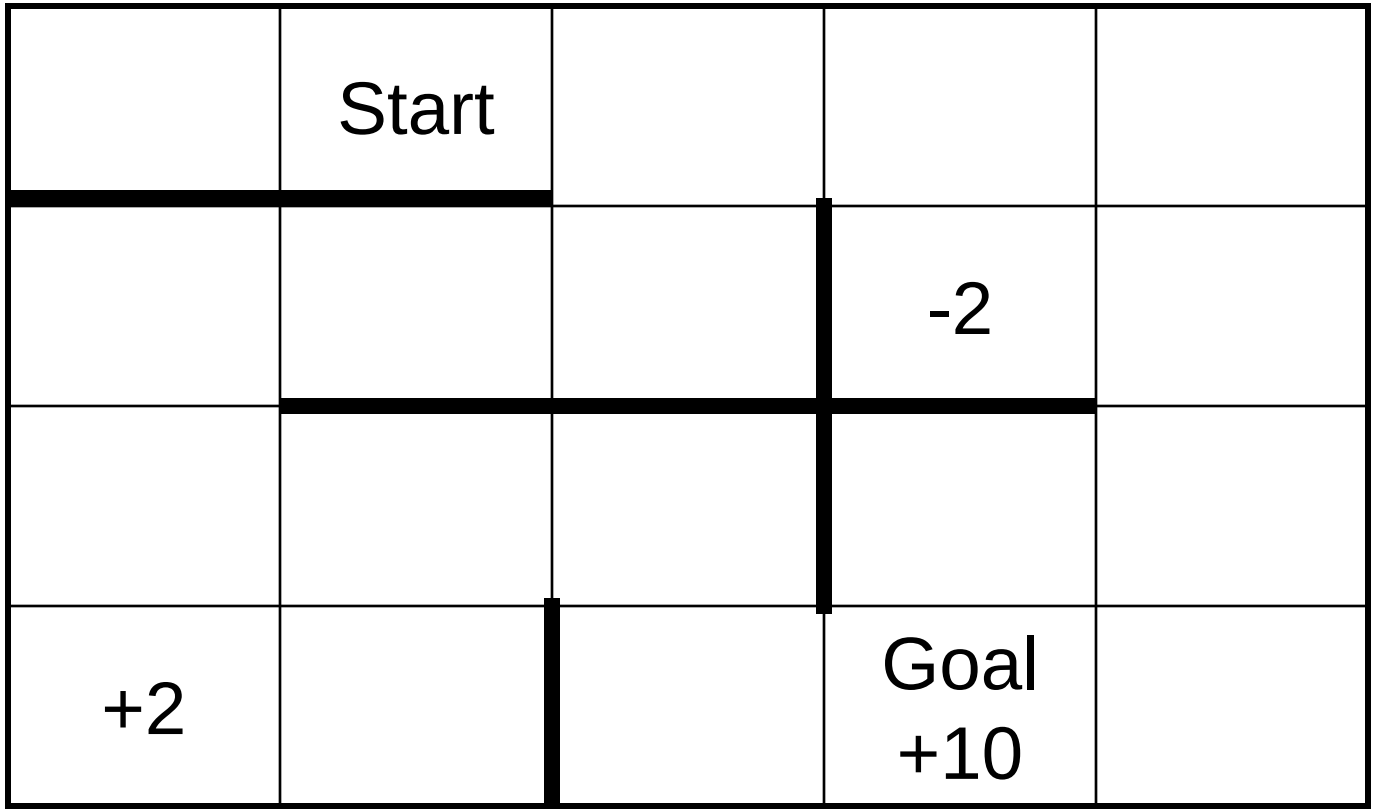
New plan ontology (more a "policy")

What is new?

