

# **Deep Reinforcement Learning**

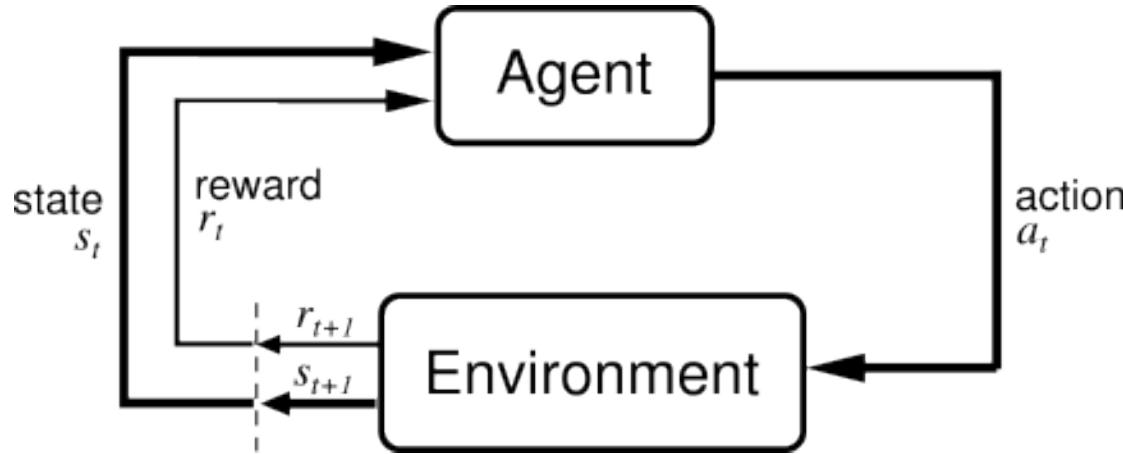
## **Lecture 1: Motivation + Overview + MDPs + Exact Solution Methods**

Pieter Abbeel

OpenAI / UC Berkeley / Gradescope

Many slides made with John Schulman, Yan (Rocky) Duan and Xi (Peter) Chen

# Markov Decision Process



Assumption: agent gets to observe the state

# Some Reinforcement Learning Success Stories



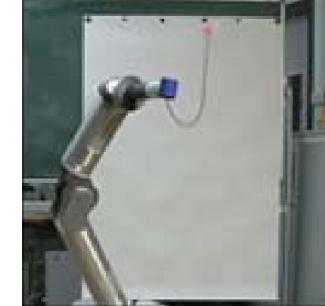
Kohl and Stone, 2004



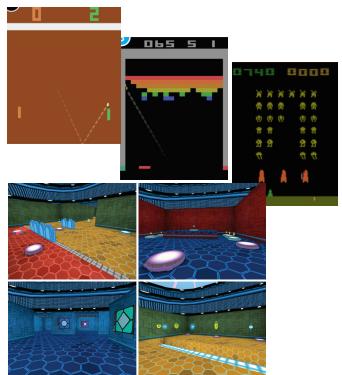
Ng et al, 2004



Tedrake et al, 2005

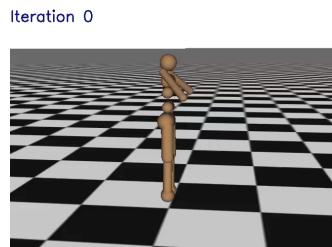


Kober and Peters, 2009



Silver et al, 2014 (DPG)  
Lillicrap et al, 2015 (DDPG)

Mnih et al 2013 (DQN)  
Mnih et al, 2015 (A3C)



Schulman et al,  
2016 (TRPO + GAE)

