MONTE CARLO AND OFF-POLICY LEARNING

NOTATIONAL NOTE:

FUNCTIONS VS PROBABILITY STATEMENTS

$$\Pr(A \mid B)$$

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

Probability statements are mathematical statements about how often events occur.

$$f\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}\ ,\ z = f(x, y)$$

Functions process inputs and return a value.

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

The $\mid$ 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: $p(s', r \mid s, a)$ — not a standard way to write a function. The typical way would be to replace '$\mid$' 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 $\pi(a \mid s)$ — 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 π_*

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 π: $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$
  $G \leftarrow 0$
  Loop for each step of episode, $t = T-1, T-2, \ldots, 0$:
    $G \leftarrow \gamma G + R_{t+1}$
    Unless the pair S_t, A_t appears in $S_0, A_0, S_1, A_1 \ldots, S_{t-1}, A_{t-1}$:
      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

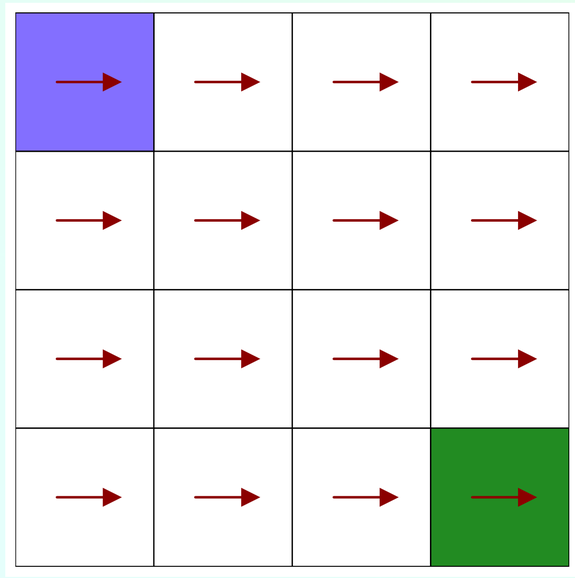
- Car starts in the driveway and other cars are “wild”

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)



