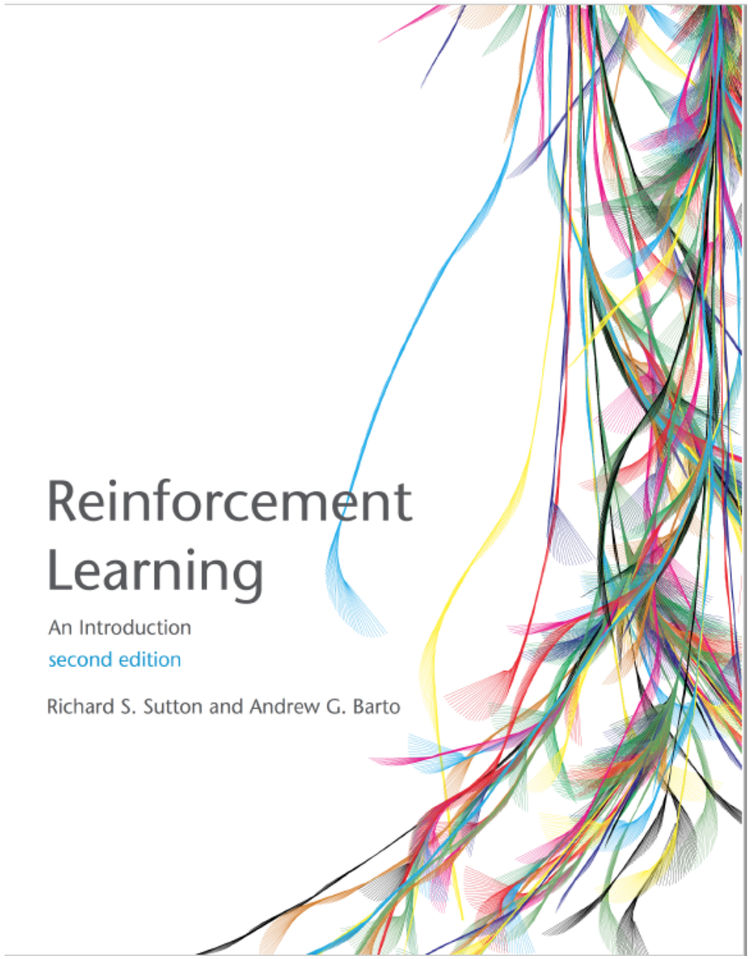
Control

OR: Find the best policy

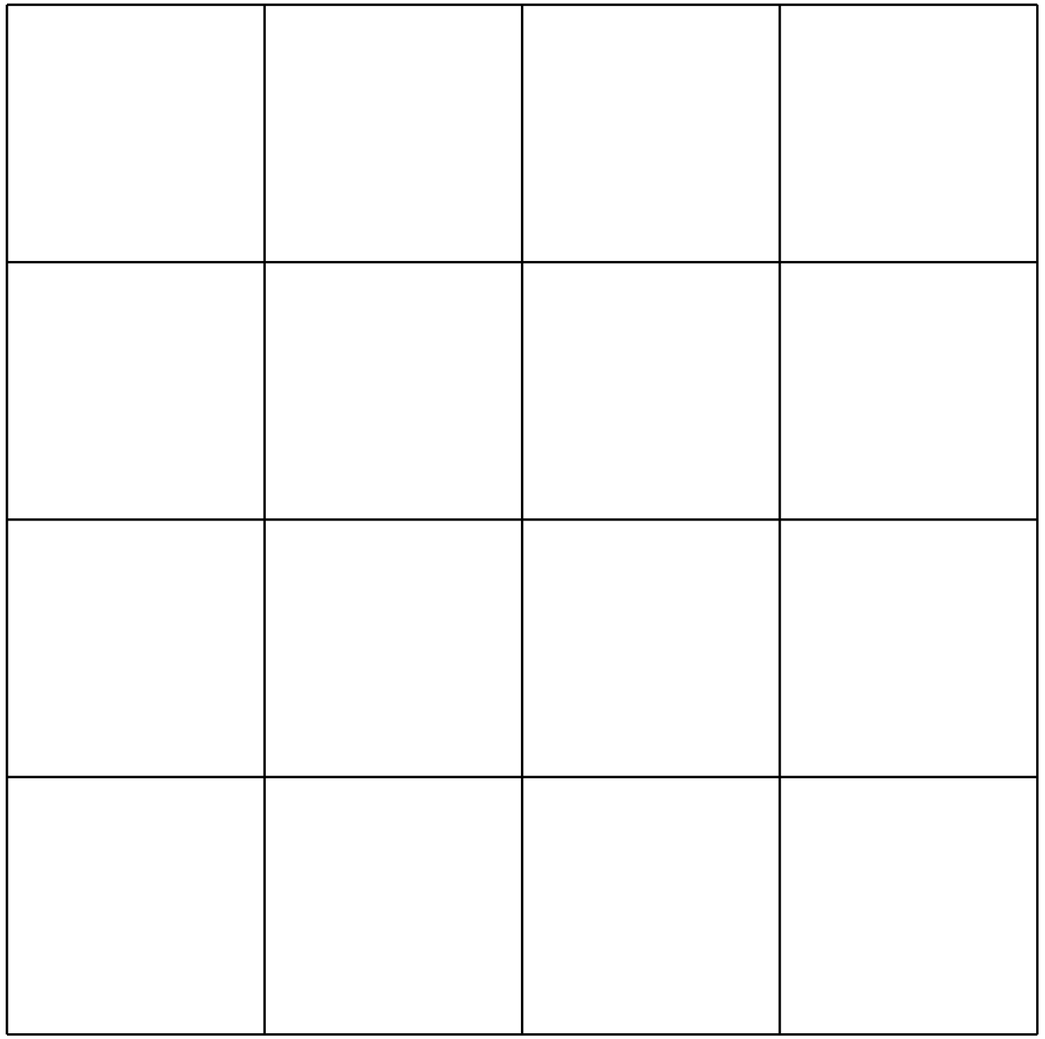


Chapter 5 & 6

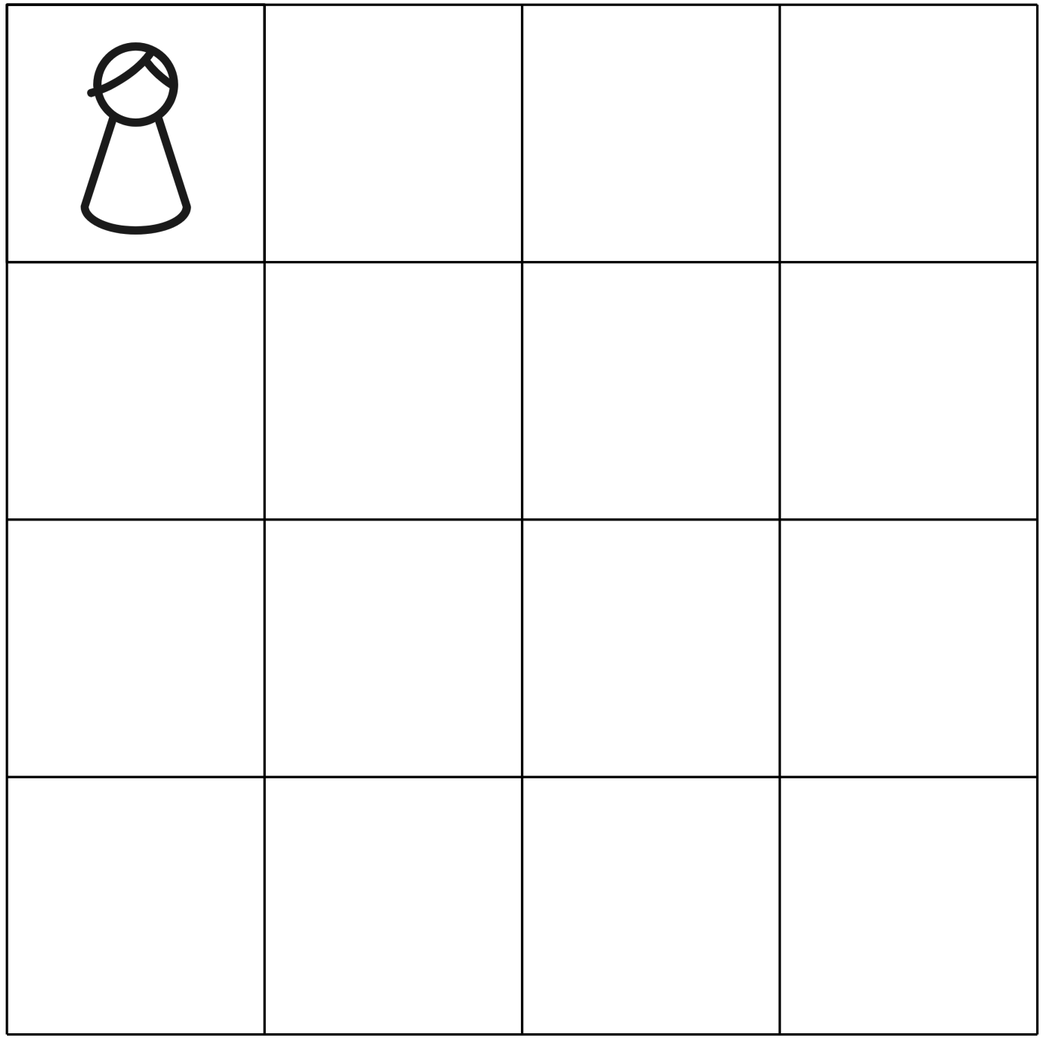
Lecture by David Silver

Grid-World

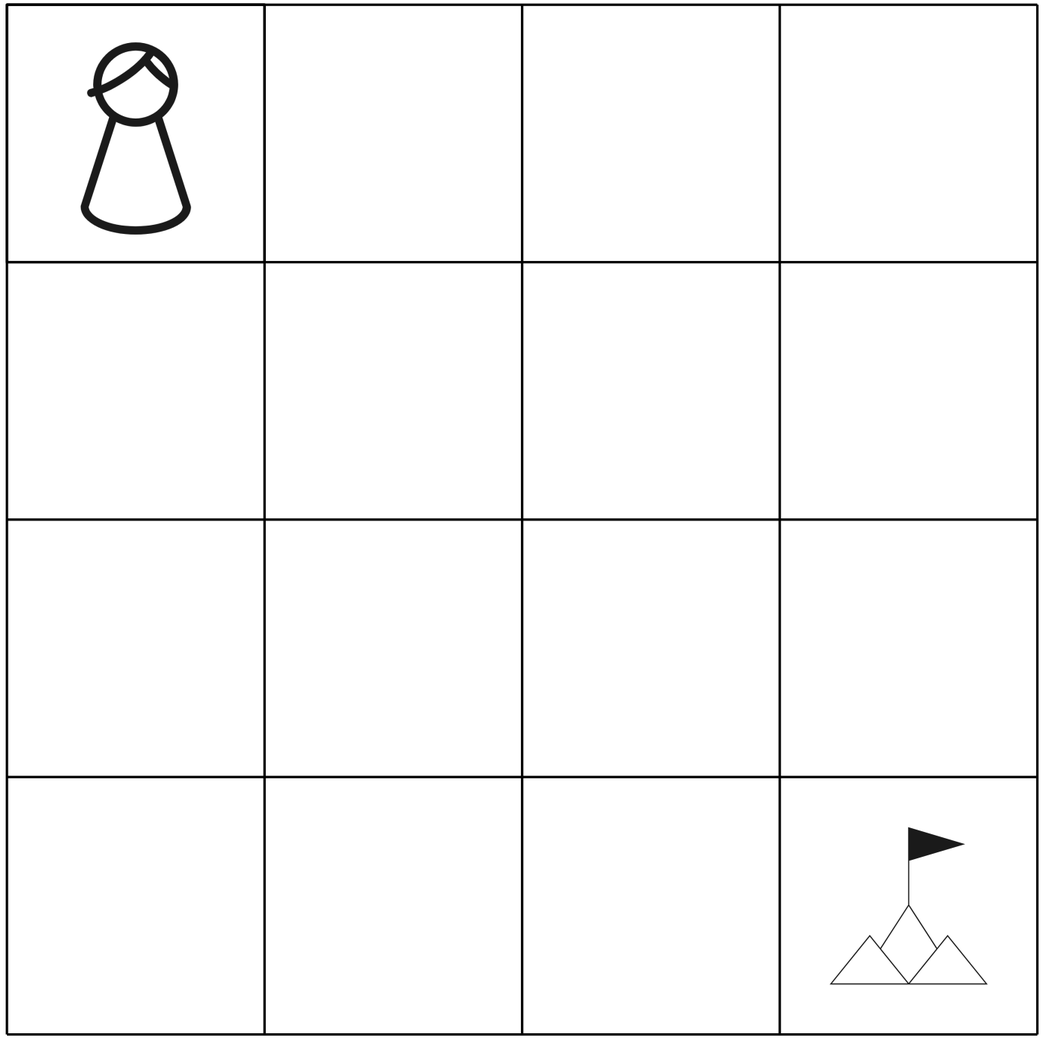
Grid-World



Grid-World



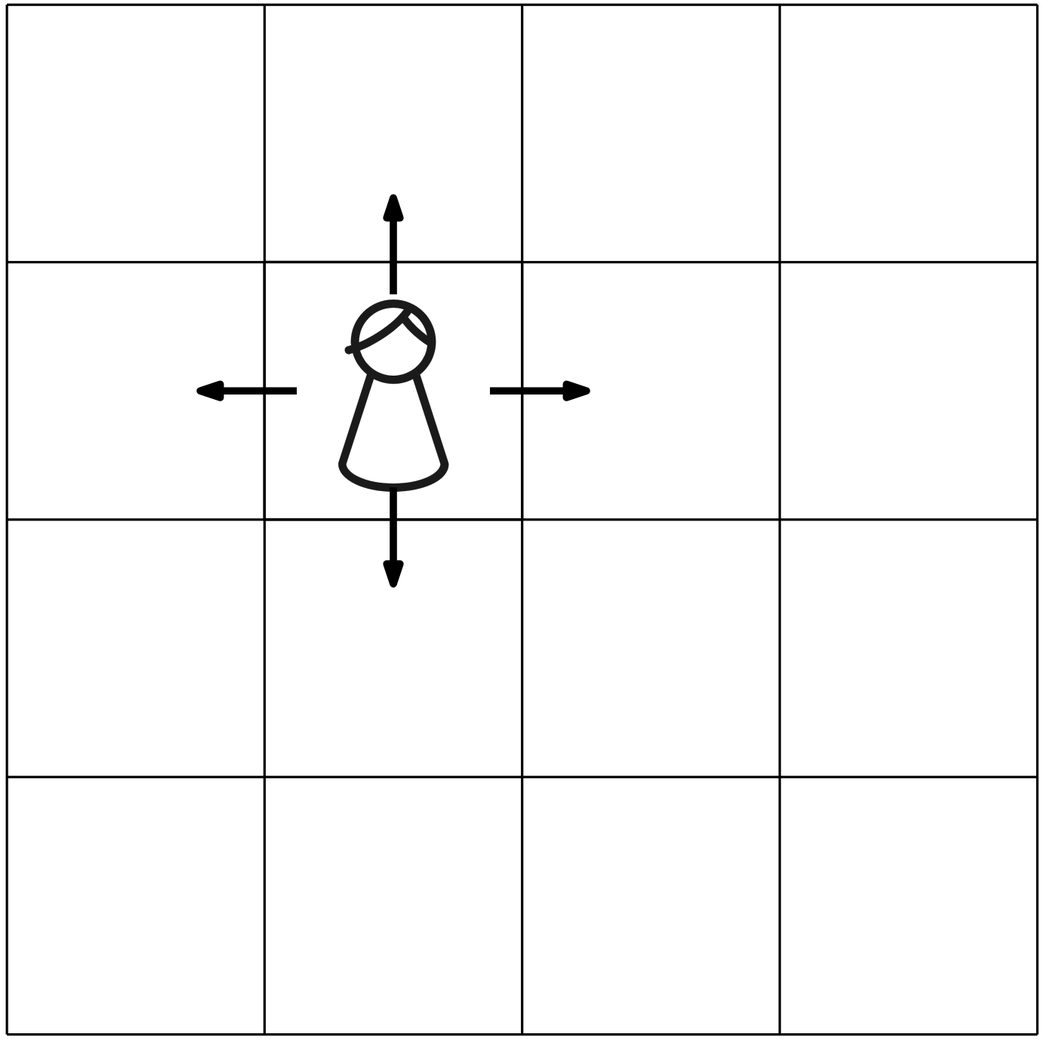
Grid-World



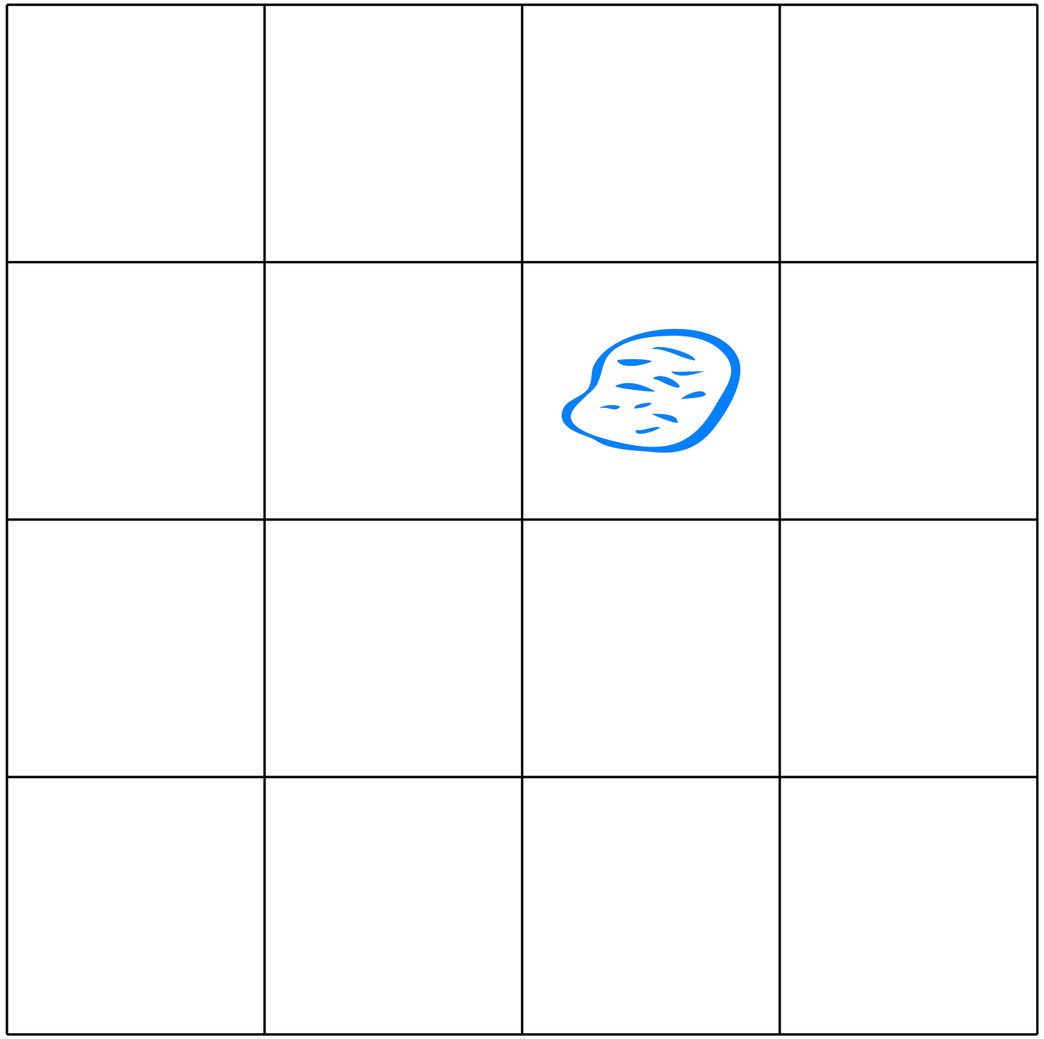
Grid-World

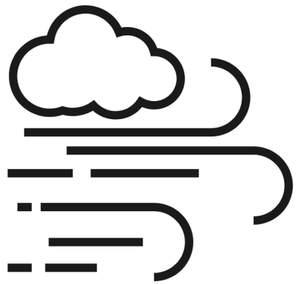
-1	-1	-1	-1
-1	-1	-10	-1
-1	-1	-1	-1
-1	-1	-1	100