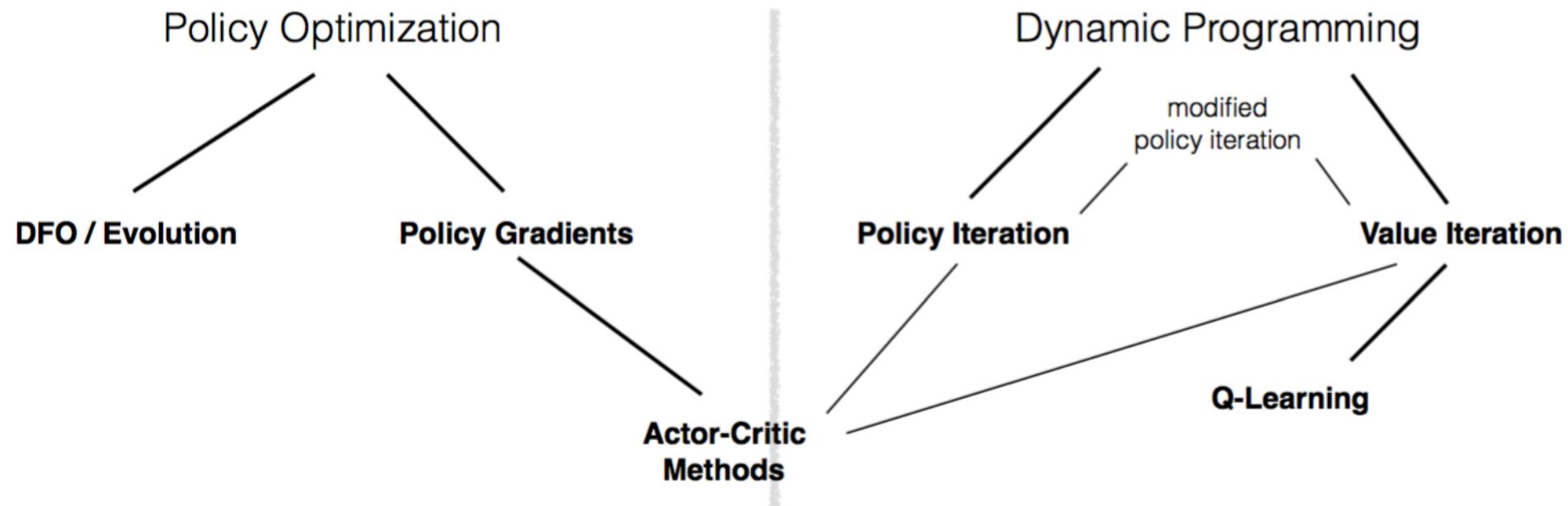


Levine\*, Finn\*, et  
al, 2016  
(GPS)

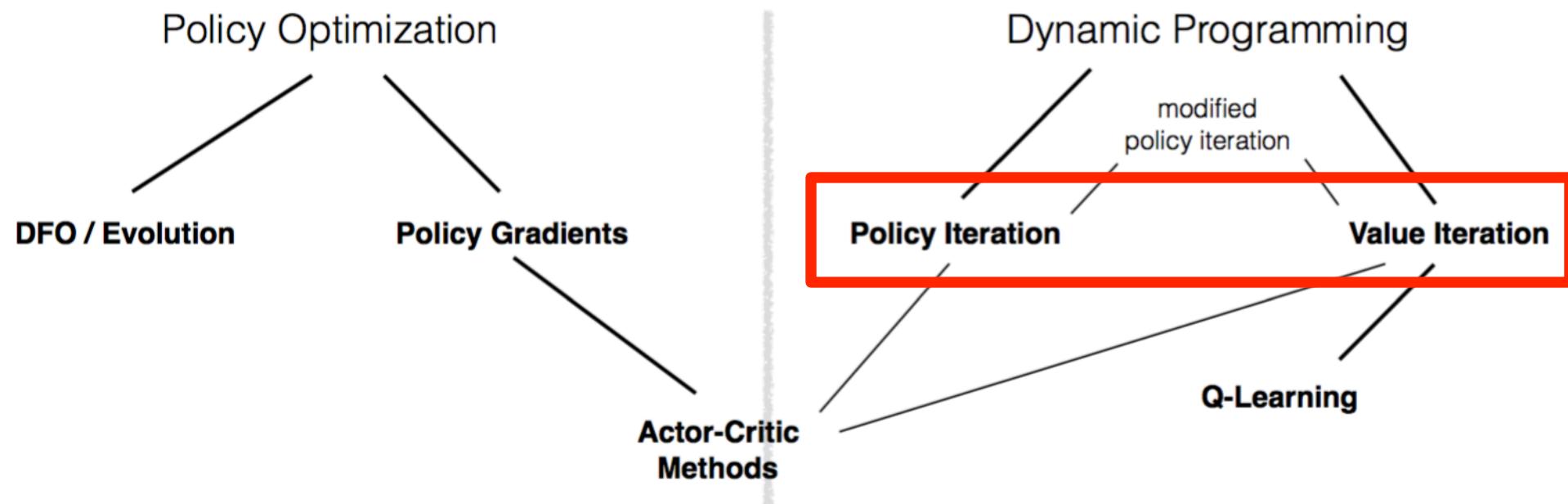


Silver\*, Huang\*, et  
al, 2016  
(AlphaGo)

# RL Algorithms Landscape



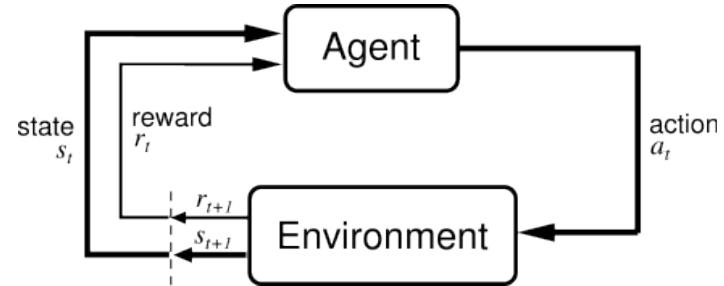
# RL Algorithms Landscape



# Markov Decision Process (MDP)

An MDP is defined by:

