

# MONTE CARLO AND OFF-POLICY LEARNING

# NOTATIONAL NOTE:

## FUNCTIONS VS PROBABILITY STATEMENTS

$$\Pr(A | B)$$

$A$  and  $B$  are events, i.e.,  $S_t=s$

Probability statements are mathematical statements about how often events occur.

$$f: \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}, z = f(x, y)$$

Functions process inputs and return a value.

What is  $f(z | x, y)$ ? — Nothing. We have not defined  $f$  to be a function of three variables.

The  $|$  symbol is not a standard way of specifying inputs to a function. Does  $f$  of  $z$  conditioned on  $x$  and  $y$  make sense?

Book's notation:  $p(s', r | s, a)$  — not a standard way to write a function. The typical way would be to replace ' $|$ ' with ','.

This is not a probability statement but a function that, given four inputs, returns a value in  $[0,1]$ .

By definition, we know that the value corresponds to a probability, but the value is just a value.

It's the same thing with  $\pi(a | s)$  — a function of two inputs, not a probability statement.

# MONTE-CARLO POLICY ITERATION

## REVIEW

No access to  $p$  or  $r$  except via sampling

Need to find  $\pi_*$

# MONTE-CARLO POLICY ITERATION

## REVIEW

Monte Carlo ES (Exploring Starts), for estimating  $\pi \approx \pi_*$

Initialize:

