TOWARDS Q-LEARNING

Scott O'Hara
Metrowest Developers Machine Learning Group
010/03/2018

REFERENCES

The material for this talk is primarily drawn from the slides, notes and lectures of these courses:

CS181 course at Harvard University:

- *CS181 Intelligent Machines: Perception, Learning and Uncertainty*, Sarah Finney, Spring 2009
- *CS181 Intelligent Machines: Perception, Learning and Uncertainty*, Prof. David C Brooks, Spring 2011
- *CS181 – Machine Learning*, Prof. Ryan P. Adams, Spring 2014. https://github.com/wihl/cs181-spring2014
- *CS181 – Machine Learning*, Prof. David Parkes, Spring 2017. https://harvard-ml-courses.github.io/cs181-web-2017/

University of California, Berkeley CS188:

- *CS188 – Introduction to Artificial Intelligence*, Profs. Dan Klein, Pieter Abbeel, et al. http://ai.berkeley.edu/home.html

Stanford course CS229 :

- *CS229 – Machine Learning, Andrew Ng.* https://see.stanford.edu/Course/CS229

UC BERKELEY CS188 IS A GREAT RESOURCE

- http://ai.berkeley.edu/home.html
- Covers:
  - Search
  - Constraint Satisfaction
  - Games
  - Reinforcement Learning
  - Bayesian Networks
  - Surveys Advanced Topics
  - And more…
- Contains: accessible, high quality YouTube videos, PowerPoint slides and homework.
- Series of projects based on the video game *PacMan*.
- Material is used in many courses around the country.

OVERVIEW

1. **Where We Have Been: MDPs**
   - Types of Machine Learning
   - Markov Decision Processes (MDPs)
   - 4 MDP Algorithms

2. **Where We Have Been: RL**
   - Reinforcement Learning
   - Model-based RL

3. **Q-Learning**
   - The Bellman Equations
   - States and Q-States
   - Exponential Smoothing

TYPES OF MACHINE LEARNING

There are (at least) 3 broad categories of machine learning problems:

Supervised Learning

$\boldsymbol{Data = \{(x_1, y_1), \ldots, (x_n, y_n)\}}$

e.g., linear regression, decision trees, SVMs

Unsupervised Learning

$\boldsymbol{Data = \{x_1, , \ldots, x_n\}}$

e.g., K-means, HAC, Gaussian mixture models

Reinforcement Learning

$\boldsymbol{Data = \{s_1, a_1, r_1, s_2, a_2, r_2 \ldots\}}$

an agent learns to act in an uncertain environment by training on data that are sequences of **state, action, reward.**

MARKOV DECISION PROCESSES

MARKOV DECISION PROCESSES

- Markov Decision Processes provide a mathematical framework for modeling decision making in situations where outcomes are partly random and partly under the control of a decision maker.

- The initial analysis of MDPs assume **complete knowledge** of states, actions, rewards, transitions, and discounts.

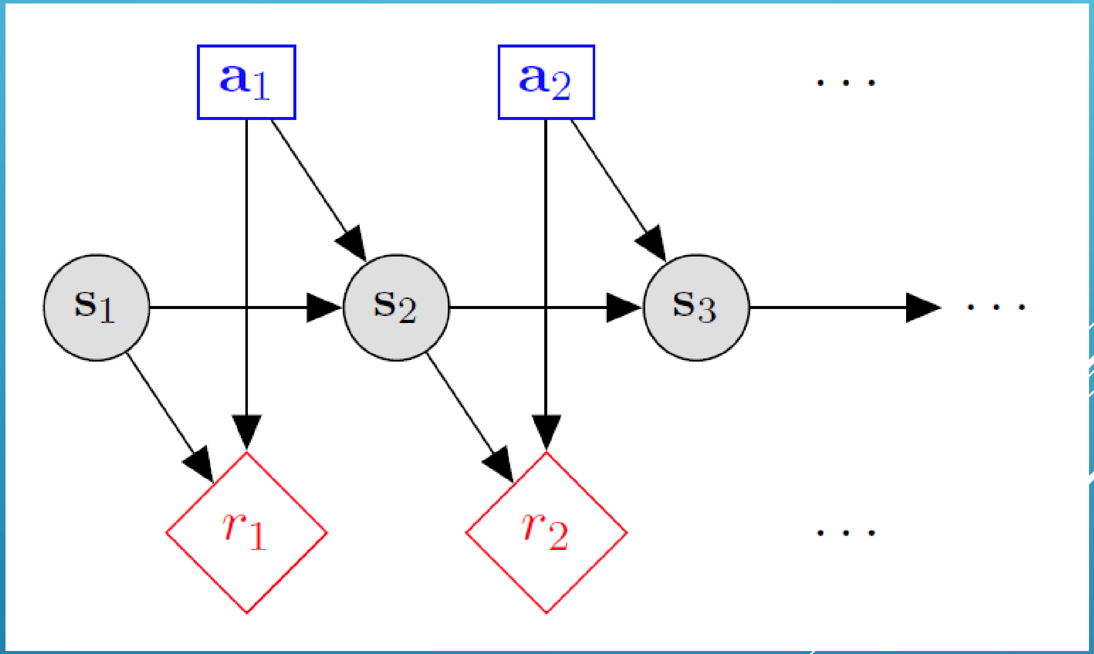
<https://en.wikipedia.org/w/index.php?title=Markov_decision_process&oldid=855934986> [accessed 11 September 2018]

MARKOV DECISION PROCESSES

- **States:** $s_1, \dots, s_n$
- **Actions:** $a_1, \dots, a_m$
- **Reward Function:**

  $r(s, a, s') \in R$
- **Transition model:**

  $T(s,a,s') = P(s'|s, a)$
- **Discount factor:** $\gamma \in [0, 1]$



<https://en.wikipedia.org/w/index.php?title=Markov_decision_process&oldid=855934986> [accessed 11 September 2018]

WHAT IS MARKOV ABOUT MDPS?

- "Markov" generally means that given the present state, the future and the past are independent
- For Markov decision processes, "Markov" means action outcomes depend only on the current state

$$P(S_{t+1} = s' | S_t = s_t, A_t = a_t, S_{t-1} = s_{t-1}, A_{t-1}, \ldots S_0 = s_0)$$

$$=$$

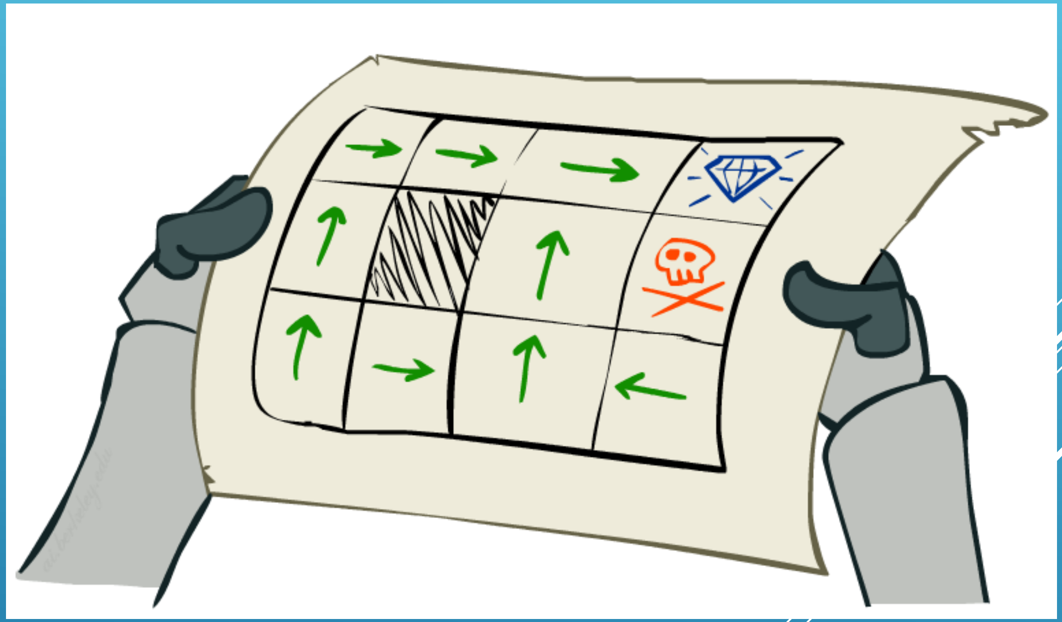
$$P(S_{t+1} = s' | S_t = s_t, A_t = a_t)$$

- This is just like search, where the successor function only depends on the current state (not the history)

Andrey Markov
(1856-1922)

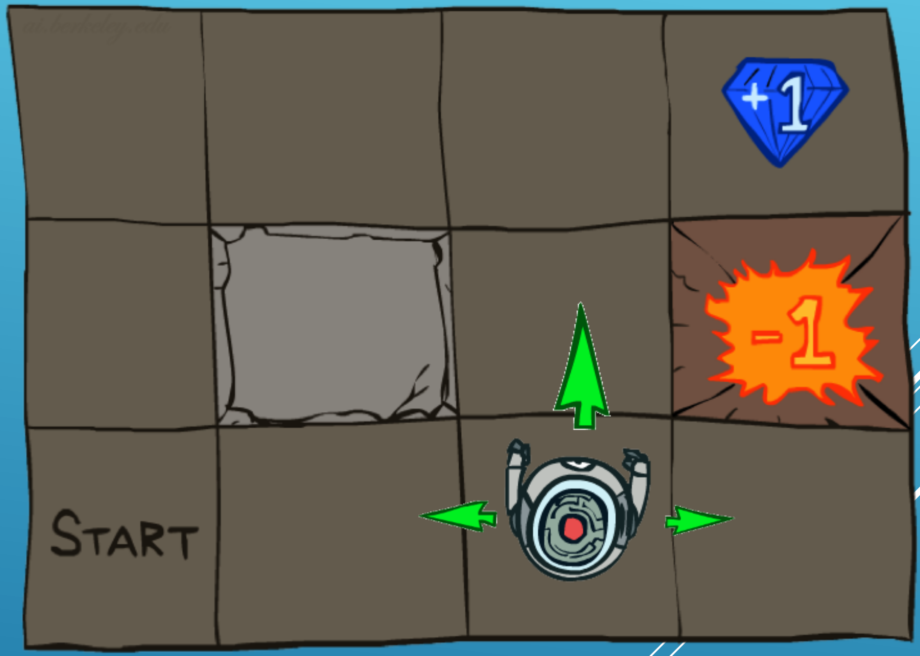
MDP GOAL: FIND AN OPTIMAL POLICY π

- In search problems, we look for an optimal plan, or sequence of actions, from start to a goal
- For MDPs, we want an optimal policy $\pi^*: S \rightarrow A$
  - A policy π gives an action for each state
  - An optimal policy is one that maximizes expected utility if followed



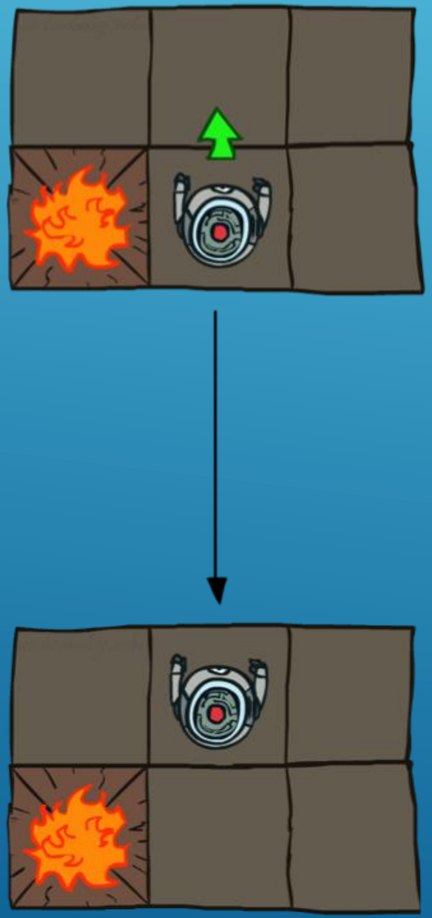
EXAMPLE: GRID WORLD

- A maze-like problem
  - The agent lives in a grid
  - Walls block the agent's path
- Noisy movement: actions do not always go as planned
  - 80% of the time, the action North takes the agent North (if there is no wall there)
  - 10% of the time, North takes the agent West; 10% East
  - If there is a wall in the direction the agent would have been taken, the agent stays put
- The agent receives rewards each time step
  - Small "living" reward each step (can be negative)
  - Big rewards come at the end (good or bad)
- Goal: maximize sum of rewards