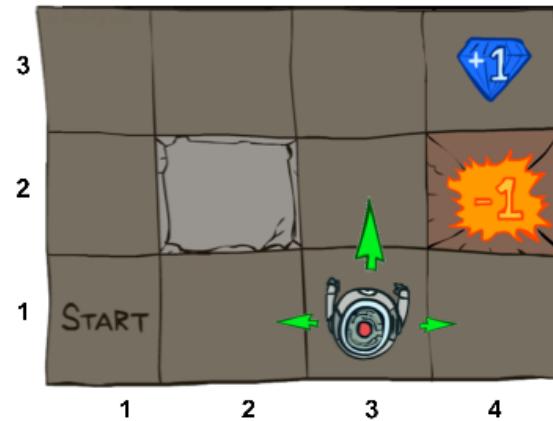
- Set of states  $S$
- Set of actions  $A$
- Transition function  $P(s' | s, a)$
- Reward function  $R(s, a, s')$
- Start state  $s_0$
- Discount factor  $\gamma$
- Horizon  $H$



# Example MDP: Gridworld

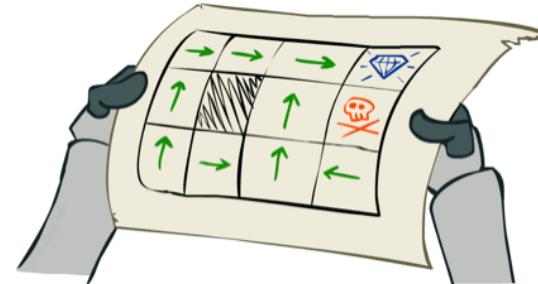
An MDP is defined by:

- Set of states  $S$
- Set of actions  $A$
- Transition function  $P(s' | s, a)$
- Reward function  $R(s, a, s')$
- Start state  $s_0$
- Discount factor  $\gamma$
- Horizon  $H$



Goal:  $\max_{\pi} \mathbb{E} \left[ \sum_{t=0}^H \gamma^t R(S_t, A_t, S_{t+1}) | \pi \right]$

$\pi$ :



# Outline

---

- Optimal Control

=

given an MDP  $(S, A, P, R, \gamma, H)$

find the optimal policy  $\pi^*$

- Exact Methods:

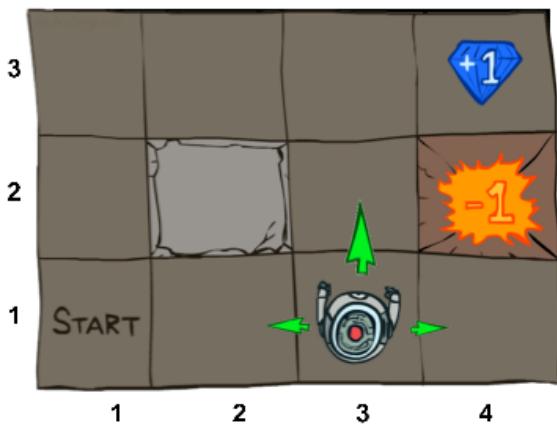
- ***Value Iteration***
- Policy Iteration

For now: small discrete state-action spaces as they are simpler to get the main concepts across. We will consider large / continuous spaces later!

# Optimal Value Function $V^*$

$$V^*(s) = \max_{\pi} \mathbb{E} \left[ \sum_{t=0}^H \gamma^t R(s_t, a_t, s_{t+1}) \mid \pi, s_0 = s \right]$$

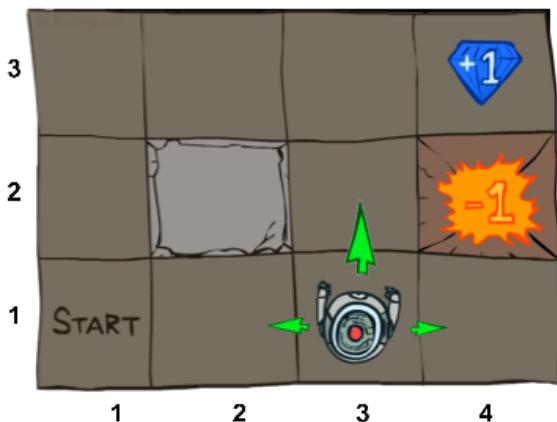
= sum of discounted rewards when starting from state  $s$  and acting optimally



# Optimal Value Function $V^*$

$$V^*(s) = \max_{\pi} \mathbb{E} \left[ \sum_{t=0}^H \gamma^t R(s_t, a_t, s_{t+1}) \mid \pi, s_0 = s \right]$$