$\pi(s) \in \mathcal{A}(s)$  (arbitrarily), for all  $s \in \mathcal{S}$

$Q(s, a) \in \mathbb{R}$  (arbitrarily), for all  $s \in \mathcal{S}, a \in \mathcal{A}(s)$

$Returns(s, a) \leftarrow$  empty list, for all  $s \in \mathcal{S}, a \in \mathcal{A}(s)$

Loop forever (for each episode):

Choose  $S_0 \in \mathcal{S}, A_0 \in \mathcal{A}(S_0)$  randomly such that all pairs have probability  $> 0$

Generate an episode from  $S_0, A_0$ , following  $\pi$ :  $S_0, A_0, R_1, \dots, S_{T-1}, A_{T-1}, R_T$

$G \leftarrow 0$

Loop for each step of episode,  $t = T-1, T-2, \dots, 0$ :

$G \leftarrow \gamma G + R_{t+1}$

Unless the pair  $S_t, A_t$  appears in  $S_0, A_0, S_1, A_1, \dots, S_{t-1}, A_{t-1}$ :

Append  $G$  to  $Returns(S_t, A_t)$

$Q(S_t, A_t) \leftarrow \text{average}(Returns(S_t, A_t))$

$\pi(S_t) \leftarrow \arg \max_a Q(S_t, a)$

# MONTE-CARLO POLICY ITERATION

## PROS AND CONS

Ensure exploration for all state-action pairs for every policy

What if we cannot “place” the environment in just any state:

- Self-driving cars: need to place all cars at certain positions and speeds (not practical)

Can the sample start state from the distribution defined by  $d_0$

- Car starts in the driveway and other cars are “wild”

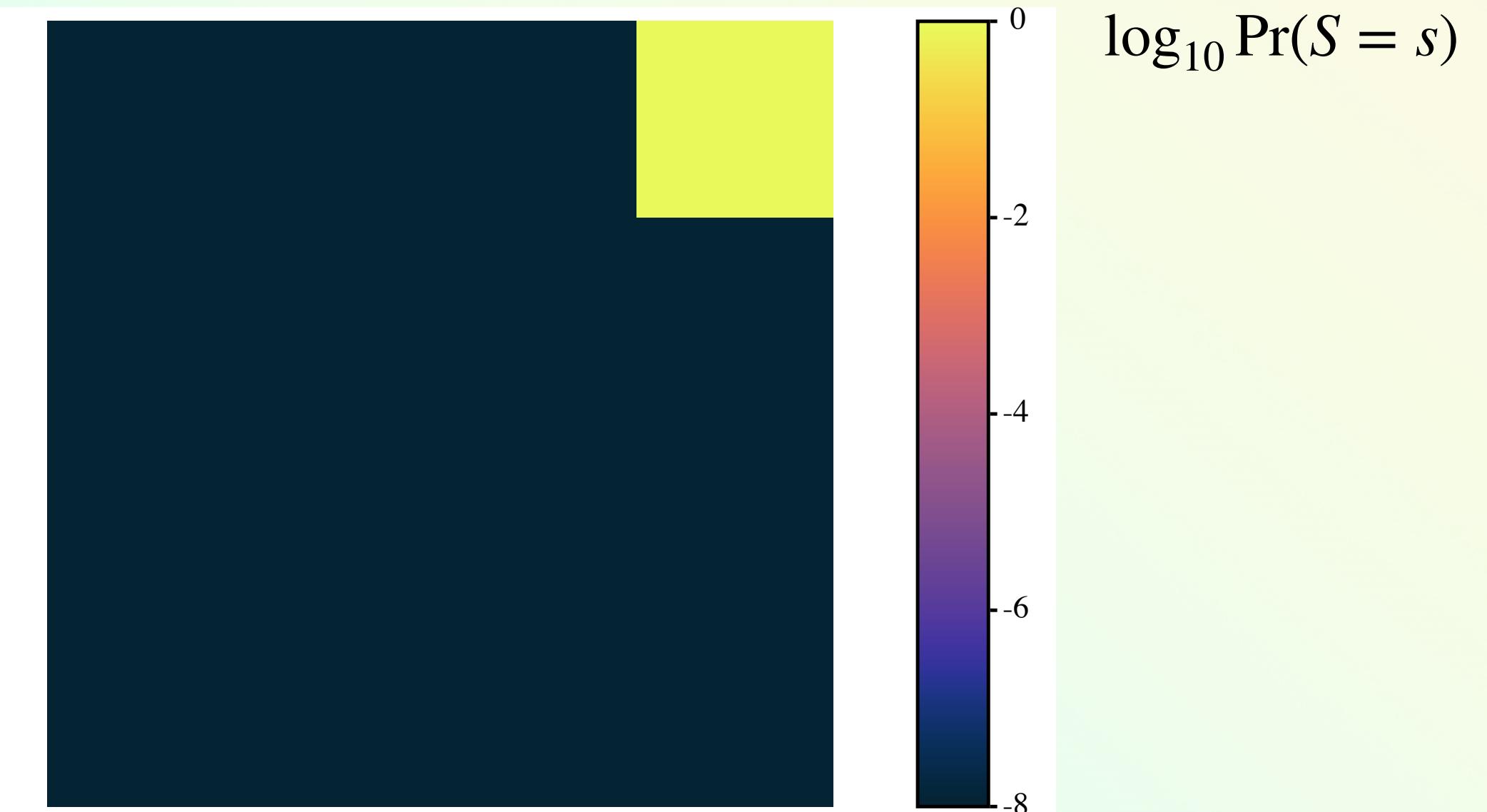
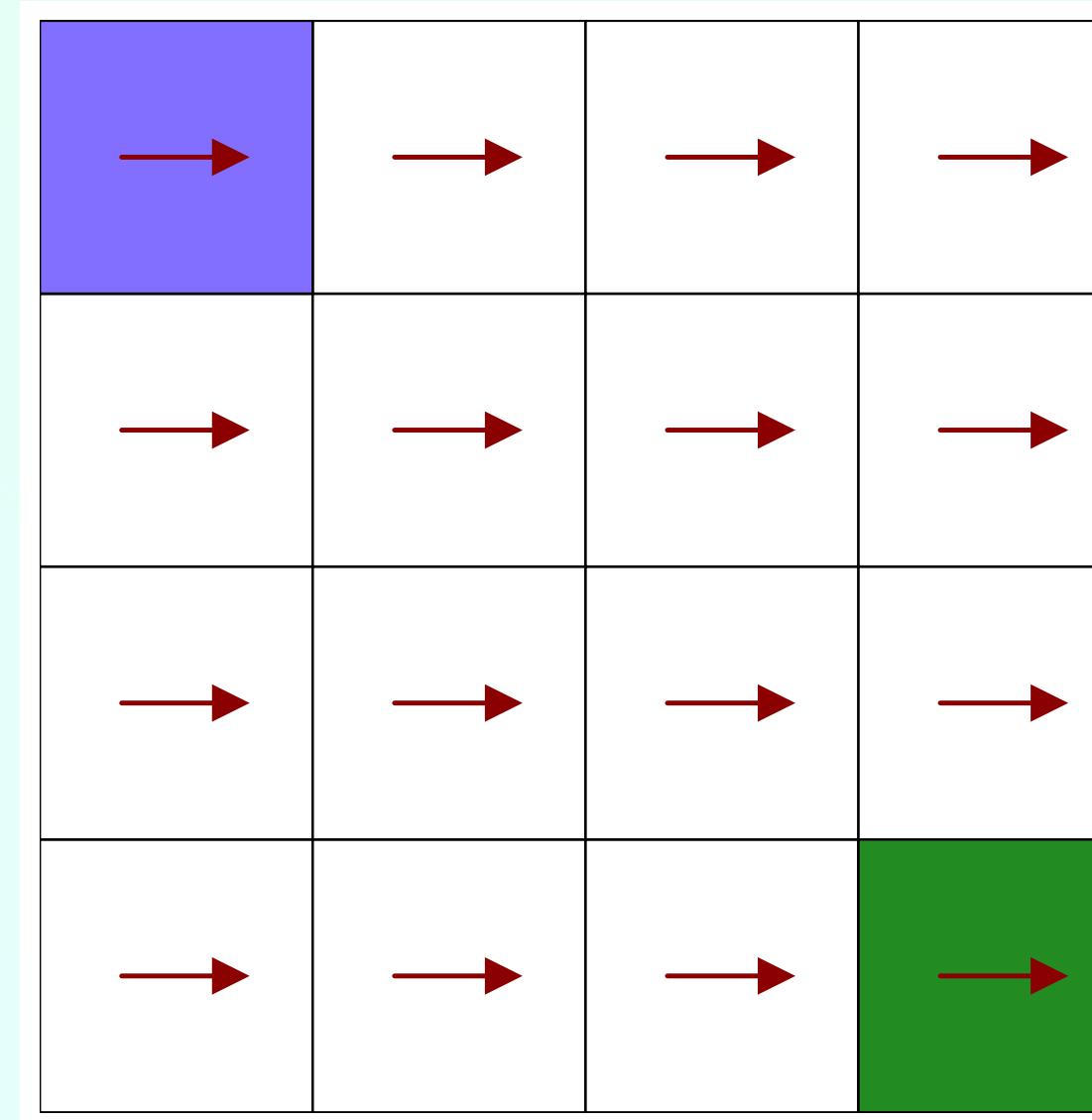
# ENSURING EXPLORATION