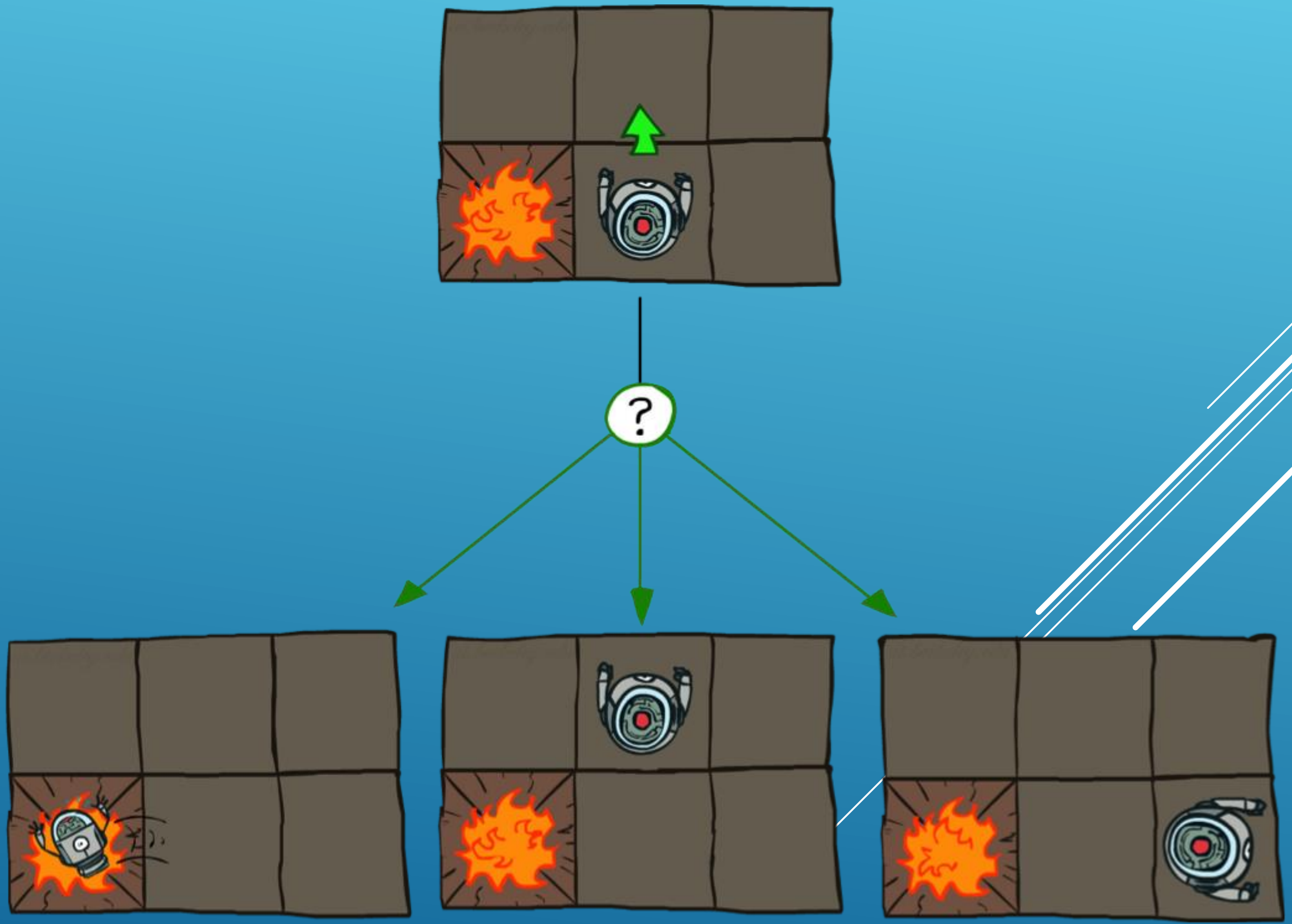


GRID WORLD ACTIONS

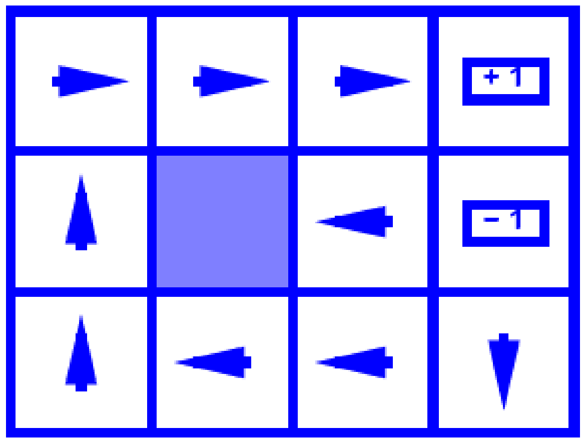
Deterministic Grid World



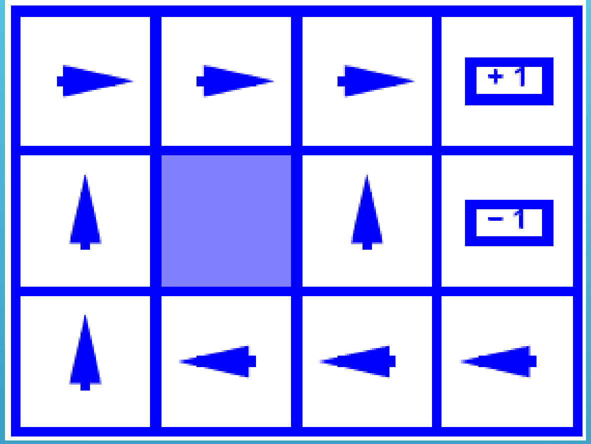
Stochastic Grid World



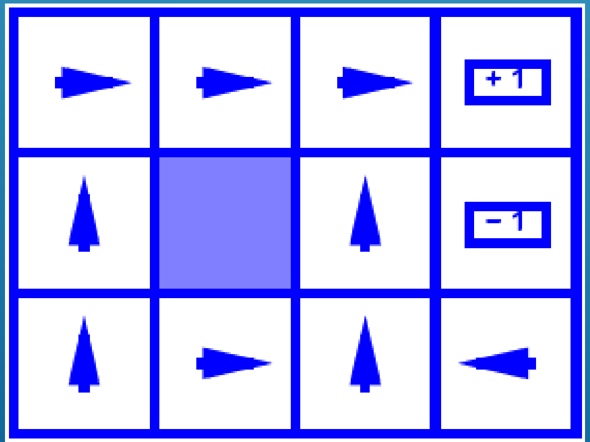
OPTIMAL POLICIES



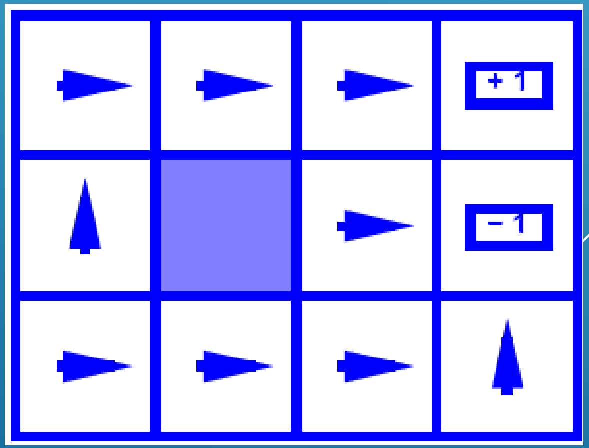
R(s) = -0.01



R(s) = -0.03



R(s) = -0.4

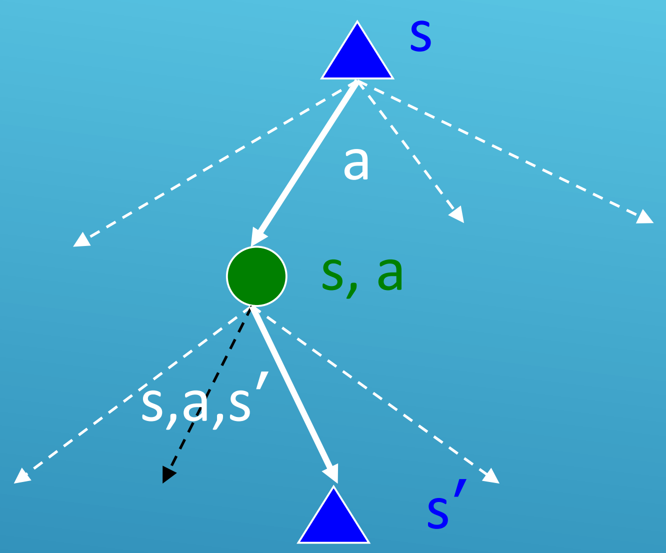


R(s) = -2.0

MDP QUANTITIES AND THE BELLMAN EQUATIONS

MDP QUANTITIES

- Markov decision processes:
  - States S
  - Actions A
  - Transitions P(s' | s,a) (or T(s,a,s'))
  - Rewards R(s,a,s') (and discount γ)
  - Start state s_0
- Quantities:
  - **Policy** = map of states to actions
  - **Utility** = sum of discounted rewards
  - **Values** = expected future utility from a state (max node)
  - **Q-Values** = expected future utility from a q-state (chance node)

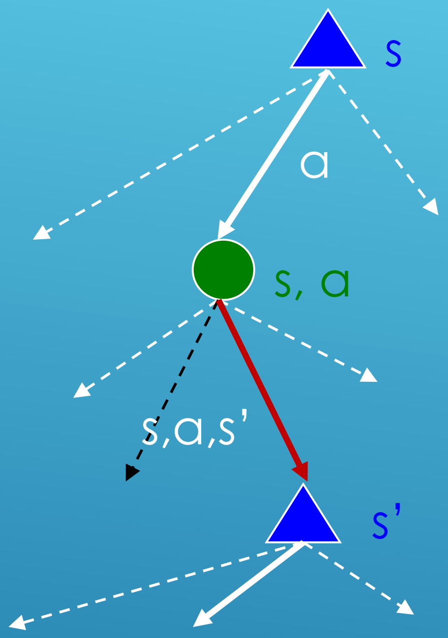


OPTIMAL QUANTITIES

- The value (utility) of a **state s**:
  $V^*(s)$ = expected utility starting in s and acting optimally

- The value (utility) of a **q-state (s,a)**:
  $Q^*(s,a)$ = expected utility starting out having taken action a from state s and (thereafter) acting optimally

- The optimal policy:
  $\pi^*(s)$ = optimal action from state s



s is a *state*

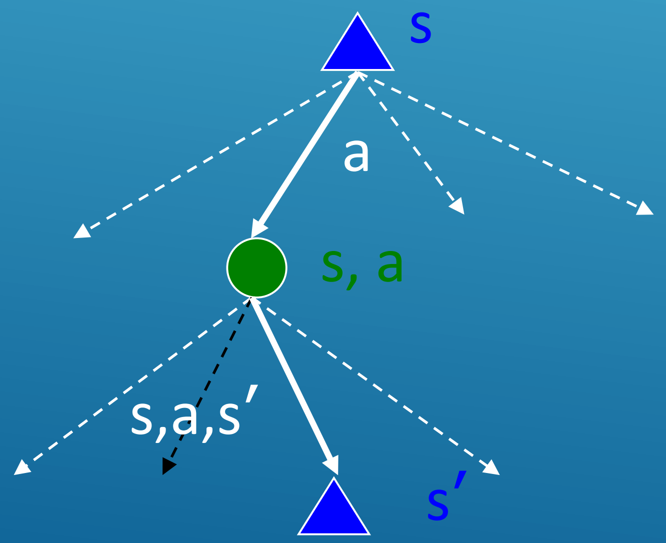
(s, a) is a *q-state*

(s,a,s') is a *transition*

[Demo: gridworld values (L9D1)]

THE BELLMAN EQUATIONS

- There is one equation $V^*(s)$ for each state s.
- There is one equation $Q^*(s,a)$ for each state s and action a.
- These are equations, not assignments. They define a relationship, which when satisfied guarantees that $V^*(s)$ and $Q^*(s,a)$ are optimal for each state and action.
- This in turn guarantees that the policy π^* is optimal.



$$V^*(s) = \max_a Q^*(s,a)$$

$$Q^*(s,a) = \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma V^*(s')\right]$$

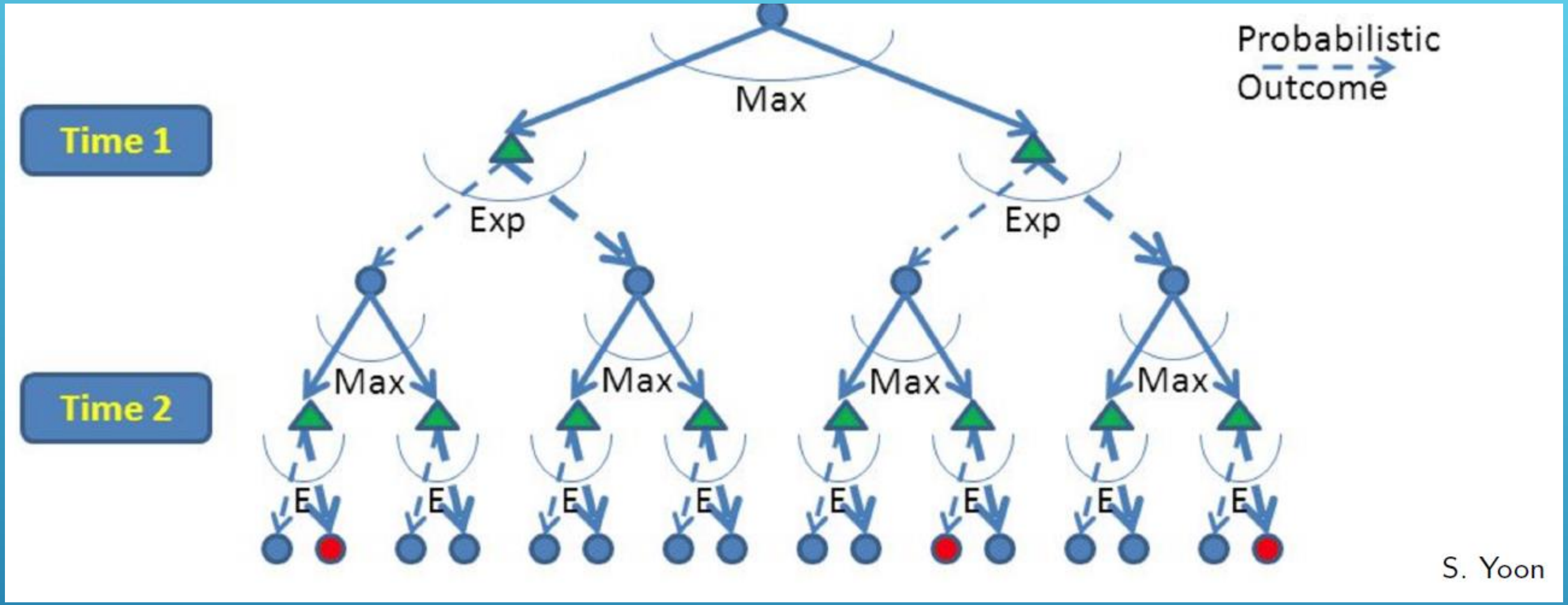
$$V^*(s) = \max_a \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma V^*(s')\right]$$

4 MDP ALGORITHMS

4 MDP ALGORITHMS

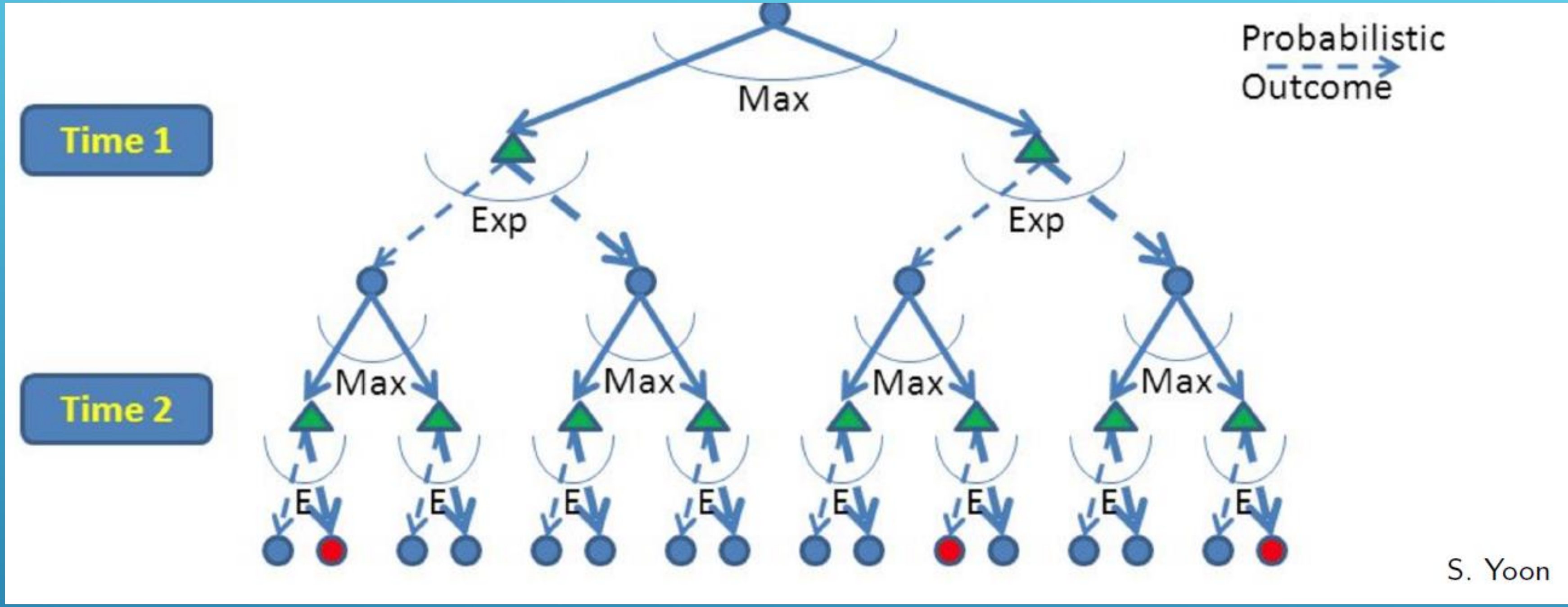
- **Expectimax** (recursive, finite horizon)
- **Value Iteration** (dynamic programming, finite horizon)
- **Value Iteration** (dynamic programming, infinite horizon)
- **Policy Iteration** (dynamic programming, infinite horizon, optimize policy)