= sum of discounted rewards when starting from state s and acting optimally



Let's assume:

actions deterministically successful, gamma = 1, H = 100

$$V^*(4,3) =$$

$$V^*(3,3) =$$

$$V^*(2,3) =$$

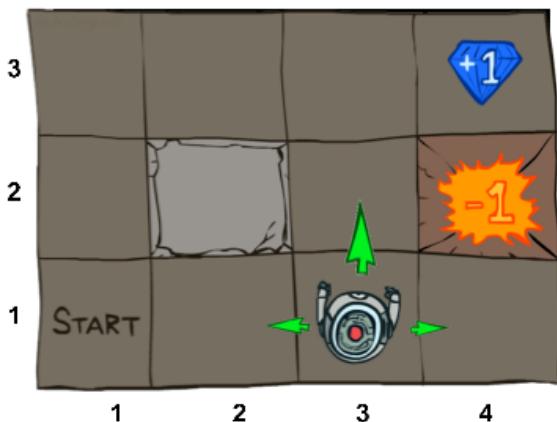
$$V^*(1,1) =$$

$$V^*(4,2) =$$

# Optimal Value Function $V^*$

$$V^*(s) = \max_{\pi} \mathbb{E} \left[ \sum_{t=0}^H \gamma^t R(s_t, a_t, s_{t+1}) \mid \pi, s_0 = s \right]$$

= sum of discounted rewards when starting from state s and acting optimally



Let's assume:

actions deterministically successful, gamma = 0.9, H = 100

$$V^*(4,3) =$$

$$V^*(3,3) =$$

$$V^*(2,3) =$$

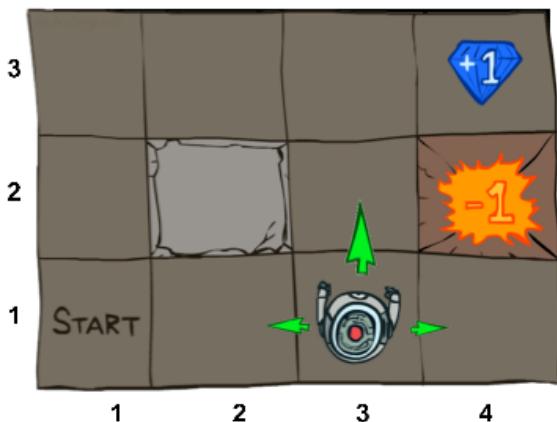
$$V^*(1,1) = 1 * 0.9 * 0.9 * 0.9 * 0.9 * 0.9$$

$$V^*(4,2) =$$

# Optimal Value Function $V^*$

$$V^*(s) = \max_{\pi} \mathbb{E} \left[ \sum_{t=0}^H \gamma^t R(s_t, a_t, s_{t+1}) \mid \pi, s_0 = s \right]$$

= sum of discounted rewards when starting from state  $s$  and acting optimally



Let's assume:

actions successful w/probability 0.8,  $\gamma = 0.9$ ,  $H = 100$

$$V^*(4,3) =$$

$$V^*(3,3) =$$

$$V^*(2,3) =$$

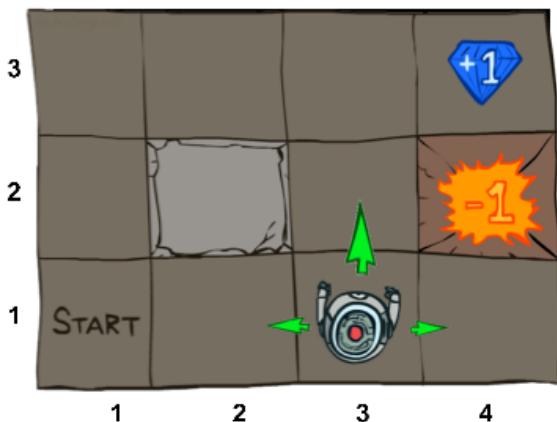
$$V^*(1,1) =$$

$$V^*(4,2) =$$

# Optimal Value Function $V^*$

$$V^*(s) = \max_{\pi} \mathbb{E} \left[ \sum_{t=0}^H \gamma^t R(s_t, a_t, s_{t+1}) \mid \pi, s_0 = s \right]$$

