GRIDWORLD + The Value function

BIOE 498/598 PJ

Markov Decision Process (MDP)

States	s	where we are
actions	a	what we choose (next state)
rewards	r	gain/loss from (s,a)

Markov Property: a state tells us everything we need to know (memoryless)

$$s_0, a_0, r_1, s_1, a_1, r_2, s_2, \ldots\ldots, s_T$$

Making Decisions in an MDP

Assume: MDP is deterministic
Process terminates (@ time T)

Let's say we are at state s_0.

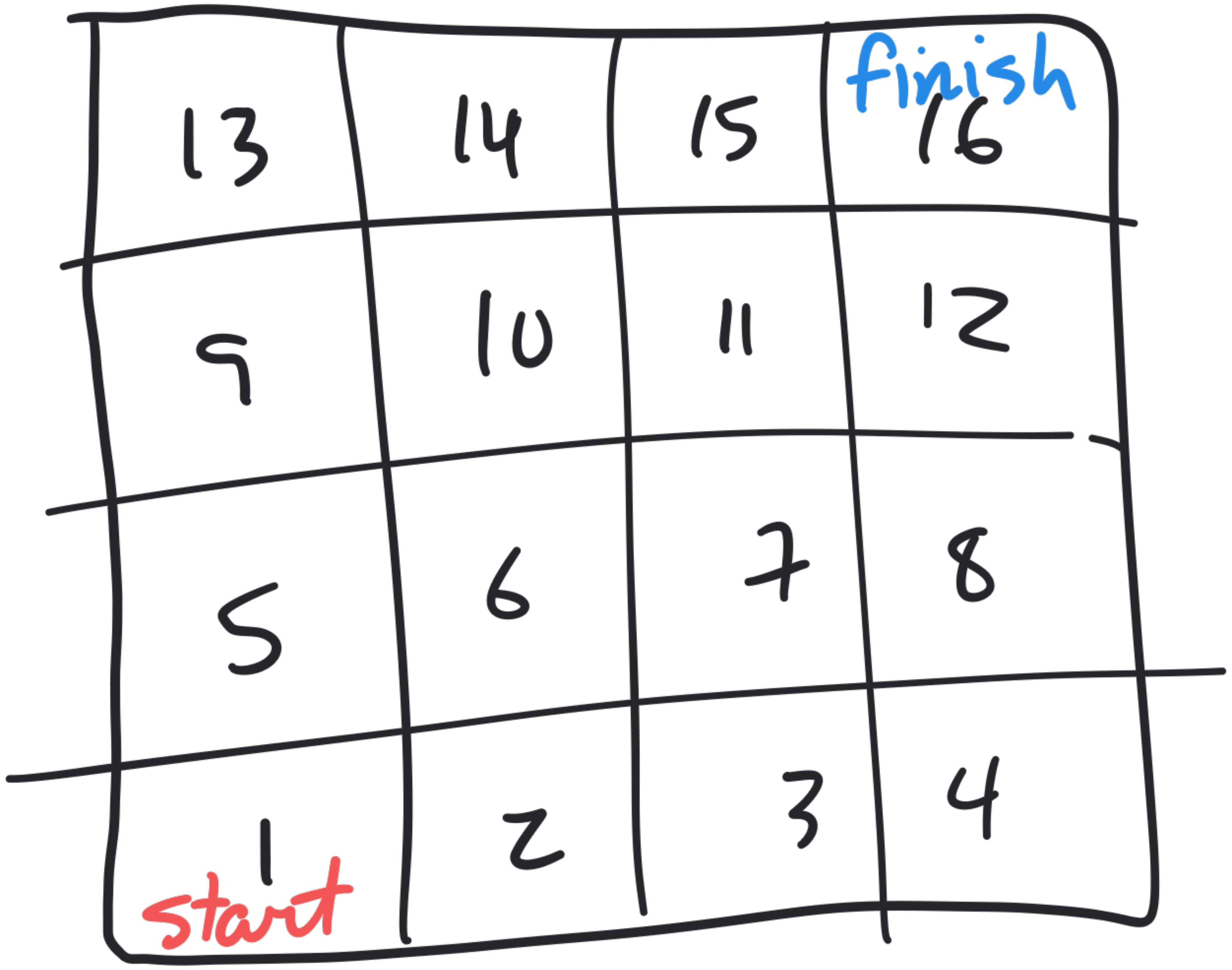
max Reward

$$= \max_a \left\{ \underset{\overset{\uparrow}{\text{immediate reward}}}{r_1} + \underbrace{\sum_{i=2}^{T} r_i}_{\overset{\uparrow}{\text{future reward}}} \right\} = \max_a \left\{ r_1 + V(s') \right\}$$

$$\uparrow$$

Value function

GRIDWORLD



13	14	15	finish 16
9	10	11	12
5	6	7	8
1 start	2	3	4

Compue V(s) for every square

If I know V(s), then I can always find the shortest path.