EXPECTIMAX: TOP-DOWN, RECURSIVE



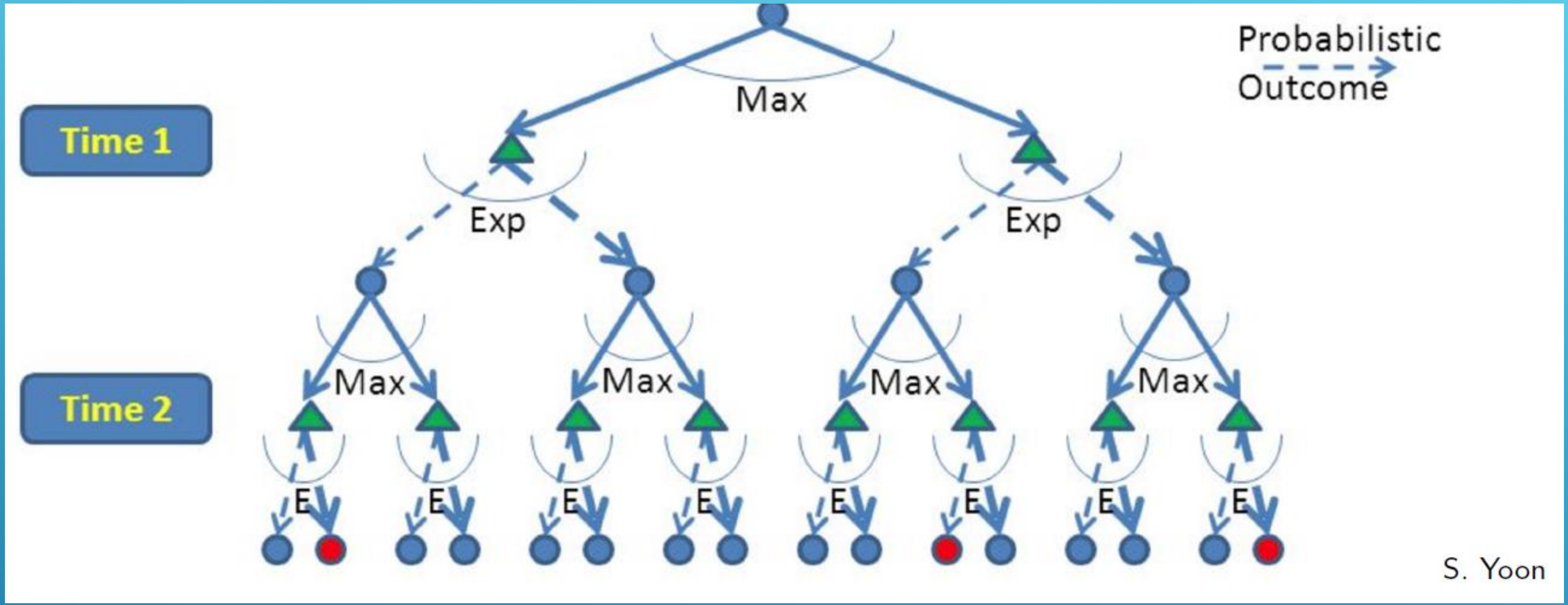
- Build out a look-ahead tree to the decision horizon; take the max over actions, expectations over next states.
- Solve from the leaves, backing-up the expectimax values.
- Finds best move for 1 state

EXPECTIMAX: A GAME AGAINST NATURE



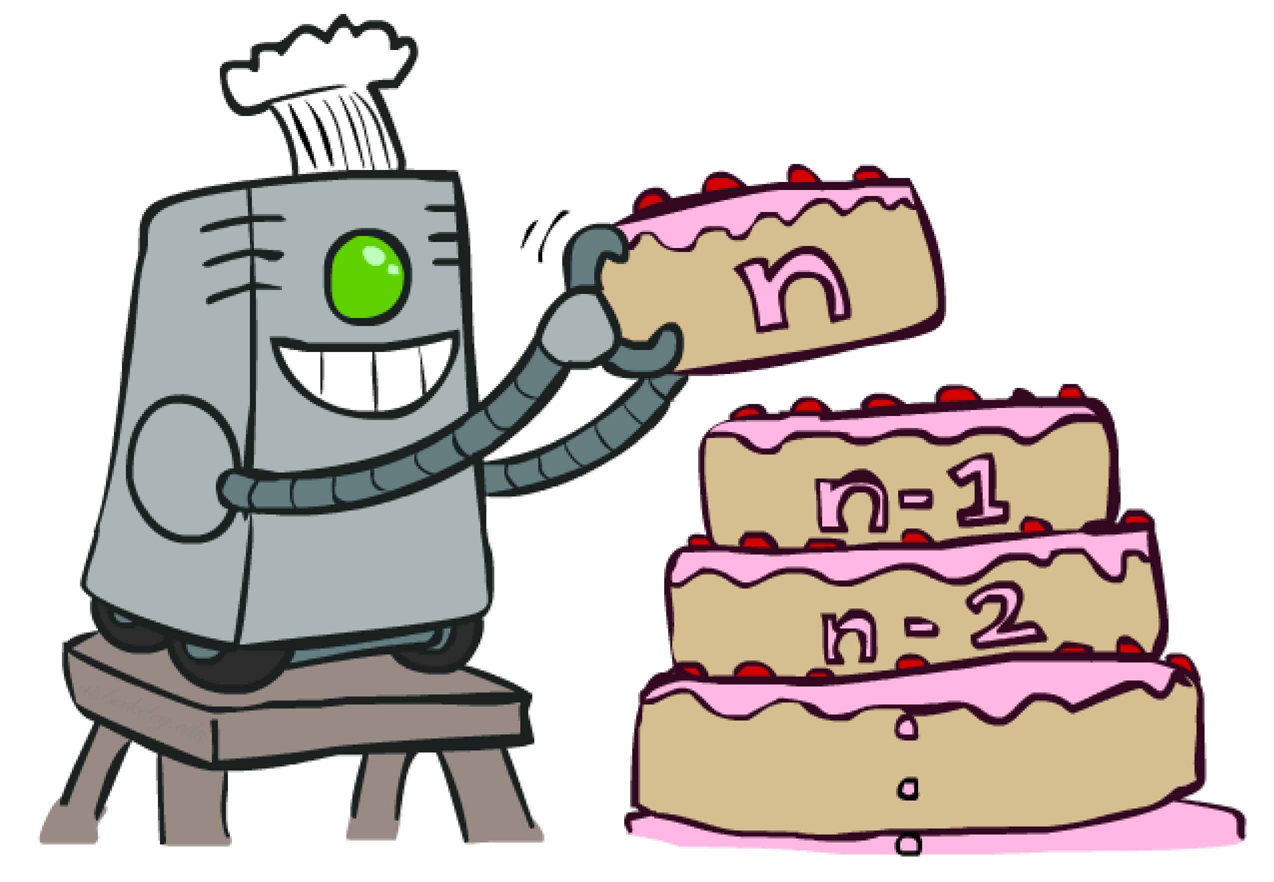
- Expectimax is like a game-playing algorithm except the opponent is nature.
- Expectimax is strongly related to the minmax algorithm used in game theory, but the response is probabilistic.
- Nodes where you move are called **states**: S (△)
- Nodes where nature moves are called **Q-states**: <S,A> (○)

EXPECTIMAX: TOP-DOWN, RECURSIVE



- Problems:
  (1) computation is exponential in the horizon
  (2) may expand the same subtree multiple times.

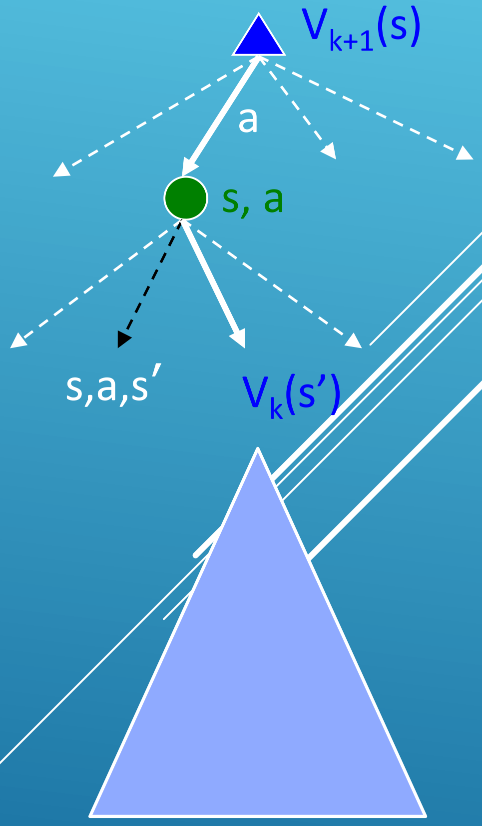
VALUE ITERATION USES DYNAMIC PROGRAMMING



VALUE ITERATION

- Start with $V_0(s) = 0$ - no time steps left means an expected reward sum of zero
- Given vector of $V_k(s)$ values, do one ply from each state:

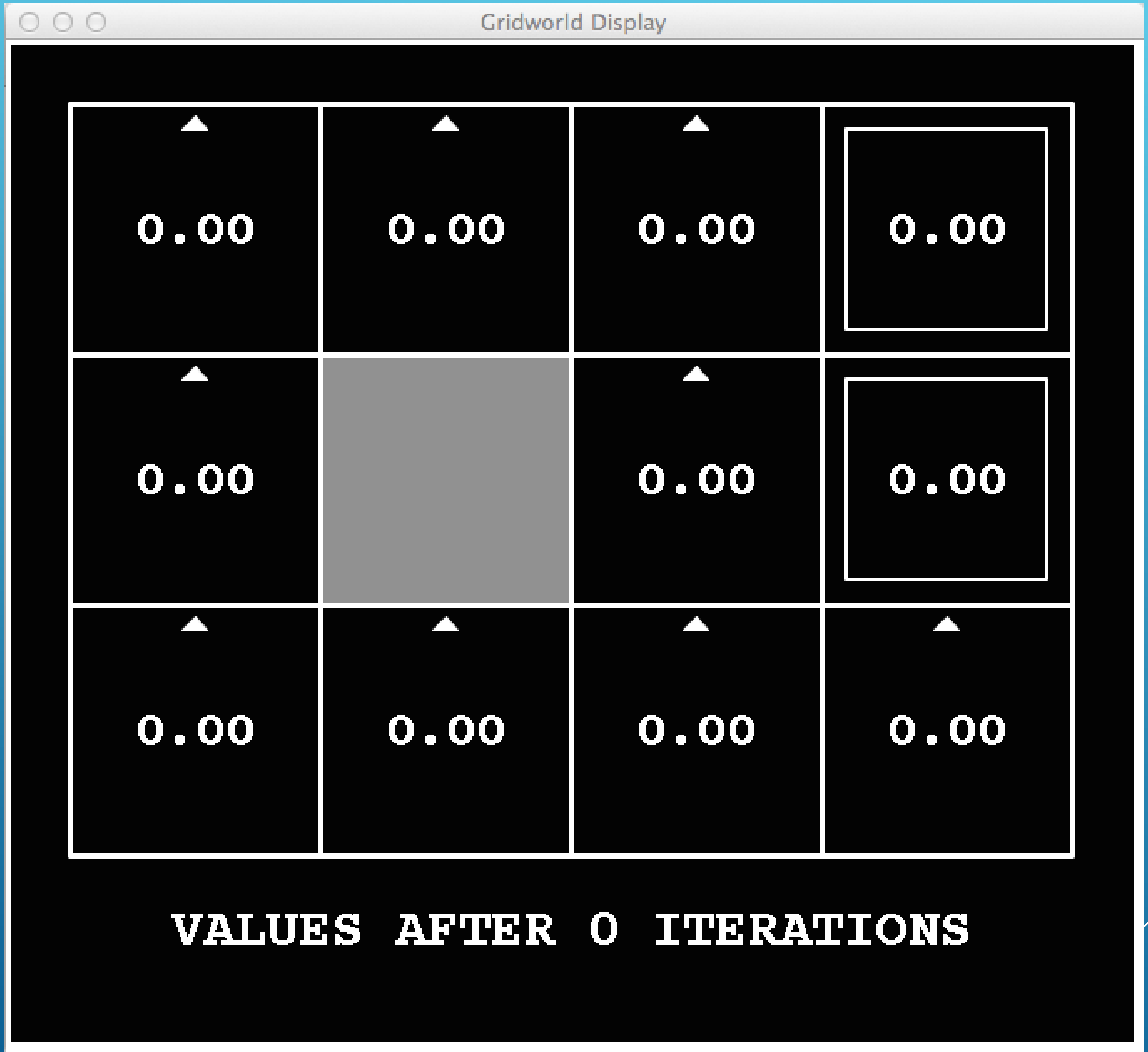
$$V_{k+1}(s) \leftarrow \max_a \left[\sum_{s'} P(s'|s,a)\left[R(s,a,s') + \gamma V_k(s')\right]\right]$$