= sum of discounted rewards when starting from state  $s$  and acting optimally



Let's assume:

actions successful w/probability 0.8, gamma = 0.9, H = 100

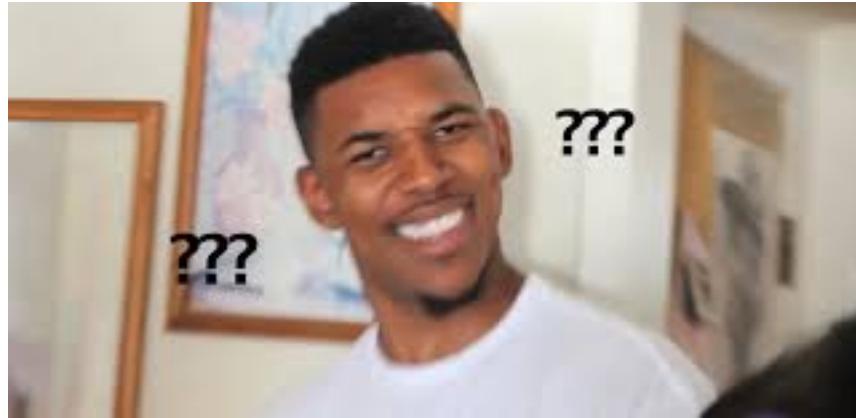
$$V^*(4,3) = 1$$

$$V^*(3,3) = 0.8 * 0.9 * V^*(4,3) + 0.1 * 0.9 * V^*(3,3) + 0.1 * 0.9 * V^*(3,2)$$

$$V^*(2,3) =$$

$$V^*(1,1) =$$

$$V^*(4,2) =$$



David Silver to the Rescue  
RL2-MDP.pdf

# Value Iteration

- $V_0^*(s)$  = optimal value for state s when H=0
  - $V_0^*(s) = 0 \quad \forall s$
- $V_1^*(s)$  = optimal value for state s when H=1
  - $V_1^*(s) = \max_a \sum_{s'} P(s'|s, a)(R(s, a, s') + \gamma V_0^*(s'))$
- $V_2^*(s)$  = optimal value for state s when H=2
  - $V_2^*(s) = \max_a \sum_{s'} P(s'|s, a)(R(s, a, s') + \gamma V_1^*(s'))$
- $V_k^*(s)$  = optimal value for state s when H = k
  - $V_k^*(s) = \max_a \sum_{s'} P(s'|s, a)(R(s, a, s') + \gamma V_{k-1}^*(s'))$

# Value Iteration

Algorithm:

Start with  $V_0^*(s) = 0$  for all s.

For  $k = 1, \dots, H$ :

For all states  $s$  in  $S$ :

$$V_k^*(s) \leftarrow \max_a \sum_{s'} P(s'|s, a) (R(s, a, s') + \gamma V_{k-1}^*(s'))$$

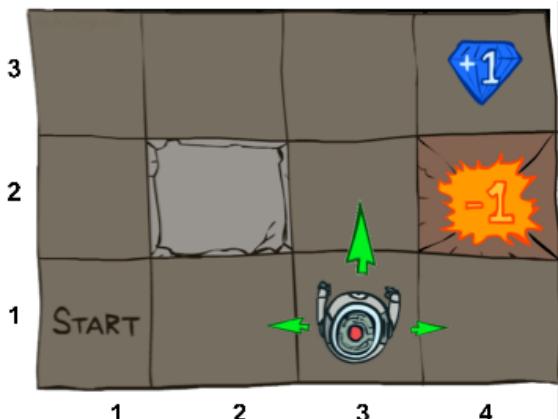
$$\pi_k^*(s) \leftarrow \arg \max_a \sum_{s'} P(s'|s, a) (R(s, a, s') + \gamma V_{k-1}^*(s'))$$

This is called a **value update** or **Bellman update/back-up**

# Value Iteration

$$V_0(s) \leftarrow 0$$

$$k = 0$$



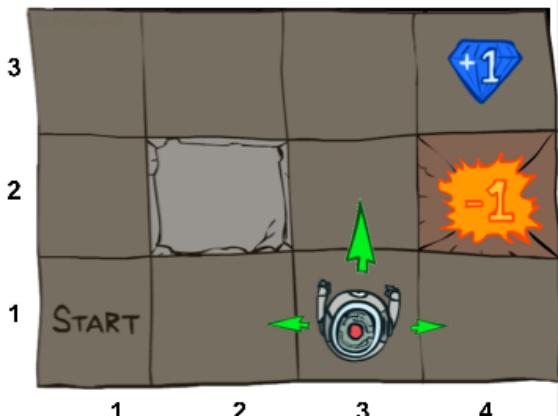
VALUES AFTER 0 ITERATIONS			
0.00	0.00	0.00	0.00
0.00		0.00	0.00
0.00	0.00	0.00	0.00

Noise = 0.2  
Discount = 0.9

# Value Iteration

$$V_1(s) \leftarrow \max_a \sum_{s'} P(s'|s, a)(R(s, a, s') + \gamma V_0(s'))$$

$k = 0$



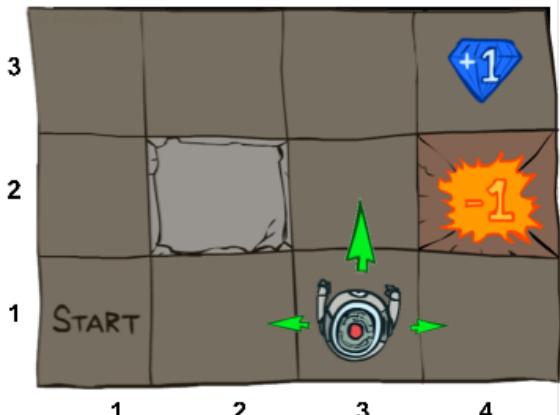
VALUES AFTER 0 ITERATIONS			
0.00	0.00	0.00	0.00
0.00		0.00	0.00
0.00	0.00	0.00	0.00