Get from start to finish in the fewest number of steps.

How do I compute $V(s)$?

$r_i = -1$ except at the finish.

States $S = \{1, 2, \ldots, 16\}$

start (1)   terminal (16)

13	14	15	16
9	10	11	12
5	6	7	8
1	2	3	4

actions are defined for every state

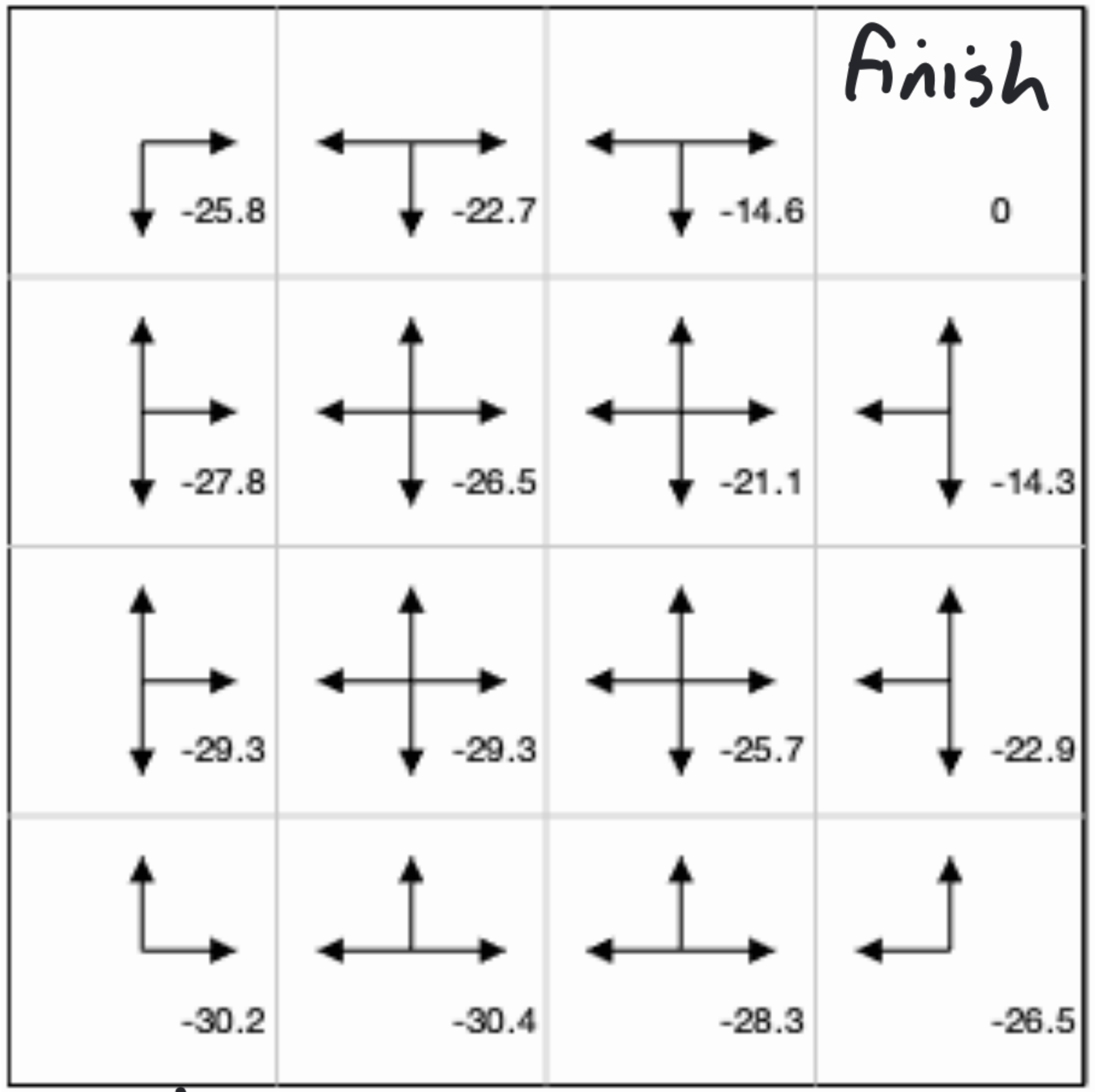
Approach: Monte Carlo Method

1. Pick a state.
2. Take a random walk until I reach the finish
3. Count steps.
4. Repeat.

5, 6, 10, 9, 5, 6, 7, 11, 15, 16

$V(5) = -9$

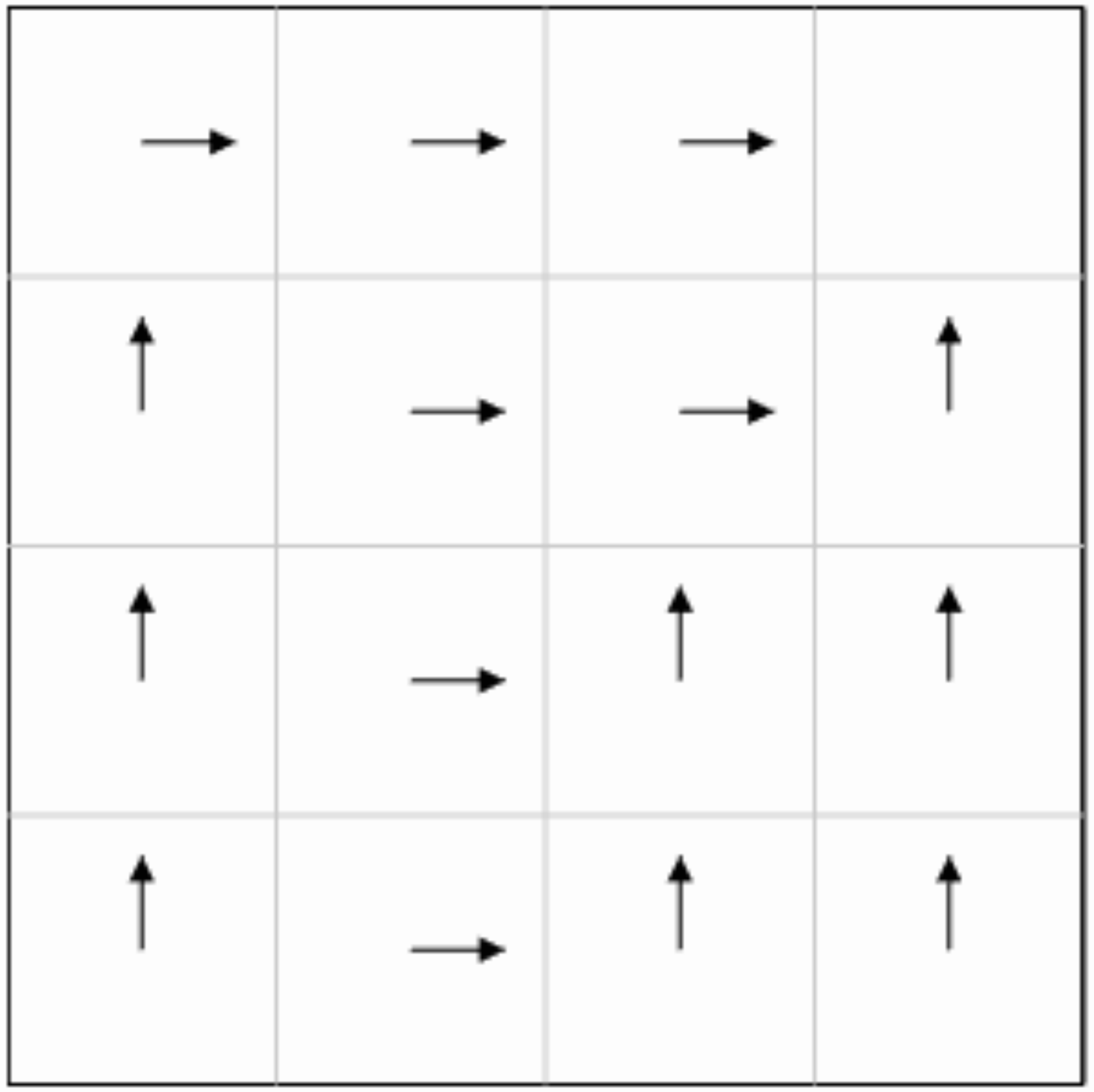
V(s)