Grid-World

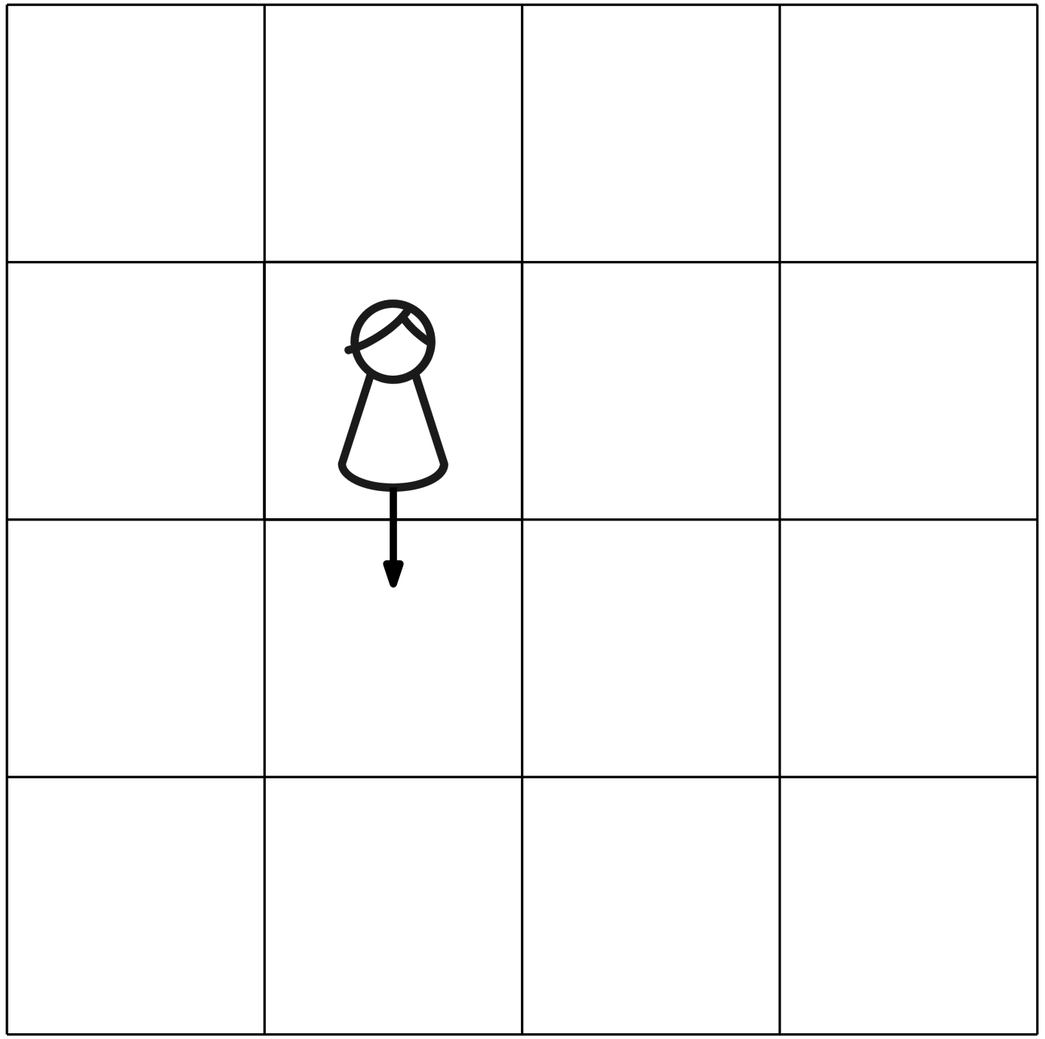


Grid-World

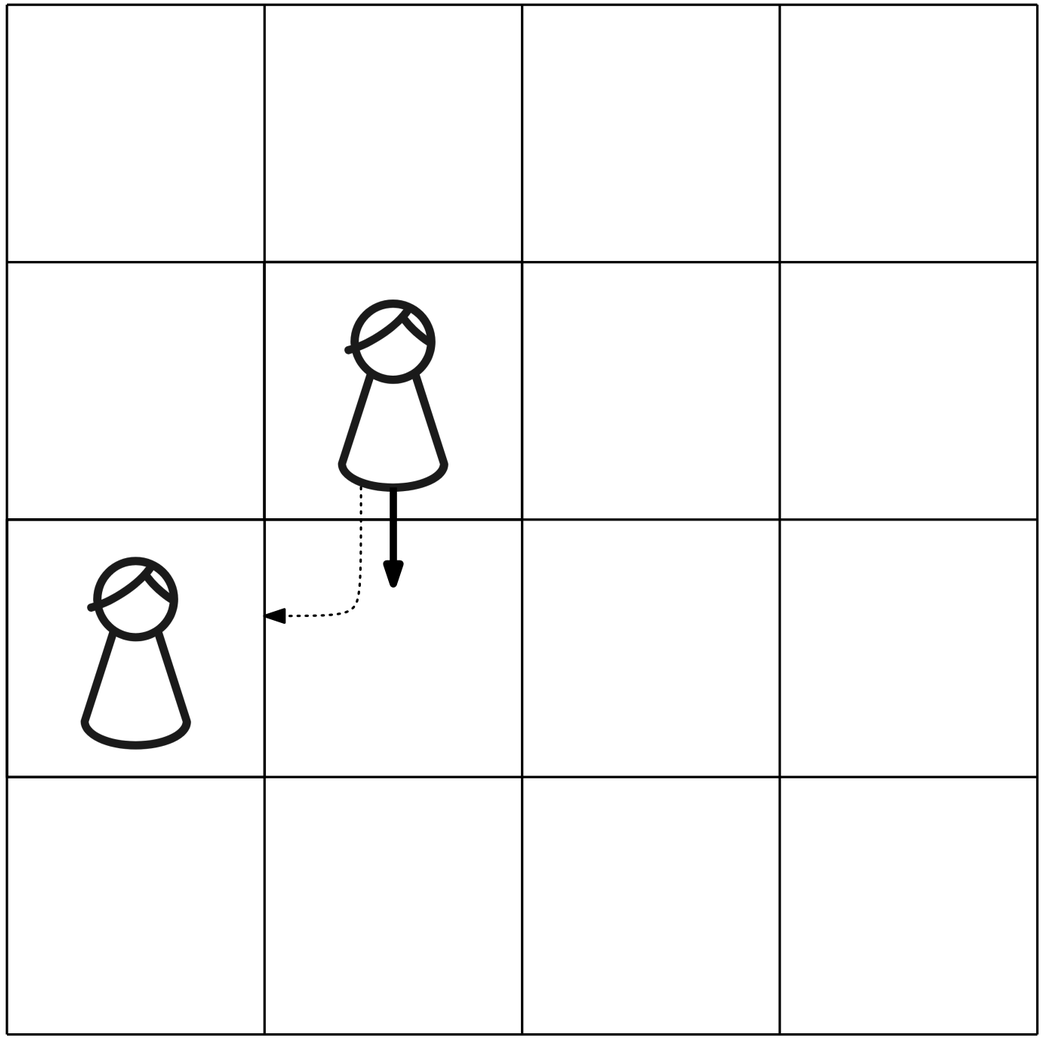




Grid-World

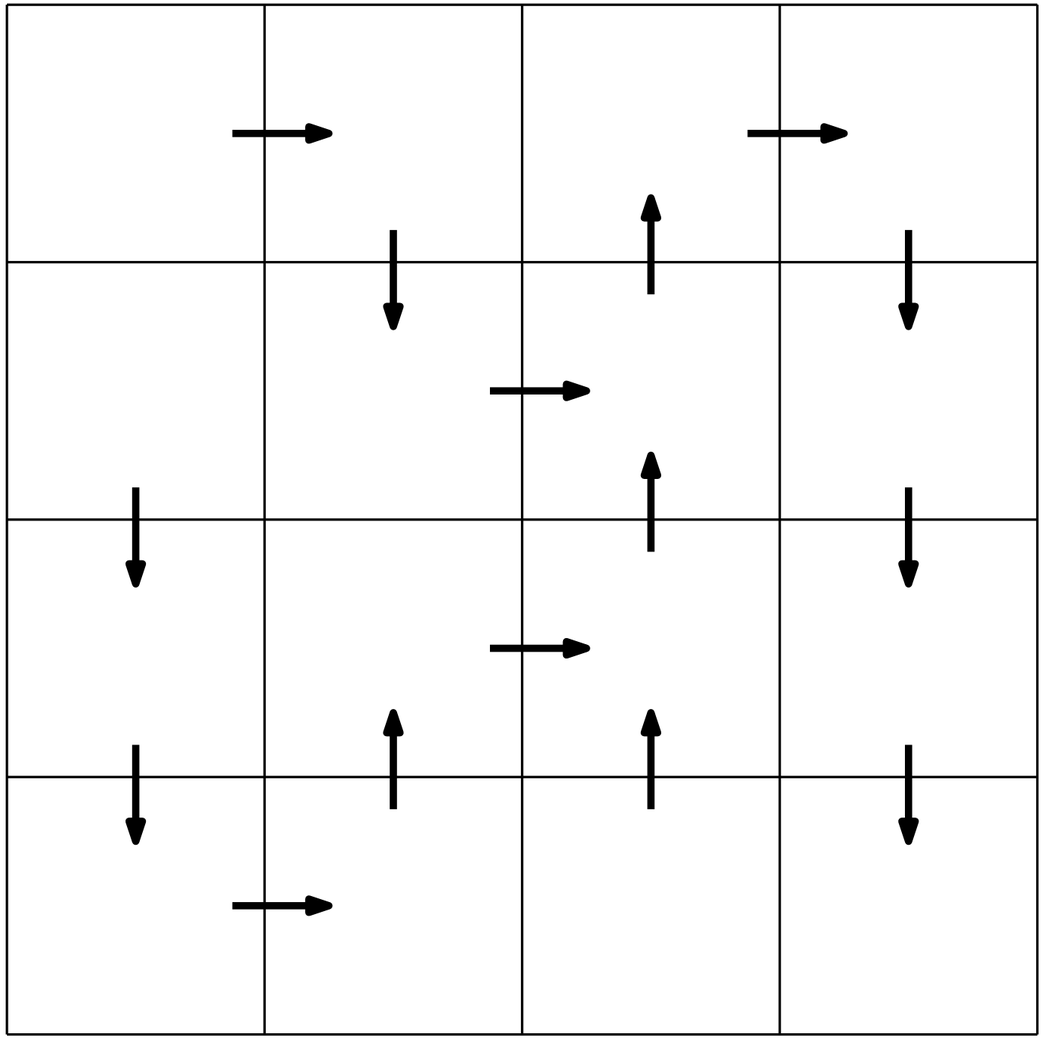


Grid-World

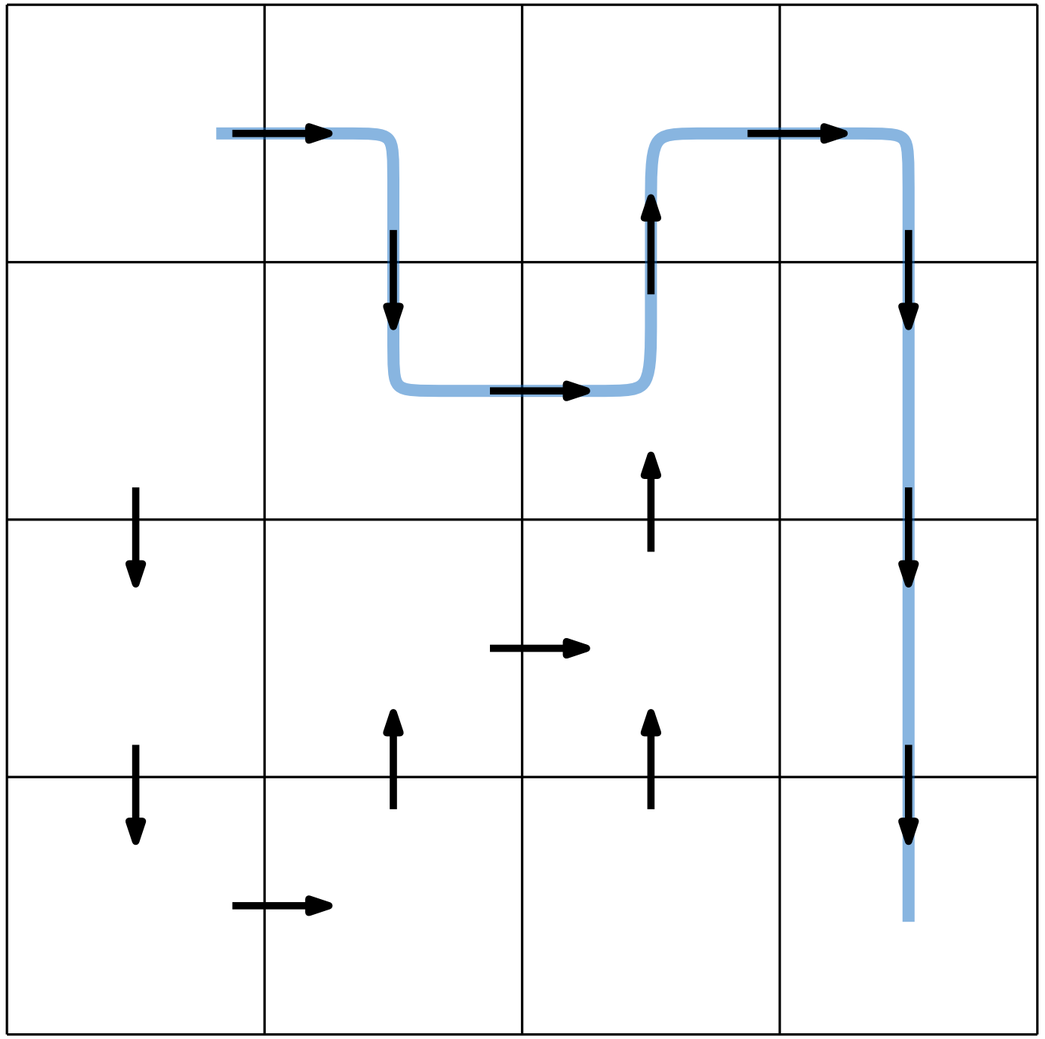


model-free

Grid-World Policy

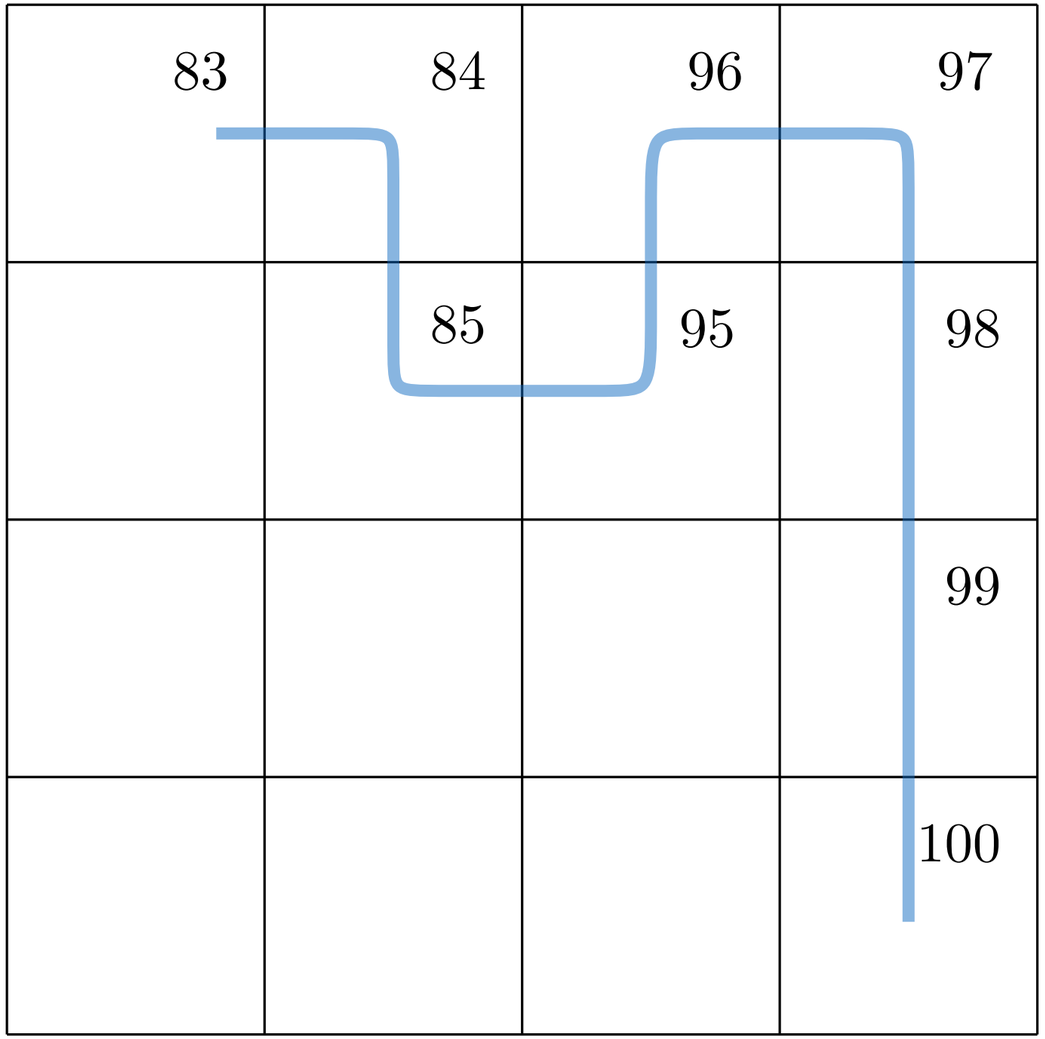


Grid-World Policy



Trajectory Sample

Grid-World

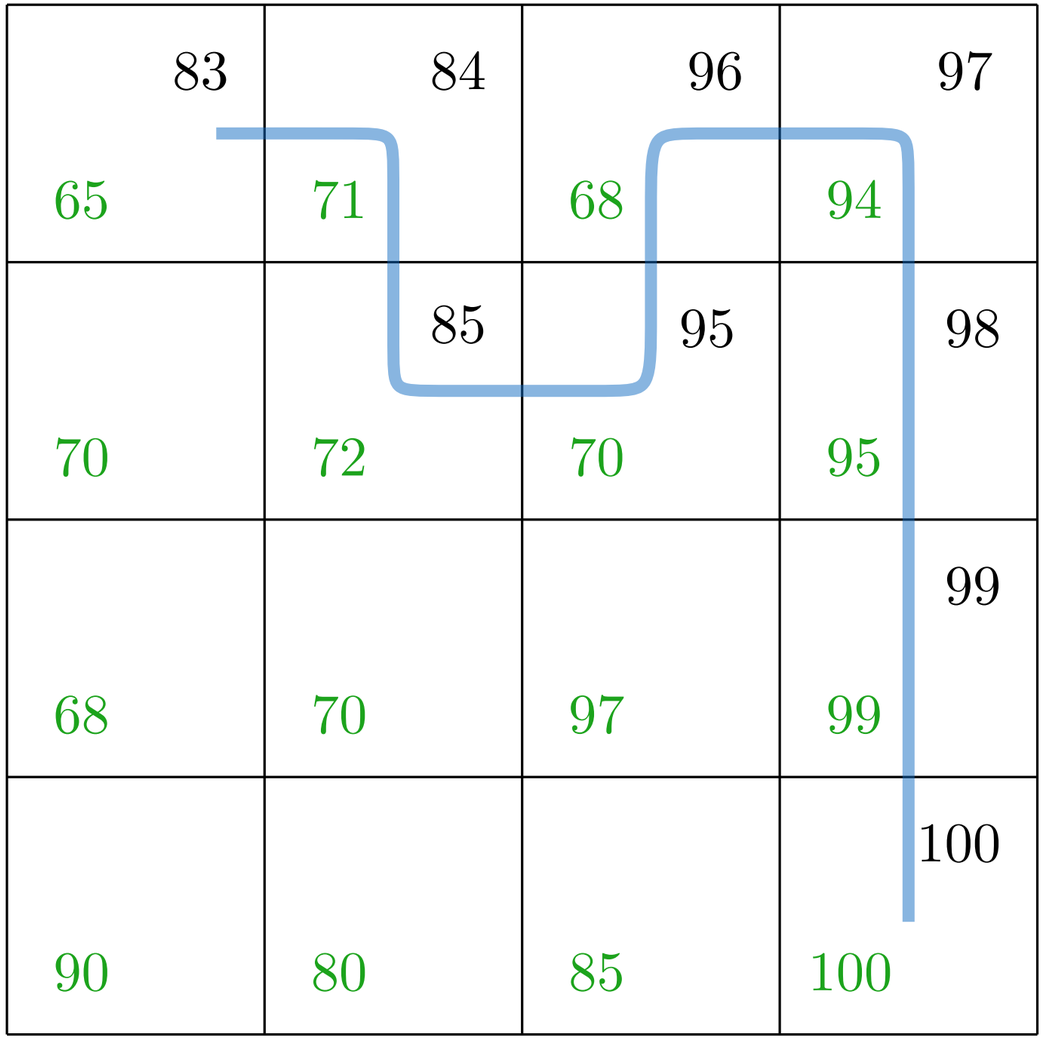


Return $\gamma = 1$

Trajectory

Sample

Grid-World



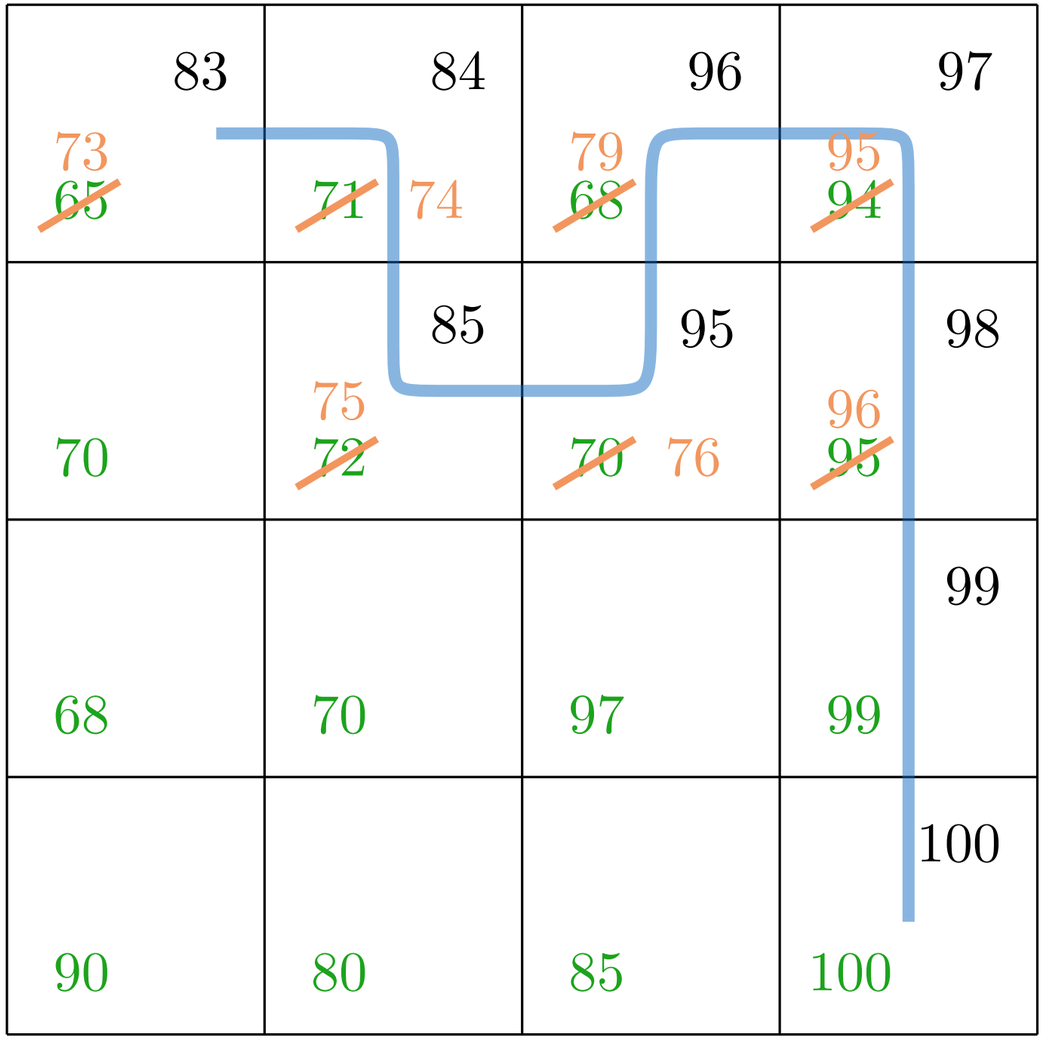
Return $\gamma = 1$

Value Estimate

Trajectory

Sample

Grid-World



Return $\gamma = 1$