Grid World

ACTIONS

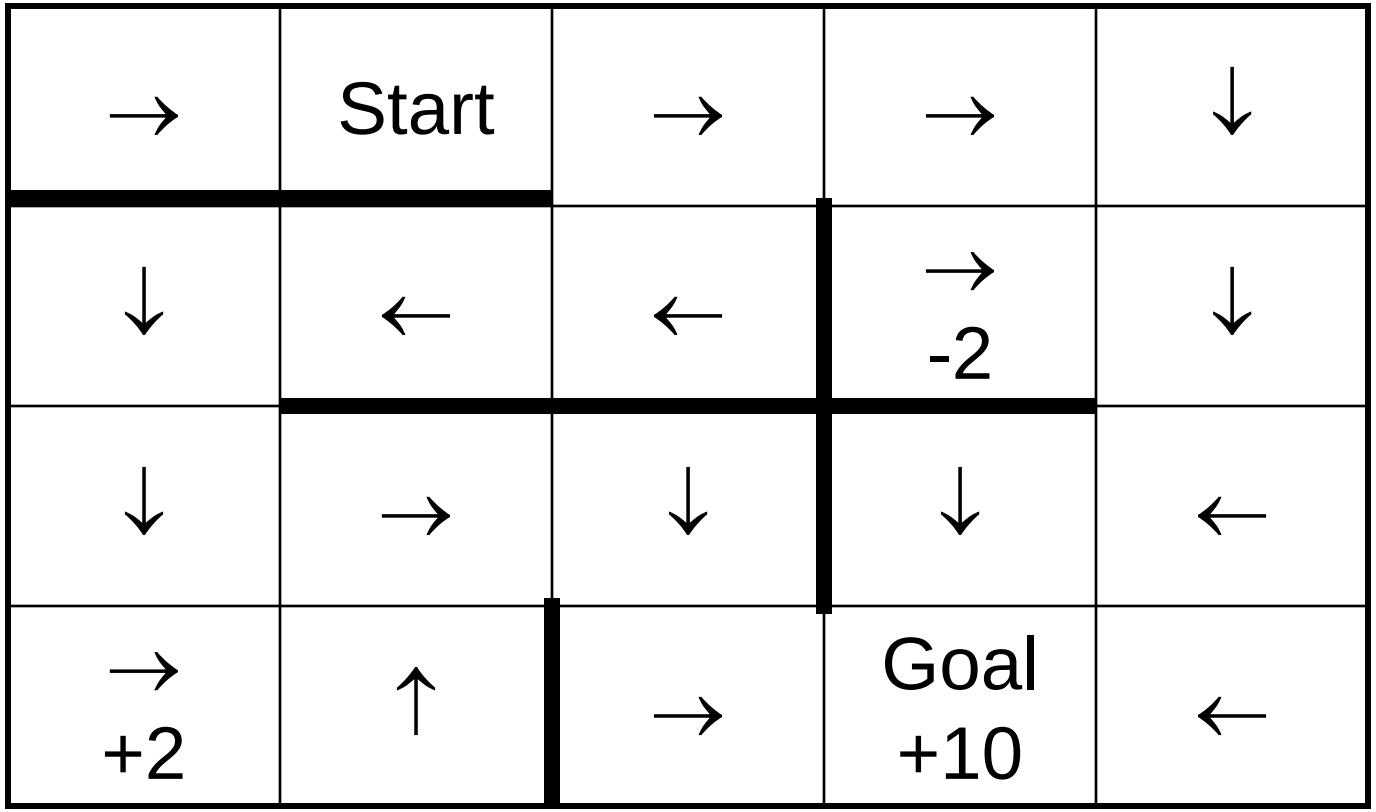
← Left
→ Right
↑ Up
↓ Down



Grid World Policy

ACTIONS

← Left
→ Right
↑ Up
↓ Down



Markov Decision Process

- Set of states $S = \{s_i\}$
- Set of actions $A = \{a_j\}$
- Initial distribution over S
  (more general than R&N)
- Transition model
- Probabilistic rewards
  (more general than R&N)
- State with no exit are “absorbing”
  (not necessary)

What is a “Distribution” ?

- More later; for now…
- Distribution over initial states
  - Whenever needed, an initial s_0 is chosen
  - Each $s \in S$ is chosen with some probability; together they sum to 1.0
- Statistical distributions also for:
  - Probabilistic rewards
  - Transition model

Rewards

(for us and R&N associated with a state)

- Rewards can be stochastic but are bounded
- Positive rewards are good
- Negative rewards are bad
- A policy that collects more rewards faster is better
- Is an immediate positive reward always desirable?
- Should an immediate negative reward always be avoided?

What is a Transition Model?

If actions were deterministic:

$T: S \times A \rightarrow S$

But actions are stochastic:

$T: S \times A \times S \rightarrow [0,1]$

A Policy chooses an action in a state

This determines a distribution over next states

The resultant state is chosen (by the world) according to the distribution

T(s,a,s')=0.2 means executing action a in state s results in state s' with a probability of 0.2

Markov Systems

- First order: current state provides all relevant information
- Second order: knowing the last two states provides all of the relevant information
- Nth order: state history of length N provides all of the relevant information
- We will assume first order Markov
  (Is this a strong assumption? What about Nth order?)