Noise = 0.2  
Discount = 0.9

# Value Iteration

$$V_2(s) \leftarrow \max_a \sum_{s'} P(s'|s, a)(R(s, a, s') + \gamma V_1(s'))$$

$k = 1$



0.00	0.00	0.00	1.00
0.00		0.00	-1.00
0.00	0.00	0.00	0.00

Noise = 0.2  
Discount = 0.9

# Value Iteration

$$V_2(s) \leftarrow \max_a \sum_{s'} P(s'|s, a)(R(s, a, s') + \gamma V_1(s'))$$

k = 2

0.00	0.00	0.72	1.00
0.00		0.00	-1.00
0.00	0.00	0.00	0.00
VALUES AFTER 2 ITERATIONS			

Noise = 0.2

Discount = 0.9

# Value Iteration

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

k = 3

0.00	0.52	0.78	1.00
0.00		0.43	-1.00
0.00	0.00	0.00	0.00
VALUES AFTER 3 ITERATIONS			

Noise = 0.2

Discount = 0.9

# Value Iteration

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

k = 4



Noise = 0.2  
Discount = 0.9

# Value Iteration

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

k = 5



Noise = 0.2  
Discount = 0.9

# Value Iteration

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

k = 6



Noise = 0.2  
Discount = 0.9

# Value Iteration

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

k = 7



Noise = 0.2  
Discount = 0.9

# Value Iteration

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

k = 8



Noise = 0.2  
Discount = 0.9

# Value Iteration

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

k = 9



Noise = 0.2

Discount = 0.9

# Value Iteration

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

k = 10



Noise = 0.2  
Discount = 0.9

# Value Iteration

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

k = 11



Noise = 0.2

Discount = 0.9

# Value Iteration

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

k = 12



Noise = 0.2  
Discount = 0.9

# Value Iteration

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

k = 100



Noise = 0.2  
Discount = 0.9

# Value Iteration Convergence

**Theorem.** *Value iteration converges. At convergence, we have found the optimal value function  $V^*$  for the discounted infinite horizon problem, which satisfies the Bellman equations*

$$\forall S \in S : V^*(s) = \max_a \sum_{s'} P(s'|s, a) [R(s, a, s') + \gamma V^*(s')]$$

- Now we know how to act for infinite horizon with discounted rewards!
  - Run value iteration till convergence.
  - This produces  $V^*$ , which in turn tells us how to act, namely following:

$$\pi^*(s) = \arg \max_a \sum_{s'} P(s'|s, a) [R(s, a, s') + \gamma V^*(s')]$$

- Note: the infinite horizon optimal policy is **stationary**, i.e., the optimal action at a state  $s$  is the same action at all times. (Efficient to store!)

# Convergence: Intuition

- $V^*(s)$  = expected sum of rewards accumulated starting from state  $s$ , acting optimally for  $\infty$  steps
- $V_H^*(s)$  = expected sum of rewards accumulated starting from state  $s$ , acting optimally for  $H$  steps
- Additional reward collected over time steps  $H+1, H+2, \dots$

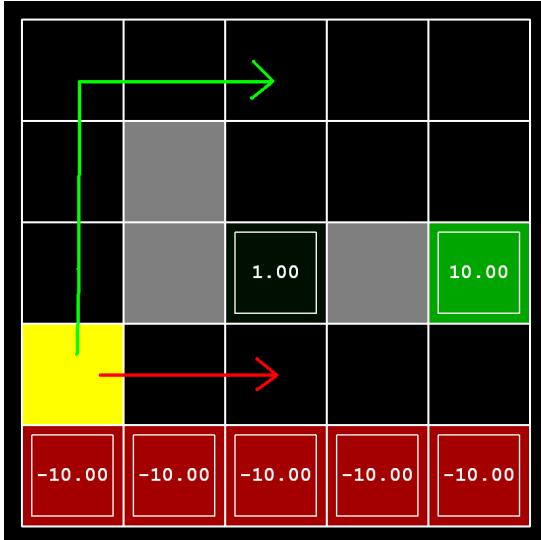
$$\gamma^{H+1} R(s_{H+1}) + \gamma^{H+2} R(s_{H+2}) + \dots \leq \gamma^{H+1} R_{max} + \gamma^{H+2} R_{max} + \dots = \frac{\gamma^{H+1}}{1 - \gamma} R_{max}$$

goes to zero as  $H$  goes to infinity

Hence  $V_H^* \xrightarrow{H \rightarrow \infty} V^*$

For simplicity of notation in the above it was assumed that rewards are always greater than or equal to zero. If rewards can be negative, a similar argument holds, using  $\max |R|$  and bounding from both sides.