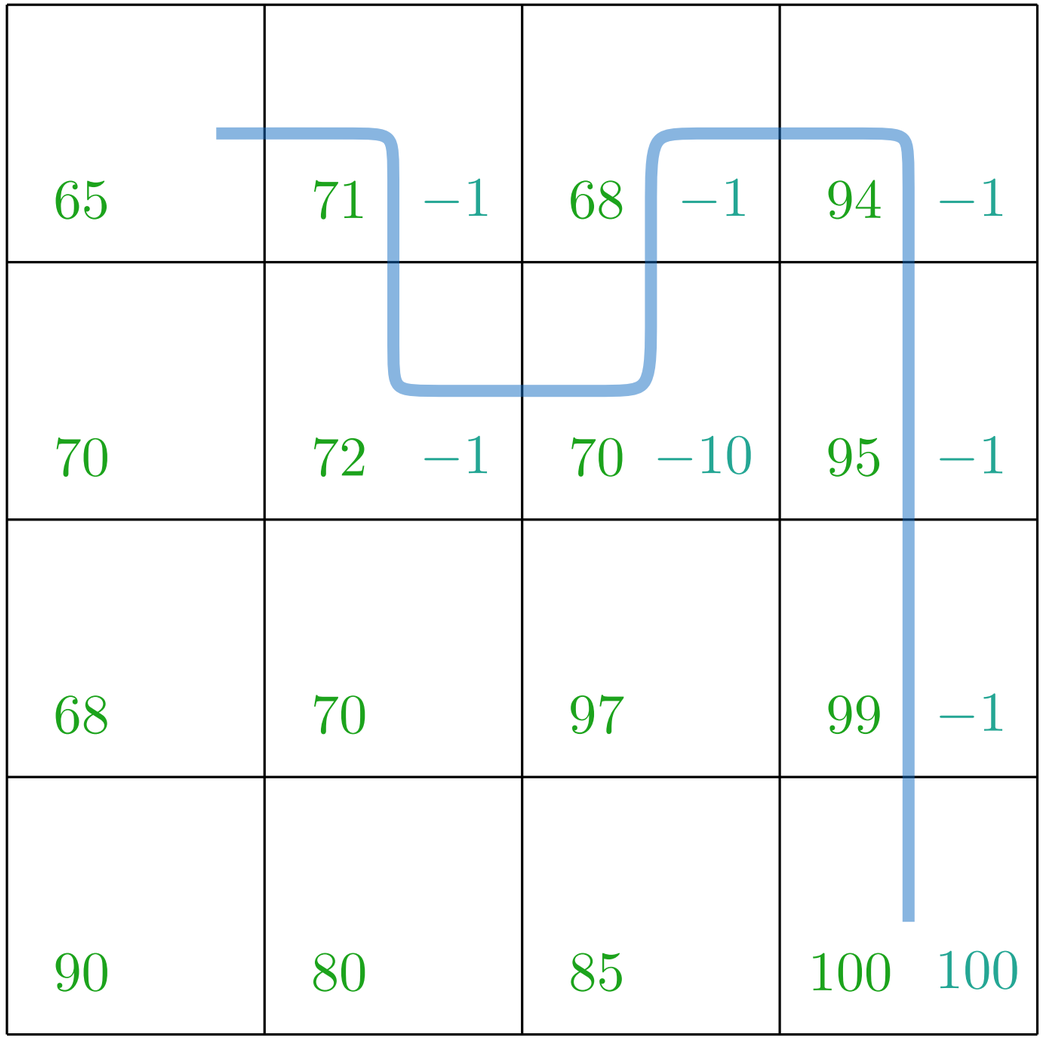
Value Estimate

MC-Update

Trajectory

Sample

Grid-World



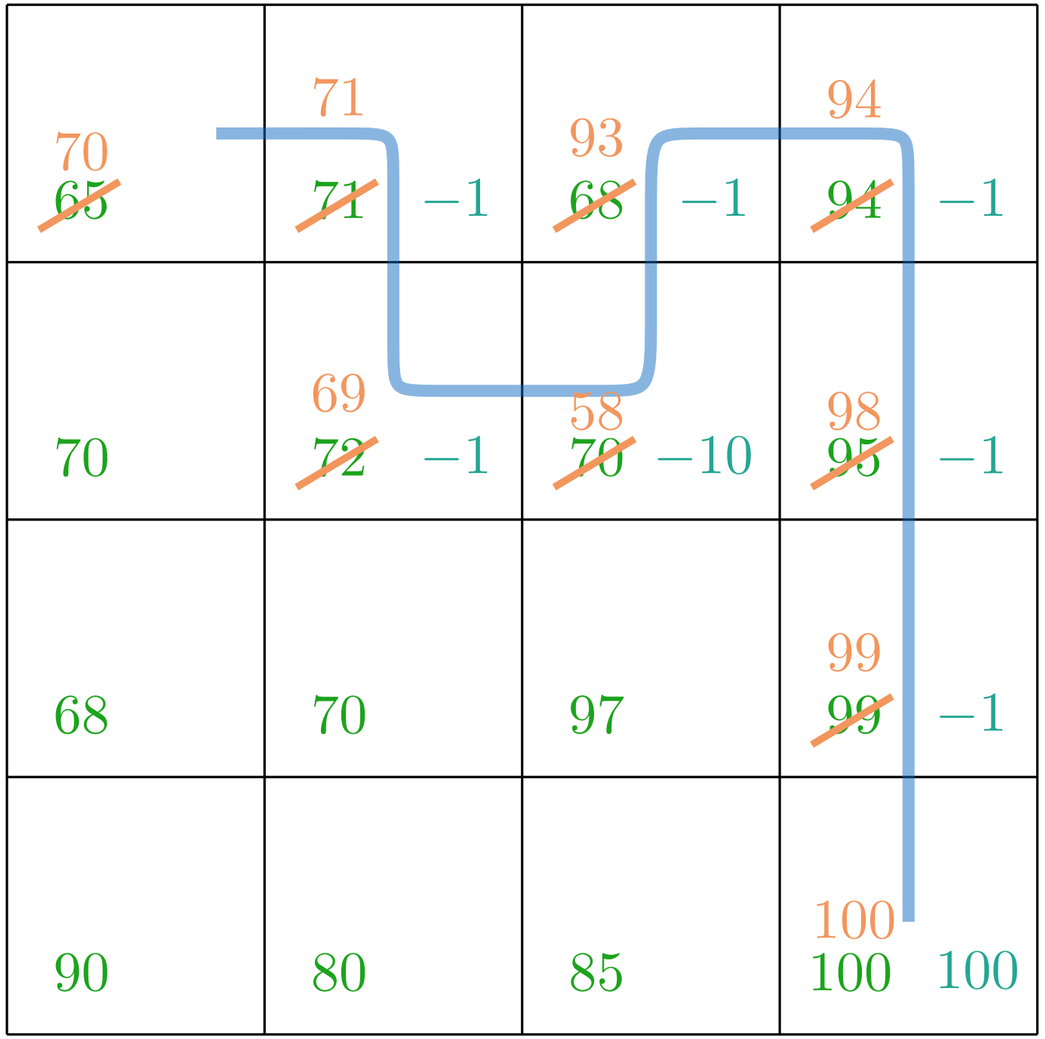
Rewards

Value Estimate

Trajectory

Sample

Grid-World



Rewards

Value Estimate

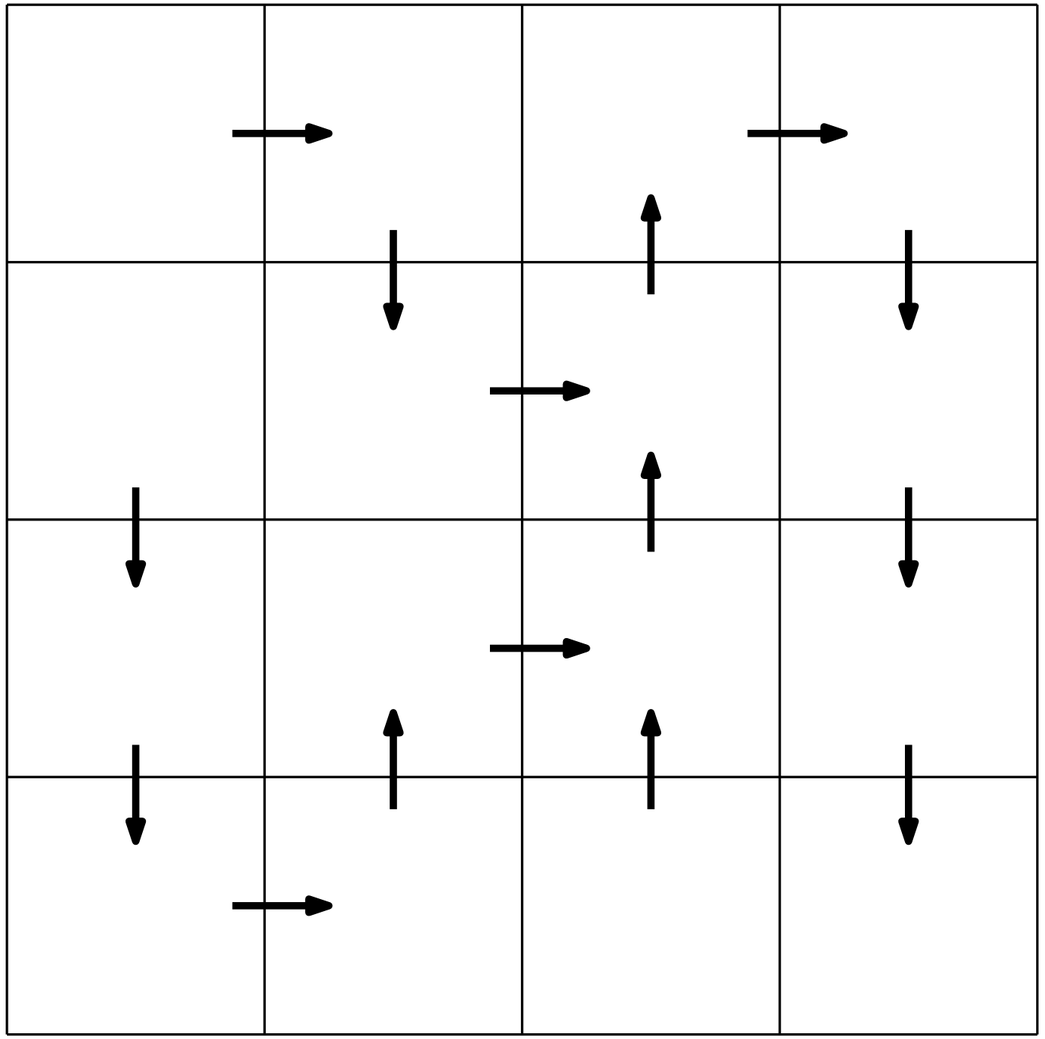
TD-Update

$\alpha = 1$

Trajectory

Sample

Grid-World Policy

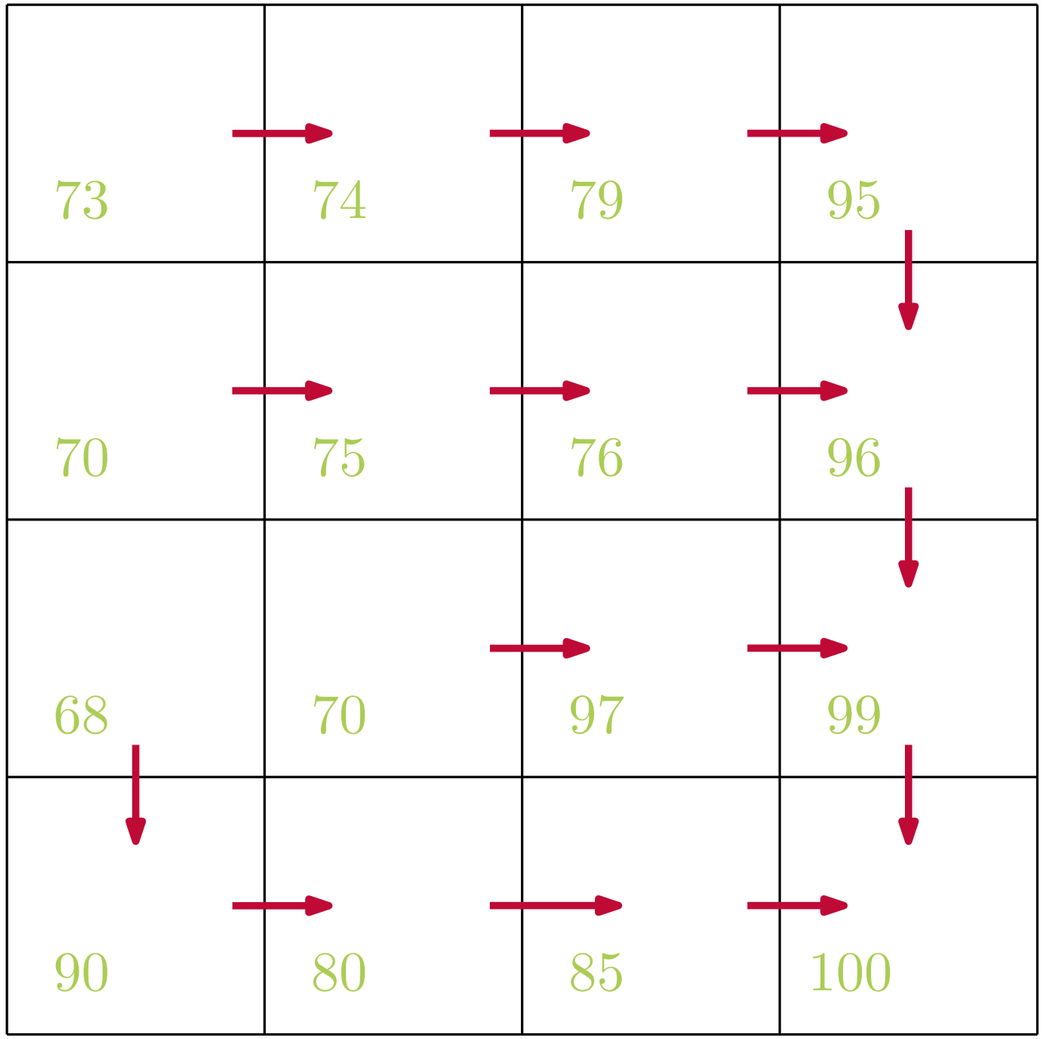


Grid-World

73	74	79	95
70	75	76	96
68	70	97	99
90	80	85	100

New Estimate

Grid-World

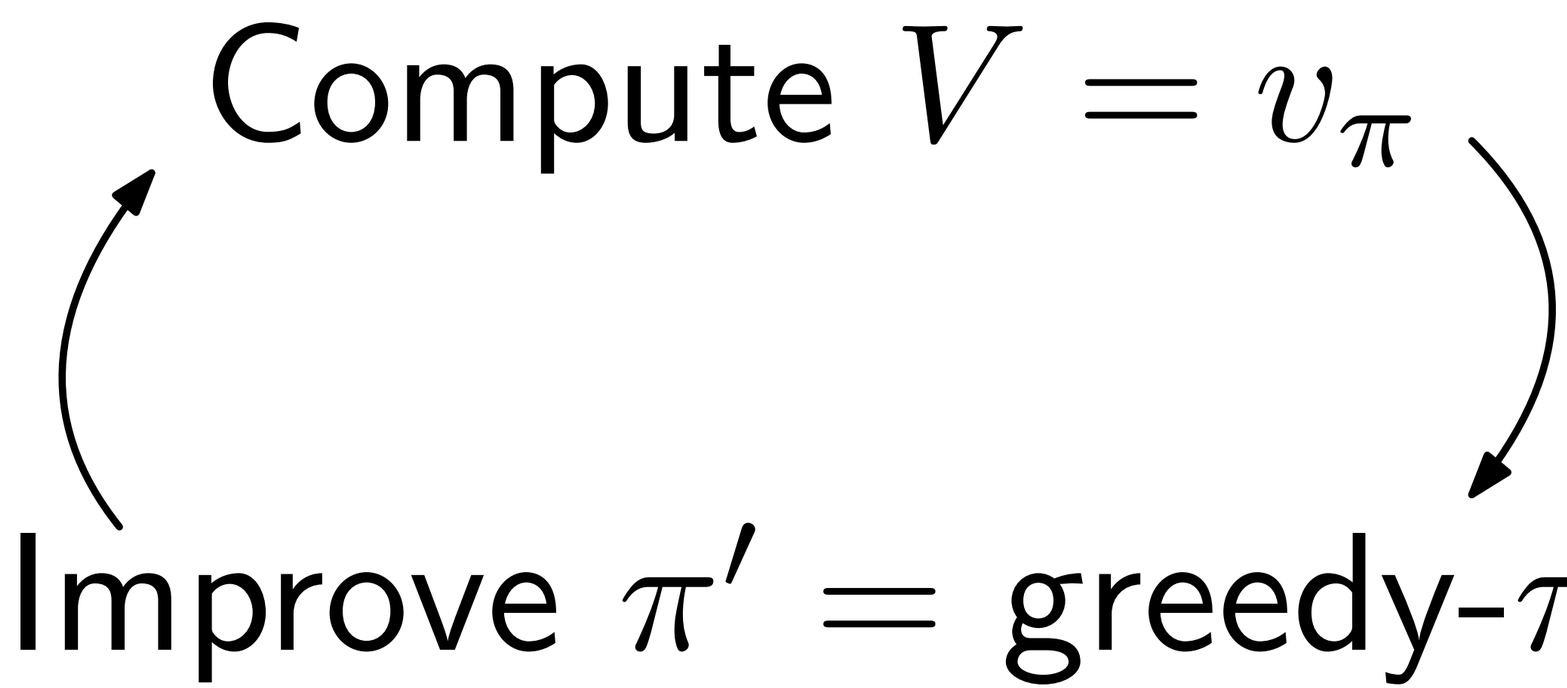


New Estimate

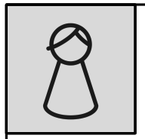
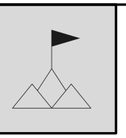
Greedy - Policy