Partially Observable Markov Systems

- First order Markov: there is NO information beyond the current state
- Suppose there IS information but you cannot see it…what does that mean?
- Partially Observable Markov System
- Belief State = distribution over system states
- With more states, many more belief states
- Reinforcement Learning becomes more complex
  - MDP vs. POMDP
  - In general, intractable / undecidable / unlearnable
  - There are many special restrictions

Classical Planning Problems

Precise Logic-based World Model

- Frame Problem
- Qualification Problem
- Uncertainties
- Multiple agents
- Time & incomplete actions
- …

Brittleness of a precise analytic model

Reinforcement Learning: Find a (nearly) optimal policy given

- Set of states S = {s}
- Set of actions A = {a}
- ~~Initial distribution over S~~
- ~~Transition model~~
- ~~Reward function~~

These properties of the world are denied to the reinforcement learner

Can we still acquire the optimal policy?

Reinforcement Learning

- Underlying Markov process
- Finite set of distinguishable states
  (each with a fixed but unknown reward distribution)
- Finite set of known actions
  (with fixed but unknown probabilistic effects)
- Learn an optimal policy by experiencing the world
- Contrast with Classical Planning
- Surprising? Markov assumption(!)

What is “Optimal”?

- Expected discounted reward
- What does “expected” mean?
- What is discounting?
- Why discount?
  - Pre-theoretic intuitions of rationality
  - Immortal beings
  - Relation to action penalties (optimism under uncertainty…)

Some Important Distinctions

- Active / Passive
- ❖ On Policy / Off Policy
- ❖ Direct (model free)
  / Indirect (model based)
- ❖ Foreshadowing: first exposure to
  Generative vs. Discriminative

- ❖ (not highlighted in text)

Passive vs. Active Learning

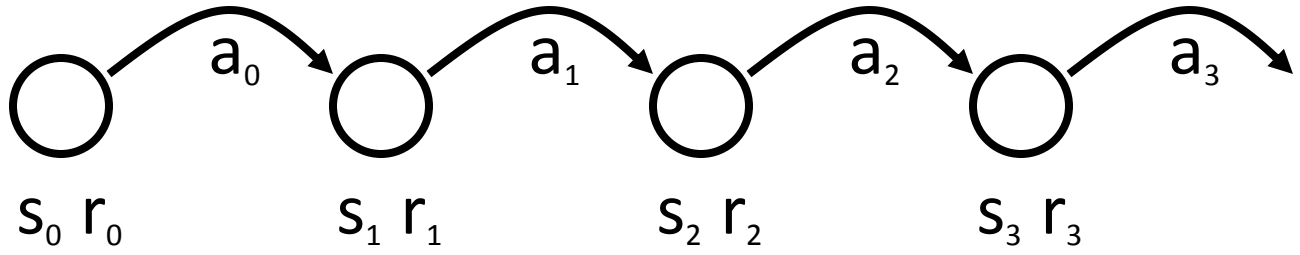
- Passive
  - Hold the policy constant (or watch someone else choose actions)
  - Learn about the world
- Active: change the policy to reflect new information (while learning)
  - Convergence complications?
  - On Policy vs. Off Policy
- Exploration: excitation of world dynamics
- Exploitation: gaining higher rewards
- Statistics – Expected value & confidence
  Central Limit Theorem???

World / Model as Functions

- Transition function
  - T: S x A x S $\rightarrow$ [0,1]
  - T(s,a,$\cdot$) denotes a probability distribution over next states
- Reward function
  - Rw: S x $\Re$ $\rightarrow$ [0,1]
  - Each Rw(s ,$\cdot$) denotes a probability distribution over rewards
- What do we care about?
- R: S $\rightarrow$ $\Re$
- R maps states to expected rewards

If we know T and R…

- We know enough to act optimally (although the algorithm is inefficient)
- We can estimate T and R from data



T(i,j,k) can be estimated as the ratio:

times action j takes us from state i to state k
divided by # times action j is tried in state i

R(i) can be estimated as the sample average reward in state i

Why Inefficient?

- Rewards are local
- Prefer a global notion of state goodness including discounted future rewards
- This will be policy dependent (why?)
- Utility of a state s given a policy π with discount γ

$$U^{\pi}(s) = E\left[\sum_{t=0}^{\infty} \gamma^{t} R(s_t)\right]$$

Given that s_0=s and we follow policy π, R&N eqn 21.1 (also 17.3)

If we knew U^{π^*} and T…

then the optimal policy is obvious.

In state s choose the action with the highest expected utility:

$$\pi^*(s) = \arg\max_a \sum_{s'} T(s,a,s') \cdot U^{\pi^*}(s')$$

This is equation 17.4 in R&N

Can we estimate U^{π} ?

Recall we can already estimate T (how?)

- Initialize U(s) arbitrarily
- Iteratively improve it:

$$U(s) \leftarrow U(s) + \alpha \cdot error$$
$$new \qquad old$$

- α is the learning rate $0 < \alpha < 1$
- What is the error?