$\log_{10} \Pr(S = s)$

ENSURING EXPLORATION

FROM START STATE

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

Then all reachable state action pairs will be tried infinitely often

ENSURING EXPLORATION

FROM START STATE

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

Then all reachable state action pairs will be tried infinitely often

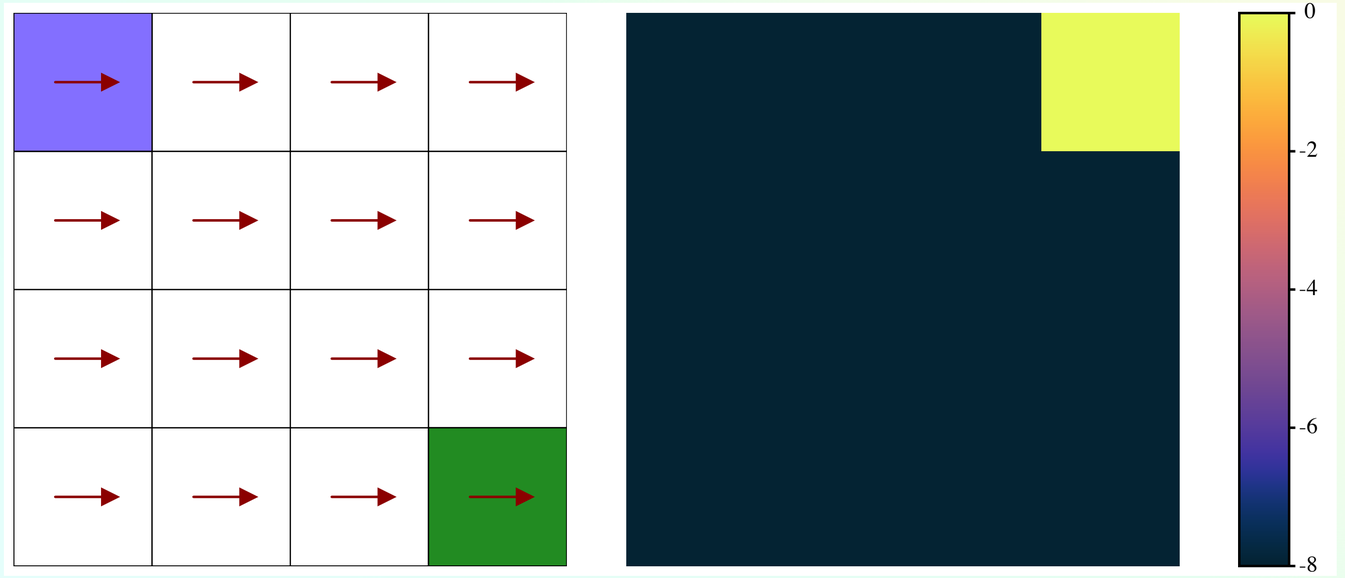
$\epsilon-$greedy policy

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

ENSURING EXPLORATION

FROM START STATE

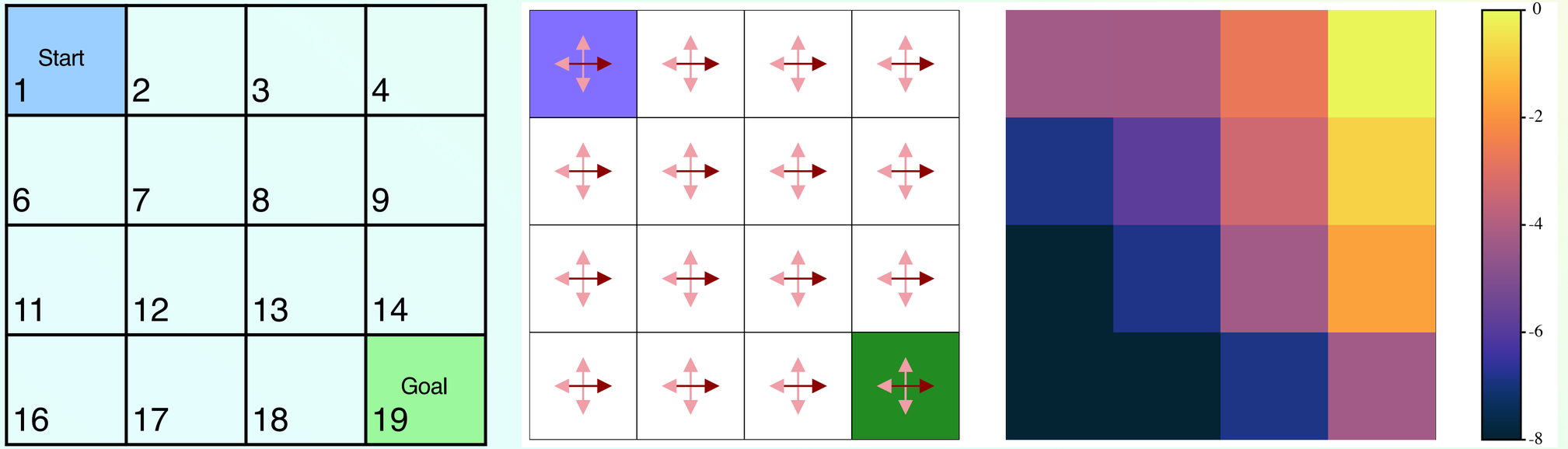
Go right policy with $\epsilon = 0$



ENSURING EXPLORATION

FROM START STATE

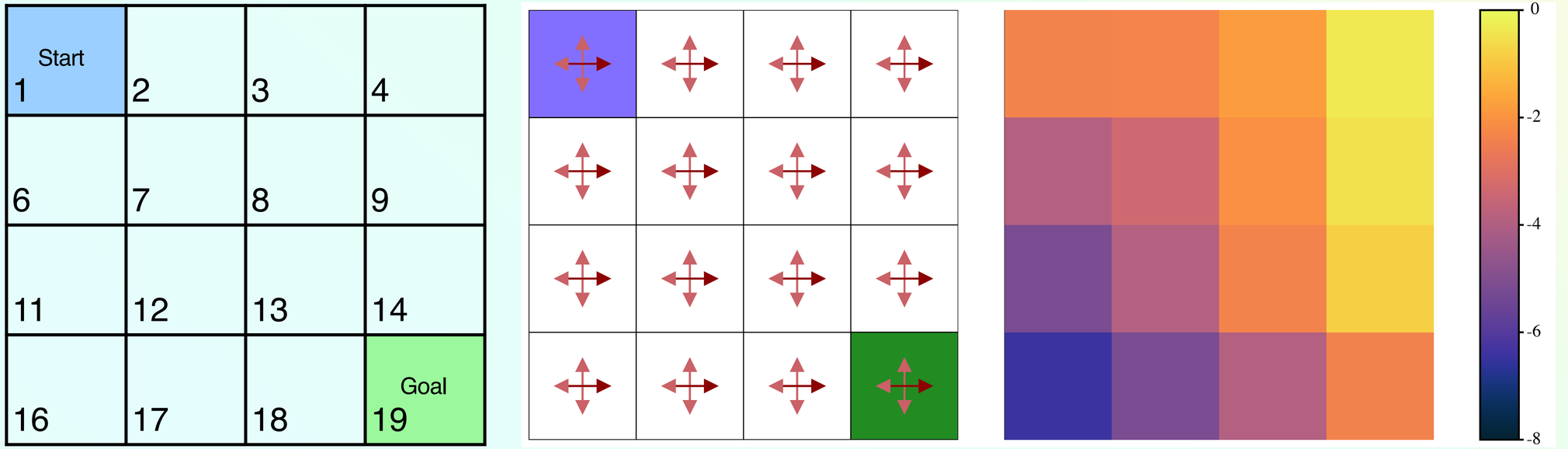
Go right policy with $\epsilon = 0.01$



ENSURING EXPLORATION

FROM START STATE

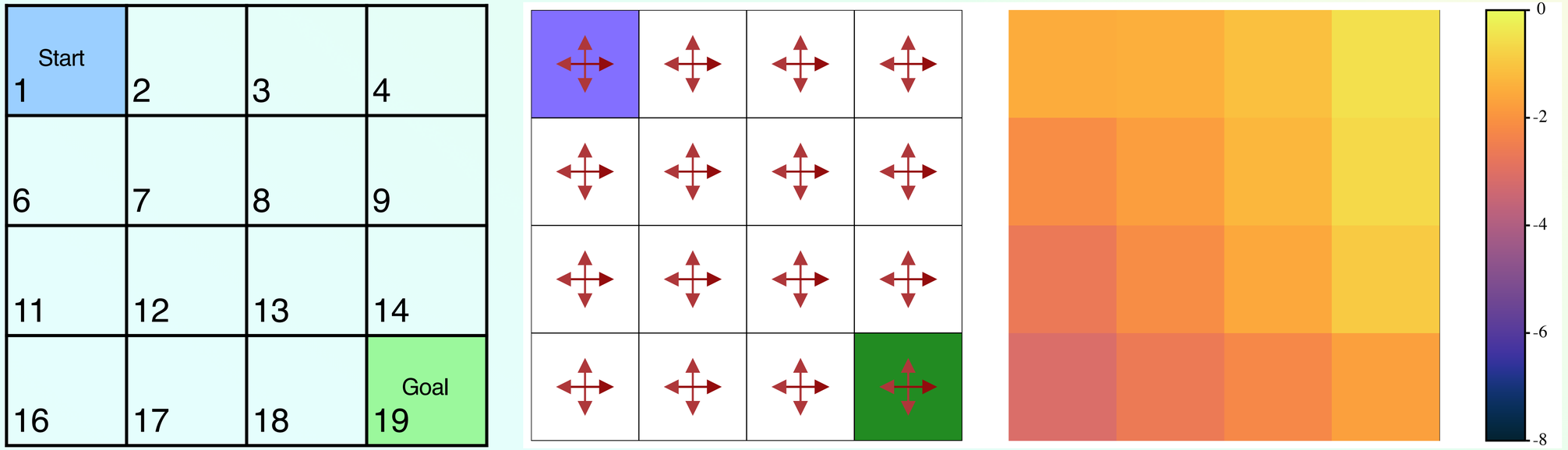
Go right policy with $\epsilon = 0.1$



ENSURING EXPLORATION

FROM START STATE