- Repeat until convergence

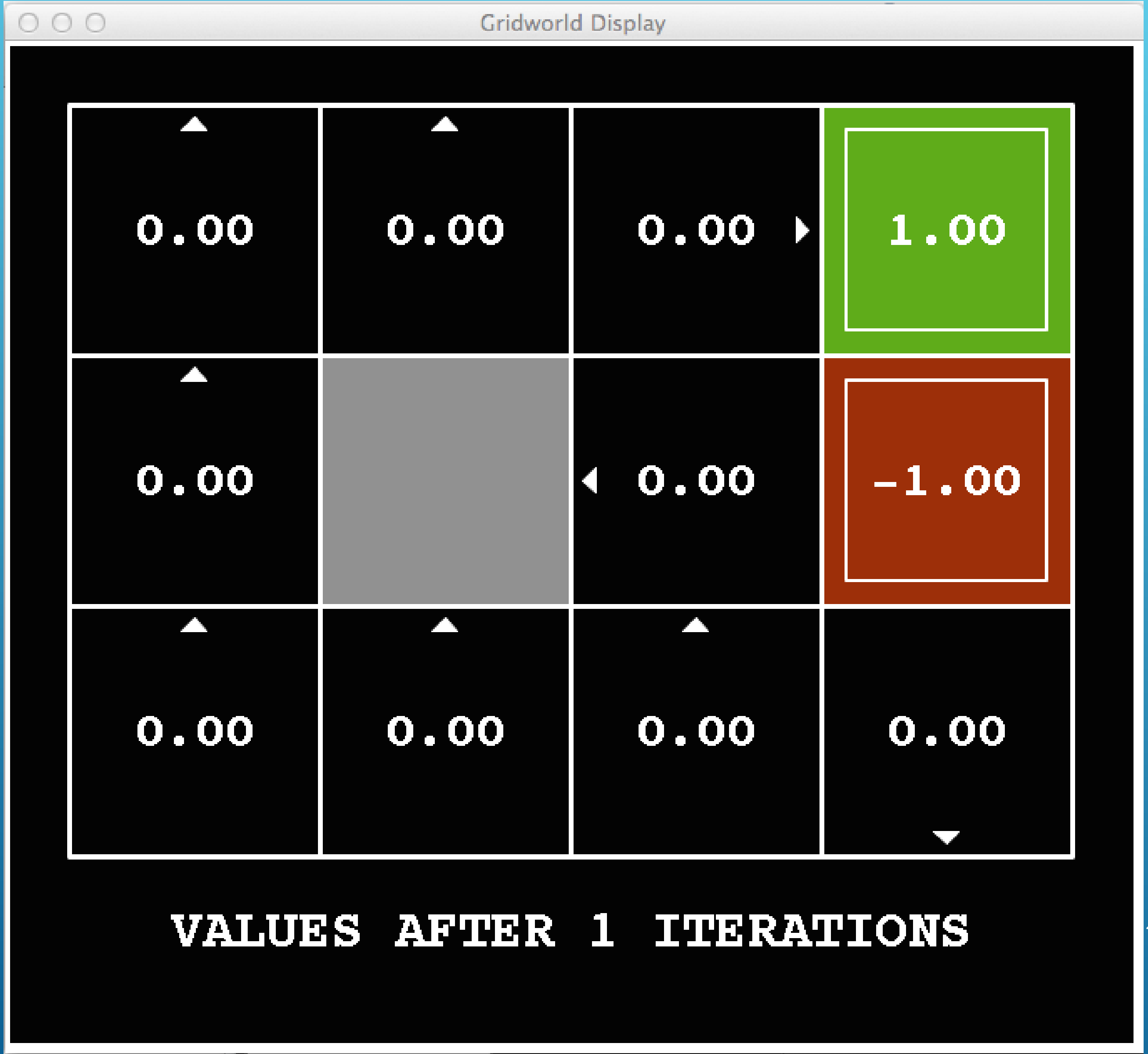
- Complexity of each iteration: $O(S^2A)$
  - For every state s, there are |A| actions
  - For every state s and action a, there are |S| possible states s'
- Theorem: will converge to unique optimal values
  - Basic idea: approximations get refined towards optimal values
  - Policy may converge long before values do

K=0



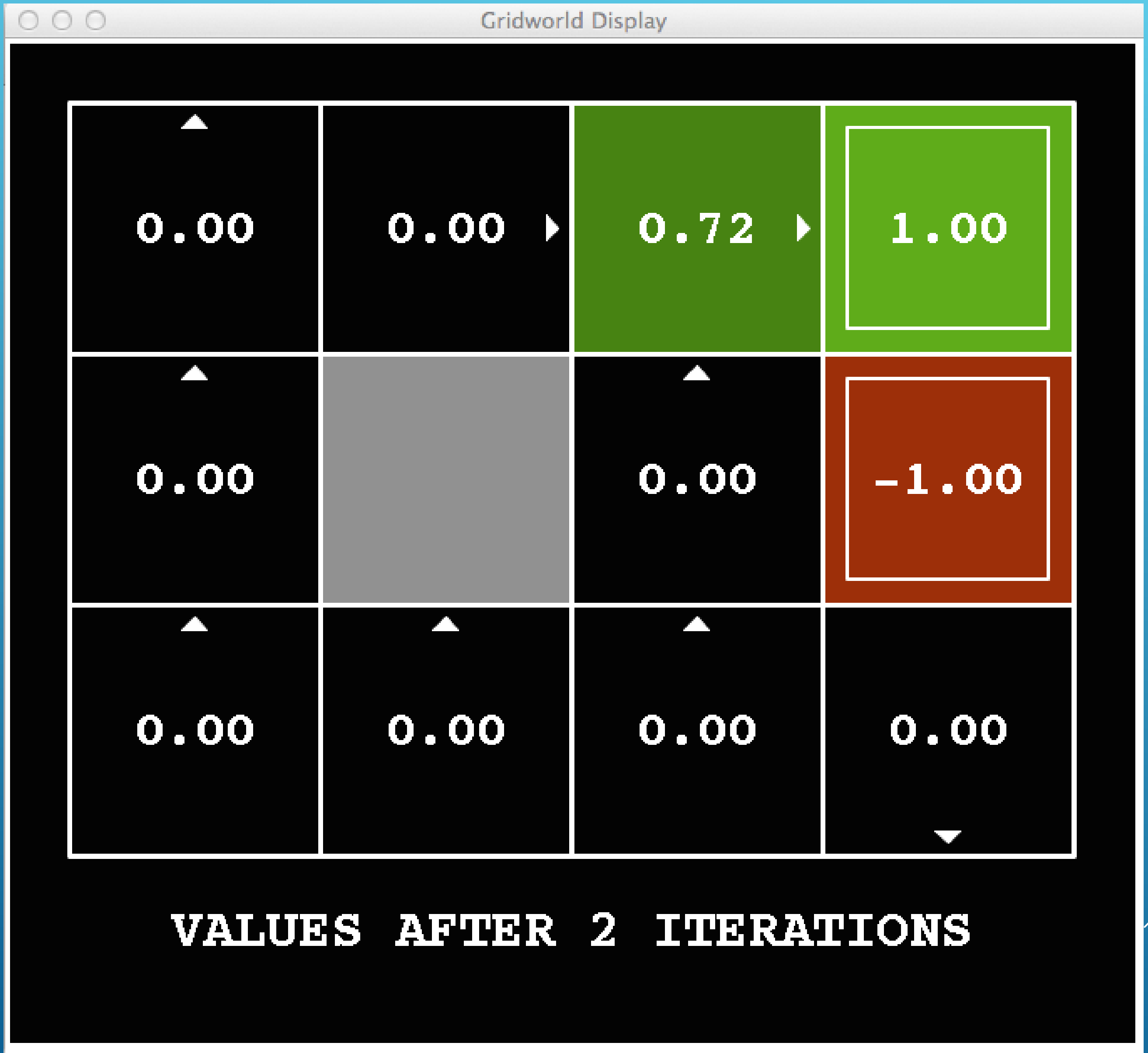
Noise = 0.2
Discount = 0.9
Living reward = 0

K=1



Noise = 0.2
Discount = 0.9
Living reward = 0

K=2

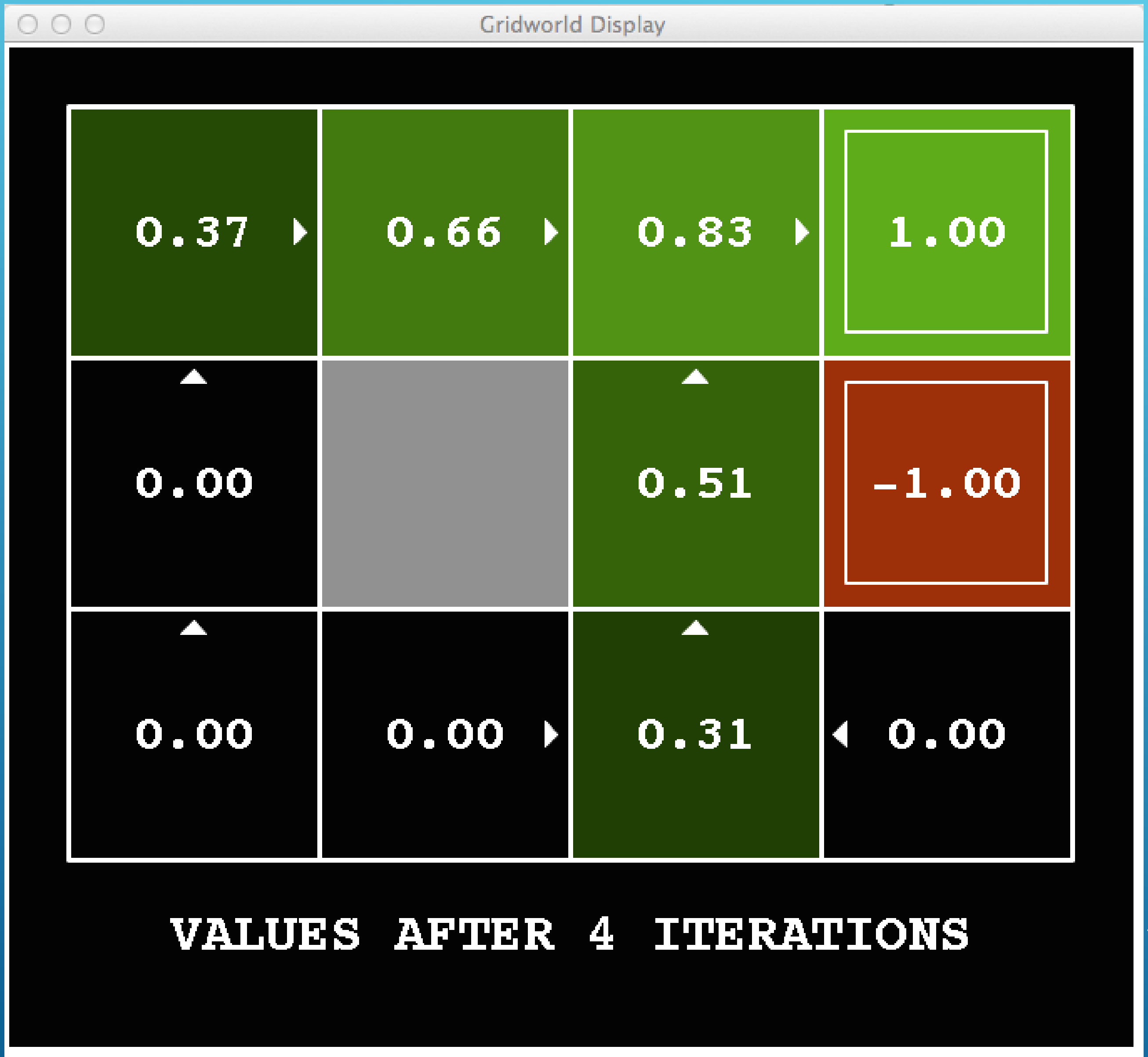


Noise = 0.2
Discount = 0.9
Living reward = 0

K=3

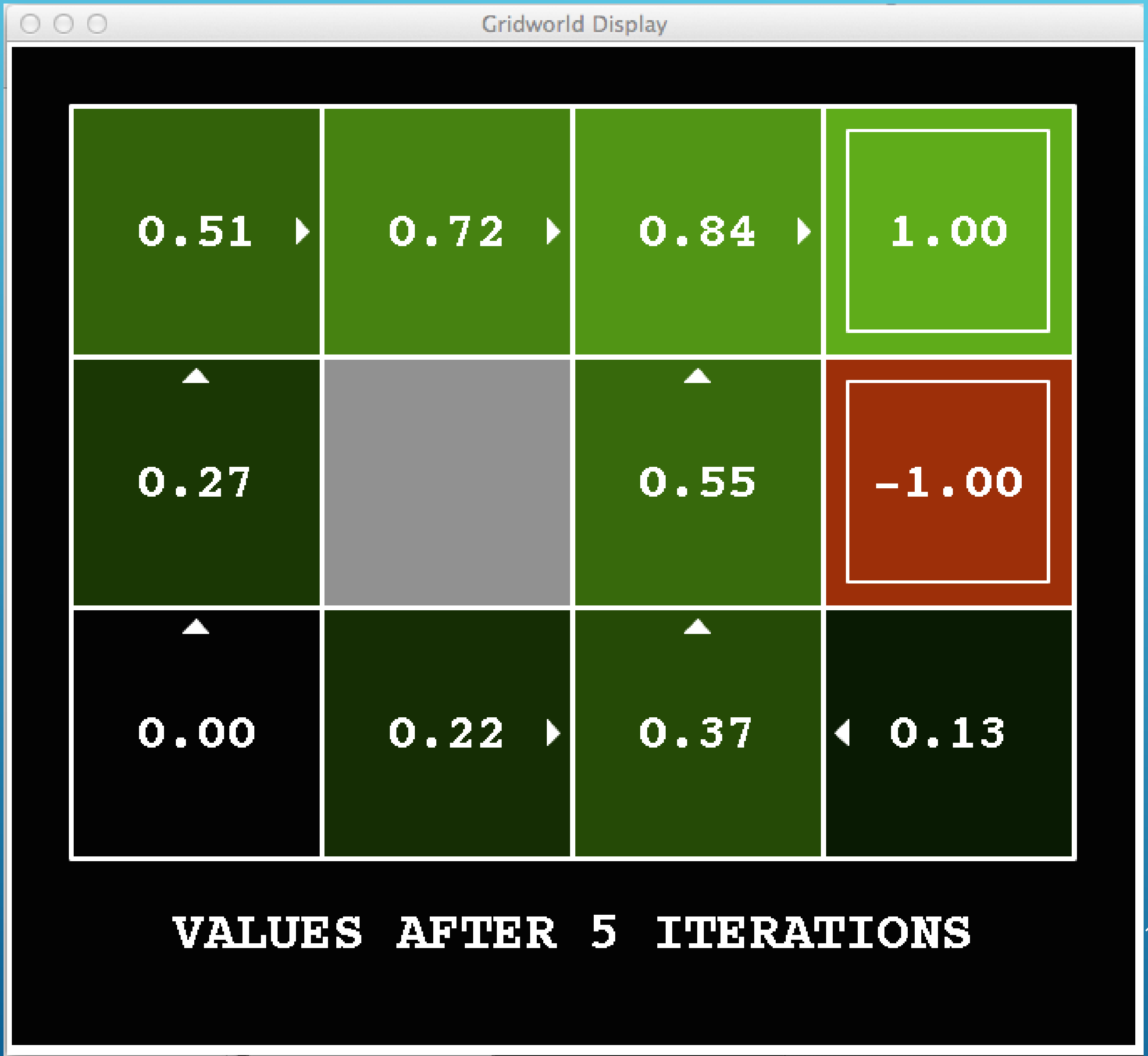
Noise = 0.2
Discount = 0.9
Living reward = 0