FROM START STATE

Need to evaluate  $q_{\pi}(s, a)$  for all  $s, a$  pairs

A deterministic policy may not reach all states (also does not try each action)

Start	2	3	4
1	2	3	4
6	7	8	9
11	12	13	14
16	17	18	Goal 19



# ENSURING EXPLORATION

FROM START STATE

If  $\pi(a | s) > 0$  for all  $a$

Then all reachable state action pairs will be tried infinitely often

# ENSURING EXPLORATION

FROM START STATE

If  $\pi(a | s) > 0$  for all  $a$

Then all reachable state action pairs will be tried infinitely often

$\epsilon$ -greedy policy

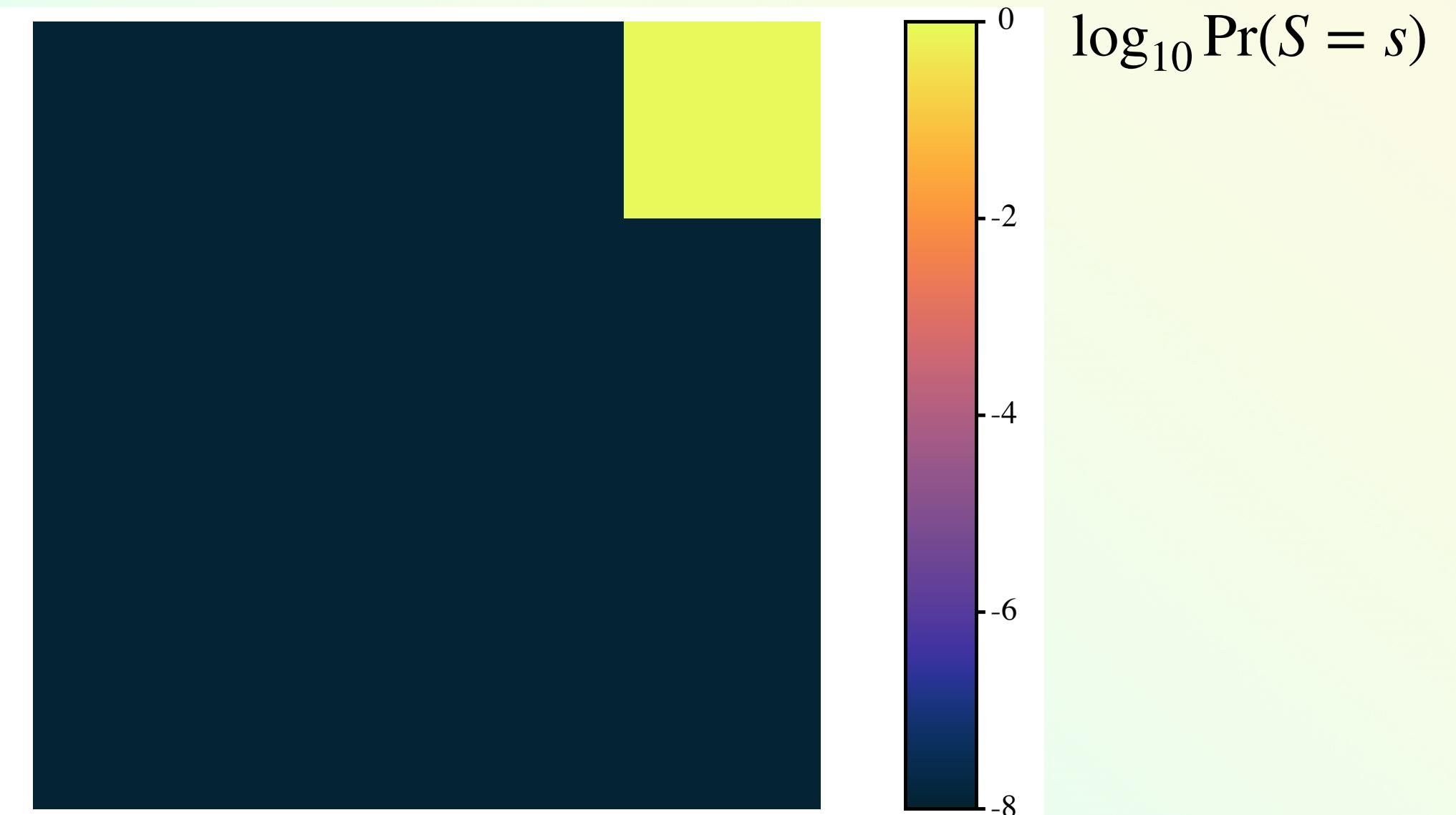
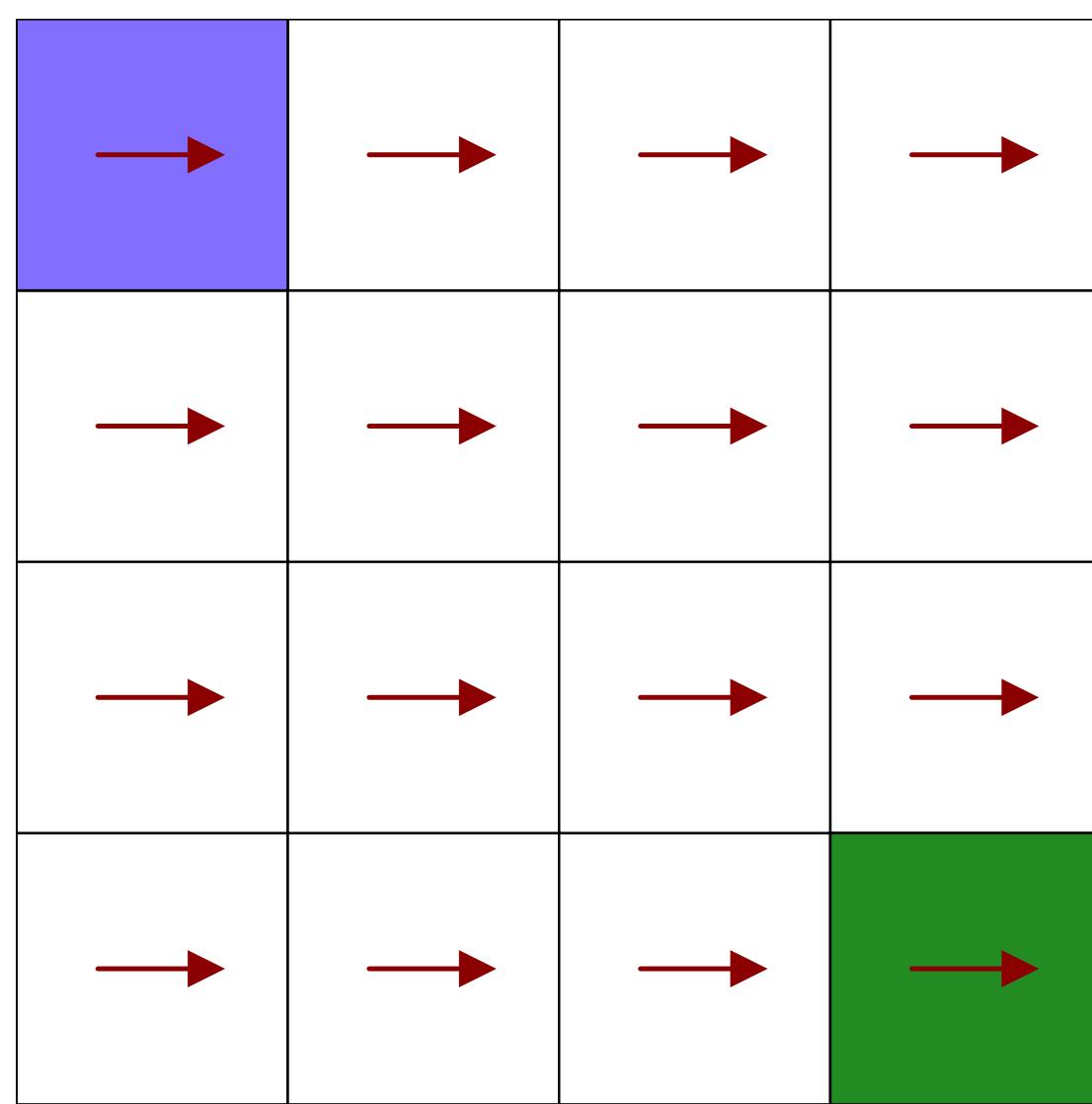
$$\pi(a | s) = \begin{cases} \frac{(1 - \epsilon)}{|\mathcal{A}^*|} + \frac{\epsilon}{|\mathcal{A}|} & \text{if } a \in \mathcal{A}^* \\ \frac{\epsilon}{|\mathcal{A}|} & \text{otherwise} \end{cases}, \quad \mathcal{A}^* = \arg \max_a q(s, a)$$