Policy-Iteration

Policy-Iteration



Assignment 1

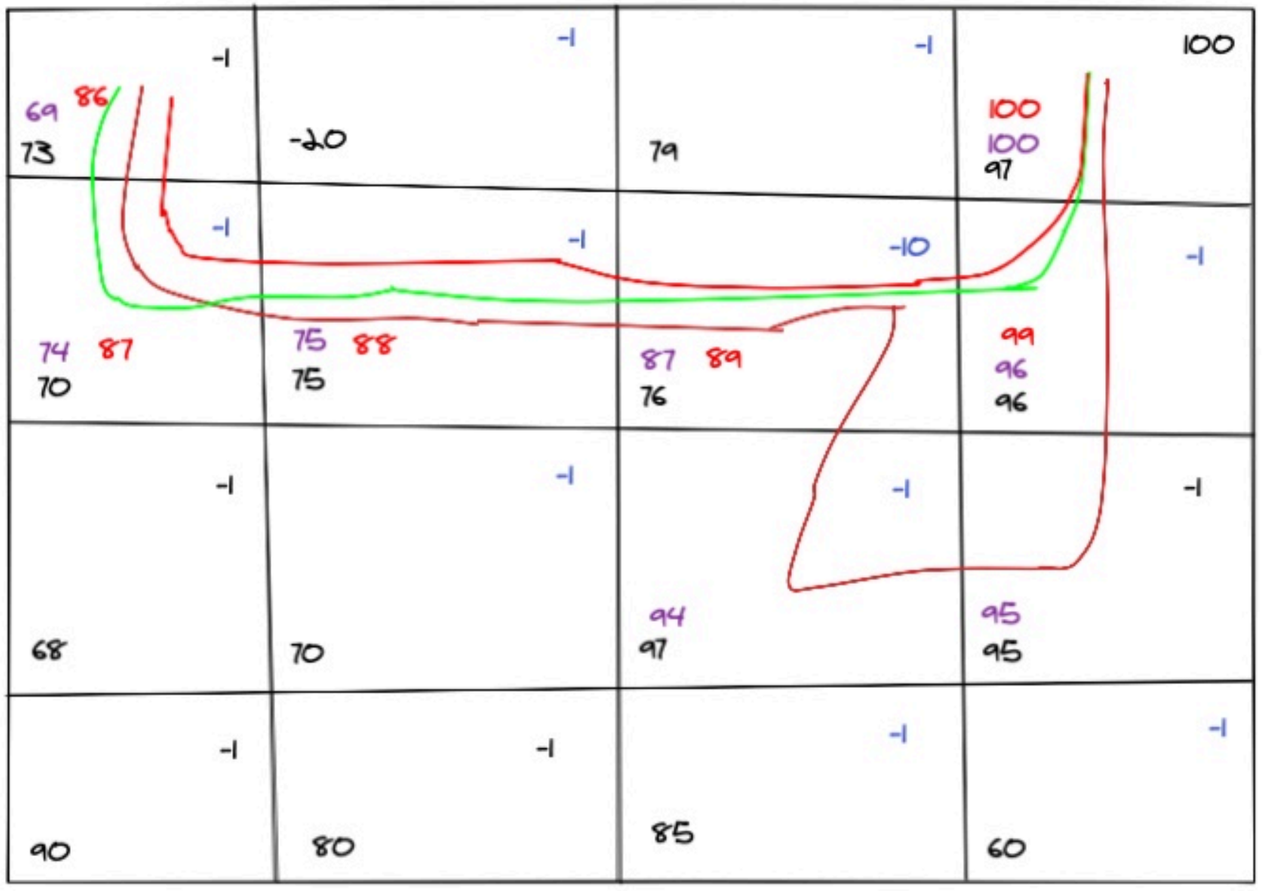
 -1	-1	-1	 100
73	-20	79	97
-1	-1	-10	-1
70	75	76	96
-1	-1	-1	-1
68	70	97	95
-1	-1	-1	-1
90	80	85	60

Value-Estimate

Rewards

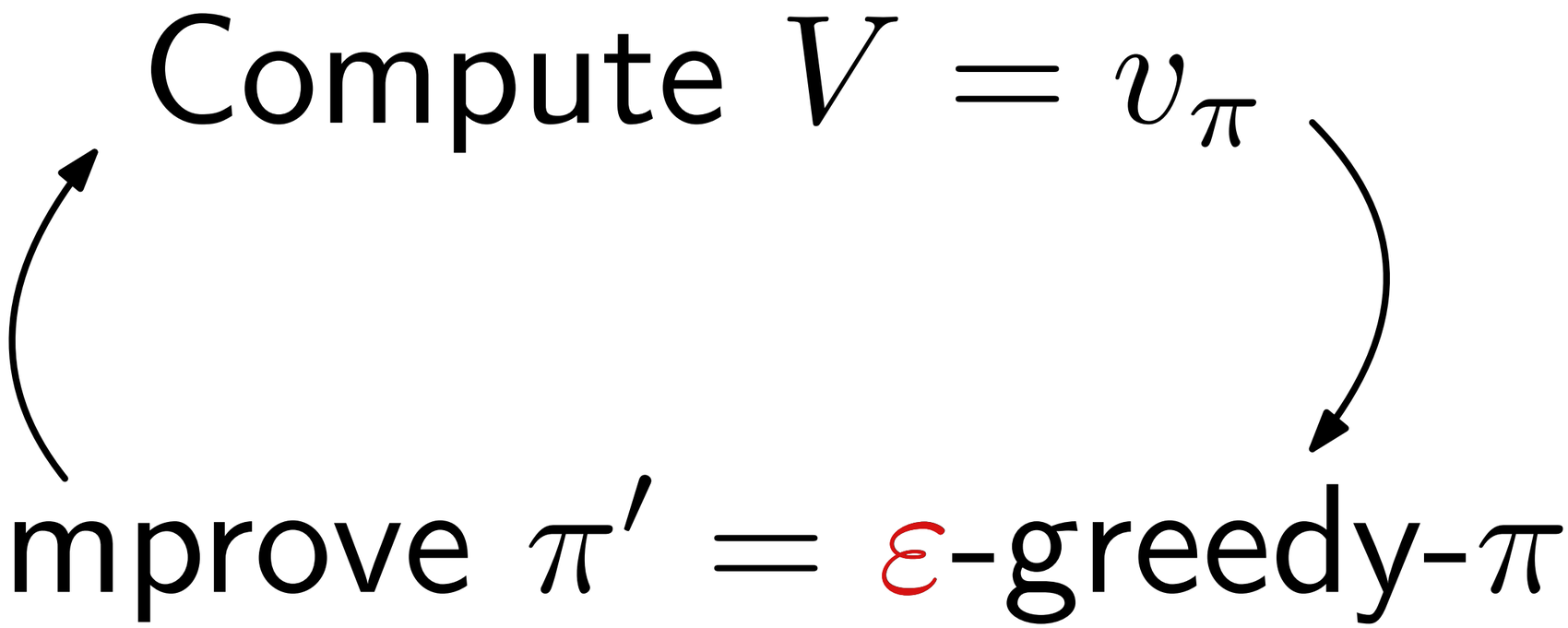
Where do we converge to?
Is it π^*?
How can you fix this?

Post on Teams


-1
69 86
73
-1
-20
-1
79
100
100
97
-1
74 87
70
-1
75 88
75
-10
87 89
76
-1
99
96
96
-1
68
-1
70
-1
94
97
-1
95
95
-1
90
-1
80
-1
85
-1
60

Policy-Iteration

Policy-Iteration



Assignment 2

-1	-10	-10	100
-1	-1	-1	-1
-1	-1	-1	-1
-1	-1	-1	-1

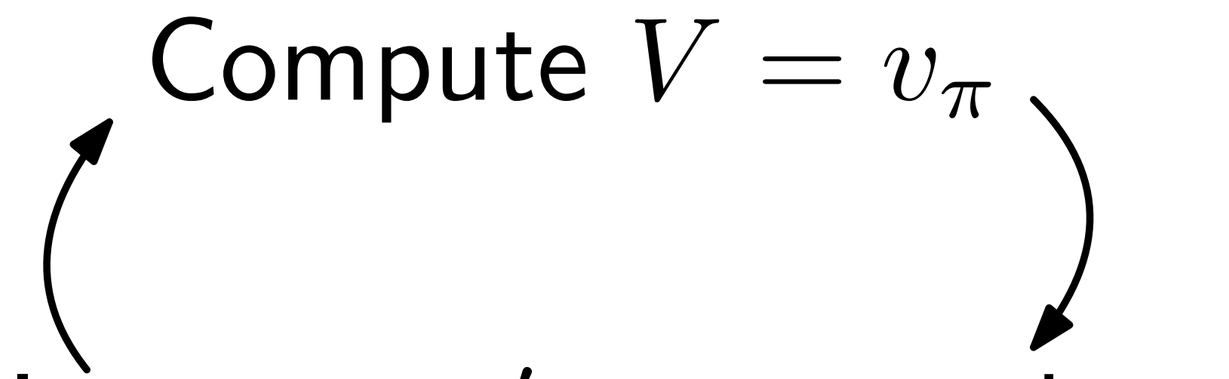
Rewards

Where do we converge to?
Is it π^*?
How can you fix this?

Post on Teams

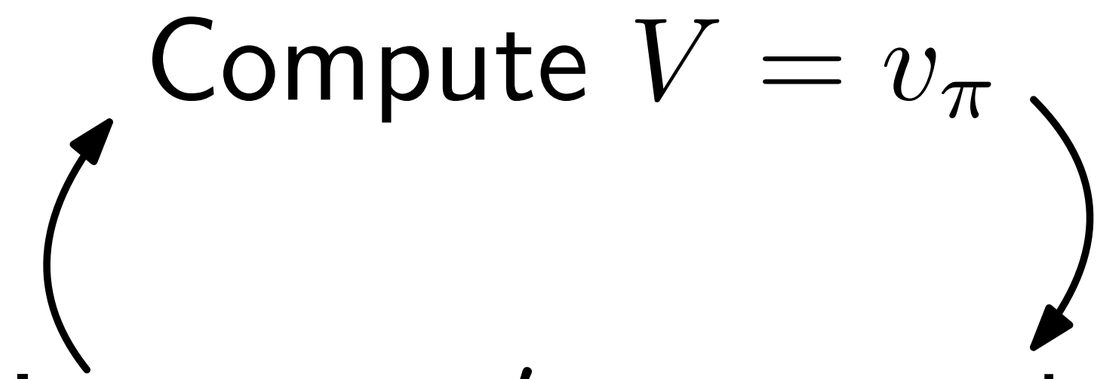
Policy-Iteration

Policy-Iteration



ε-decay

Policy-Iteration



ε-decay

too low $\rightarrow$ not exploring

too high $\rightarrow$ optimize objective

Assignment 3

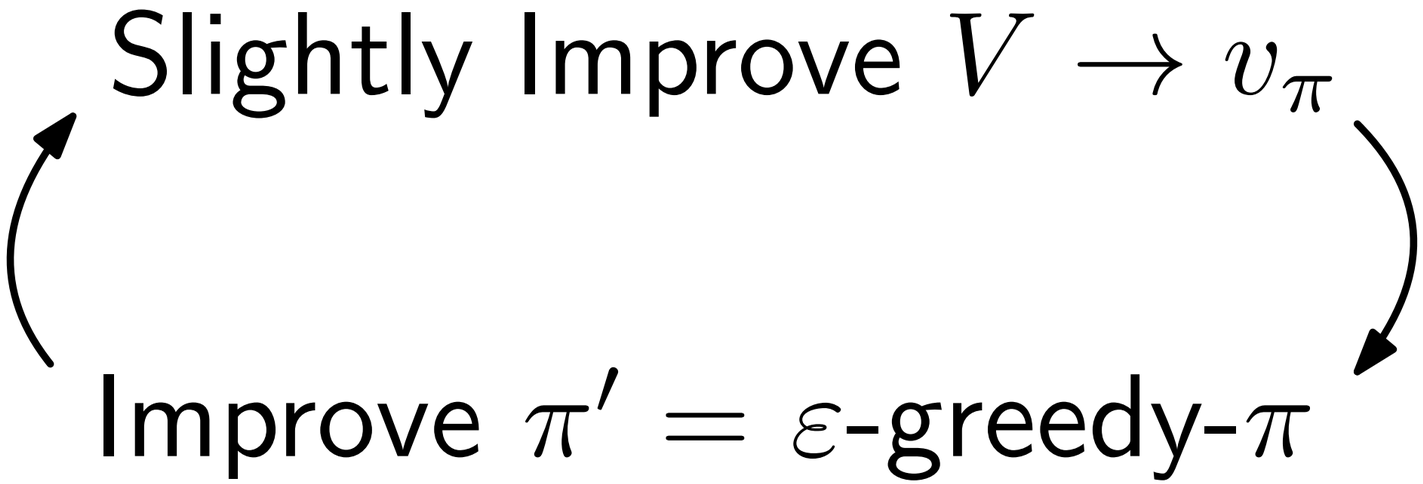
Converging to v_π takes very long.

How can we speed-up the policy iteration?