K=4



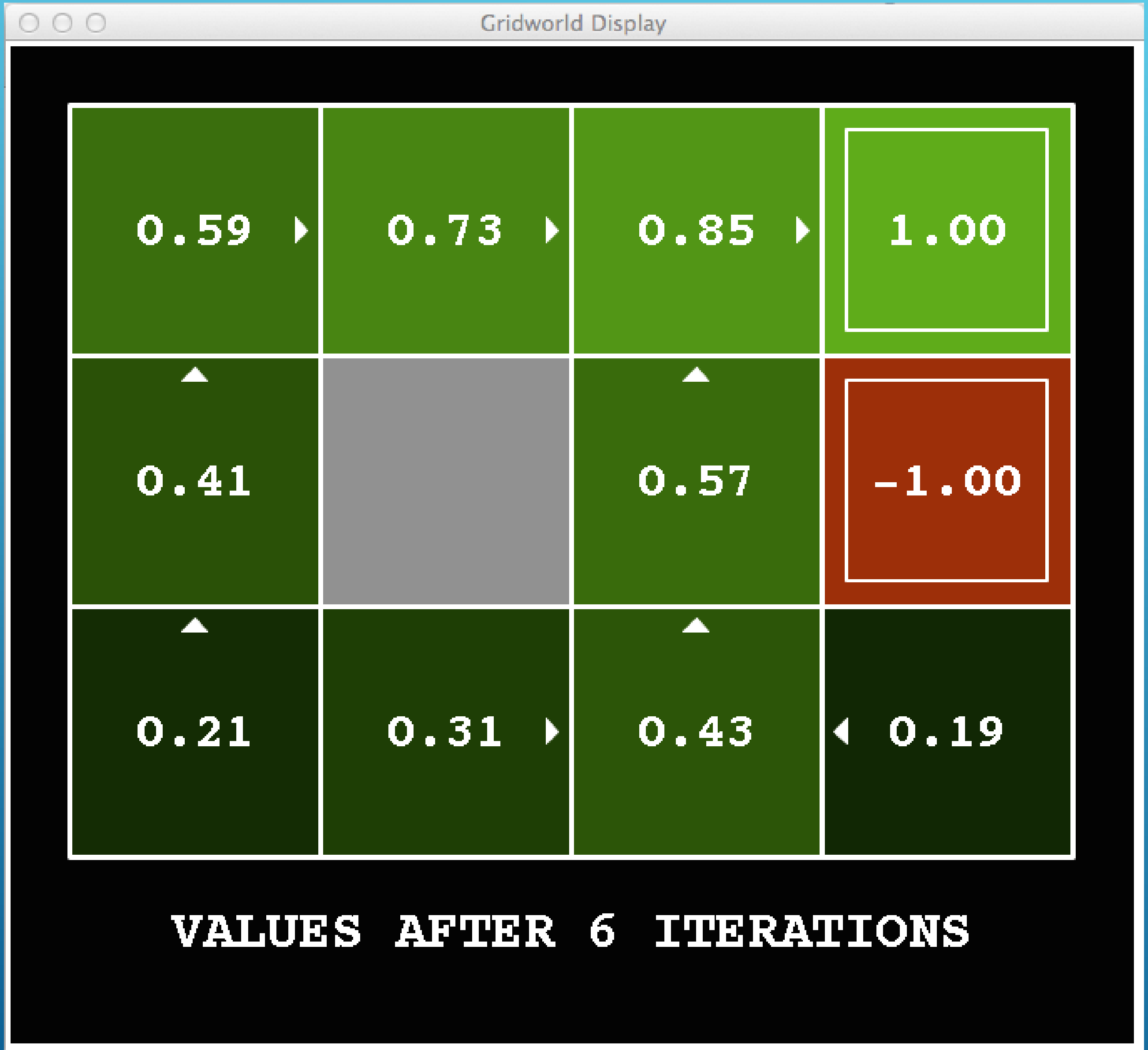
Noise = 0.2
Discount = 0.9
Living reward = 0

K=5



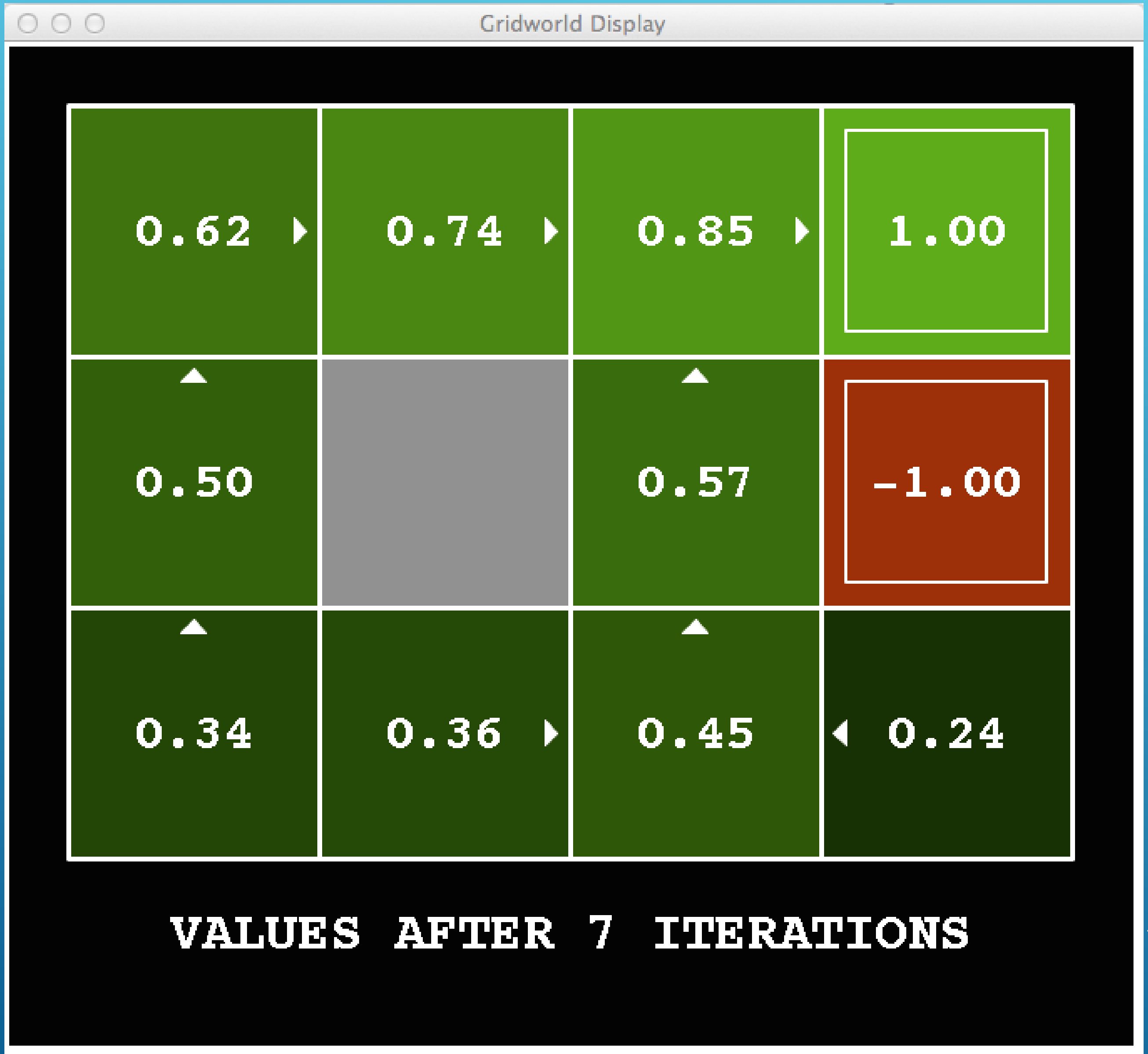
Noise = 0.2
Discount = 0.9
Living reward = 0

K=6



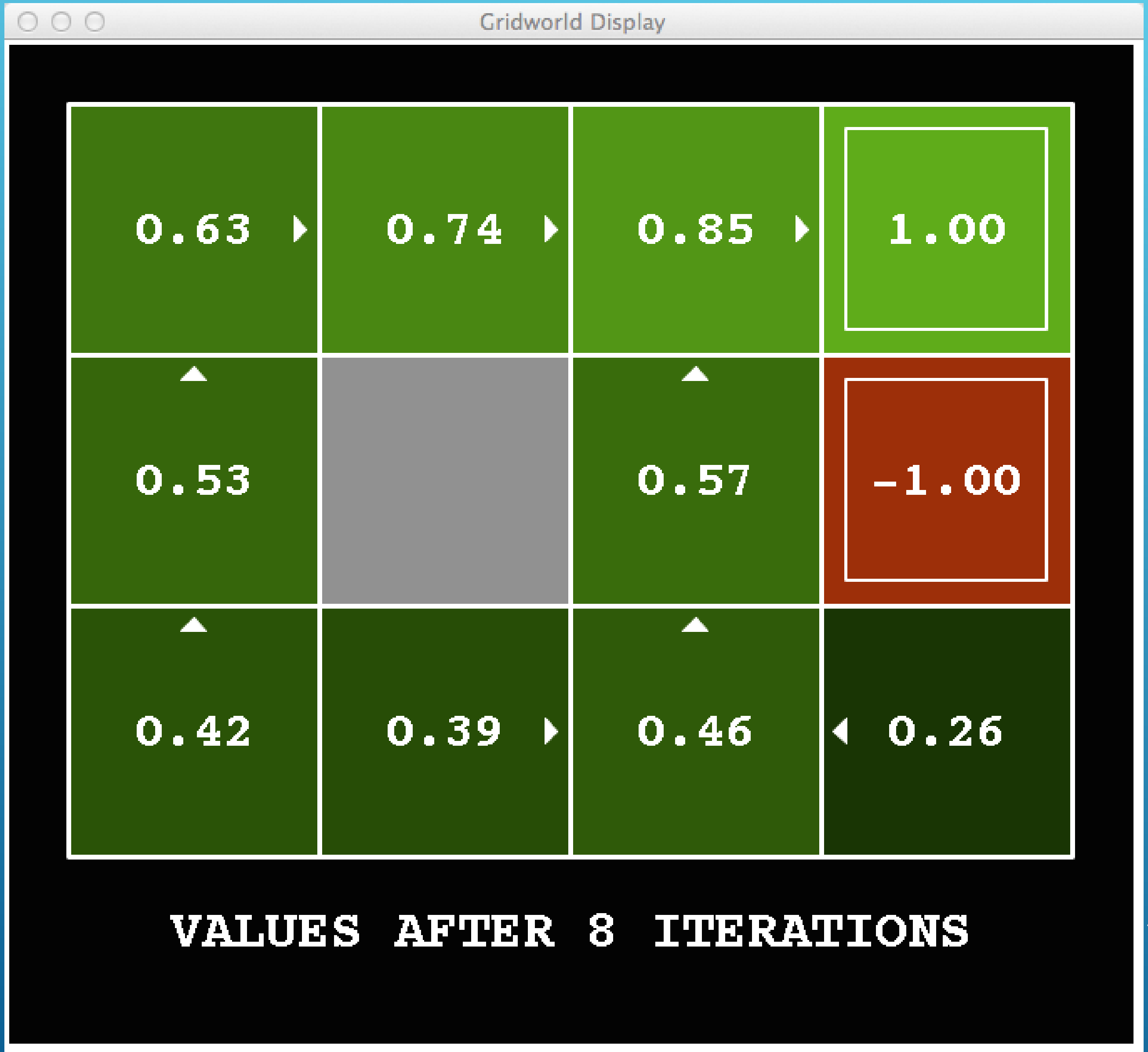
Noise = 0.2
Discount = 0.9
Living reward = 0