# ENSURING EXPLORATION

FROM START STATE

Go right policy with  $\epsilon = 0$

Start	2	3	4
1	6	7	8
11	12	13	14
16	17	18	Goal 19

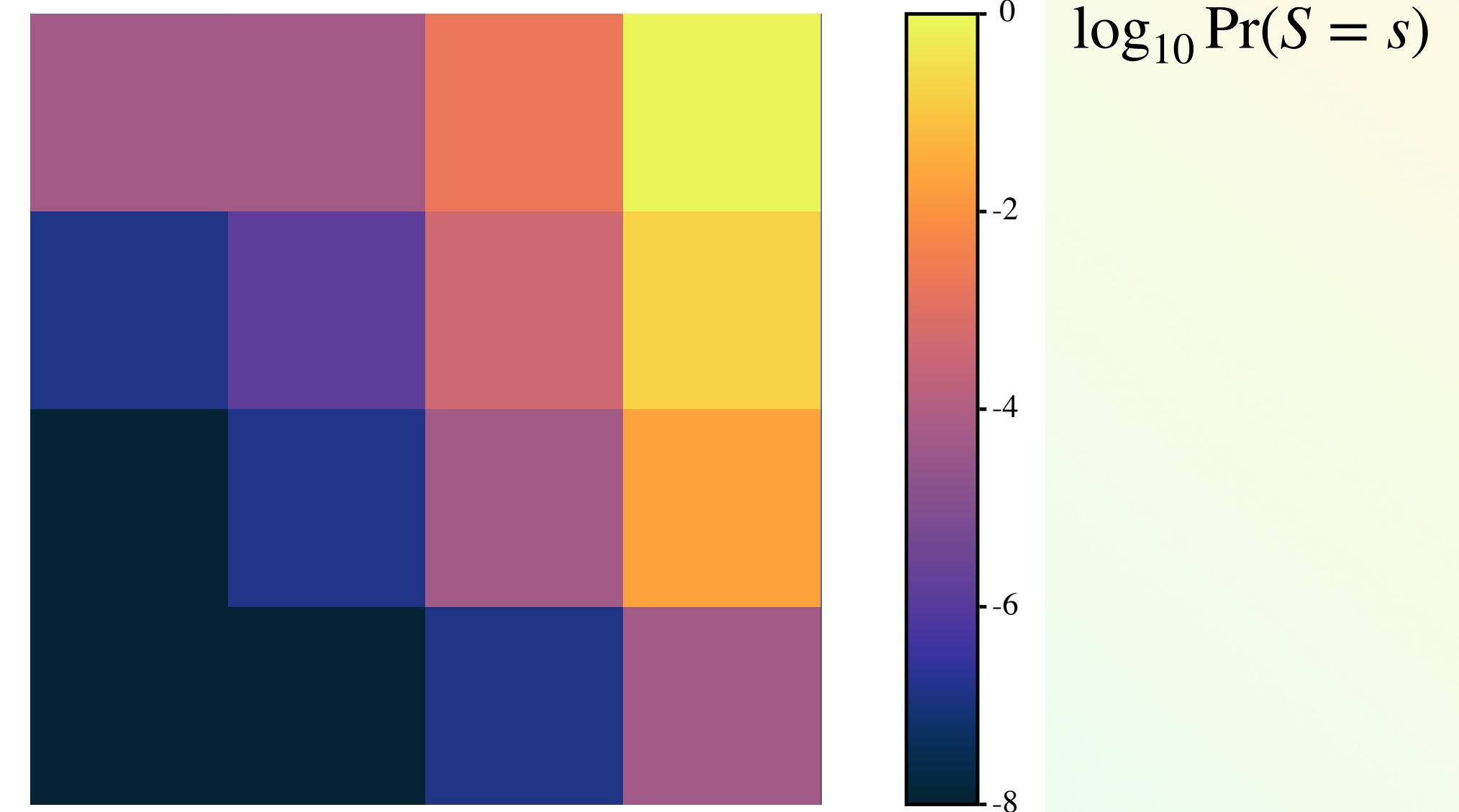
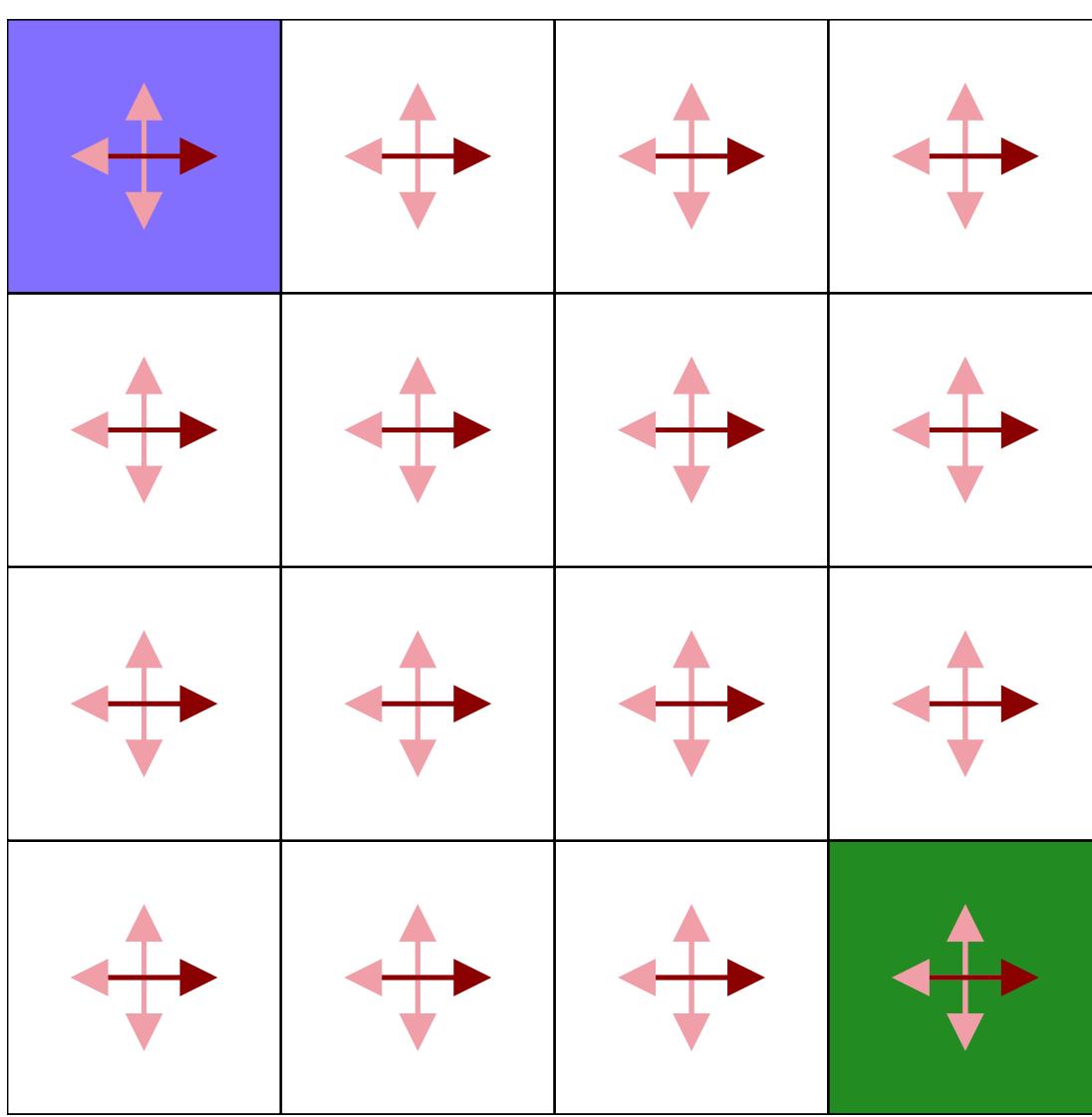