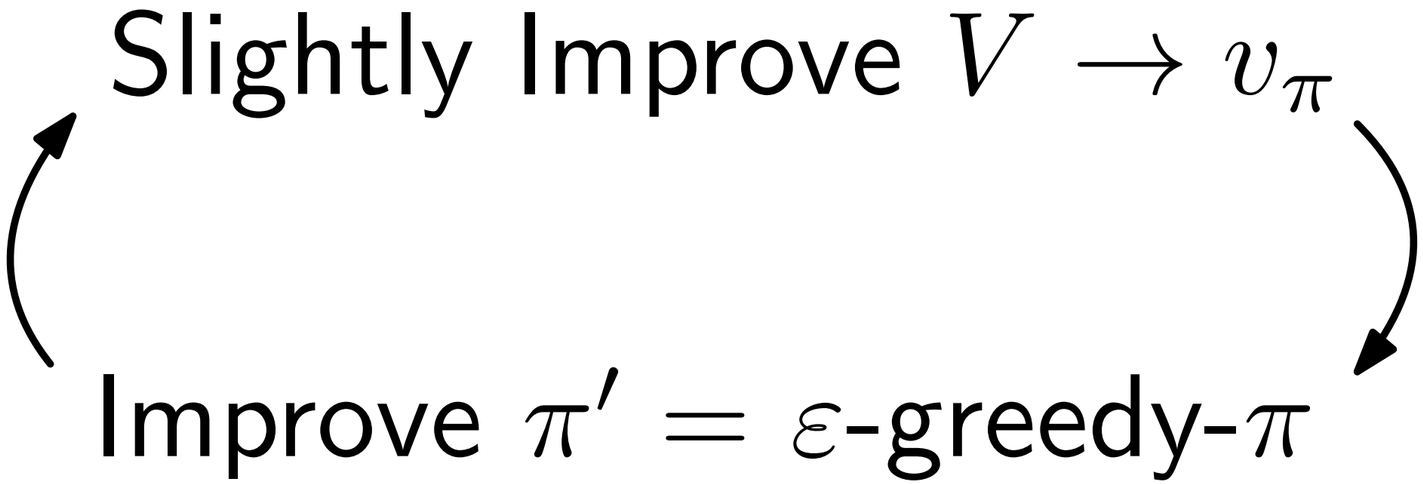
Post on Teams

Policy-Iteration

Policy-Iteration

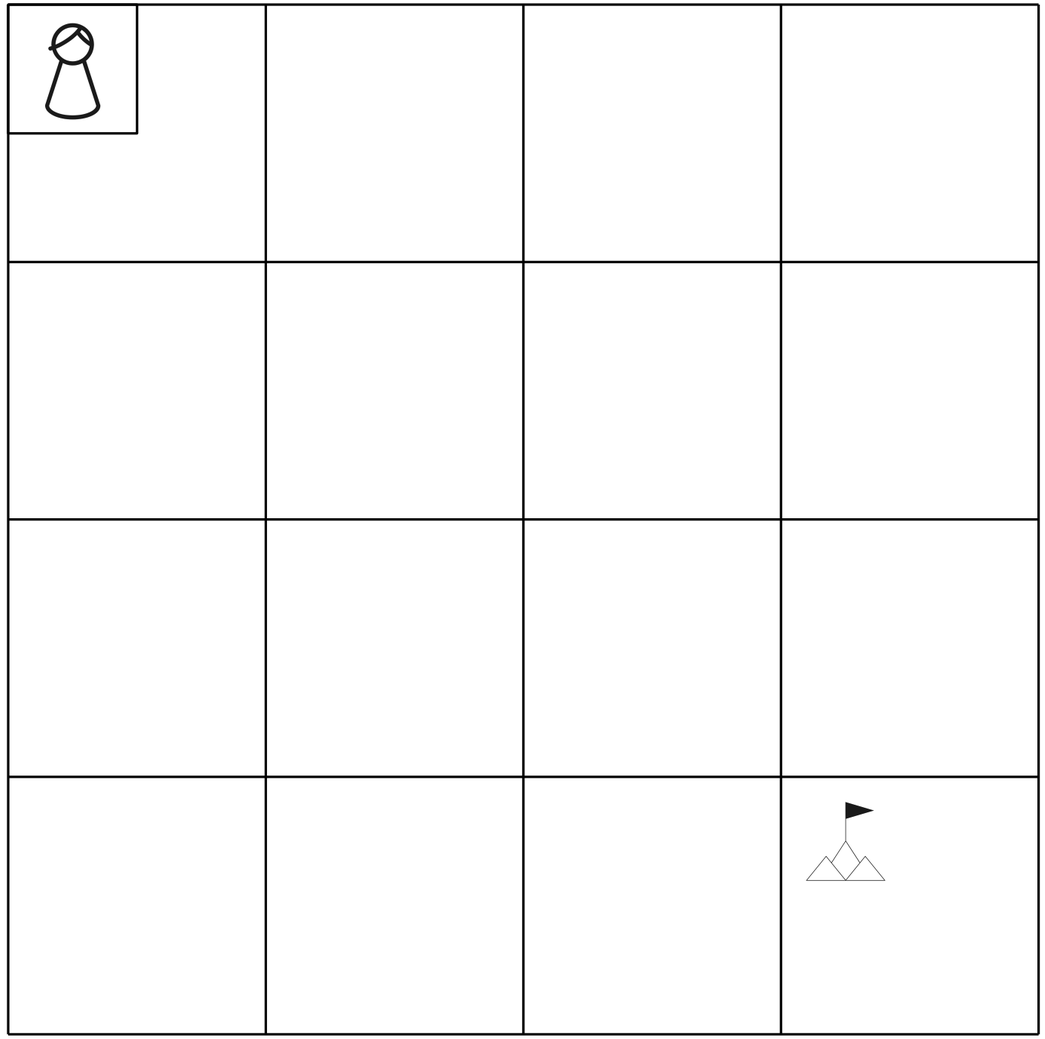


Policy-Iteration



Only one step using TD-learning

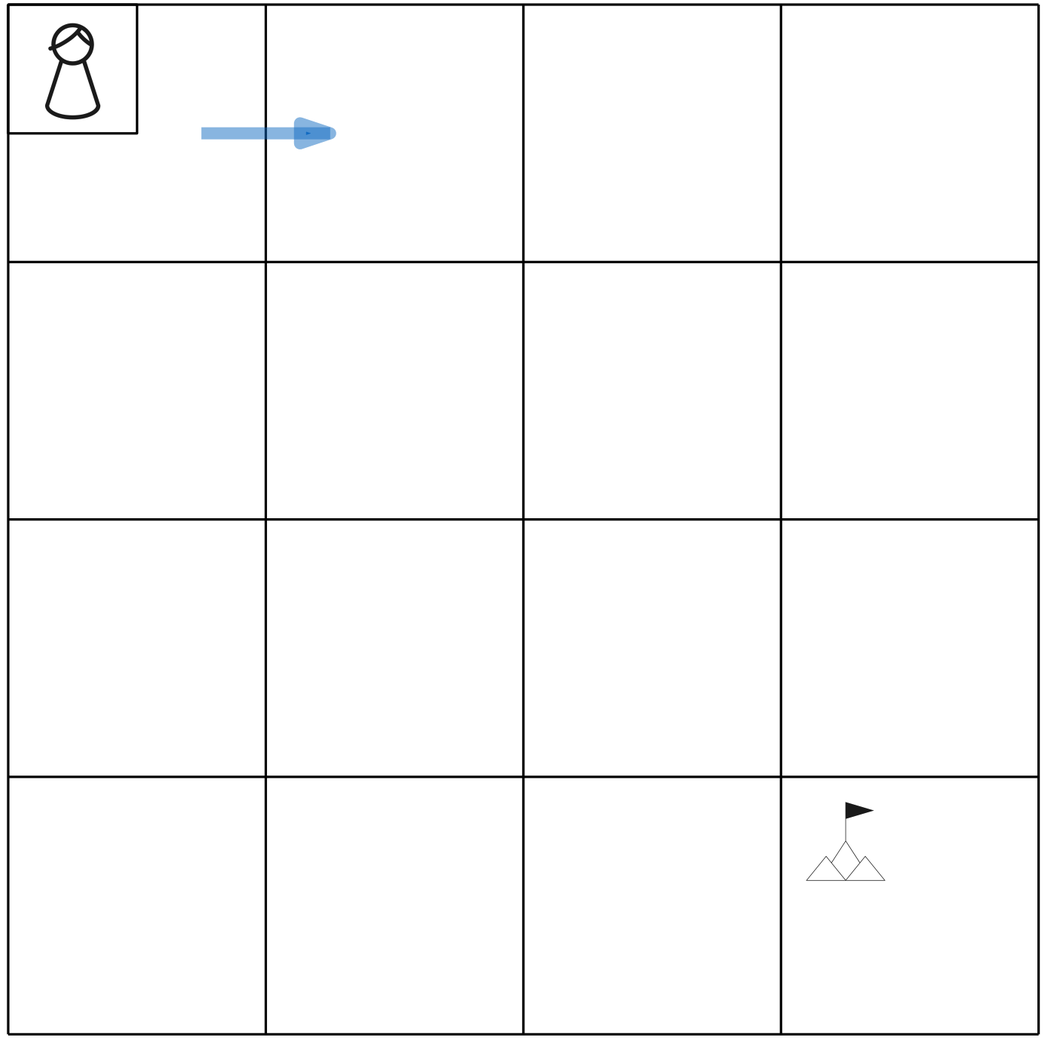
Policy-Iteration



Sample

Update

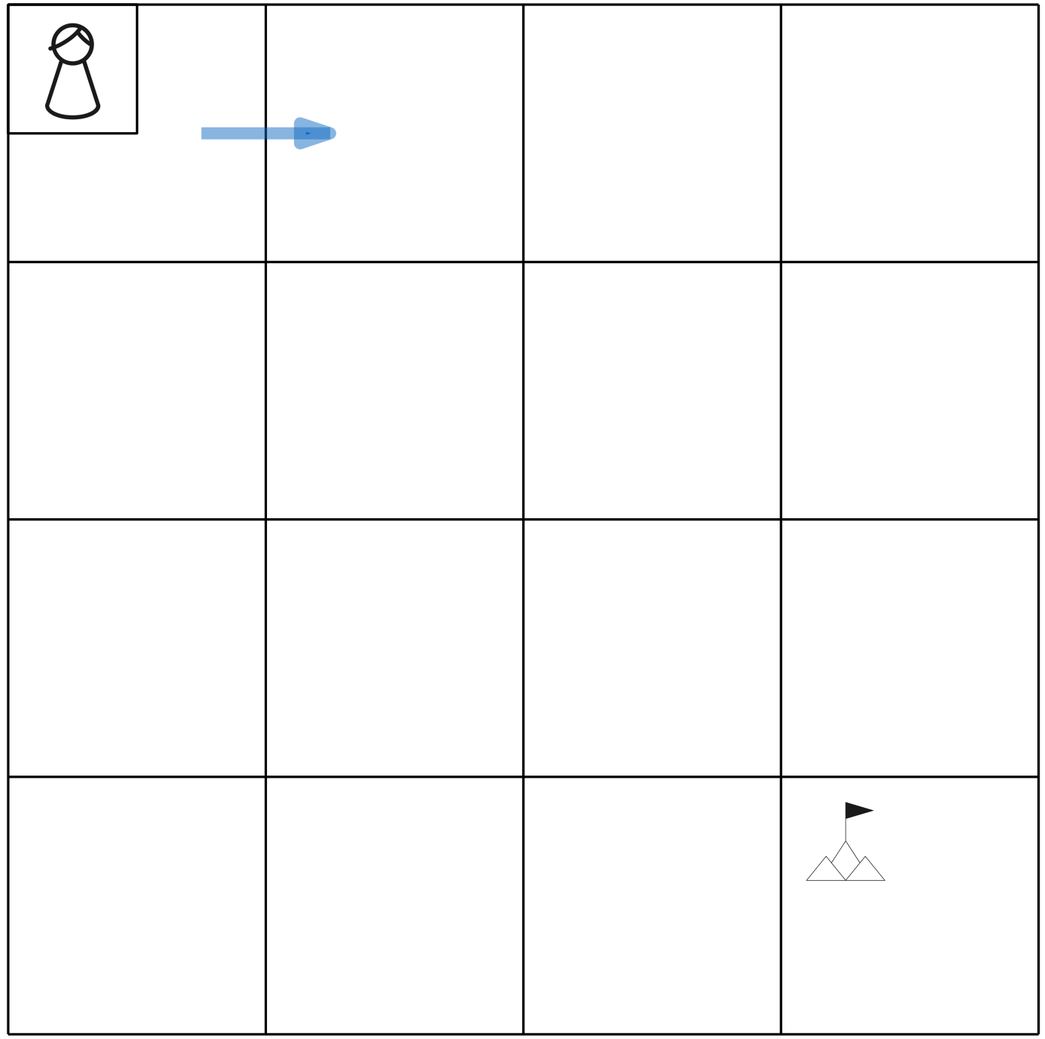
Policy-Iteration



Sample

Update

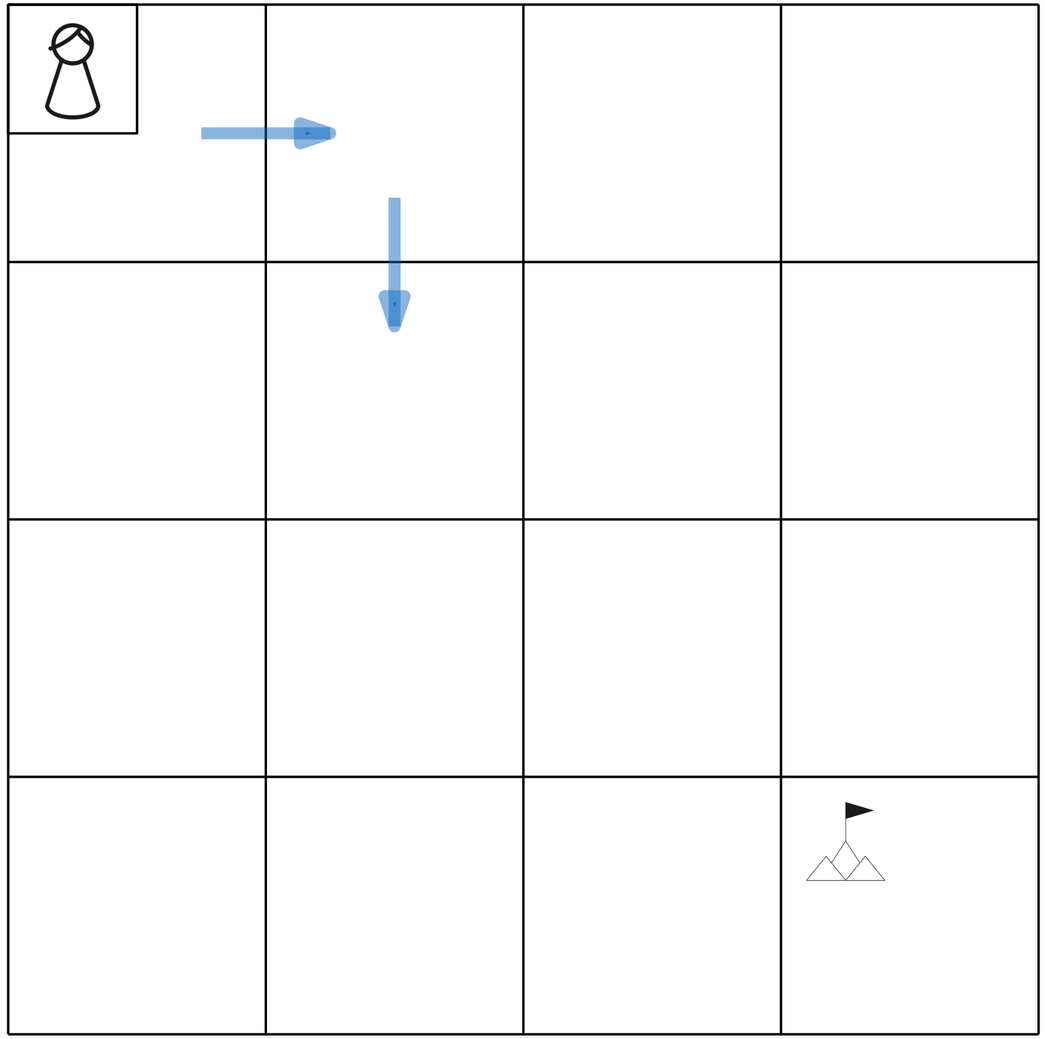
Policy-Iteration



Sample

Update

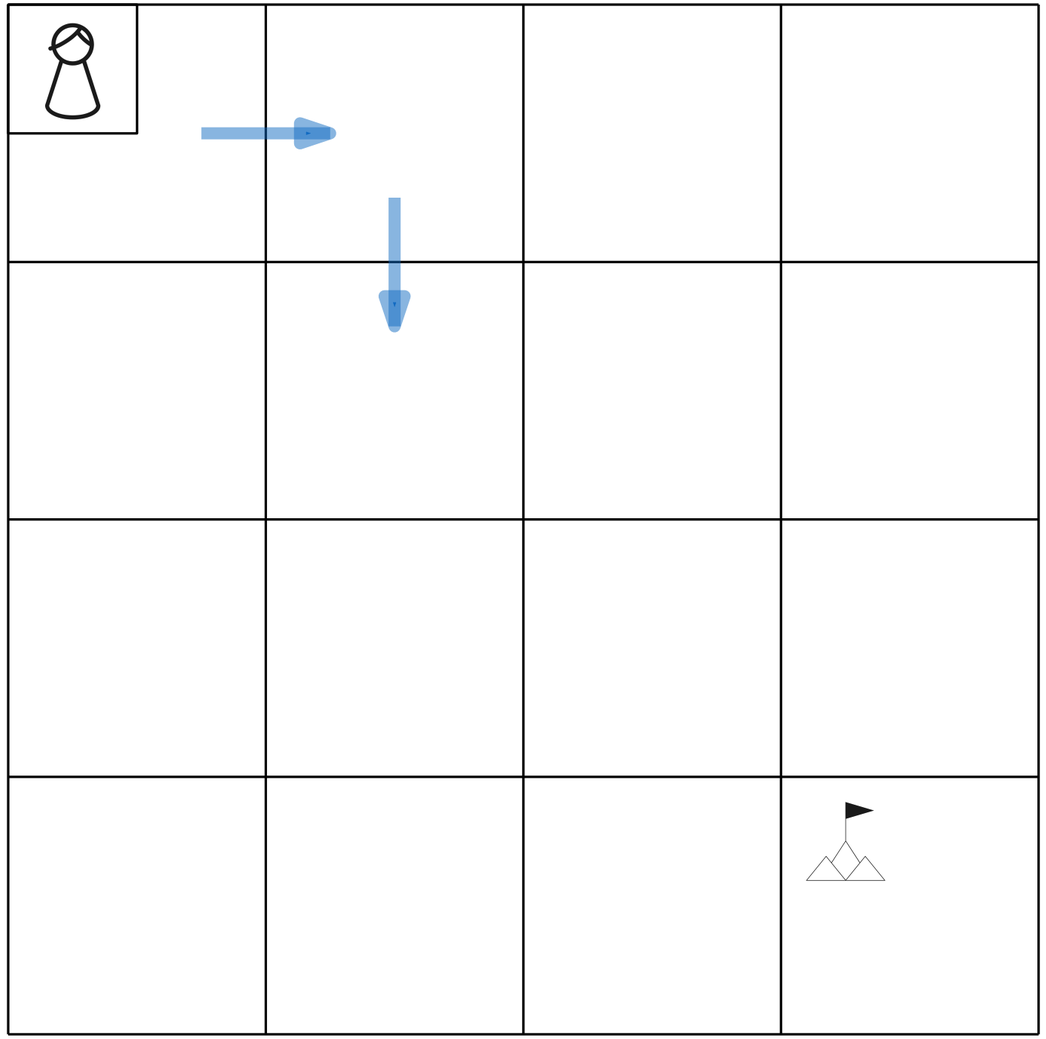
Policy-Iteration



Sample

Update

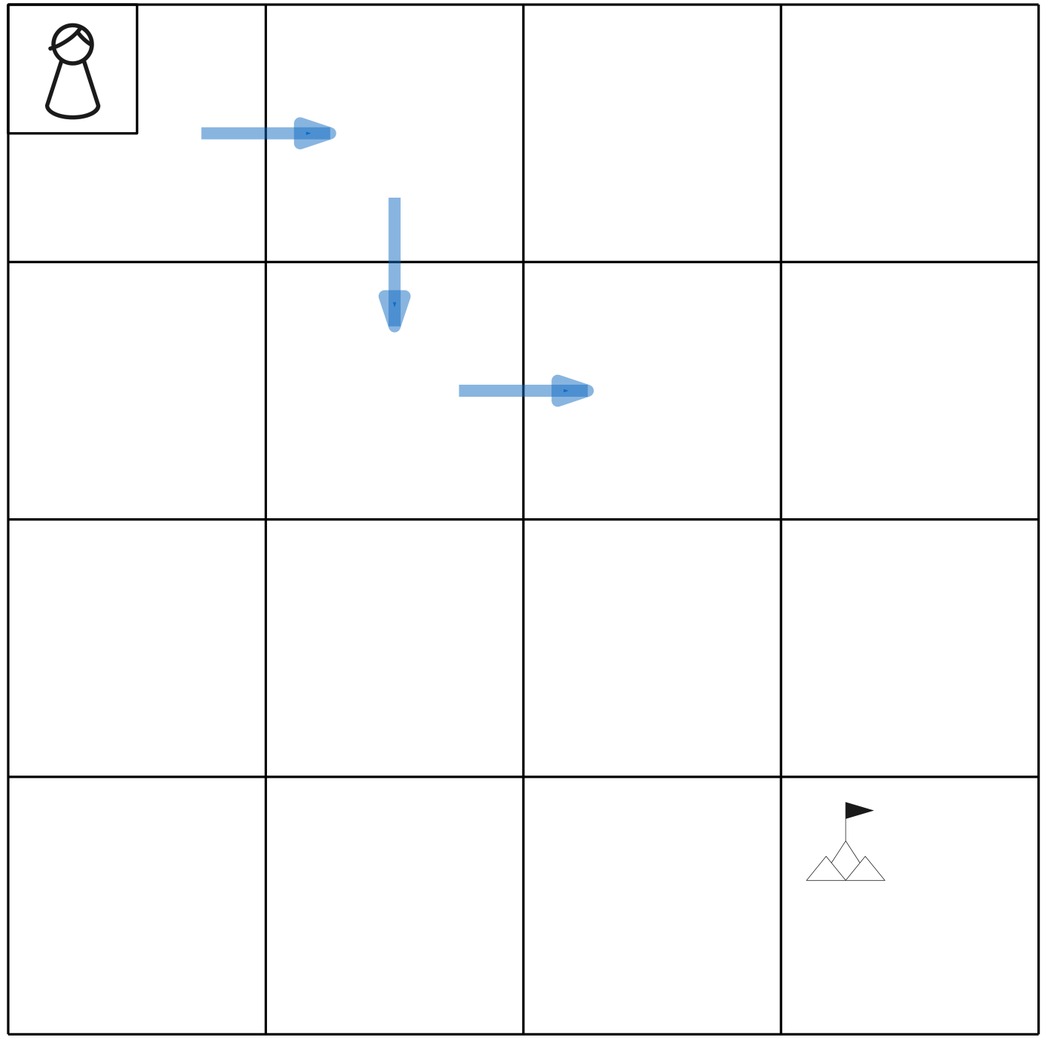
Policy-Iteration



Sample

Update

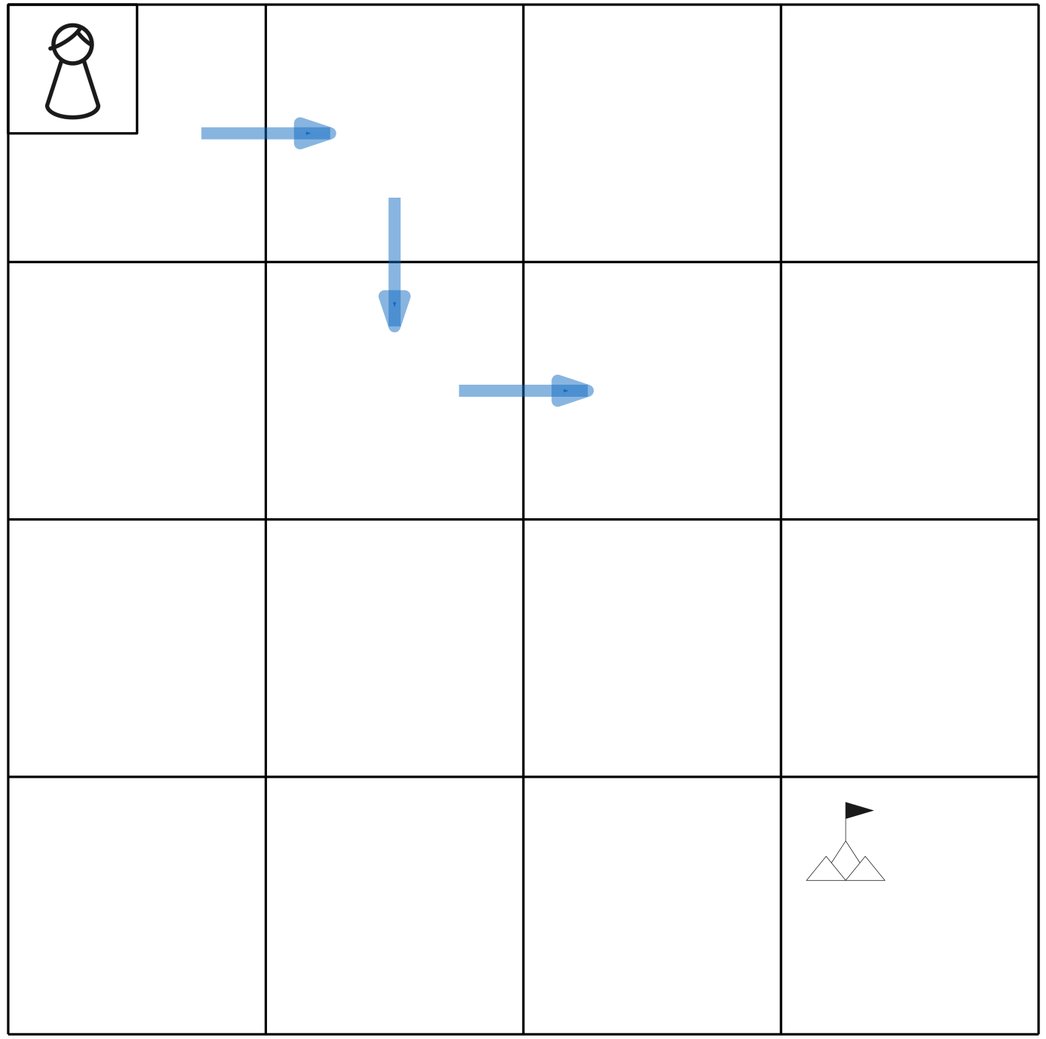