# Exercise 1: Effect of Discount and Noise



- (a) Prefer the close exit (+1), risking the cliff (-10) (1)  $\gamma = 0.1$ , noise = 0.5
- (b) Prefer the close exit (+1), but avoiding the cliff (-10) (2)  $\gamma = 0.99$ , noise = 0
- (c) Prefer the distant exit (+10), risking the cliff (-10) (3)  $\gamma = 0.99$ , noise = 0.5
- (d) Prefer the distant exit (+10), avoiding the cliff (-10) (4)  $\gamma = 0.1$ , noise = 0

# Exercise 1 Solution

0.00 ↗	0.00 ↗	0.01 ↓	0.01 ↗	0.10 ↓
0.00		0.10	0.10 ↗	1.00
↓		↓		↓
0.00		1.00		10.00
↓		↑		↑
0.00 ↗	0.01 ↗	0.10	0.10 ↗	1.00
-10.00	-10.00	-10.00	-10.00	-10.00

(a) Prefer close exit (+1), risking the cliff (-10)

---

(4)  $\gamma = 0.1$ , noise = 0

# Exercise 1 Solution

0.00	0.00	0.00	0.00	0.03
0.00		0.05	0.03	0.51
0.00		1.00		10.00
0.00	0.00	0.05	0.01	0.51
-10.00	-10.00	-10.00	-10.00	-10.00

(b) Prefer close exit (+1), avoiding the cliff (-10)

---

(1)  $\gamma = 0.1$ , noise = 0.5

# Exercise 1 Solution

9.41 ↗	9.51 ↗	9.61 ↗	9.70 ↗	9.80 ▼
9.32		9.70 ↗	9.80 ↗	9.90 ▼
9.41 ▼		1.00		10.00
9.51 ↗	9.61 ↗	9.70 ↗	9.80 ↗	9.90 ▲
-10.00	-10.00	-10.00	-10.00	-10.00

(c) Prefer distant exit (+1), risking the cliff (-10)

---

(2)  $\gamma = 0.99$ , noise = 0

# Exercise 1 Solution

8.67 ▶	8.93 ▶	9.11 ▶	9.30 ▶	9.42 ▼
▲		▲		
8.49		9.09	9.42 ▶	9.68 ▼
▲				
8.33		1.00		10.00
▲	▲	▲	▲	▲
7.13	5.04	3.15	5.68	8.45
-10.00	-10.00	-10.00	-10.00	-10.00

(d) Prefer distant exit (+1), avoid the cliff (-10)

---

(3)  $\gamma = 0.99$ , noise = 0.5

# Q-Values

---

$Q^*(s, a)$  = expected utility starting in  $s$ , taking action  $a$ , and (thereafter) acting optimally

Bellman Equation:

$$Q^*(s, a) = \sum_{s'} P(s'|s, a)(R(s, a, s') + \gamma \max_{a'} Q^*(s', a'))$$

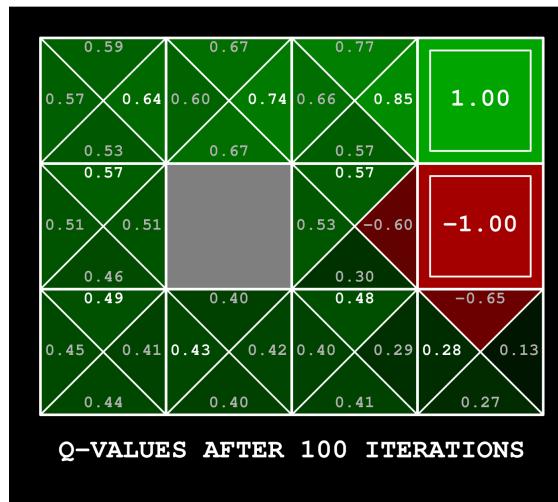
Q-Value Iteration:

$$Q_{k+1}^*(s, a) \leftarrow \sum_{s'} P(s'|s, a)(R(s, a, s') + \gamma \max_{a'} Q_k^*(s', a'))$$

# Q-Value Iteration

$$Q_{k+1}^*(s, a) \leftarrow \sum_{s'} P(s'|s, a)(R(s, a, s') + \gamma \max_{a'} Q_k^*(s', a'))$$

$k = 100$



Noise = 0.2  
Discount = 0.9

# Outline

---

- Optimal Control

=

given an MDP  $(S, A, P, R, \gamma, H)$

find the optimal policy  $\pi^*$

- Exact Methods:



***Value Iteration***

- Policy Iteration

For now: small discrete state-action spaces as they are simpler to get the main concepts across. We will consider large / continuous spaces later!