# ENSURING EXPLORATION

FROM START STATE

Go right policy with  $\epsilon = 0.01$

Start 1	2	3	4
6	7	8	9
11	12	13	14
16	17	18	Goal 19

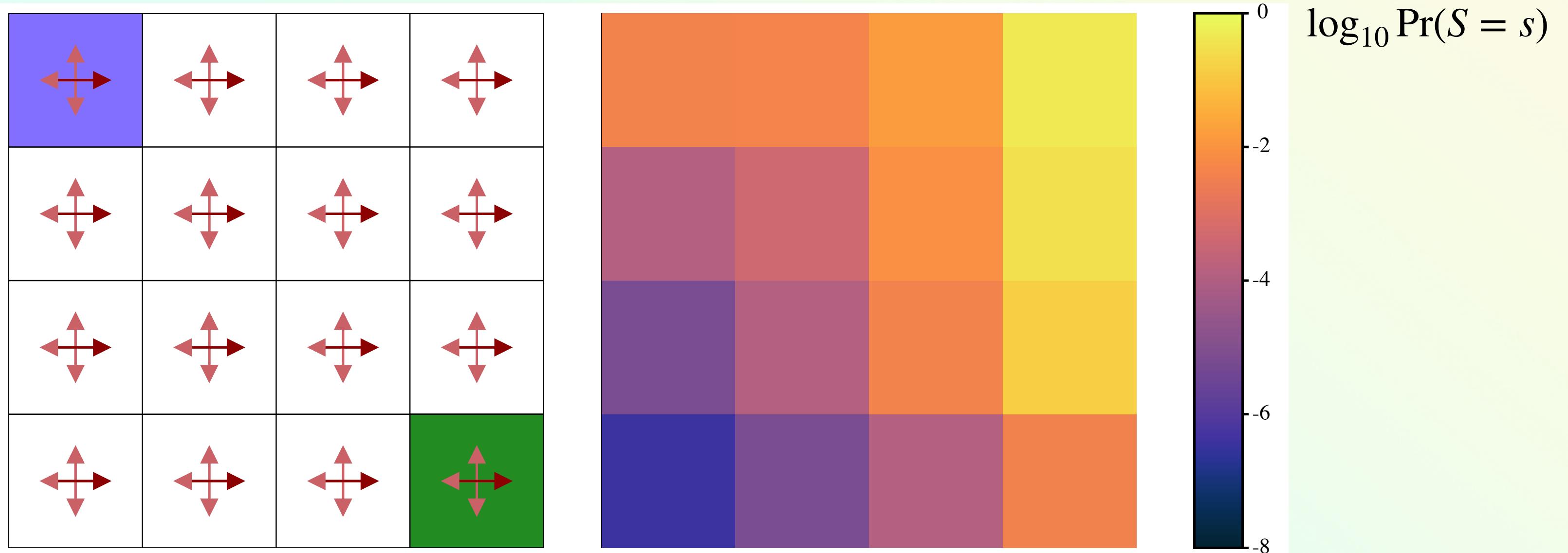


# ENSURING EXPLORATION

FROM START STATE

Go right policy with  $\epsilon = 0.1$

Start 1	2	3	4
6	7	8	9
11	12	13	14
16	17	18	Goal 19

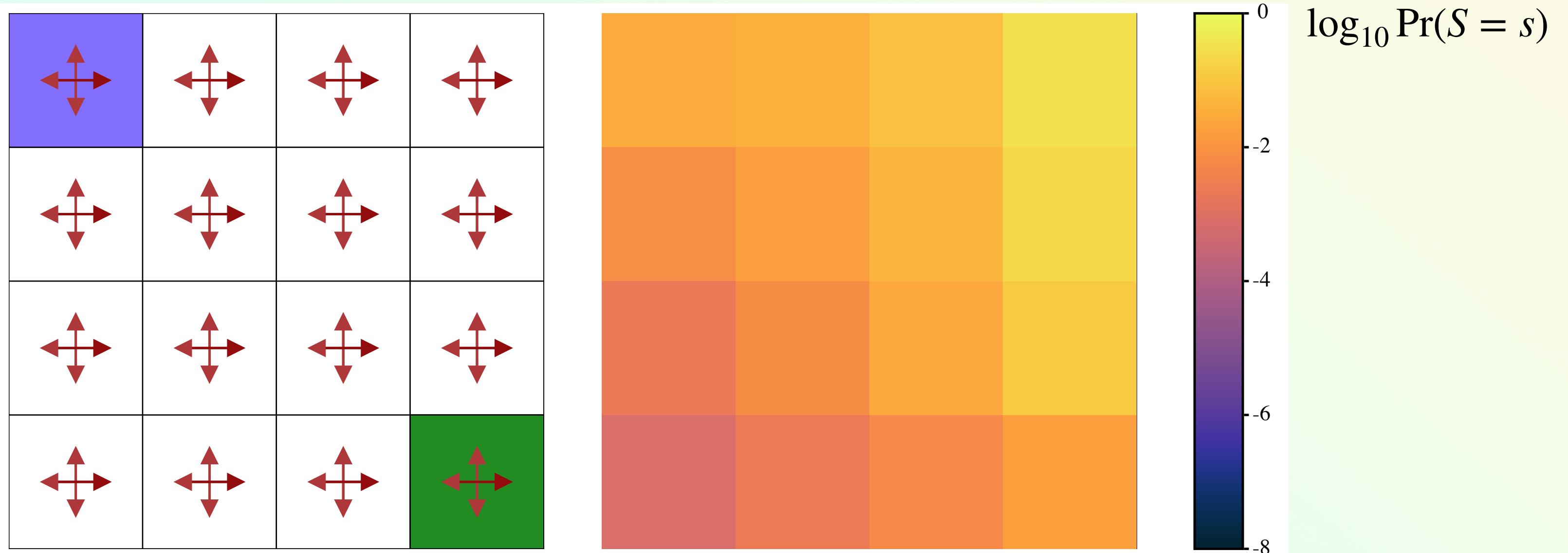