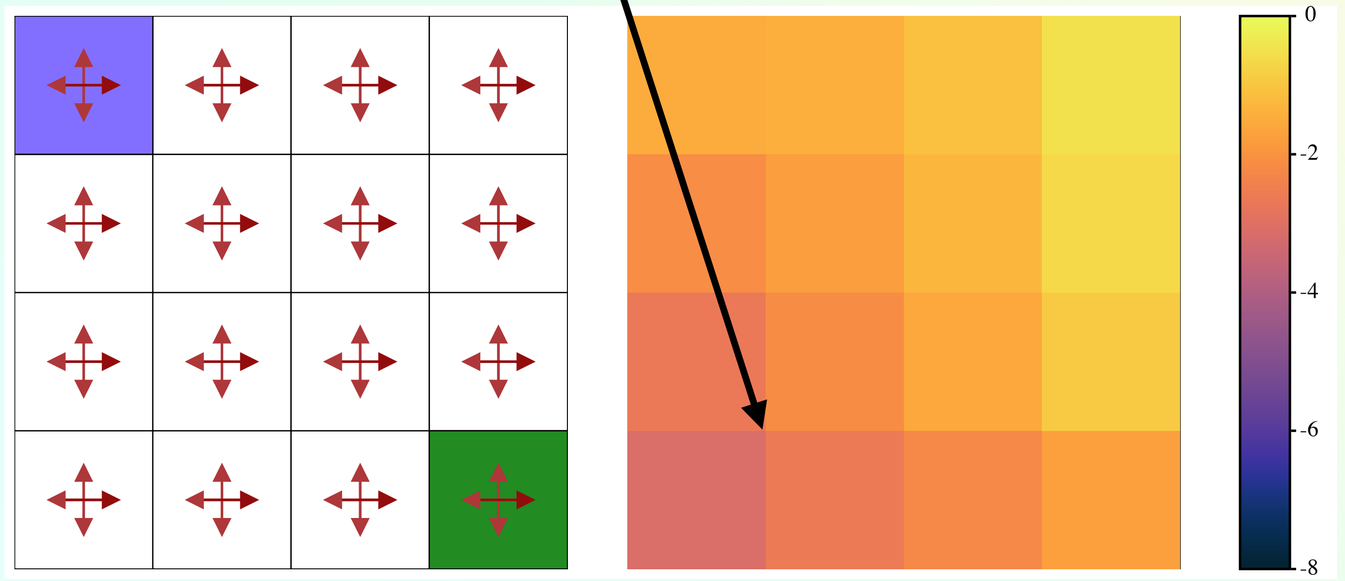
Go right policy with $\epsilon = 0.5$



ENSURING EXPLORATION

FROM START STATE

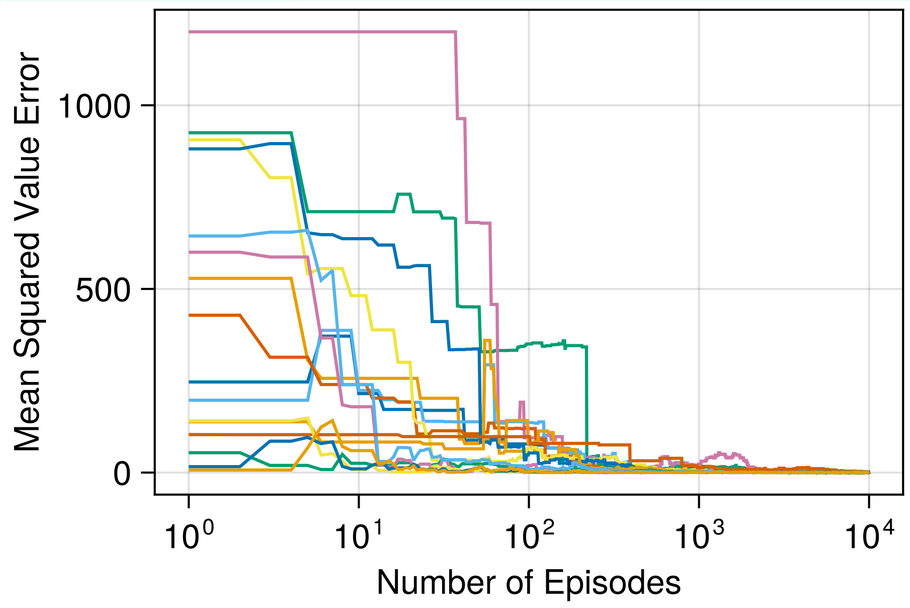
Go right policy with $\epsilon = 0.5$



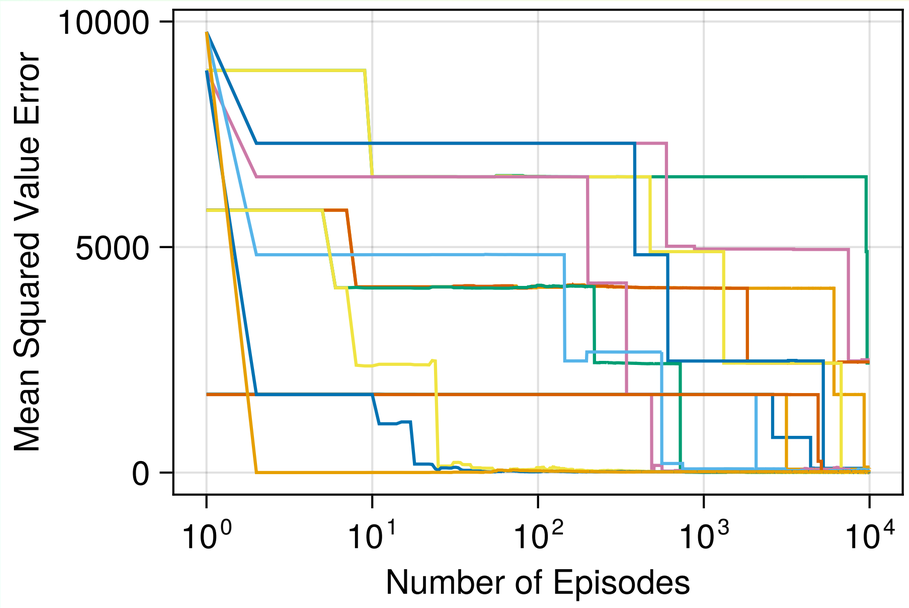
EVERY-VISIT EVALUATION

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_π estimates

OFF-POLICY LEARNING

LEARNING FROM ANOTHER POLICY

Instead of sampling actions from the policy π 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 π 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

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

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

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

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

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

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

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

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

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

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

$H'_{t:T-1} = A'_t, R'_{t+1}, S'_{t+1}, \ldots, R'_T, S'_T$ 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

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

IMPORTANCE SAMPLING

DERIVATION FOR RETURN

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

$$
\Pr(H'_{t:T-1} = h \,|\, S'_t = s) = \prod_{k=0}^{T-1-t} b\left(a_{t+k} \,|\, s_{t+k}\right) p\left(s_{t+1+k}, r_{t+1+k} \,|\, s_{t+k}, a_{t+k}\right)
$$

IMPORTANCE SAMPLING

DERIVATION FOR RETURN

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

IMPORTANCE SAMPLING

LIMITATIONS

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

Cannot learn about an action under π if it won't be taken under b

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

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

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

$\prod_{k=0}^{T-1-t} \frac{\pi(a_{t+k} \mid s_{t+k})}{b(a_{t+k} \mid s_{t+k})}$ 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 π 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 π if gets at least β return in expectation

β 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 β

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

Let π_{candiate} 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., $\Pr\left(J(\pi) \geq l(D, \pi)\right) \geq 1 - \alpha$

If $l\left(D_{test}, \pi_{\text{canditate}}\right) \geq \beta$

Return π_{candiate}

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 ∞ 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