Policy-Iteration



Sample

Update

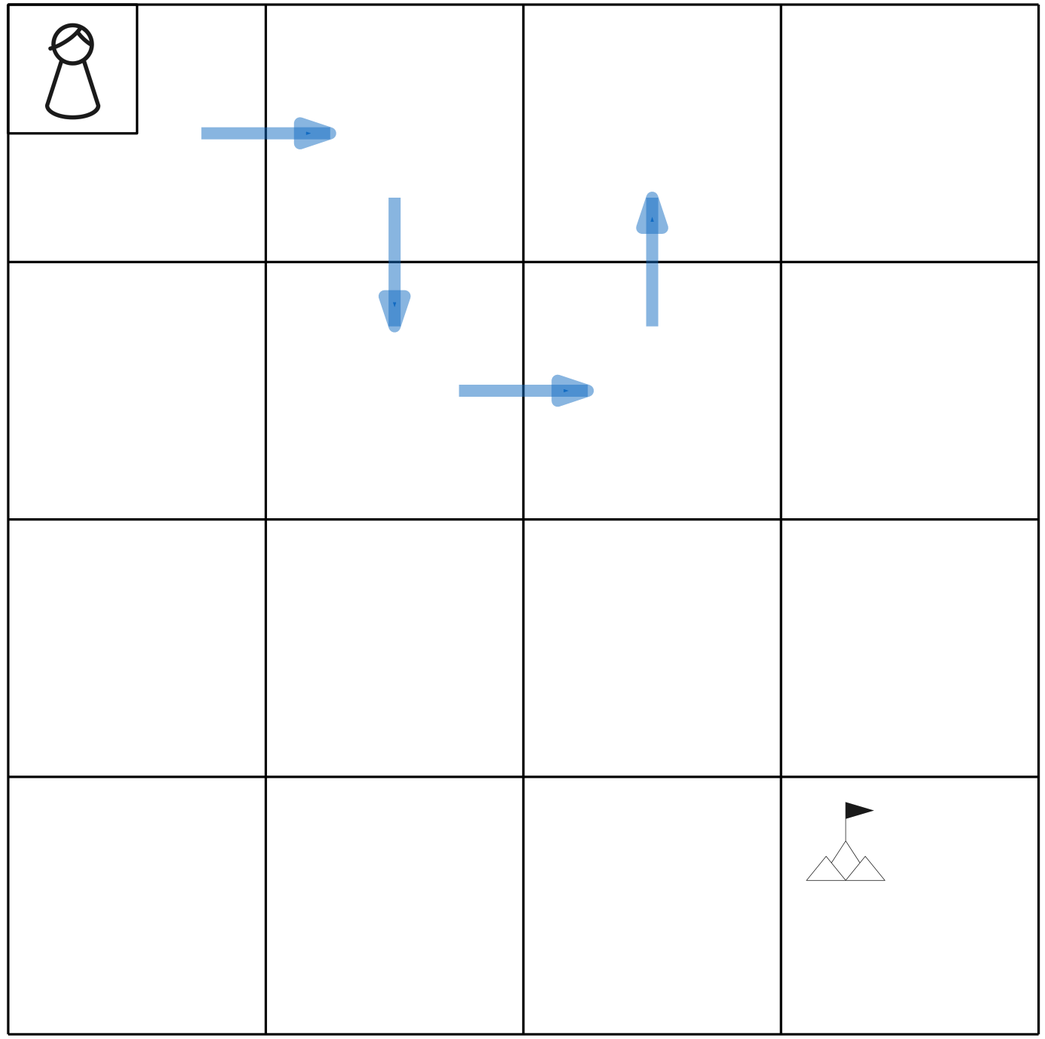
Policy-Iteration



Sample

Update

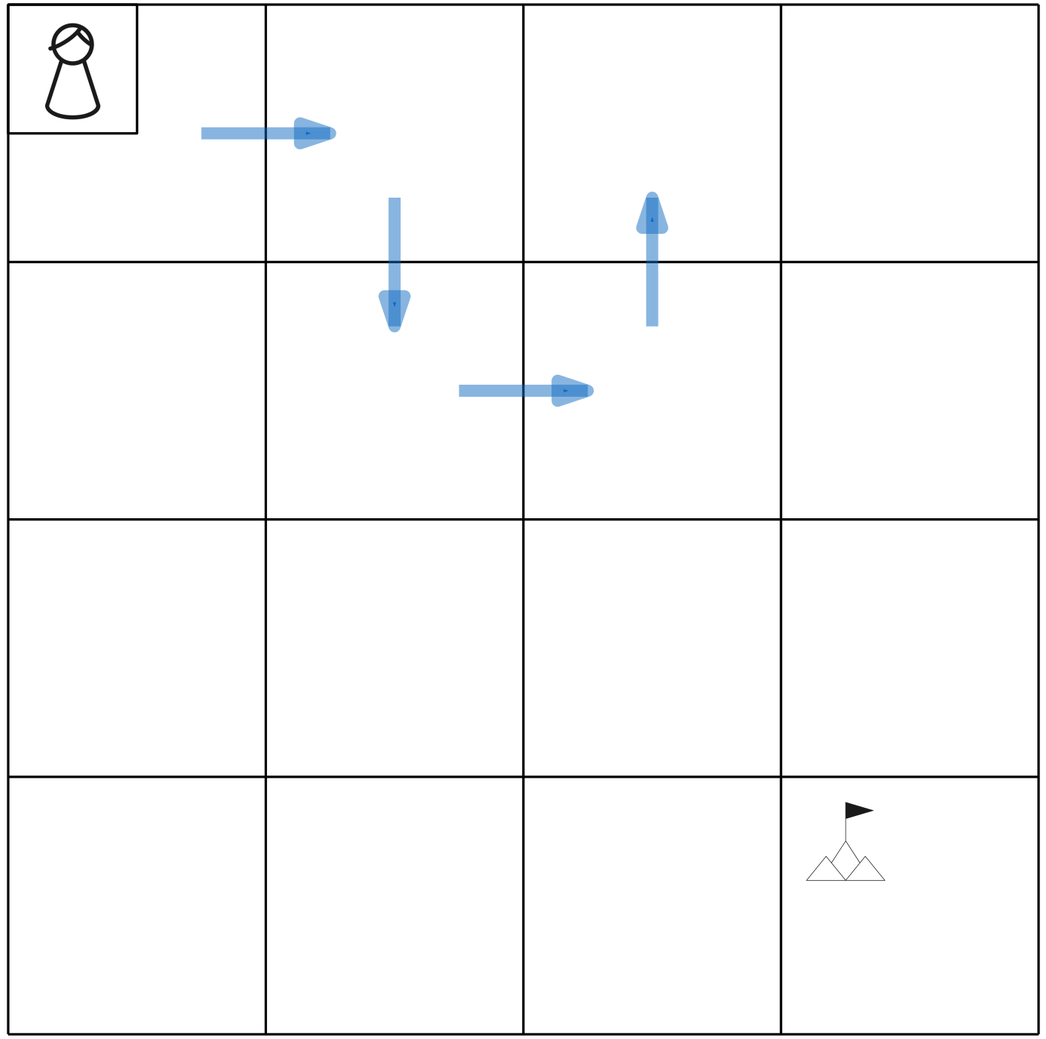
Policy-Iteration



Sample

Update

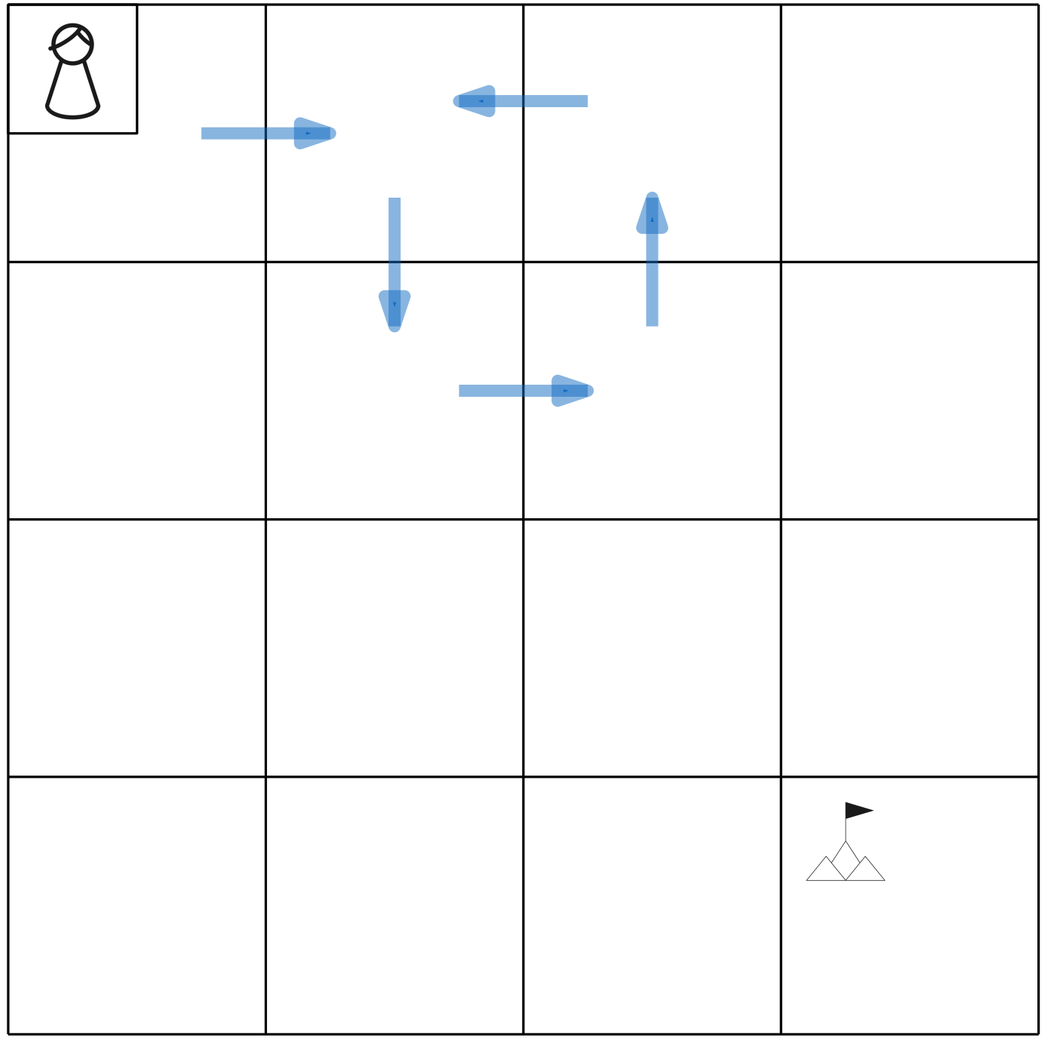
Policy-Iteration



Sample

Update

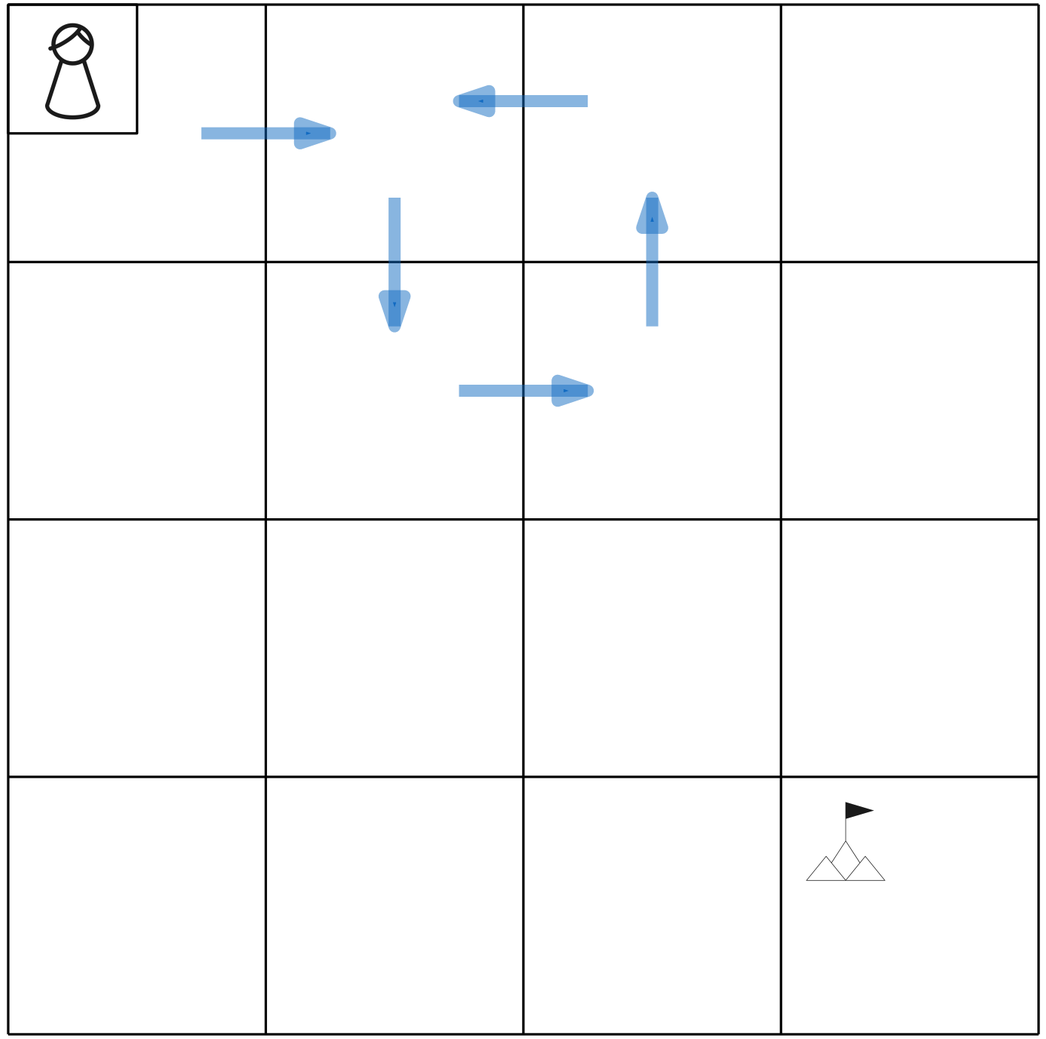
Policy-Iteration



Sample

Update

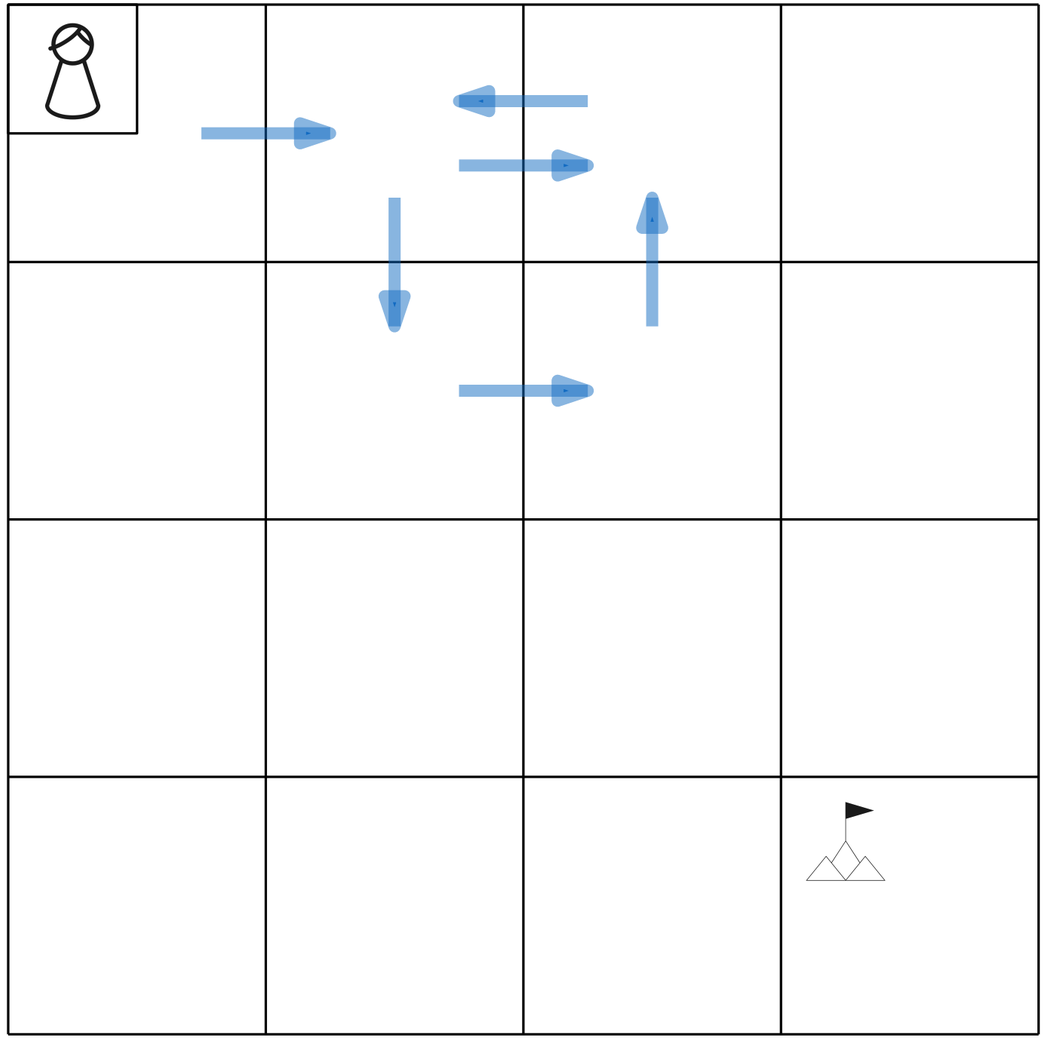
Policy-Iteration



Sample

Update

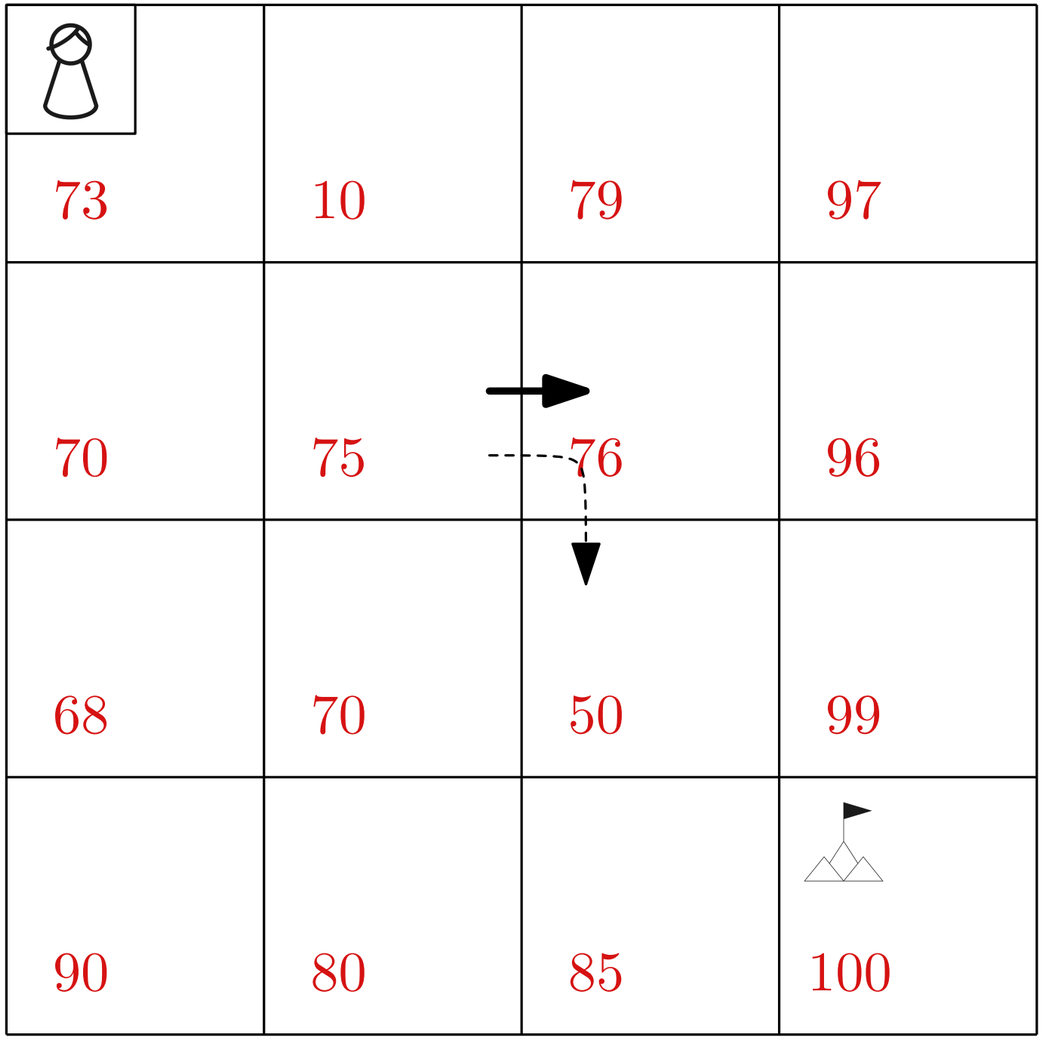
Policy-Iteration



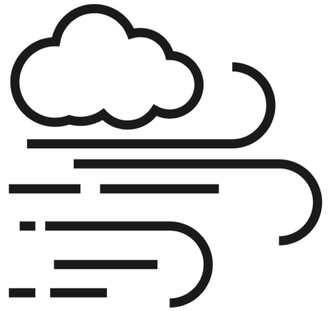
Sample

Update

Assignment 4



True-Value $v*$



model-free

How to recover π^*?