# ENSURING EXPLORATION

FROM START STATE

Go right policy with  $\epsilon = 0.5$

Start	2	3	4
1	6	7	8
11	12	13	14
16	17	18	Goal 19

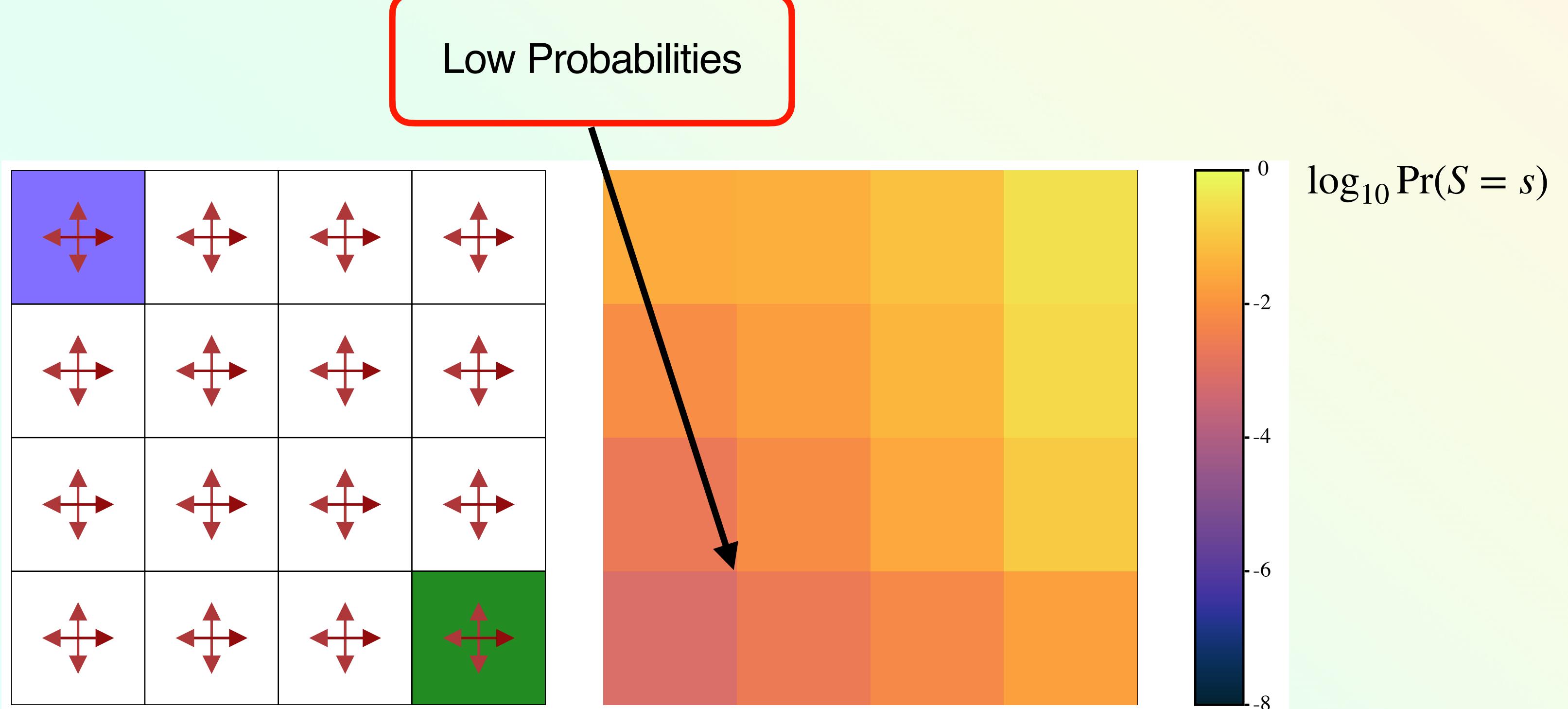


# ENSURING EXPLORATION

FROM START STATE

Go right policy with  $\epsilon = 0.5$

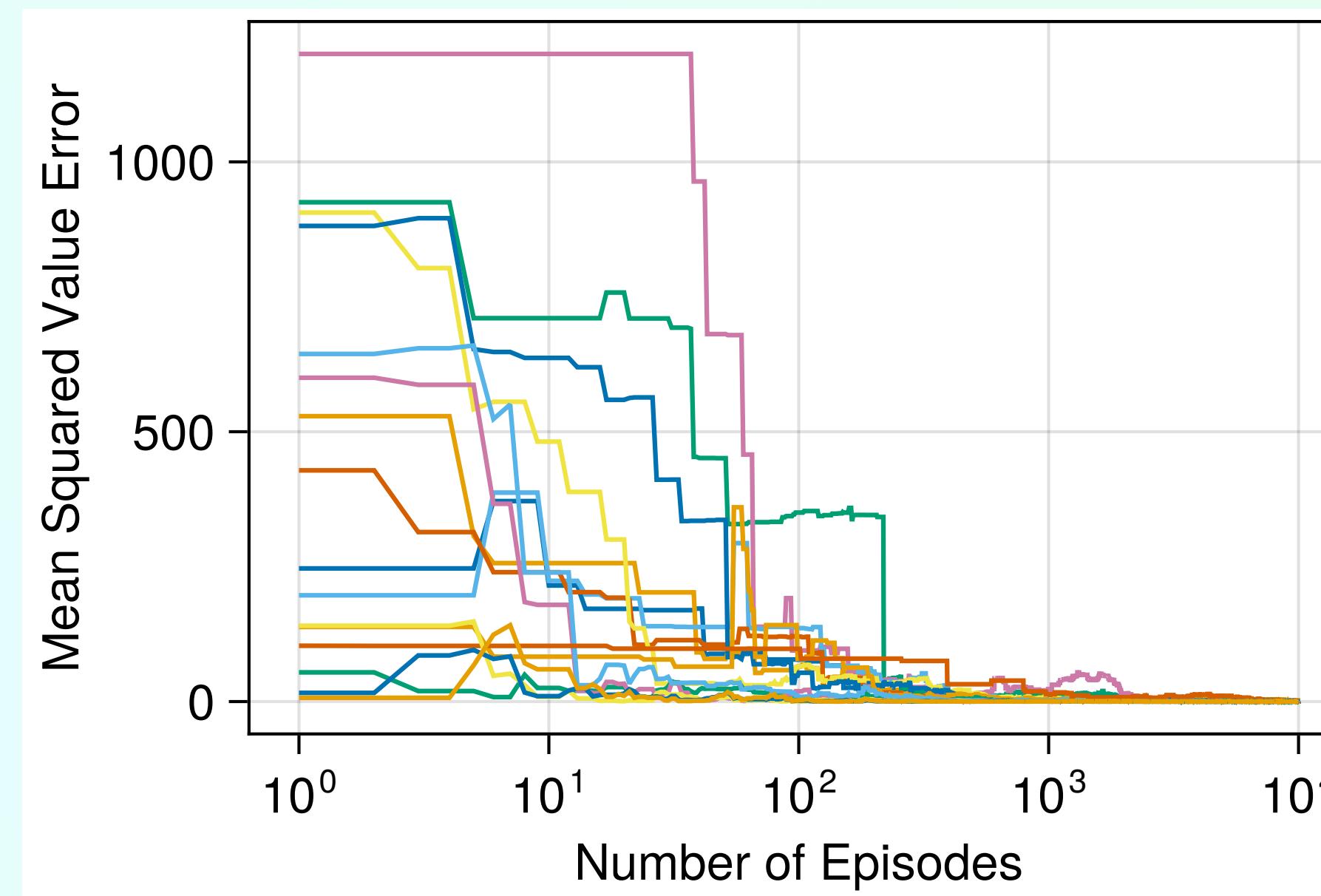
Start	2	3	4
1	6	7	8
11	12	13	14
16	17	18	Goal 19