K=7



Noise = 0.2
Discount = 0.9
Living reward = 0

K=8

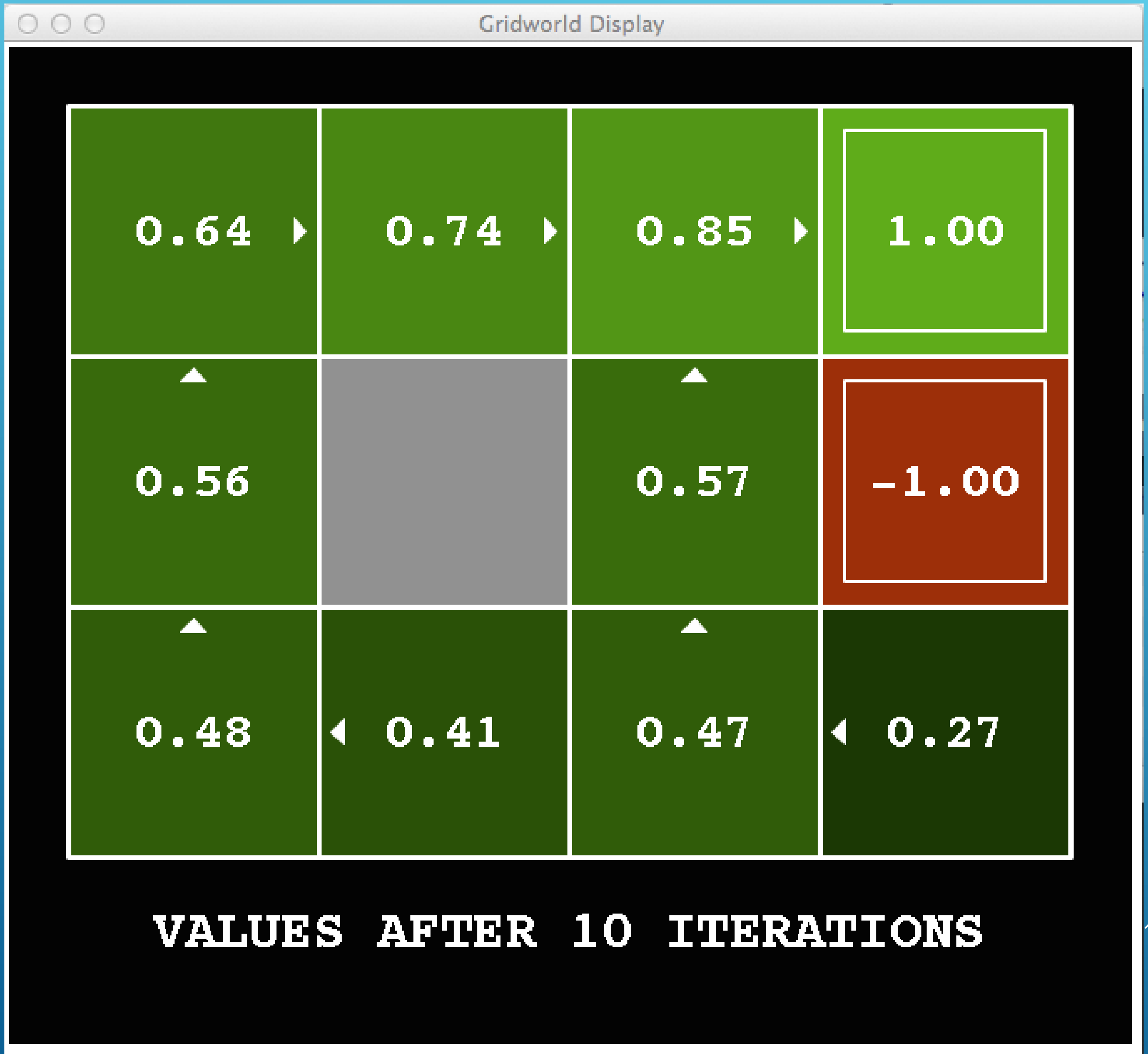


Noise = 0.2
Discount = 0.9
Living reward = 0

K=9

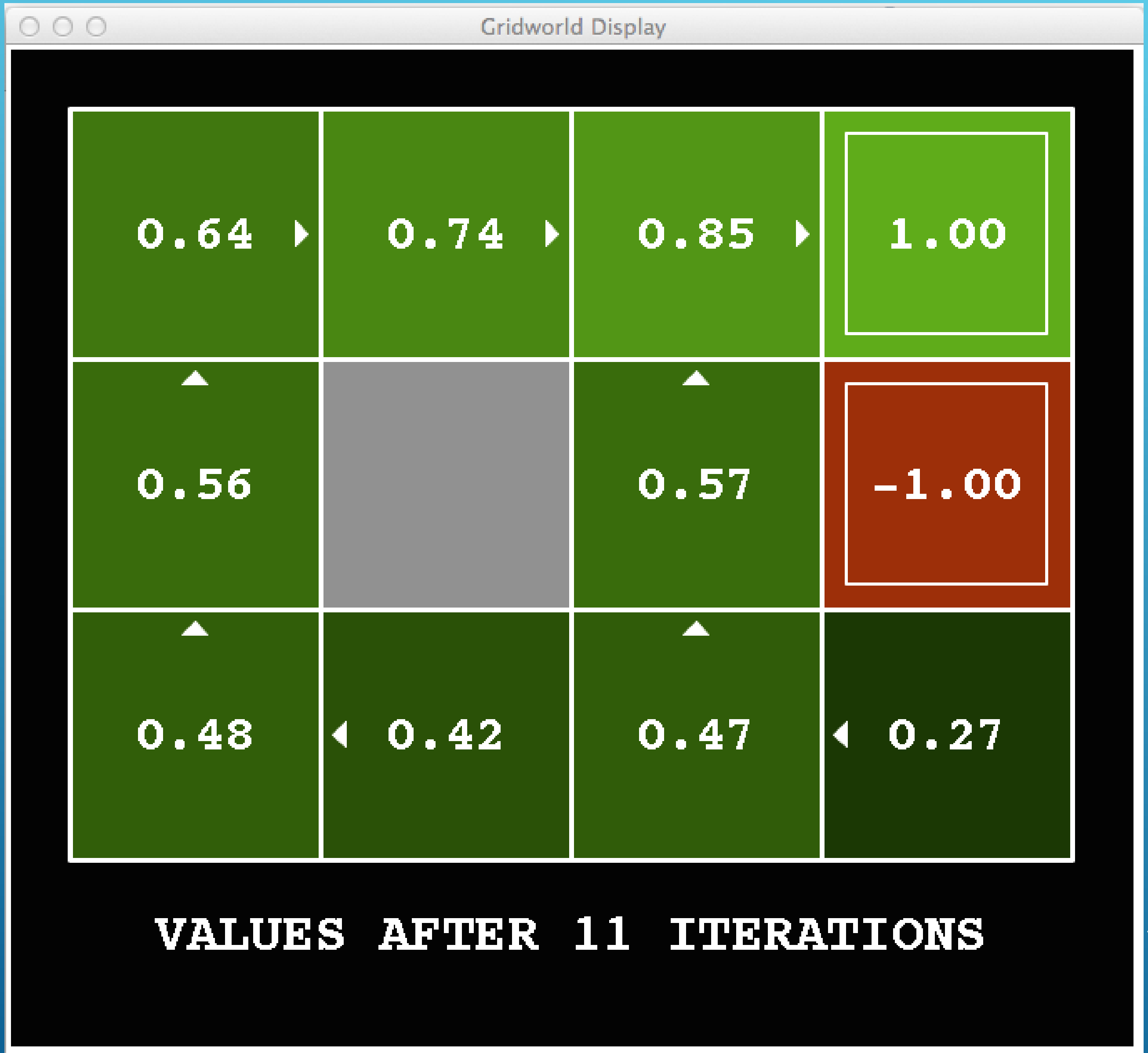
Noise = 0.2
Discount = 0.9
Living reward = 0

K=10



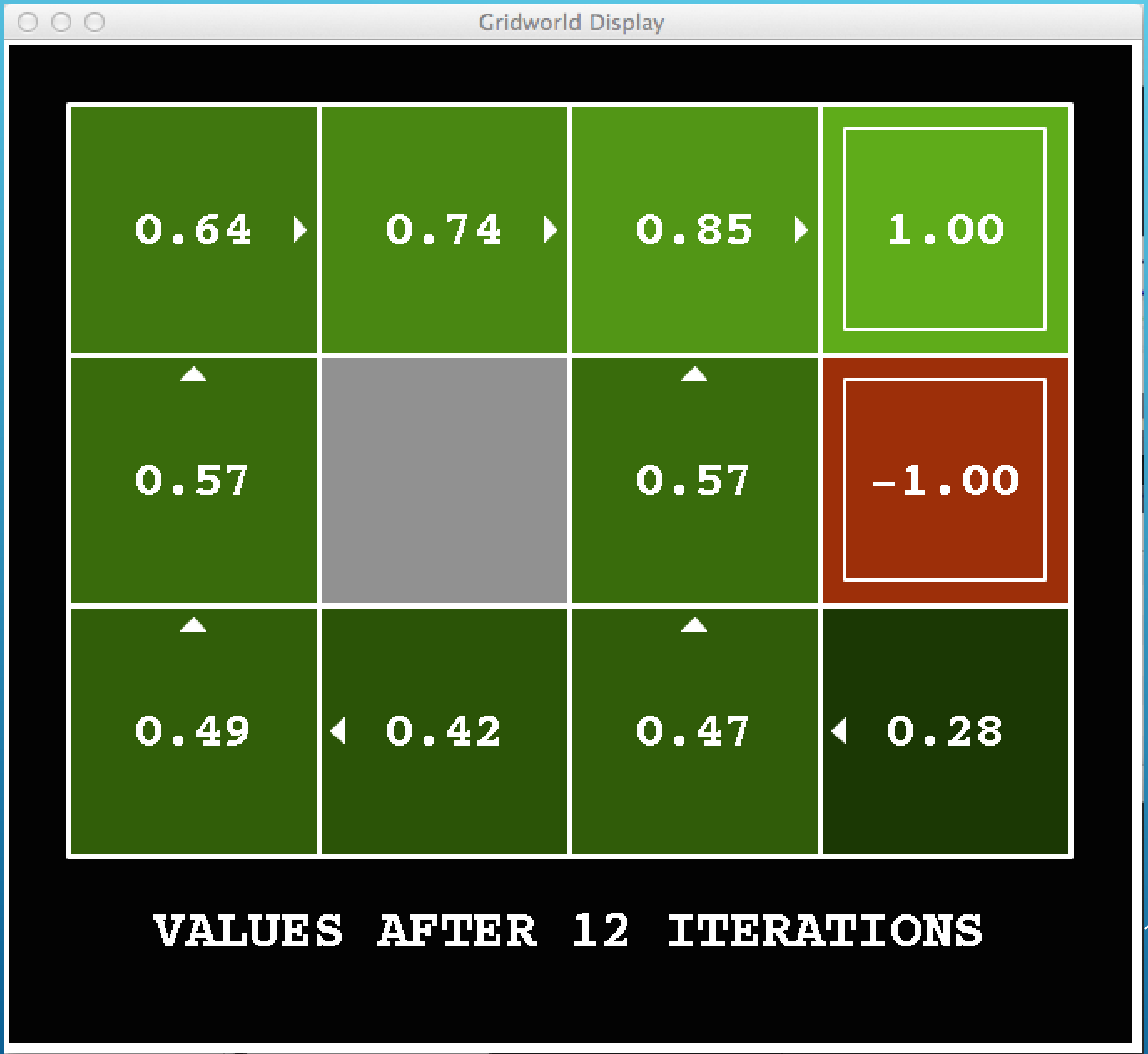
Noise = 0.2
Discount = 0.9
Living reward = 0

K=11



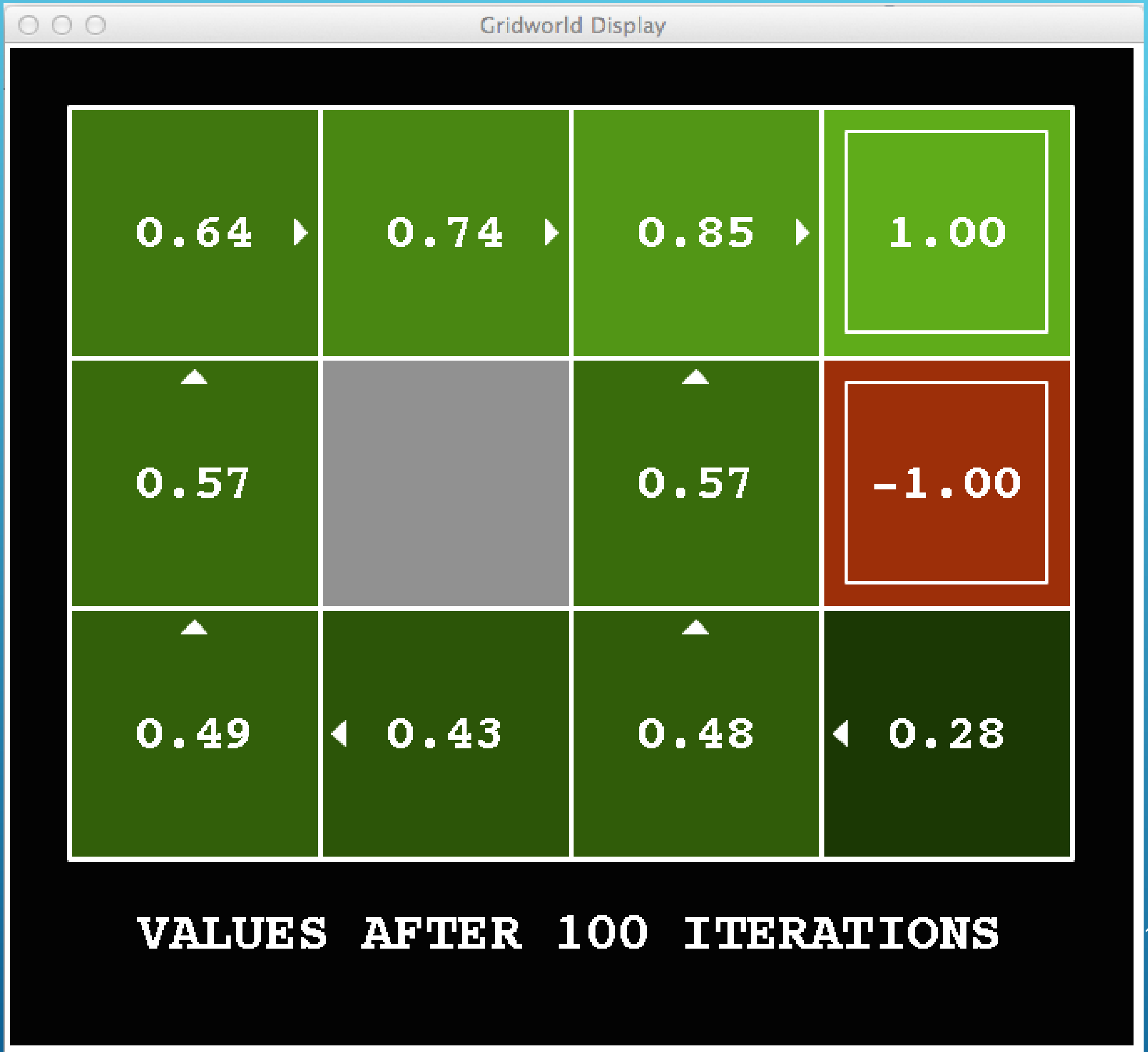
Noise = 0.2
Discount = 0.9
Living reward = 0

K=12



Noise = 0.2
Discount = 0.9
Living reward = 0

K=100



Noise = 0.2
Discount = 0.9
Living reward = 0

POLICY ITERATION

- Alternative approach for optimal values:
  - **Step 1: Policy evaluation:** calculate utilities for some fixed policy (not optimal utilities!) until convergence
  - **Step 2: Policy improvement:** update policy using one-step look-ahead with resulting converged (but not optimal!) utilities as future values
  - **Repeat** steps until policy converges
- This is **policy iteration**
  - It's still optimal!
  - Can converge (much) faster under some conditions

POLICY ITERATION

- **Step 1: Policy Evaluation**: For fixed current policy π, find values with policy evaluation:
  - Iterate until values converge:

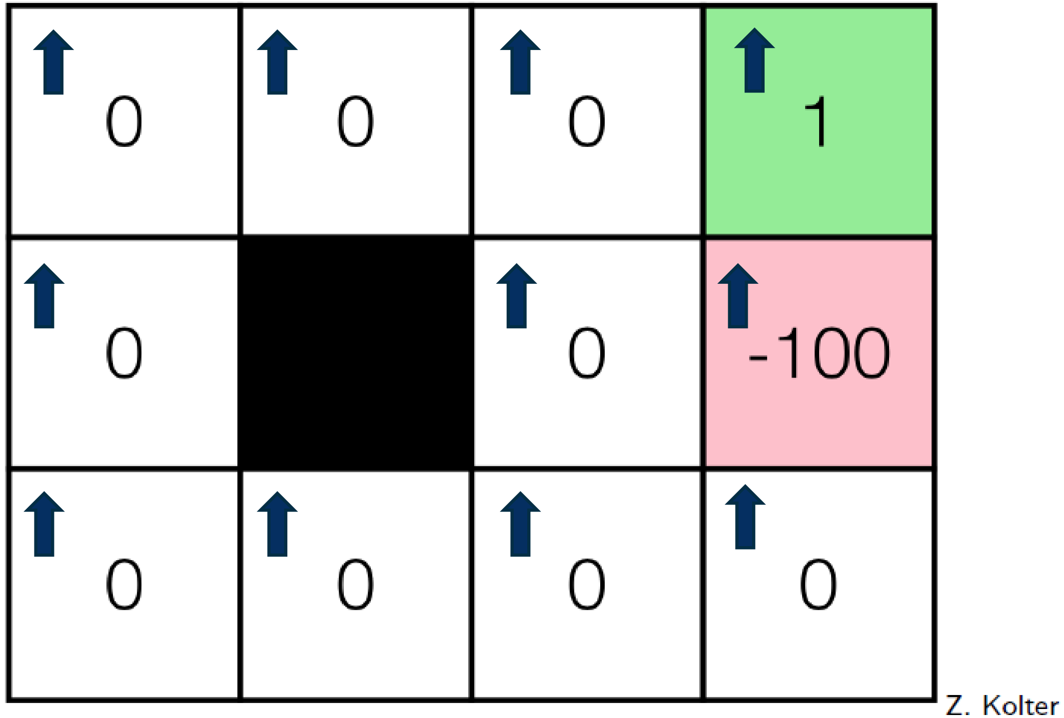
$$V_{k+1}^{\pi_i}(s) \leftarrow \sum_{s'} T(s, \pi_i(s), s') \left[R(s, \pi_i(s), s') + \gamma V_k^{\pi_i}(s')\right]$$

- **Step 2: Improvement:** For fixed values, get a better policy using policy extraction:
  - One-step look-ahead:

$$\pi_{i+1}(s) = \arg\max_a \sum_{s'} T(s, a, s') \left[R(s, a, s') + \gamma V^{\pi_i}(s')\right]$$