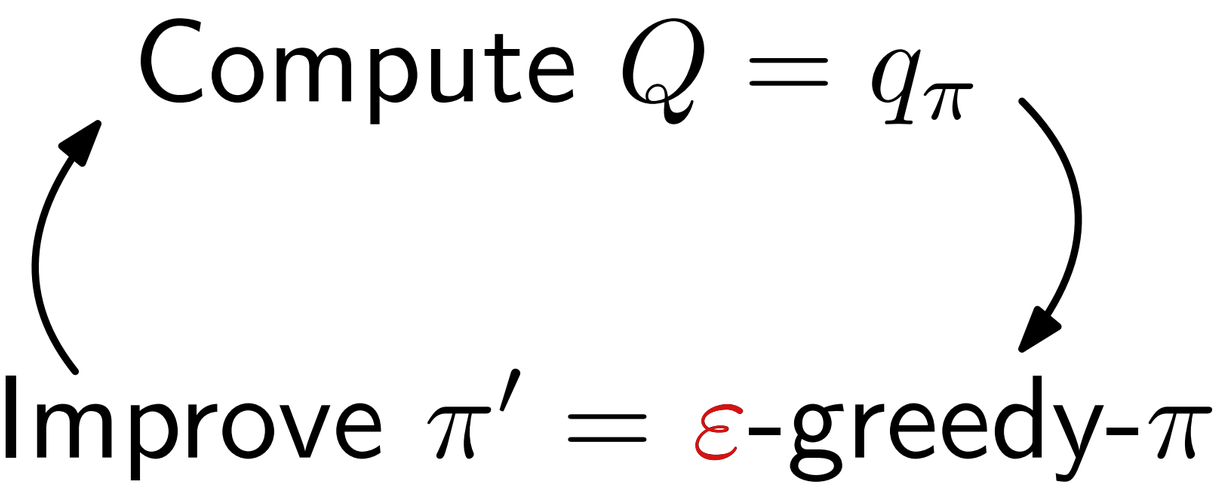
What to learn instead of state-value function?

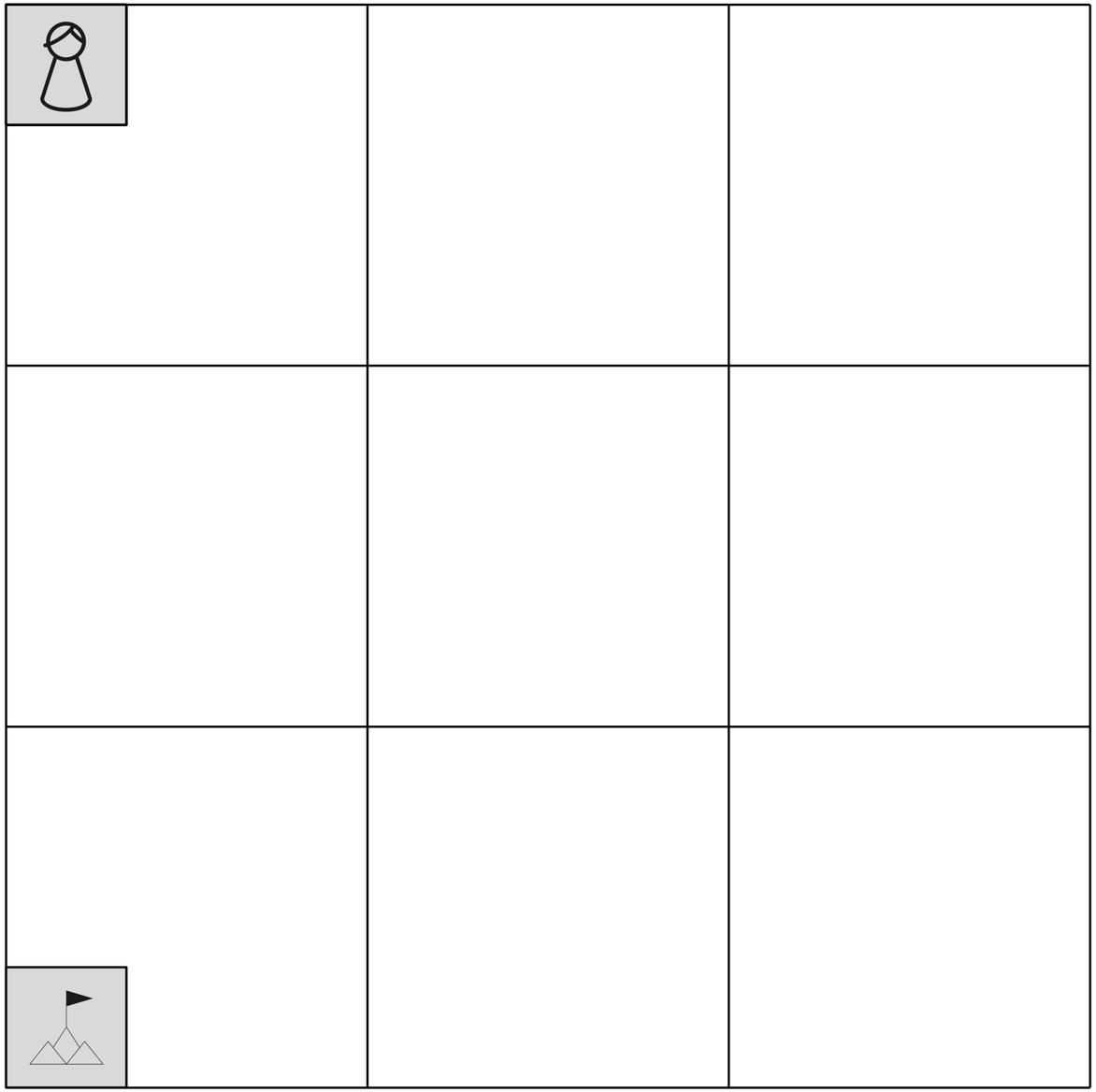
Post on Teams

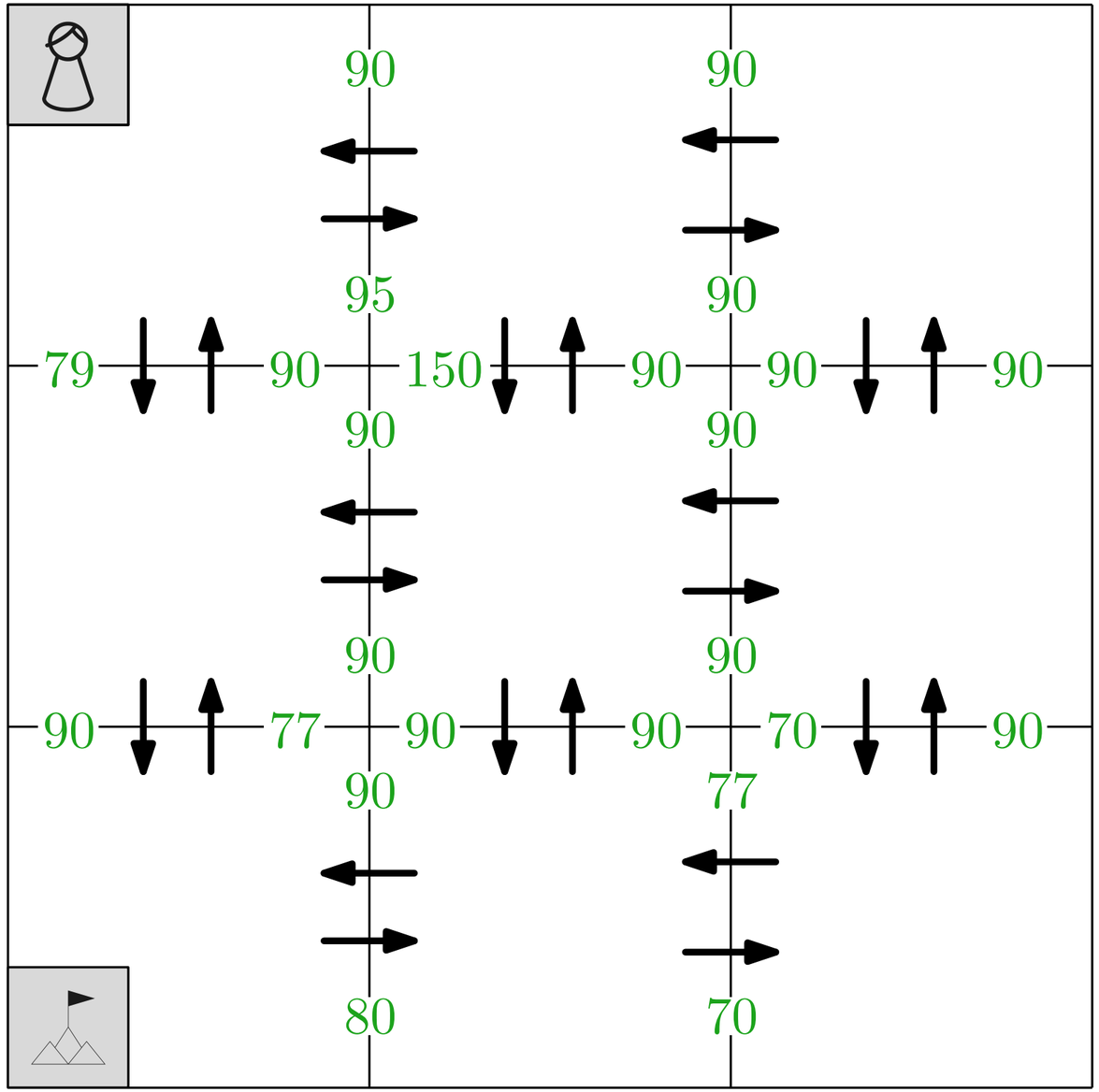
State-Action-Value Function

State-Action-Value Function

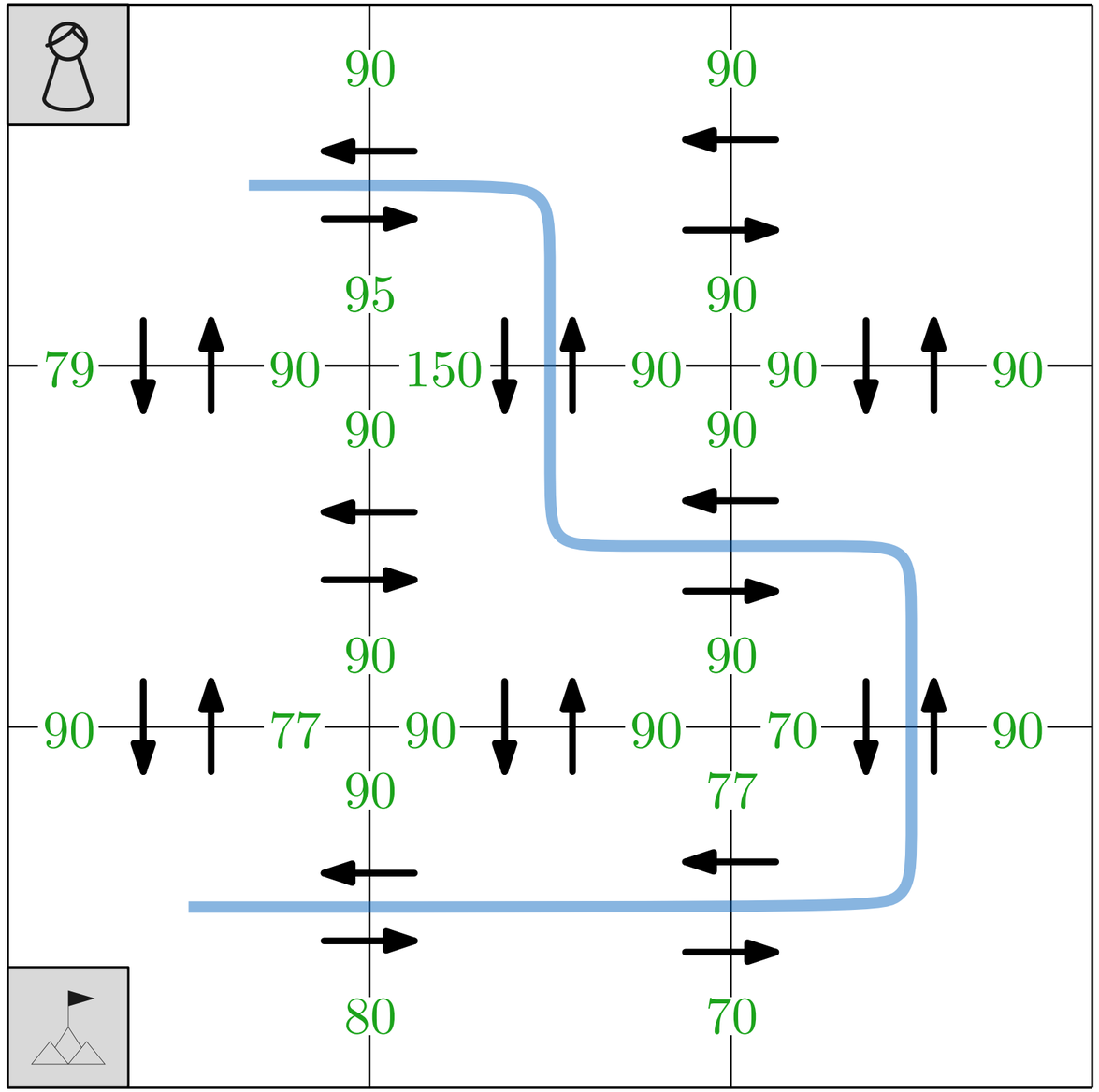
$$q_\pi : S \times A \to \mathbb{R} \qquad q_\pi : (s, a) = \mathbb{E}\ G(s, a)$$