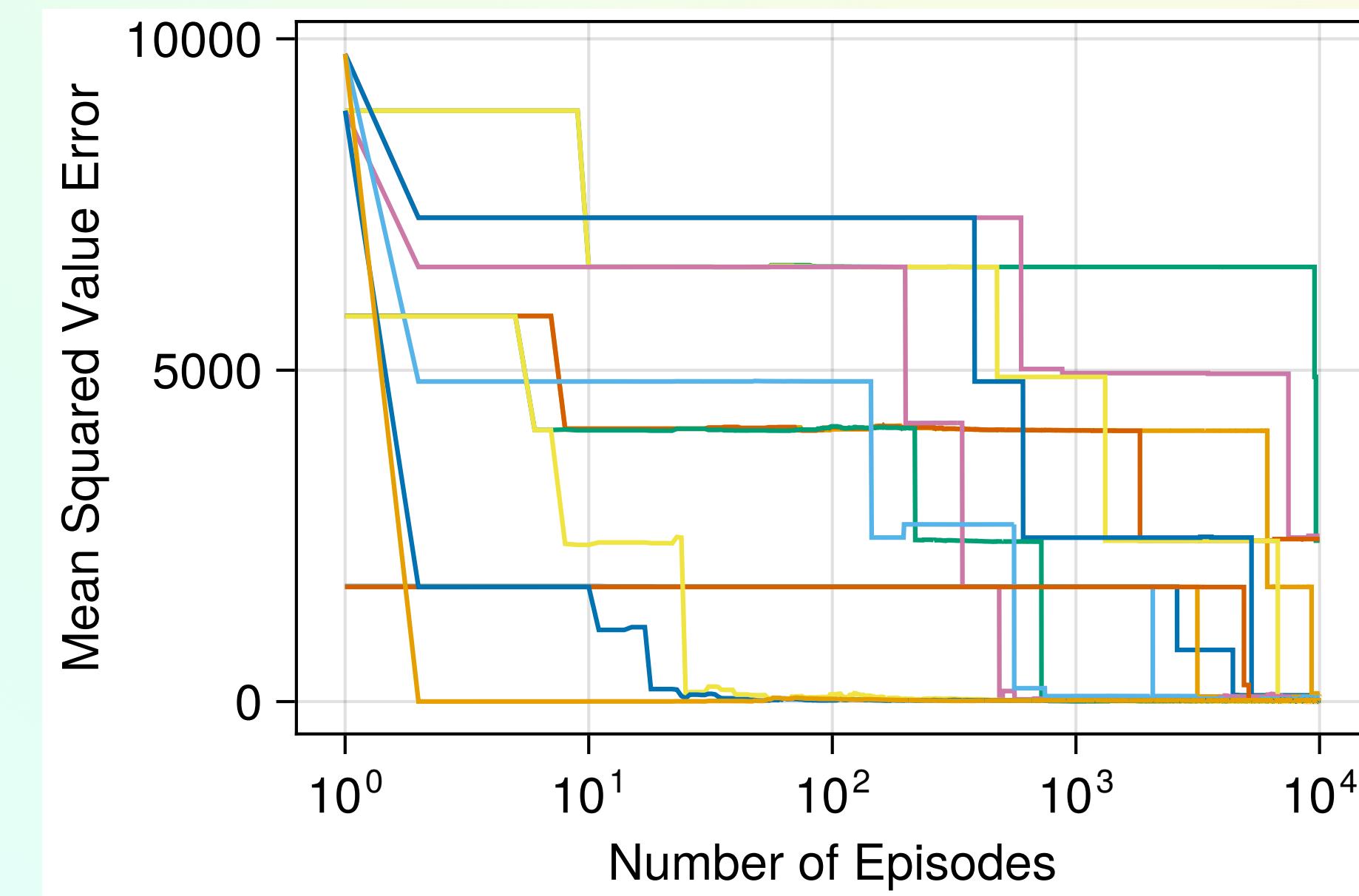
# EVERY-VISIT EVALUATION

HOW MUCH DATA?

$\epsilon = 0.5$



$\epsilon = 0.01$



# EVERY-VISIT EVALUATION

HOW MUCH DATA?

Some policies will require many thousands (millions?) of episodes to have accurate  $q_\pi$  estimates

# OFF-POLICY LEARNING

## LEARNING FROM ANOTHER POLICY

Instead of sampling actions from the policy  $\pi$  that we want to evaluate (the *evaluation policy*), sample actions from some policy  $b$  (the *behavior policy*).

Off-policy methods are any method that try to evaluate or improve policy  $\pi$  using data collected from some other policy

# OFF-POLICY LEARNING

LEARNING FROM ANOTHER POLICY

Distribution mismatch, averaging returns with  $b$ , leads to  $v_b$  and  $v_b(s) \neq v_\pi(s)$

Need to reweight the returns so that in expectation, they lead to  $v_\pi(s)$

# IMPORTANCE SAMPLING

## DERIVATION

$$X \in \mathcal{X}, X' \in \mathcal{X}$$

$$p(x) \doteq \Pr(X = x)$$

$$q(x) \doteq \Pr(X' = x)$$

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} \Pr(X = x)x = \sum_{x \in \mathcal{X}} p(x)x$$

From the definition

# IMPORTANCE SAMPLING

## DERIVATION

$$X \in \mathcal{X}, X' \in \mathcal{X}$$

$$p(x) \doteq \Pr(X = x)$$

$$q(x) \doteq \Pr(X' = x)$$

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} p(x)x$$

$$= \sum_{x \in \mathcal{X}} \frac{q(x)}{q(x)} p(x)x$$

Multiply by 1

# IMPORTANCE SAMPLING

## DERIVATION

$$X \in \mathcal{X}, X' \in \mathcal{X}$$

$$p(x) \doteq \Pr(X = x)$$

$$q(x) \doteq \Pr(X' = x)$$

$$\begin{aligned}\mathbb{E}[X] &= \sum_{x \in \mathcal{X}} p(x)x \\ &= \sum_{x \in \mathcal{X}} \frac{q(x)}{q(x)} p(x)x \\ &= \sum_{x \in \mathcal{X}} q(x) \frac{p(x)}{q(x)} x\end{aligned}$$

Rearrange terms

# IMPORTANCE SAMPLING

## DERIVATION

$$X \in \mathcal{X}, X' \in \mathcal{X}$$

$$p(x) \doteq \Pr(X = x)$$

$$q(x) \doteq \Pr(X' = x)$$

$$\begin{aligned}\mathbb{E}[X] &= \sum_{x \in \mathcal{X}} p(x)x \\ &= \sum_{x \in \mathcal{X}} \frac{q(x)}{q(x)} p(x)x \\ &= \sum_{x \in \mathcal{X}} q(x) \frac{p(x)}{q(x)} x \\ &= \sum_{x \in \mathcal{X}} \Pr(X' = x) \frac{p(x)}{q(x)} x = \mathbb{E} \left[ \frac{p(X')}{q(X')} X' \right]\end{aligned}$$

$q(x)$  implies  $x$  is distributed like  $X'$

# IMPORTANCE SAMPLING

DERIVATION FOR RETURN

$$b(a | s) \doteq \Pr(A'_t = a | S'_t = s)$$

$$G'_t = \sum_{k=0} \gamma^k R'_{t+1+k}$$

$H_{t:T-1} = A_t, R_{t+1}, S_{t+1}, \dots, R_T, S_T$  Random variable for the episode using  $\pi$  ignoring  $S_t$

$H'_{t:T-1} = A'_t, R'_{t+1}, S'_{t+1}, \dots, R'_T, S'_T$  Random variable for the episode using  $b$  ignoring  $S'_t$

# IMPORTANCE SAMPLING

DERIVATION FOR RETURN

$$\begin{aligned} v_{\pi}(s) &= \mathbb{E}[G_t | S_t = s] \\ &= \sum_g \Pr(G_t = g | S_t = s) g \\ &= \sum_h \Pr(H_{t:T-1} = h | S_t = s) \sum_{k=0}^{\infty} \gamma^k r_{t+1+k} \\ &= \sum_h \frac{\Pr(H'_{t:T-1} = h | S'_t = s)}{\Pr(H'_{t:T-1} = h | S'_t = s)} \Pr(H_{t:T-1} = h | S_t = s) \sum_{k=0}^{\infty} \gamma^k r_{t+1+k} \\ &= \sum_h \Pr(H'_{t:T-1} = h | S'_t = s) \frac{\Pr(H_{t:T-1} = h | S_t = s)}{\Pr(H'_{t:T-1} = h | S'_t = s)} \sum_{k=0}^{\infty} \gamma^k r_{t+1+k} \end{aligned}$$

