MONTE CARLO METHODS

SOLVING MDPS

WHERE WE ARE

Dynamic Programming:

- Taught us how to evaluate and find optimal policy when we know p and r
- No interaction required

We rarely have perfect estimates of p and sometimes we know r

Storing p can be expensive $|p| = |\mathcal{S}|^2 \times |\mathcal{A}|$

We cannot always do dynamics programming, so what can we do?

SOLVING MDPS

WHERE WE ARE

Dynamic Programming:

- Taught us how to evaluate and find optimal policy when we know p and r
- No interaction required

We rarely have perfect estimates of p and sometimes we know r

Storing p can be expensive $|p| = |\mathcal{S}|^2 \times |\mathcal{A}|$

We cannot always do dynamics programming, so what can we do?

Learn from interaction!

MONTE CARLO METHODS

GENERAL POLICY ITERATION

Input: n number of samples to estimate q_π

π_1 — Initialize the first policy (random?)

For i in $\{1,2,3,\ldots\}$

$Q \leftarrow \text{evaluate}(\pi, n)$

$\pi_{i+1} \leftarrow \text{new\_policy}(Q)$

MONTE CARLO METHODS

GENERAL POLICY ITERATION

Input: n number of samples to estimate q_π

π_1 — Initialize the first policy (random?)

For i in {1,2,3,…}

$\boxed{Q \leftarrow \text{evaluate}(\pi, n)}$

$\pi_{i+1} \leftarrow \text{new\_policy}(Q)$

MONTE CARLO POLICY EVALUATION

ESTIMATING q_π

We want $\forall s, a,\ Q(s,a) \approx q_\pi(s,a)$

Then being greedy or ϵ-greedy on q will lead to an improved π

$q_\pi(s,a) = \mathbb{E}[G_t | S_t = s, A_t = a]$

- Need to start in state s and take action a and then follow π to compute one G_0
- Repeat this many times to get $G_{0,1}, G_{0,2}, \ldots, G_{0,n}$

$$Q(s,a) = \frac{1}{n}\sum_{i=1}^{n} G_{0,i}$$

Repeat for all s, a

MONTE CARLO POLICY EVALUATION

ESTIMATING q_π

We want $\forall s, a,\ Q(s, a) \approx q_\pi(s, a)$

Then being greedy or ϵ-greedy on q will lead to an improved π

$q_\pi(s, a) = \mathbb{E}[G_t \,|\, S_t = s, A_t = a]$

Need to start in state s and take action a and then follow π to compute one G_0

Repeat this many times to get $G_{0,1}, G_{0,2}, \ldots, G_{0,n}$

$$Q(s, a) = \frac{1}{n}\sum_{i=1}^{n} G_{0,i}$$

Not efficient! We encounter many stay in an episode and could get more estimates of Q

Repeat for all s, a

FIRST-VISIT MONTE CARLO

ESTIMATING q_π

The first time the agent is in state s and takes action a during an episode, save the return G_t

Update Q iteratively:

$$Q_{n+1}(s,a) = Q_n(s,a) + \frac{1}{n}(G_t - Q_n(s,a))$$

FIRST-VISIT MONTE CARLO

ESTIMATING q_π

The first time the agent is in state s and takes action a during an episode, save the return G_t

Update Q iteratively:

$$Q_{n+1}(s,a) = Q_n(s,a) + \frac{1}{n}(G_t - Q_n(s,a))$$

Consider the first episode:

$$s_1, a_2, R_1, s_3, a_1, R_2, s_1, a_1, R_3, s_1, a_2, R_4, s_3, a_1, R_5, s_\infty$$

FIRST-VISIT MONTE CARLO

ESTIMATING q_π

The first time the agent is in state s and takes action a during an episode, save the return G_t

Update Q iteratively:

$$Q_{n+1}(s,a) = Q_n(s,a) + \frac{1}{n}(G_t - Q_n(s,a))$$

Consider the first episode:

$$s_1, a_2, R_1, s_3, a_1, R_2, s_1, a_1, R_3, s_1, a_2, R_4, s_3, a_1, R_5, s_\infty$$

$$Q(s_1,a_2) = Q_2(s_1,a_2) = Q_1(s_1,a_2) + \frac{1}{1}(G_0 - Q_1(s_1,a_2)) = G_0$$

FIRST-VISIT MONTE CARLO

ESTIMATING q_π

The first time the agent is in state s and takes action a during an episode, save the return G_t