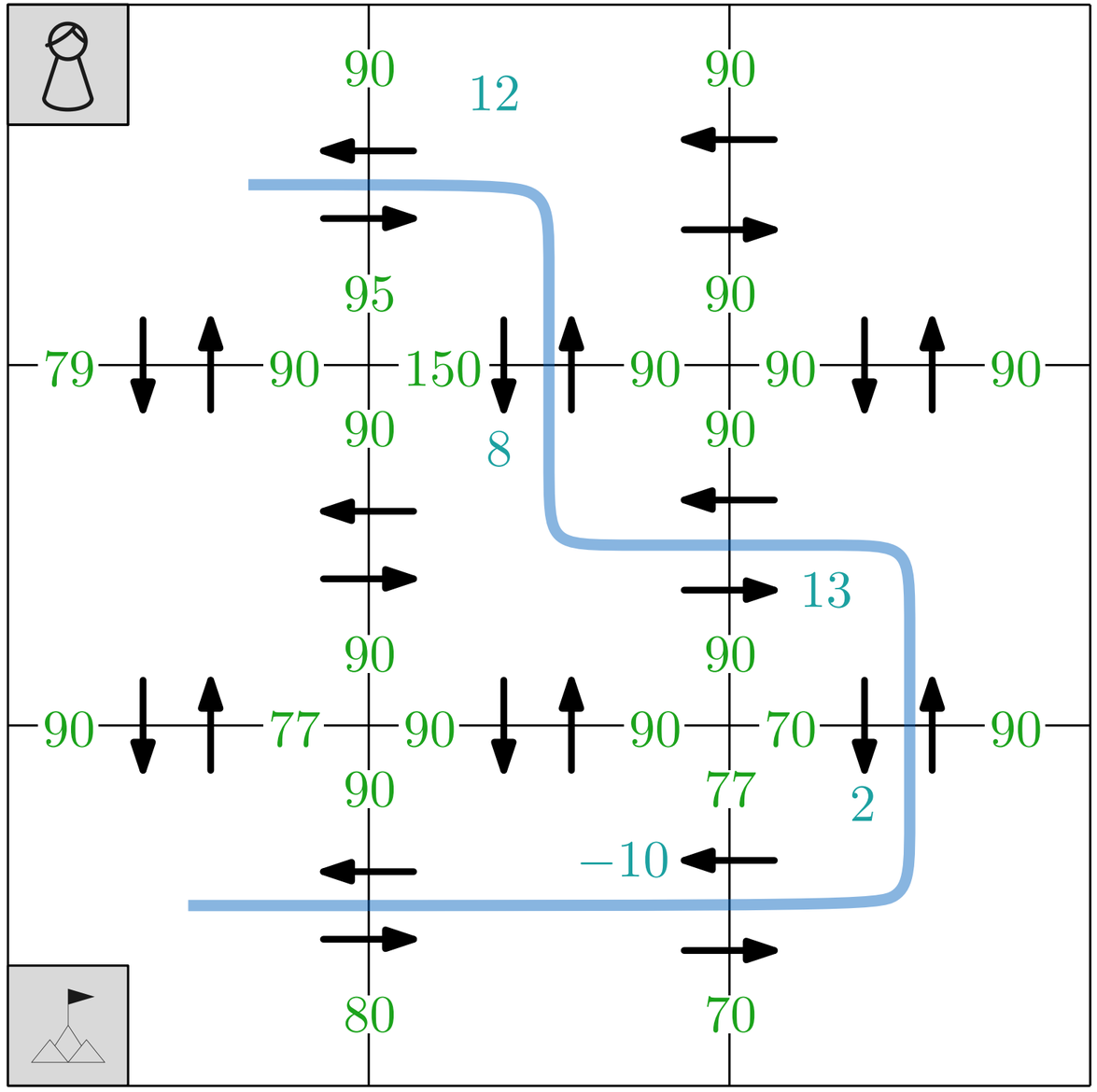
Q-Estimate



Q-Estimate

Sample

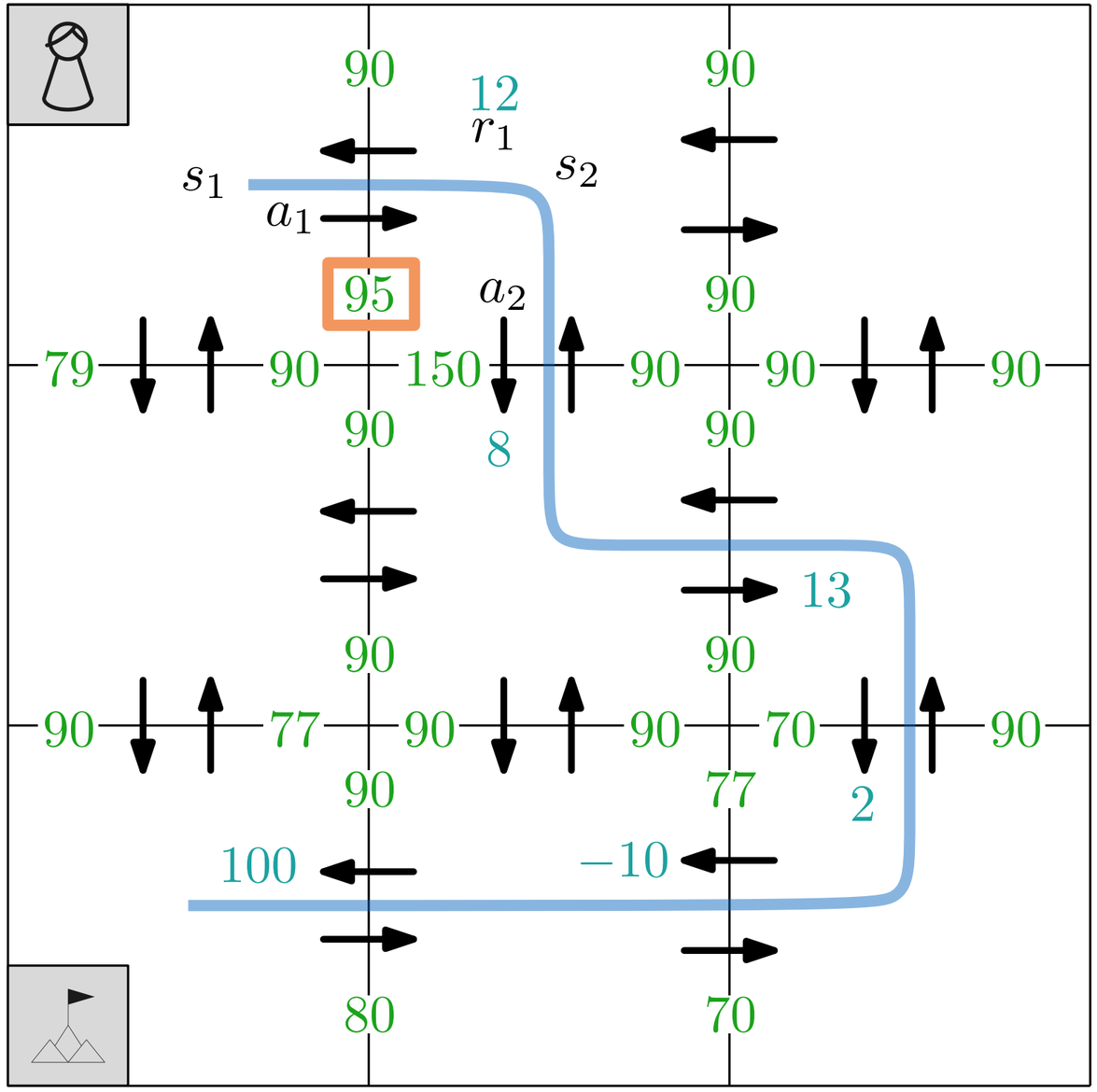
Rewards



Q-Estimate

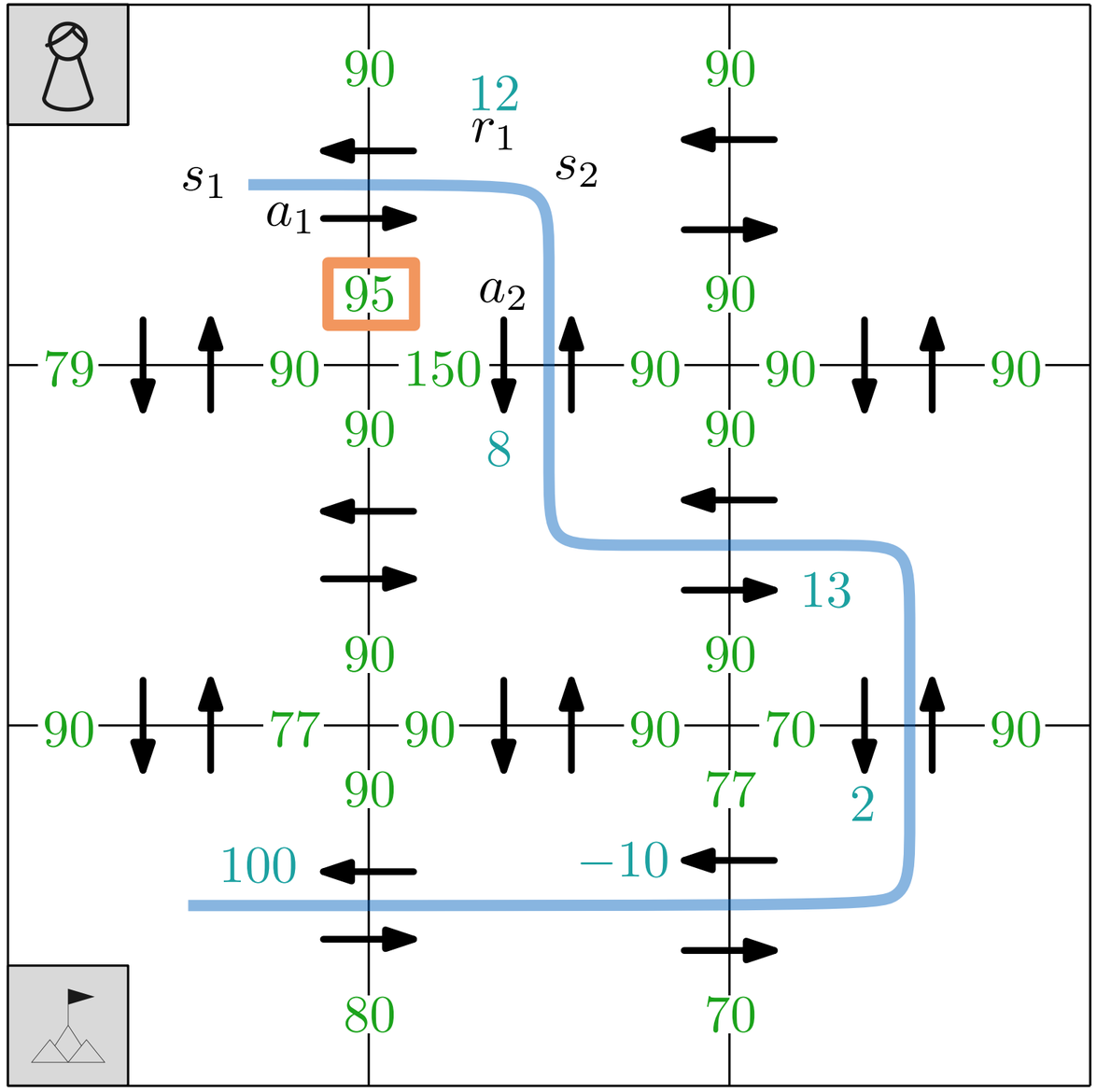
Sample

Rewards



Q-Estimate
Sample
Rewards
TD-Update

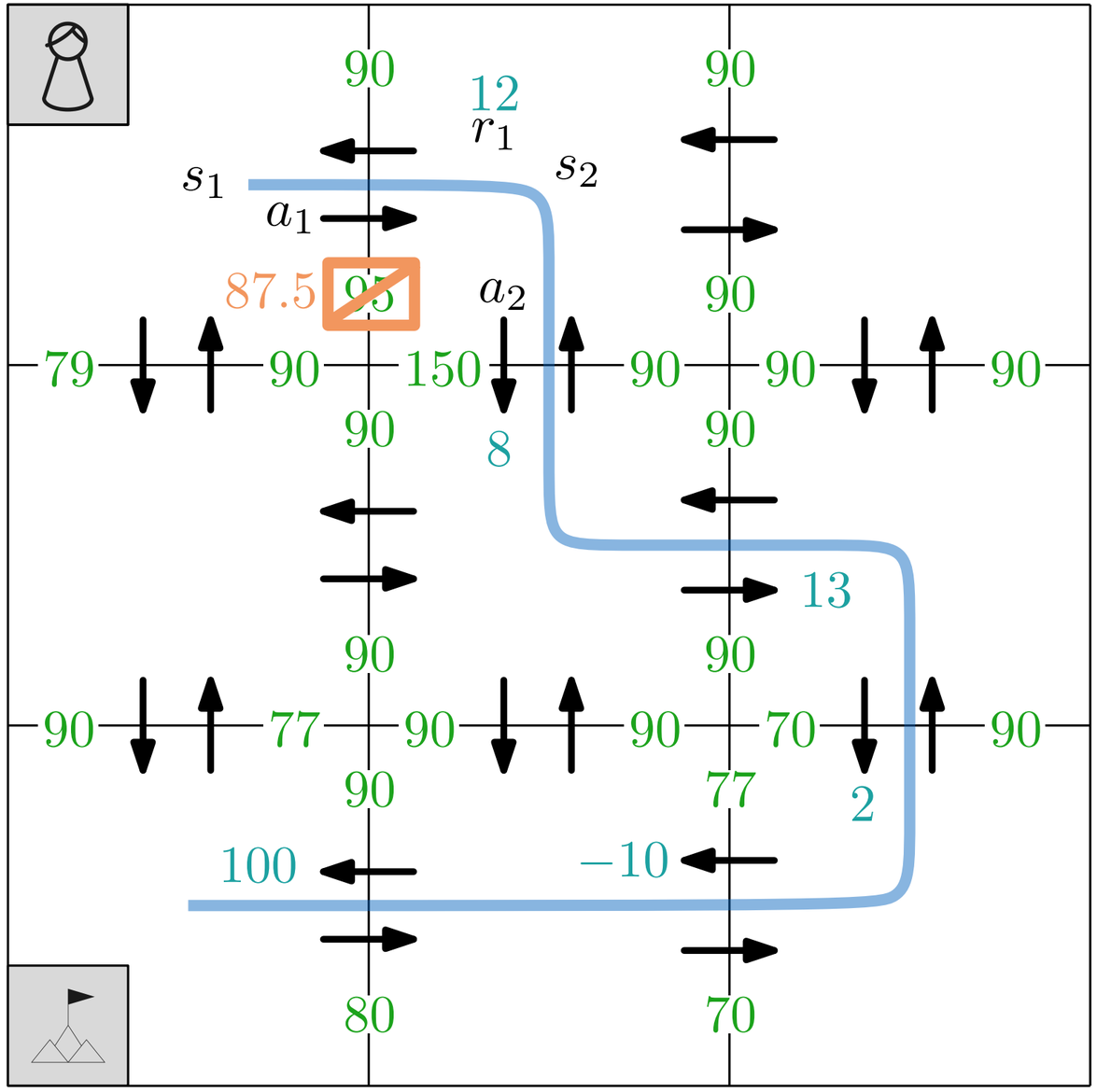
$\alpha = 0.1, \quad \gamma = 1$



Q-Estimate
Sample
Rewards
TD-Update

$$\alpha = 0.1, \quad \gamma = 1$$

$$Q(s_1, a_1) = (0.9)\ 95 + (0.1)\ (12 + 153) = 87.5$$



Q-Estimate
Sample
Rewards
TD-Update

$$\alpha = 0.1, \quad \gamma = 1$$

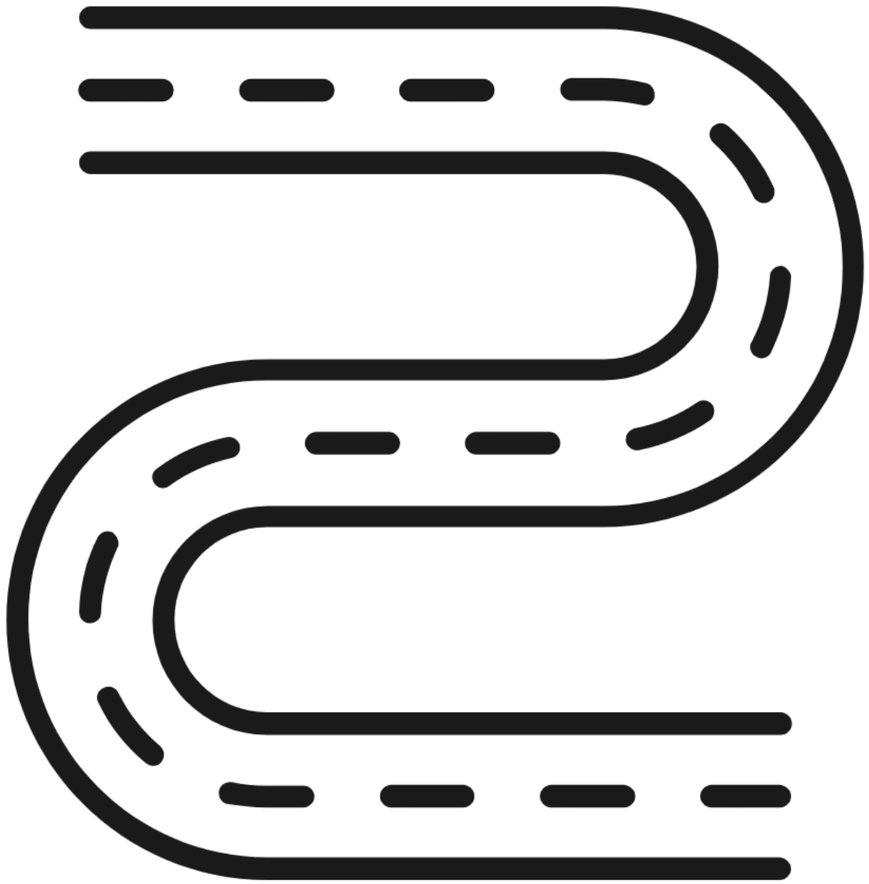
$$Q(s_1, a_1) = (0.9)\ 95 + (0.1)\ (12 + 153) = 87.5$$

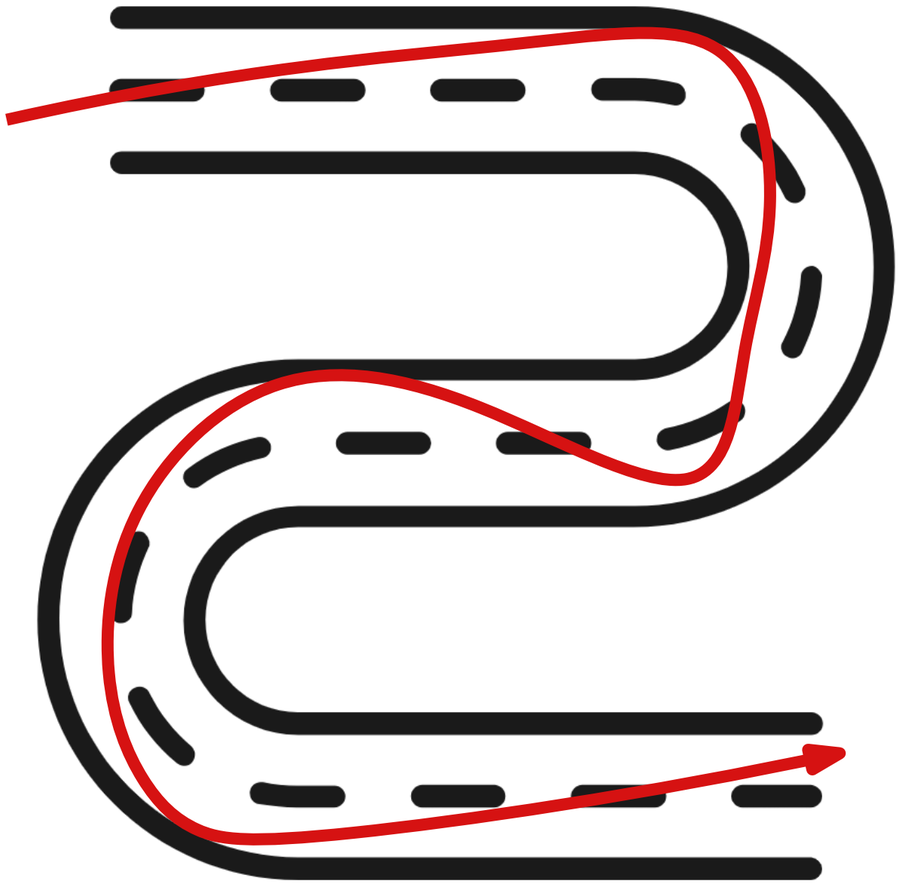
Assignment 5

1 Given q_π, and the dynamics p,
give a formula that describes v_π.

2 Given are s,a.
We sample r ,s' from the environment.
We sample a' from our policy.
Assume that α, γ are given.
Give a formula that describes the update rule
for Q(s,a).

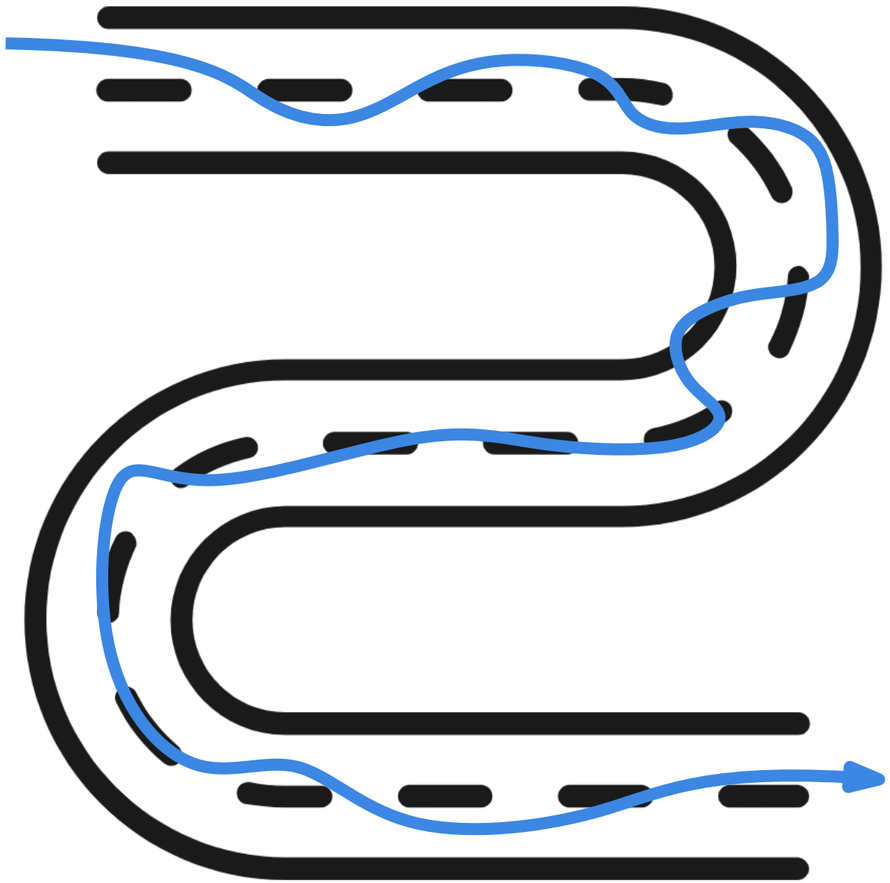
Post on Teams





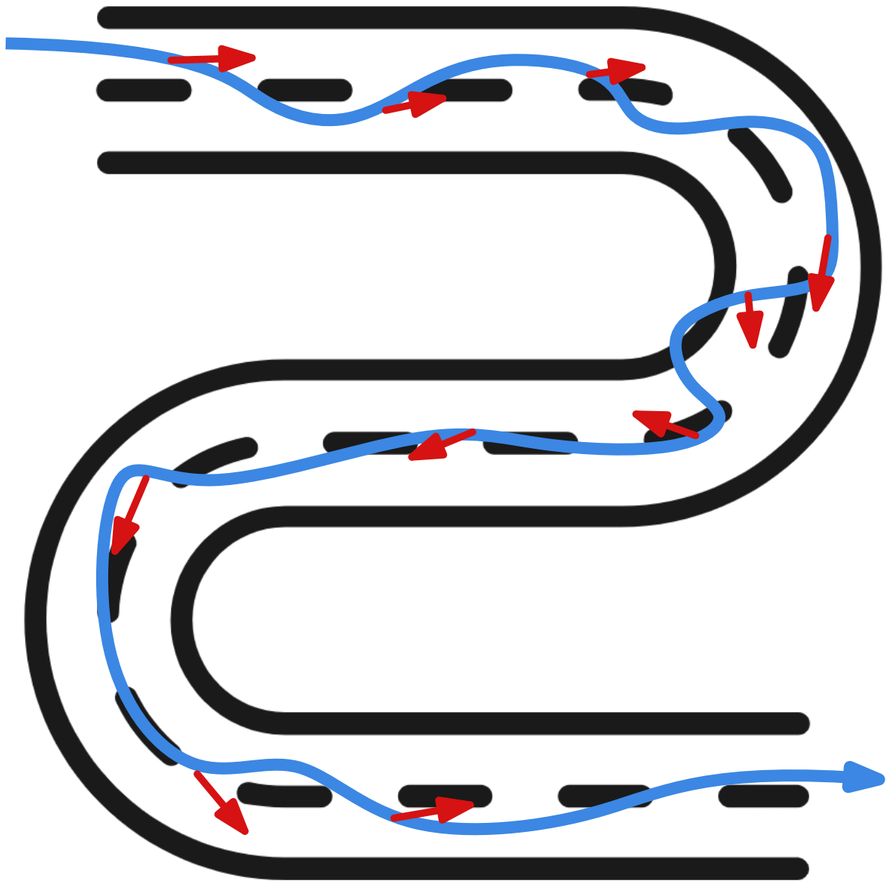
ideal route

Learn from the best!



exploratory route

See many states.



ideal route

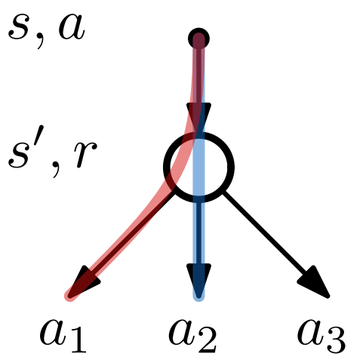
Learn from the best!

exploratory route

See many states.

Decouple exploration exploitation

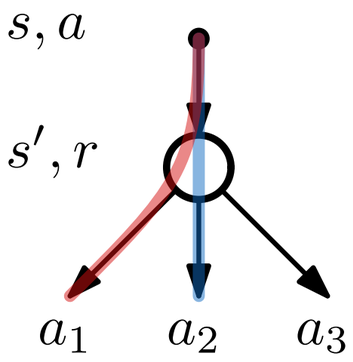
Q-Learning



$a_1 = \max_{a'} Q(s'.a')$

a_2 determines where we go next.

Q-Learning



$a_1 = \max_{a'} Q(s'.a')$

a_2 determines where we go next.

$$Q(s,a) = Q(s,a) + \alpha(r + \gamma Q(s', a_1) - Q(s,a))$$

Assignment 6

-1	-10	-10	100
-1	-1	-1	-1
-1	-1	-1	-1
-1	-1	-1	-1

Rewards

Where does Q-learning converge to?
Is it π^*?

Post on Teams