# IMPORTANCE SAMPLING

DERIVATION FOR RETURN

$$\begin{aligned}\Pr(H_{t:T-1} = h \mid S_t = s) &= \sum_{a_t, s_{t+1}, r_{t+1}, \dots} \Pr(A_t = a_t, S_{t+1} = s_{t+1}, R_{t+1} = r_{t+1}, \dots \mid S_t = s) \\ &= \pi(a_t \mid s_t) p(s_{t+1}, r_{t+1} \mid s_t, a_t) \pi(a_{t+1} \mid s_{t+1}) p(s_{t+2}, r_{t+2} \mid s_{t+1}, a_{t+1}) \dots \\ &= \prod_{k=0}^{T-1} \pi(a_{t+k} \mid s_{t+k}) p(s_{t+1+k}, r_{t+1+k} \mid s_{t+k}, a_{t+k})\end{aligned}$$

# IMPORTANCE SAMPLING

DERIVATION FOR RETURN

$$\begin{aligned}\Pr(H_{t:T-1} = h \mid S_t = s) &= \sum_{a_t, s_{t+1}, r_{t+1}, \dots} \Pr(A_t = a_t, S_{t+1} = s_{t+1}, R_{t+1} = r_{t+1}, \dots \mid S_t = s) \\ &= \pi(a_t \mid s_t) p(s_{t+1}, r_{t+1} \mid s_t, a_t) \pi(a_{t+1} \mid s_{t+1}) p(s_{t+2}, r_{t+2} \mid s_{t+1}, a_{t+1}) \dots \\ &= \prod_{k=0}^{T-1} \pi(a_{t+k} \mid s_{t+k}) p(s_{t+1+k}, r_{t+1+k} \mid s_{t+k}, a_{t+k}) \\ \Pr(H'_{t:T-1} = h \mid S'_t = s) &= \prod_{k=0}^{T-1} b(a_{t+k} \mid s_{t+k}) p(s_{t+1+k}, r_{t+1+k} \mid s_{t+k}, a_{t+k})\end{aligned}$$

# IMPORTANCE SAMPLING

## DERIVATION FOR RETURN

$$\begin{aligned} v_\pi(s) &= \sum_h \Pr(H'_{t:T-1} = h | S'_t = s) \frac{\Pr(H'_{t:T-1} = h | S'_t = s)}{\Pr(H'_{t:T-1} = h | S'_t = s)} \sum_{k=0}^{\infty} \gamma^k r_{t+1+k} \\ &= \sum_h \Pr(H'_{t:T-1} = h | S'_t = s) \frac{\prod_{k=0}^{T-1} \pi(a_{t+k} | s_{t+k}) p(s_{t+1+k}, r_{t+1+k} | s_{t+k}, a_{t+k})}{\prod_{k=0}^{T-1} b(a_{t+k} | s_{t+k}) p(s_{t+1+k}, r_{t+1+k} | s_{t+k}, a_{t+k})} \sum_{k=0}^{\infty} \gamma^k r_{t+1+k} \\ &= \sum_h \Pr(H'_{t:T-1} = h | S'_t = s) \prod_{k=0}^{T-1} \frac{\pi(a_{t+k} | s_{t+k}) p(s_{t+1+k}, r_{t+1+k} | s_{t+k}, a_{t+k})}{b(a_{t+k} | s_{t+k}) p(s_{t+1+k}, r_{t+1+k} | s_{t+k}, a_{t+k})} \sum_{k=0}^{\infty} \gamma^k r_{t+1+k} \\ &= \sum_h \Pr(H'_{t:T-1} = h | S'_t = s) \underbrace{\prod_{k=0}^{T-1} \frac{\pi(a_{t+k} | s_{t+k})}{b(a_{t+k} | s_{t+k})}}_{=\rho_{t:T-1}} \sum_{k=0}^{\infty} \gamma^k r_{t+1+k} \\ &= \sum_h \Pr(H'_{t:T-1} = h | S'_t = s) \rho_{t:T-1} \sum_{k=0}^{\infty} \gamma^k r_{t+1+k} \\ &= \mathbb{E} [\rho_{t:T-1} G'_t | S'_t = s] \end{aligned}$$

# IMPORTANCE SAMPLING

## LIMITATIONS

If  $\pi(a | s) > 0$  then  $b(a | s) > 0$  for all  $s, a$

Cannot learn about an action under  $\pi$  if it won't be taken under  $b$

Need to pick  $b$  well or variance can become very large (meaning more samples are needed)

$$\frac{\pi(a | s)}{b(a | s)} > 1 \text{ if } b(a | s) < \pi(a | s)$$

$$\frac{\pi(a | s)}{b(a | s)} < 1 \text{ if } b(a | s) > \pi(a | s)$$

$\prod_{k=0}^{T-1-t} \frac{\pi(a_{t+k} | s_{t+k})}{b(a_{t+k} | s_{t+k})}$  will be near 0 or VERY large as  $T$  increases, unless  $\pi = b$

# REAL WORLD USE CASES

## GATHERING DATA FROM AN EXPERT

Have access to an expert decision-maker or a good heuristic, e.g., doctor.

Leverage this other policy to collect data, then try and find a good policy

- Decisions need to be stochastic
- (usually) need to know the probability of that decision for importance sampling
- Episode length needs to be short (<10 steps) or we need LOTS of data.

# REAL WORLD USE CASES

## HIGH CONFIDENCE OFF-POLICY EVALUATION

If you use  $\pi$  it in the real world it can be disastrous if it performs poorly

- Ad recommendation – lose lots of money
- Medicine – could lead to more deaths or worse health outcomes

We want a guarantee that you do not use a policy if it is bad.

$J(\pi) = \mathbb{E}[G_0] \geq \beta$  – Only use  $\pi$  if gets at least  $\beta$  return in expectation

$\beta$  could be  $J(b)$ , performance of the expert or another policy already in use.

We cannot guarantee this 100% of the time!

# REAL WORLD USE CASES

## HIGH CONFIDENCE OFF-POLICY EVALUATION

We want to lower-bound the performance of a policy and check and see if it is better than  $\beta$

Collect two data sets  $D_{train}$  and  $D_{test}$ .

Let  $\pi_{\text{candidate}}$  be the output of an algorithm that finds a policy using  $D_{train}$

Let  $l(D, \pi)$  returns lower confidence interval on  $J(\pi)$  using data set  $D$ , i.e.,  $\Pr(J(\pi) \geq l(D, \pi)) \geq 1 - \alpha$

If  $l(D_{test}, \pi_{\text{candidate}}) \geq \beta$

Return  $\pi_{\text{candidate}}$

Else return NO SOLUTION FOUND

No solution found means you need to collect more data and retest

# SUMMARY

## LEARNING FROM THE START STATE

Improving in every state can take a lot of data

Only need to improve/evaluate from start state distribution

- Easier if the optimal policy from the start state can ignore lots of states
- In the worst case, this can mean having to be optimal in every state.

Off-Policy

Can use behavior policy to search for an optimal policy without ever trying the optimal policy

Importance sampling can reduce variance by using a policy to reach states not likely to be sampled from the evaluation policy

Importance sampling ratios will go towards 0 or  $\infty$  quickly as episode length increases.

Generally, a start state distribution with high coverage ( $d_0(s) > 0$  for many  $s$ ) is desirable.

Can start in states that would be unlikely to be reached if there was only one or a few states included in  $S_0$

# NEXT CLASS

## WHAT YOU SHOULD DO

1. Quiz Due Wednesday
2. Watch week 2 of sampling methods material for Wednesday

Wednesday: Temporal Difference Learning for Prediction