→↓ -25.8	←→↓ -22.7	←→↓ -14.6	finish 0
↑→↓ -27.8	←↑→↓ -26.5	←↑→↓ -21.1	←↑↓ -14.3
↑→↓ -29.3	←↑→↓ -29.3	←↑→↓ -25.7	←↑↓ -22.9
↑→ -30.2	←↑→ -30.4	←↑→ -28.3	←↑ -26.5

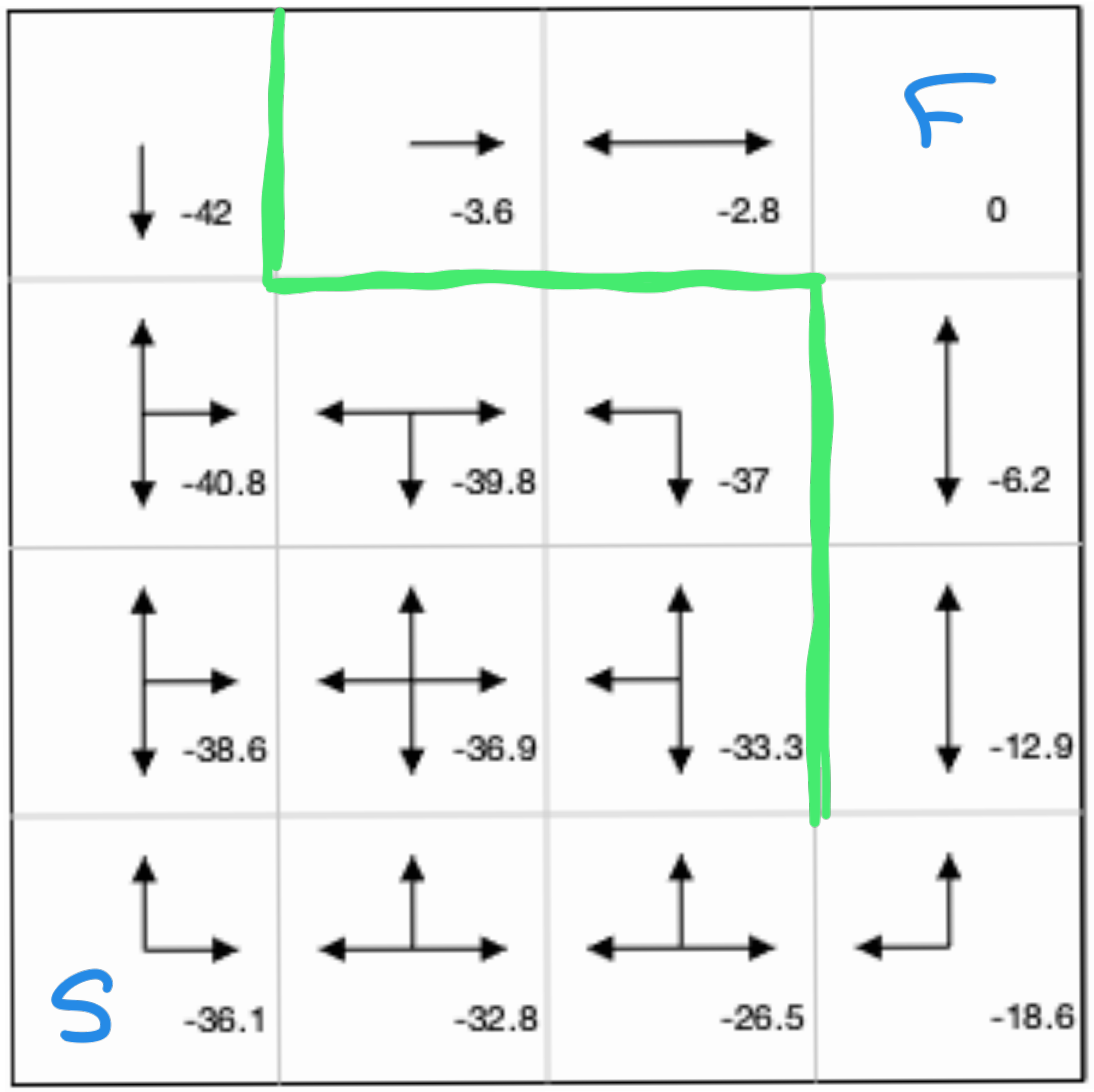
start

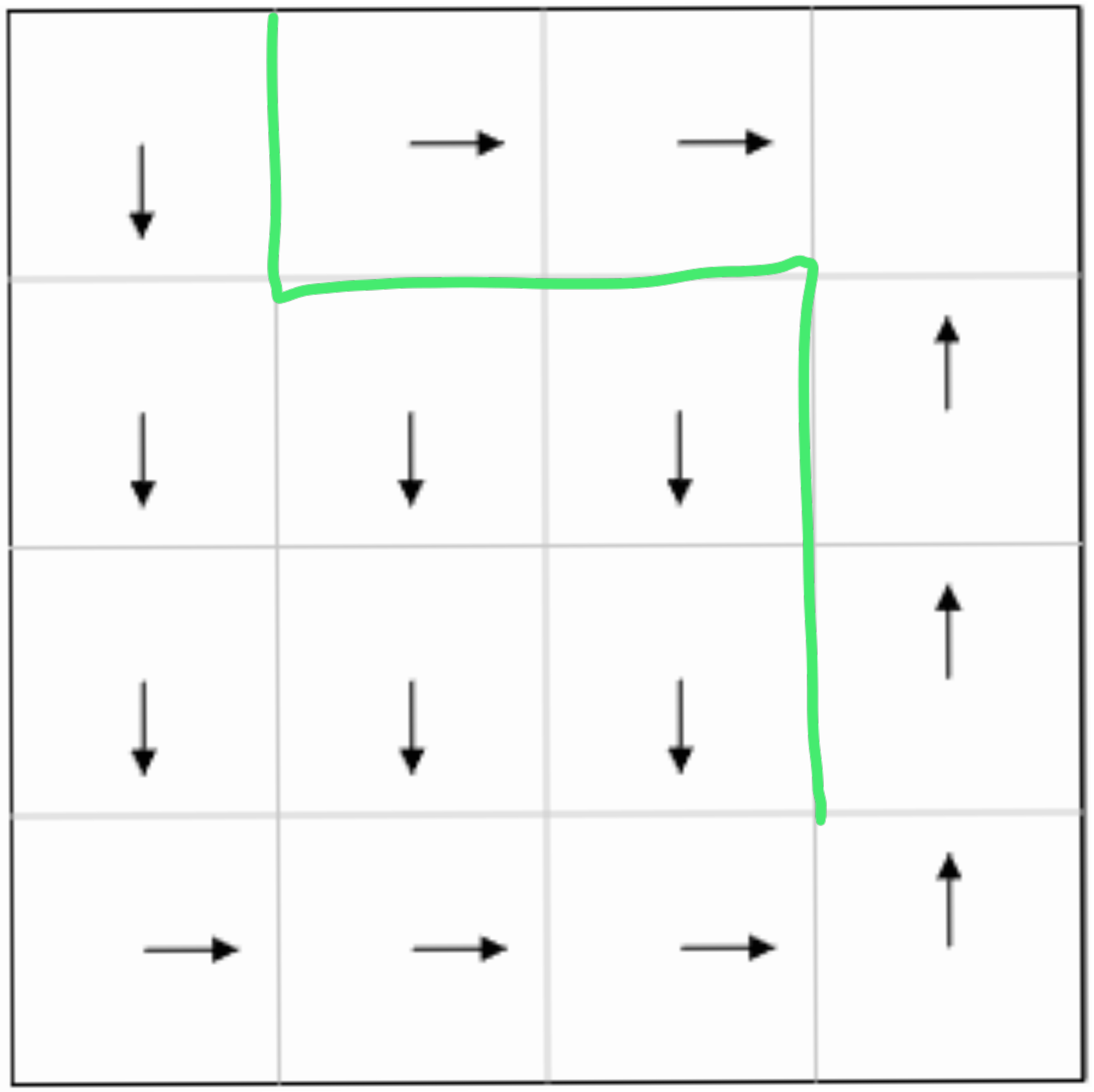
Policy Improvement



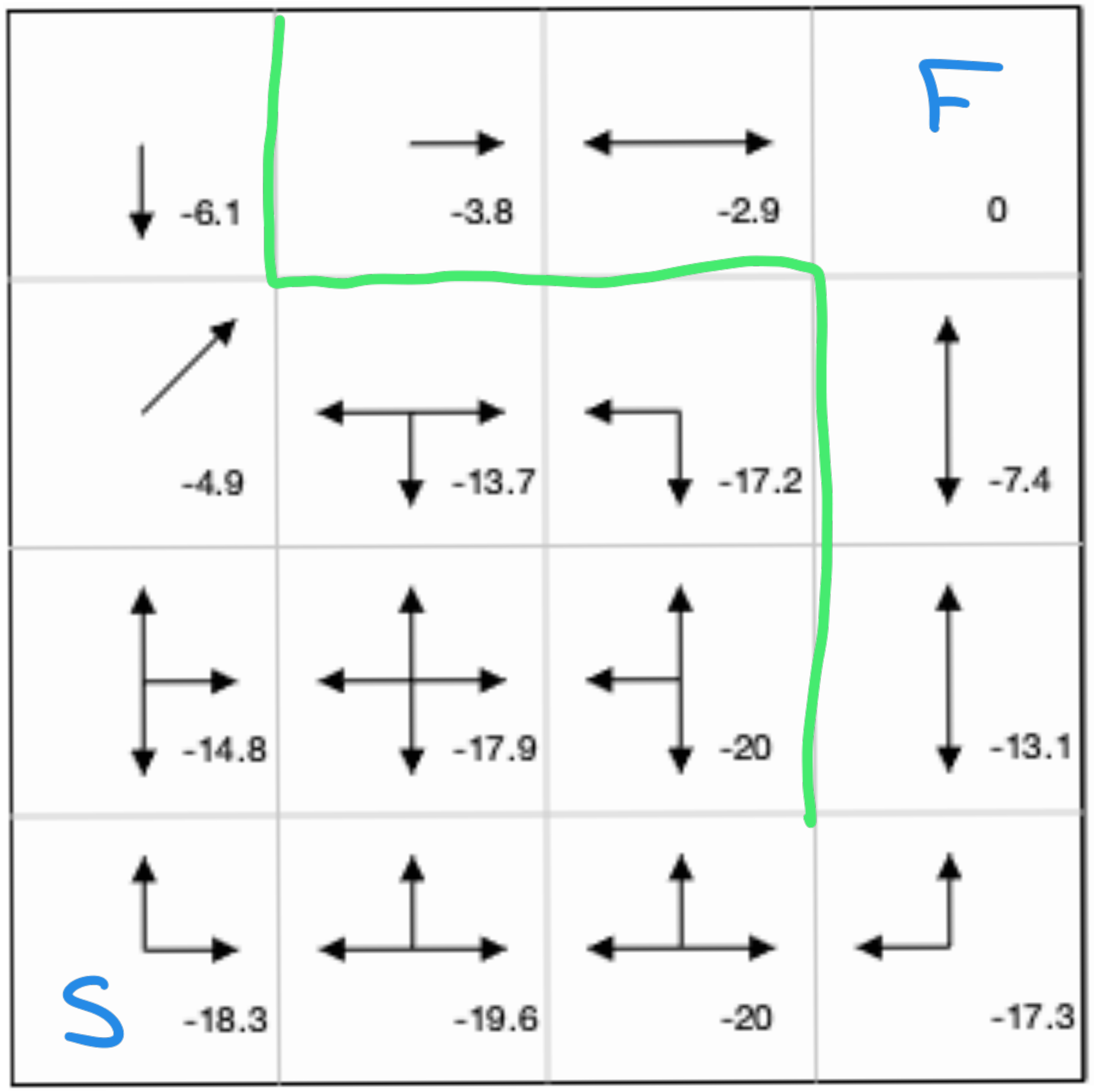
→	→	→	
↑	→	→	↑
↑	→	↑	↑
↑	→	↑	↑

V(s)





V(s)


F
-6.1
-3.8
-2.9
0
-4.9
-13.7
-17.2
-7.4
-14.8
-17.9
-20
-13.1
S
-18.3
-19.6
-20
-17.3