POLICY ITERATION EXAMPLE (0)

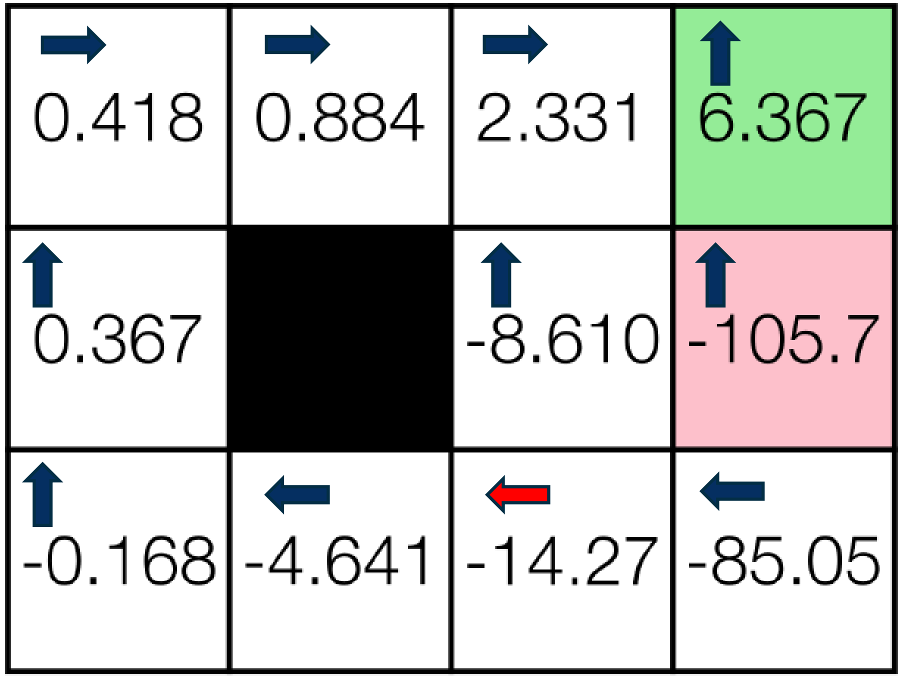
Example on a different grid world, initialized with $\pi(s) = \uparrow$ (all states). NOTE: don't stop in goal states in this grid world, thus MDP value can be < -100 when in -100 state.



Original reward function

POLICY ITERATION EXAMPLE (1)

Example on a different grid world, initialized with $\pi(s) = \uparrow$ (all states). NOTE: don't stop in goal states in this grid world, thus MDP value can be < -100 when in -100 state.

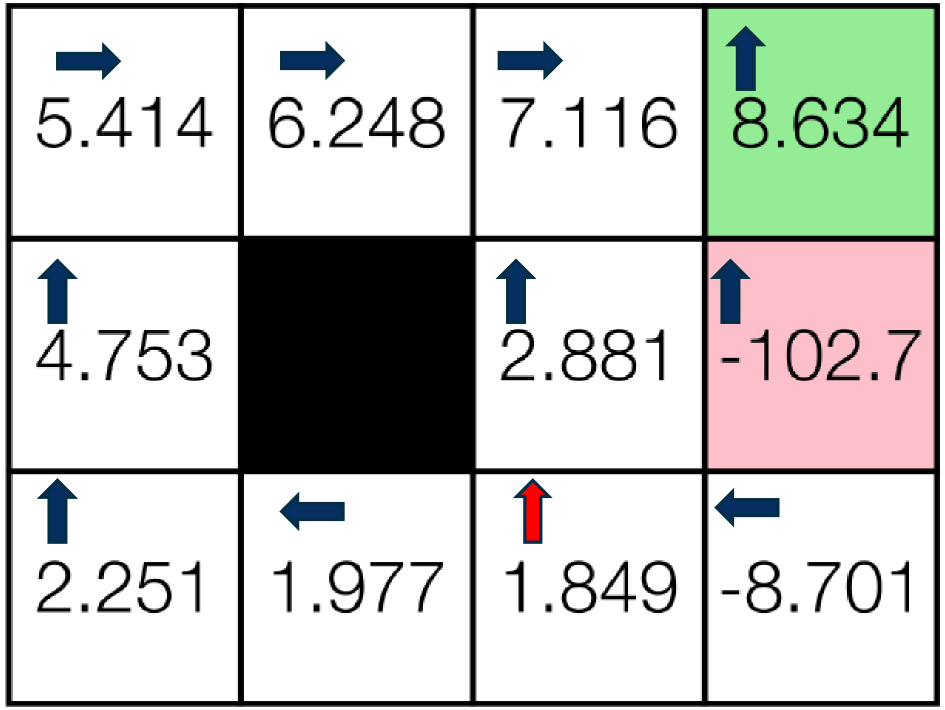


Z. Kolter

V^{π} at one iteration

POLICY ITERATION EXAMPLE (2)

Example on a different grid world, initialized with $\pi(s) = \uparrow$ (all states). NOTE: don't stop in goal states in this grid world, thus MDP value can be < -100 when in -100 state.

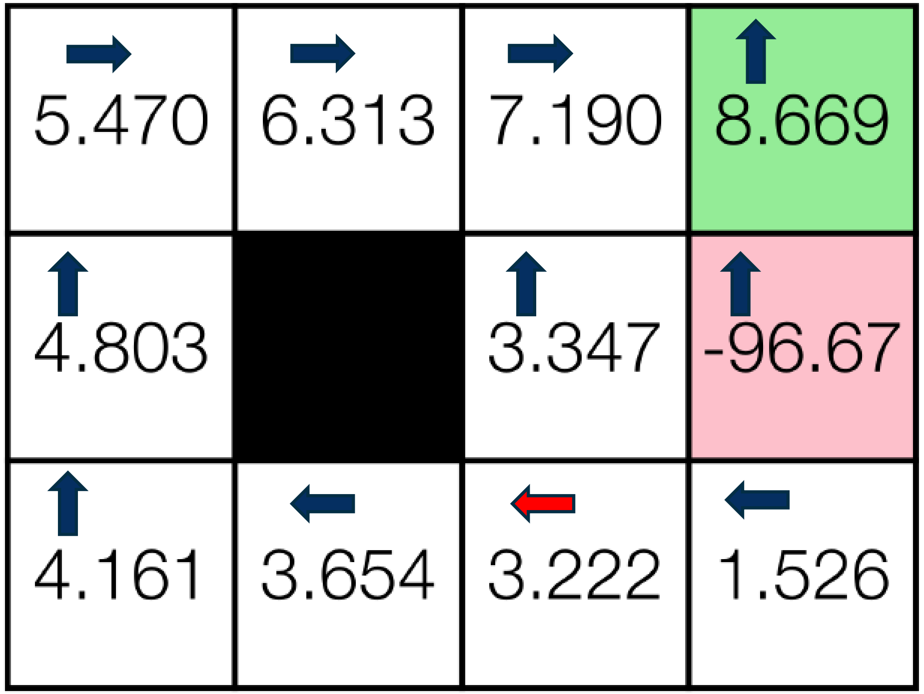


Z. Kolter

V^{π} at two iterations

POLICY ITERATION EXAMPLE (3)

Example on a different grid world, initialized with $\pi(s) = \uparrow$ (all states). NOTE: don't stop in goal states in this grid world, thus MDP value can be < -100 when in -100 state.



Z. Kolter

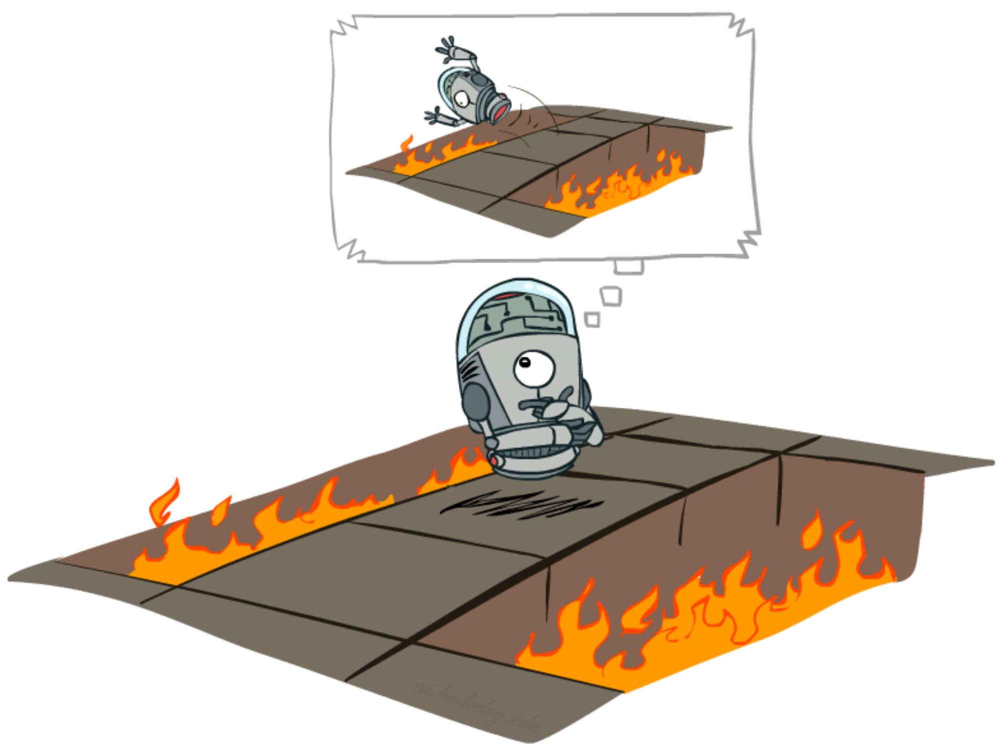
V^{π} at three iterations (converged!)

REINFORCEMENT LEARNING

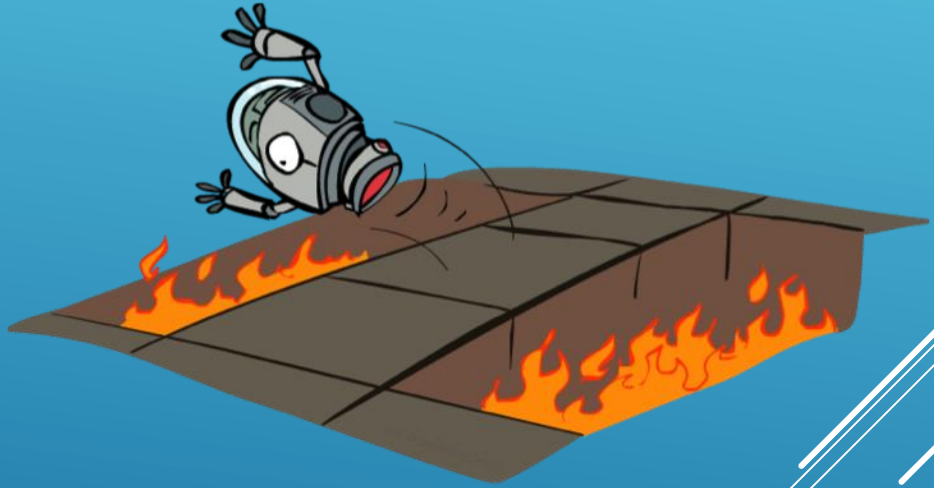
REINFORCEMENT LEARNING: THE BASIC IDEA

- Select an action
- If action leads to reward, reinforce that action
- If action leads to punishment, avoid that action
- Basically, a computational form of Behaviorism (Pavlov, B. F. Skinner)

OFFLINE VS. ONLINE



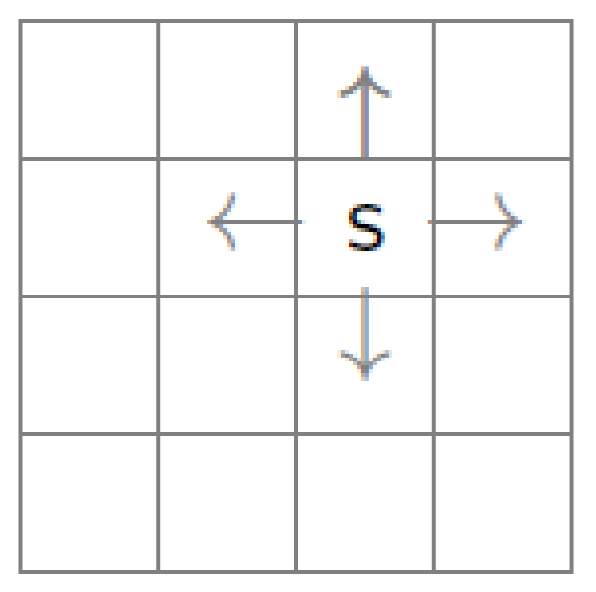
Offline Solution



Online Learning

THE LEARNING FRAMEWORK

- Learning is performed **online**, learn as we interact with the world

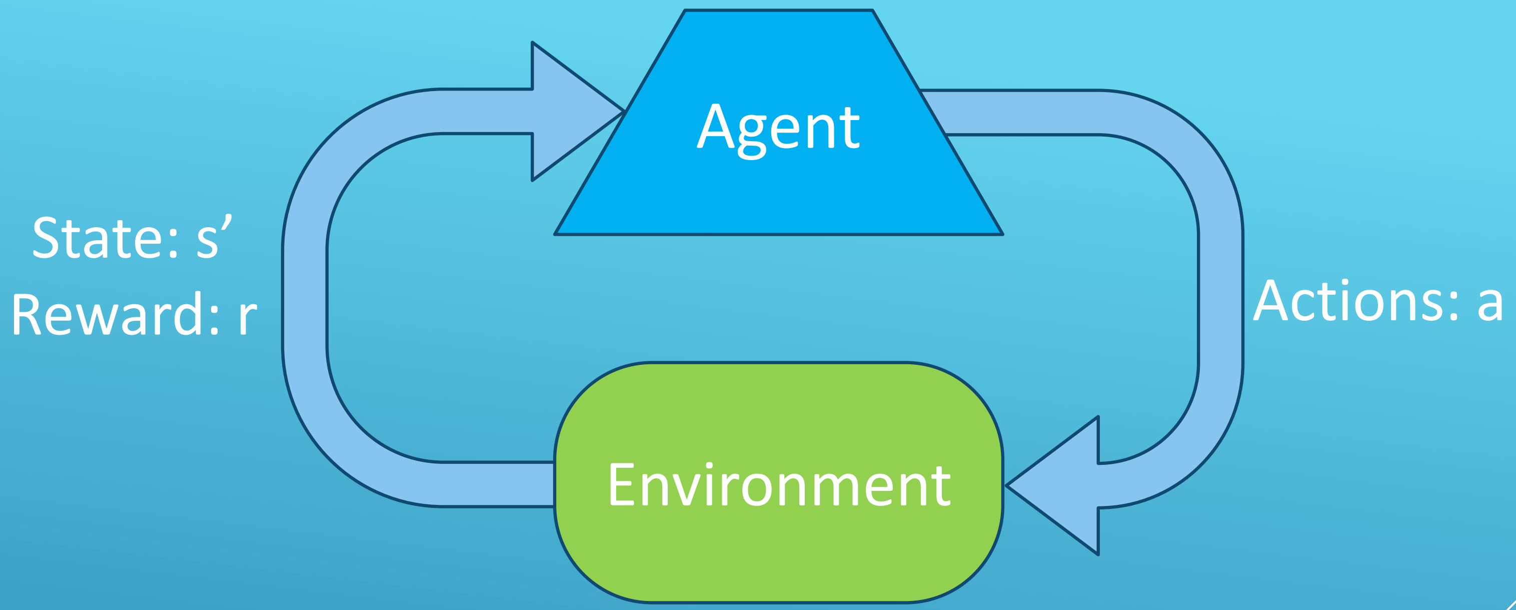


- In contrast with supervised learning, there are no training or test sets. The reward is accumulated over interactions with the environment.