# Policy Evaluation

- Recall value iteration:

$$V_k^*(s) \leftarrow \max_a \sum_{s'} P(s'|s, a) (R(s, a, s') + \gamma V_{k-1}^*(s'))$$

- Policy evaluation for a given  $\pi(s)$ :

$$V_k^\pi(s) \leftarrow \sum_{s'} P(s'|s, \pi(s)) (R(s, \pi(s), s') + \gamma V_{k-1}^\pi(s))$$

At convergence:

$$\forall s \quad V^\pi(s) \leftarrow \sum_{s'} P(s'|s, \pi(s)) (R(s, \pi(s), s') + \gamma V^\pi(s))$$

# Exercise 2

---

Consider a *stochastic* policy  $\pi(a|s)$ , where  $\pi(a|s)$  is the probability of taking action  $a$  when in state  $s$ . Which of the following is the correct update to perform policy evaluation for this stochastic policy?

1.  $V_{k+1}^{\pi}(s) \leftarrow \max_a \sum_{s'} P(s'|s, a) (R(s, a, s') + \gamma V_k^{\pi}(s'))$
2.  $V_{k+1}^{\pi}(s) \leftarrow \sum_{s'} \sum_a \pi(a|s) P(s'|s, a) (R(s, a, s') + \gamma V_k^{\pi}(s'))$
3.  $V_{k+1}^{\pi}(s) \leftarrow \sum_a \pi(a|s) \max_{s'} P(s'|s, a) (R(s, a, s') + \gamma V_k^{\pi}(s'))$

# Policy Iteration

**One iteration of policy iteration:**

- Policy evaluation for current policy  $\pi_k$  :
  - Iterate until convergence

$$V_{i+1}^{\pi_k}(s) \leftarrow \sum_{s'} P(s'|s, \pi_k(s)) [R(s, \pi(s), s') + \gamma V_i^{\pi_k}(s')]$$

- Policy improvement: find the best action according to one-step look-ahead

$$\pi_{k+1}(s) \leftarrow \arg \max_a \sum_{s'} P(s'|s, a) [R(s, a, s') + \gamma V^{\pi_k}(s')]$$

- Repeat until policy converges
- At convergence: optimal policy; and converges faster than value iteration under some conditions

# Policy Evaluation Revisited

- Idea 1: modify Bellman updates

$$V_0^\pi(s) = 0$$

$$V_{i+1}^\pi(s) \leftarrow \sum_{s'} P(s'|s, \pi(s)) [R(s, \pi(s), s') + \gamma V_i^\pi(s')]$$

- Idea 2: it is just a linear system, solve with Numpy (or whatever)

variables:  $V^\pi(s)$

constants:  $P, R$

$$\forall s \quad V^\pi(s) = \sum_{s'} P(s'|s, \pi(s)) [R(s, \pi(s), s') + \gamma V^\pi(s')]$$

# Policy Iteration Guarantees

Policy Iteration iterates over:

- Policy evaluation for current policy  $\pi_k$  :

- Iterate until convergence

$$V_{i+1}^{\pi_k}(s) \leftarrow \sum_{s'} P(s'|s, \pi_k(s)) [R(s, \pi(s), s') + \gamma V_i^{\pi_k}(s')]$$

- Policy improvement: find the best action according to one-step look-ahead

$$\pi_{k+1}(s) \leftarrow \arg \max_a \sum_{s'} P(s'|s, a) [R(s, a, s') + \gamma V^{\pi_k}(s')]$$

**Theorem.** Policy iteration is guaranteed to converge and at convergence, the current policy and its value function are the optimal policy and the optimal value function!

Proof sketch:

- (1) *Guarantee to converge:* In every step the policy improves. This means that a given policy can be encountered at most once. This means that after we have iterated as many times as there are different policies, i.e.,  $(\text{number actions})^{(\text{number states})}$ , we must be done and hence have converged.
- (2) *Optimal at convergence:* by definition of convergence, at convergence  $\pi_{k+1}(s) = \pi_k(s)$  for all states  $s$ . This means  $\forall s \quad V^{\pi_k}(s) = \max_a \sum_{s'} T(s, a, s') [R(s, a, s') + \gamma V_i^{\pi_k}(s')]$   
Hence  $V^{\pi_k}$  satisfies the Bellman equation, which means  $V^{\pi_k}$  is equal to the optimal value function  $V^*$ .

# Outline

- Optimal Control

=

given an MDP  $(S, A, P, R, \gamma, H)$

find the optimal policy  $\pi^*$

- Exact Methods:

-  ***Value Iteration***
-  ***Policy Iteration***

*Limitations:*

- Iteration over / storage for all states and actions: requires small, discrete state-action space
- Update equations require access to dynamics model