- Data is not fixed, more information is acquired as you go.

- The training distribution can be influenced by action decisions.

REINFORCEMENT LEARNING

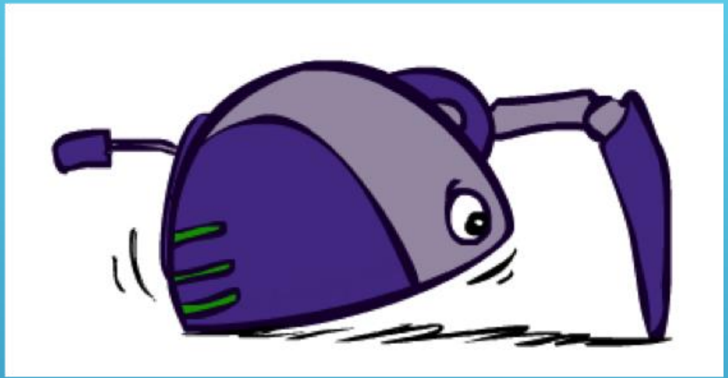


- Agent knows the current state *s*, takes action *a*, receives a reward *r* and observes the next state *s'*
- Agent has **no access** to reward model *r(s,a)* or transition model *p(s' | s,a)*
- Agent must learn to act so as to maximize expected rewards.
- All learning is based on observed samples of outcomes!
- Under these conditions, it is a very challenging problem to learn the policy π.

MODEL-BASED REINFORCEMENT LEARNING

MODEL-BASED LEARNING

- Model-Based Idea:
  - Learn an approximate model based on experiences
  - Solve for values as if the learned model were correct
- Step 1: Learn empirical MDP model
  - Count outcomes s' for each s, a
  - Normalize to give an estimate of $\hat{T}(s, a, s')$
  - Discover each $\hat{R}(s, a, s')$ when we experience (s, a, s')
- Step 2: Solve the learned MDP
  - For example, use value iteration, as before



LEARN THE REWARD AND TRANSITION DISTRIBUTIONS

- Try every action in each state a number of times
- $RTotal(s, a, s')$ =total reward for taking action a in state s and transitioning to state s'
- $N(a, s)$ = number of times action a is taken in state s
- $N(s, a, s')$ = number of times s transitions to s' on action a
- $\hat{R}(s, a, s') = RTotal(s, a, s')/N(s, a, s')$
- $\hat{T}(s, a, s') = N(s, a, s')/N(a, s)$

TRANSITION/REWARD PARAMETER TABLE

For every state s:

State s'

Action a

$\hat{T}(s, a0, s0)$ $\hat{R}(s, a0, s0)$	$\hat{T}(s, a0, s1)$ $\hat{R}(s, a0, s1)$	$\hat{T}(s, a0, s2)$ $\hat{R}(s, a0, s2)$	$\hat{T}(s, a0, s3)$ $\hat{R}(s, a0, s3)$
$\hat{T}(s, a1, s0)$ $\hat{R}(s, a1, s0)$	$\hat{T}(s, a1, s1)$ $\hat{R}(s, a1, s1)$	$\hat{T}(s, a1, s2)$ $\hat{R}(s, a1, s2)$	$\hat{T}(s, a1, s3)$ $\hat{R}(s, a1, s3)$
$\hat{T}(s, a2, s0)$ $\hat{R}(s, a2, s0)$	$\hat{T}(s, a2, s1)$ $\hat{R}(s, a2, s1)$	$\hat{T}(s, a2, s2)$ $\hat{R}(s, a2, s2)$	$\hat{T}(s, a2, s3)$ $\hat{R}(s, a2, s3)$

MODEL-BASED RL

Let π^0 be arbitrary

$k \leftarrow 0$

Experience $\leftarrow \varnothing$

Repeat

- $k \leftarrow k + 1$
- Begin in state i
- For a while:
  - Choose action a based on π^{k-1}
  - Receive reward r and transition to j
  - Experience $\leftarrow$ Experience $\cup < i, a, r, j >$
  - $i \leftarrow j$
- Learn MDP M from Experience
- Solve M to obtain π^k

MODEL-BASED RL: PROS AND CONS

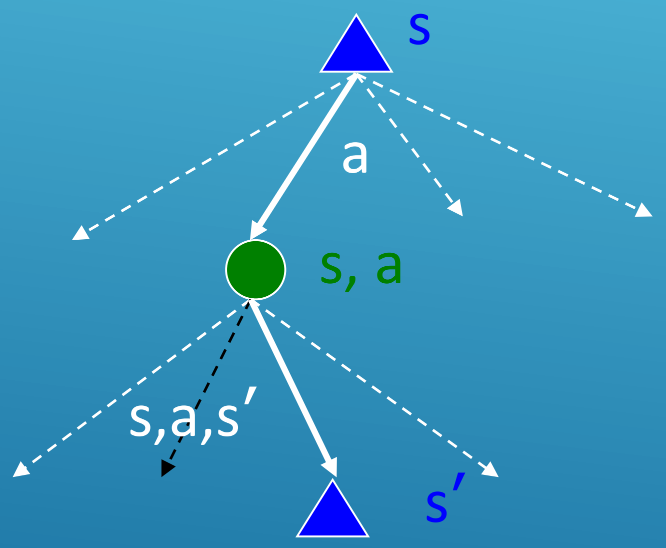
- **Pros:**
  - Makes maximal use of experience
  - Solves model optimally, given enough experience
- **Cons:**
  - Assumes model is small enough to solve
  - Requires expensive solution procedure

MODEL-FREE REINFORCEMENT LEARNING: Q-LEARNING

Q-LEARNING

- Don't learn a model, learn the Q function directly
- Appropriate when model is too large to store, solve or learn
  - size of transition of model: $O(|S^2|)$
  - value iteration cost: $O(|A||S^2|)$
  - size of Q function $O(|A||S|)$

RECALL THE BELLMAN EQUATIONS



$$V^*(s) = \max_a Q^*(s,a)$$

$$Q^*(s,a) = \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma V^*(s')\right]$$

$$V^*(s) = \max_a \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma V^*(s')\right]$$

FROM VALUE ITERATION TO Q-VALUE ITERATION

- Value iteration: find successive (depth-limited) values
  - Start with $V_0(s) = 0$, which we know is right
  - Given V_k, calculate the depth k+1 values for all states:

$$V_{k+1}(s) \leftarrow \max_a \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma V_k(s') \right]$$

- But Q-values are more useful, so compute them instead
  - Start with $Q_0(s,a) = 0$, which we know is right
  - Given Q_k, calculate the depth k+1 q-values for all q-states:

$$Q_{k+1}(s,a) \leftarrow \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma \max_{a'} Q_k(s',a') \right]$$

Q-LEARNING

- Q-Learning: sample-based Q-value iteration

$$Q_{k+1}(s,a) \leftarrow \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma \max_{a'} Q_k(s',a')\right]$$

- Learn Q(s,a) values as you go
  - Receive a sample transition (s,a,r,s')
  - Consider your old estimate: $Q(s,a)$
  - Consider your new sample estimate:

$$Q(s,a) \approx r + \gamma \max_{a'} Q(s',a')$$

  - Incorporate the new estimate into a running average:

$$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + (\alpha)\left[r + \gamma \max_{a'} Q(s',a')\right]$$

Q-LEARNING UPDATE RULE

- On transitioning from state *s* to state *s'* on action *a*, and receiving reward *r*, update:

$$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + (\alpha)\left[r + \gamma \max_{a'} Q(s',a')\right]$$

- α is the **learning rate**
- a large α results in quicker learning, but may not converge.
- α is often decreased as learning goes on.

RELATION TO EXPONENTIAL SMOOTHING

- The Q-Learning update rule is similar to a times series technique called *exponential smoothing*.
- the simplest form of exponential smoothing is given by the formulas:

$$s_0 = x_0$$
$$s_t = \alpha x_t + (1 - \alpha) s_{t-1}, \; t > 0$$

where α is the **decay rate.**

https://en.wikipedia.org/w/index.php?title=Exponential_smoothing&oldid=861463311 referenced on 10/3/2018

EXPONENTIAL SMOOTHING (2)

As time progresses, the affect on s_t of more remote terms decay exponentially as they recede into the past.

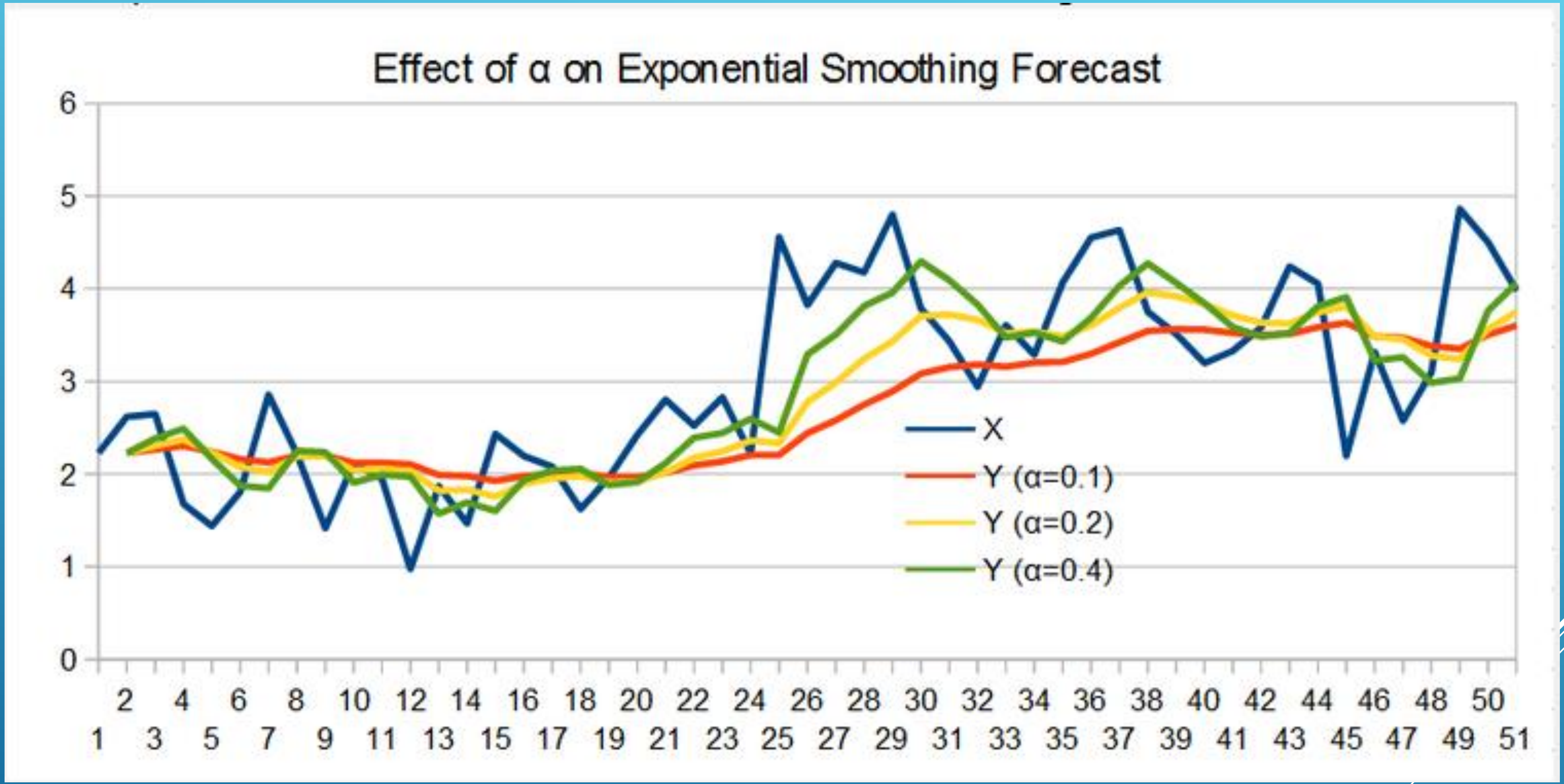
$$\begin{aligned} s_0 &= x_0 \\ s_t &= \alpha x_t + (1-\alpha) s_{t-1},\ t > 0 \end{aligned}$$

The above equations can be expanded thus:

$$\begin{aligned} s_t &= \alpha x_t + (1-\alpha) s_{t-1} \\ &= \alpha x_t + \alpha(1-\alpha) x_{t-1} + (1-\alpha)^2 s_{t-2} \\ &= \alpha \left[x_t + (1-\alpha) x_{t-1} + (1-\alpha)^2 x_{t-2} + (1-\alpha)^3 x_{t-3} + \cdots + (1-\alpha)^{t-1} x_1 \right] + (1-\alpha)^t x_0. \end{aligned}$$

https://en.wikipedia.org/w/index.php?title=Exponential_smoothing&oldid=861463311 referenced on 10/3/2018

EXPONENTIAL SMOOTHING EXAMPLE



Notice how Curves become more “wiggly” as α increases.

http://www.edscave.com/forecasting---exponential-smoothing.html referenced on 10/3/2018

Q-LEARNING ALGORITHM

For each state *s* and action *a*:

$\boxed{Q(s,a) \leftarrow 0}$

Begin in state s:

Repeat:

For all actions associated with state *s*, choose action *a* based on the Q values for state *s*

Receive reward *r* and transition to s'

$$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + (\alpha)\left[r + \gamma \max_{a'} Q(s',a')\right]$$

$\boxed{s \leftarrow s'}$

CHOOSING THE ACTION

- Learned Q function determines the policy
  - in state *s*, choose action with largest *Q(s,a)*
- Still have to worry about exploration vs. exploitation.
  - use techniques we discussed last week.

EXPLORATION RISK

- Assume we're using decreasing ϵ-exploration or simulated annealing.
- What if the optimal policy involves walking along the edge of a cliff?
- What happens during the early stages of learning?

EXPLORATION RISK

- Update rule:

$$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + (\alpha)\left[r + \gamma \max_{a'} Q(s',a')\right]$$

- Q-value is updated based on the best action.
- But if we're exploring a lot, we won't always do the best action.
- We will fall off the cliff a lot!
- We would like to take advantage of our experience on the cliff to prevent this from happening more than necessary!

NEXT TIME: MORE MODEL-FREE RL

- **more Q-Learning**
- **SARSA-Learning** – addresses problem of falling off the cliff too often.
- **TD(λ)**
- **Generalization**
- **Deep Q-Learning?**