Update Q iteratively:

$$Q_{n+1}(s,a) = Q_n(s,a) + \frac{1}{n}(G_t - Q_n(s,a))$$

Consider the first episode:

$$s_1, a_2, R_1, \boxed{s_3, a_1}, \boxed{R_2}, s_1, a_1, \boxed{R_3}, s_1, a_2, \boxed{R_4}, s_3, a_1, \boxed{R_5}, s_\infty$$

$$Q(s_1, a_2) = Q_2(s_1, a_2) = Q_1(s_1, a_2) + \frac{1}{1}(G_0 - Q_1(s_1, a_2) = G_0$$

$$Q(s_3, a_1) = \boxed{G_1}$$

FIRST-VISIT MONTE CARLO

ESTIMATING q_π

The first time the agent is in state s and takes action a during an episode, save the return G_t

Update Q iteratively:

$$Q_{n+1}(s,a) = Q_n(s,a) + \frac{1}{n}(G_t - Q_n(s,a))$$

Consider the first episode:

$$s_1, a_2, R_1, s_3, a_1, R_2, \boxed{s_1, a_1}, \boxed{R_3}, s_1, a_2, \boxed{R_4}, s_3, a_1, \boxed{R_5}, s_\infty$$

$$Q(s_1, a_2) = Q_2(s_1, a_2) = Q_1(s_1, a_2) + \frac{1}{1}(G_0 - Q_1(s_1, a_2) = G_0$$

$$Q(s_3, a_1) = G_1$$

$$Q(s_1, a_1) = \boxed{G_2}$$

FIRST-VISIT MONTE CARLO

ESTIMATING q_π

The first time the agent is in state s and takes action a during an episode, save the return G_t

Update Q iteratively:

$$Q_{n+1}(s,a) = Q_n(s,a) + \frac{1}{n}(G_t - Q_n(s,a))$$

Consider the first episode:

$$s_1, a_2, R_1, s_3, a_1, R_2, s_1, a_1, R_3, \boxed{s_1, a_2}, \boxed{R_4}, s_3, a_1, \boxed{R_5}, s_\infty$$

$$Q(s_1,a_2) = Q_2(s_1,a_2) = Q_1(s_1,a_2) + \frac{1}{1}(G_0 - Q_1(s_1,a_2) = G_0$$

Not the first time in s, a

$$Q(s_3,a_1) = G_1$$

$$Q(s_1,a_1) = G_2$$

FIRST-VISIT MONTE CARLO

ESTIMATING q_π

The first time the agent is in state s and takes action a during an episode, save the return G_t

Update Q iteratively:

$$Q_{n+1}(s,a) = Q_n(s,a) + \frac{1}{n}(G_t - Q_n(s,a))$$

Consider the first episode:

$$s_1, a_2, R_1, s_3, a_1, R_2, s_1, a_1, R_3, s_1, a_2, R_4, \boxed{s_3, a_1}, \boxed{R_5}, s_\infty$$

$$Q(s_1,a_2) = Q_2(s_1,a_2) = Q_1(s_1,a_2) + \frac{1}{1}(G_0 - Q_1(s_1,a_2) = G_0$$

Not the first time in s, a

$$Q(s_3,a_1) = G_1$$

$$Q(s_1,a_1) = G_2$$

FIRST-VISIT MONTE CARLO

ESTIMATING q_π

The first time the agent is in state s and takes action a during an episode, save the return G_t

Update Q iteratively:

$$Q_{n+1}(s,a) = Q_n(s,a) + \frac{1}{n}(G_t - Q_n(s,a))$$

Consider the first episode:

$$s_1, a_2, R_1, s_3, a_1, R_2, s_1, a_1, R_3, s_1, a_2, R_4, s_3, a_1, R_5, s_\infty$$

$Q(s_1, a_2) = Q_2(s_1, a_2) = Q_1(s_1, a_2) + \frac{1}{1}(G_0 - Q_1(s_1, a_2) = G_0$

$Q(s_3, a_1) = G_1$

$Q(s_1, a_1) = G_2$

$Q(s_2, a) = Q_1(s_2, a)$ — No data —> no update

$Q(s_3, a_2) = Q_1(s_3, a_2)$ — No data —> no update

EVERY-VISIT MONTE CARLO

ESTIMATING q_π

Every time the agent is in state s and takes action a during an episode, save the return G_t

More data = Better estimate of q_π? (Not always)

Consider the first episode:

$$s_1, a_2, R_1, s_3, a_1, R_2, s_1, a_1, R_3, s_1, a_2, R_4, s_3, a_1, R_5, s_\infty$$

$$Q(s_1, a_2) = \frac{1}{2}(G_0 + G_2)$$

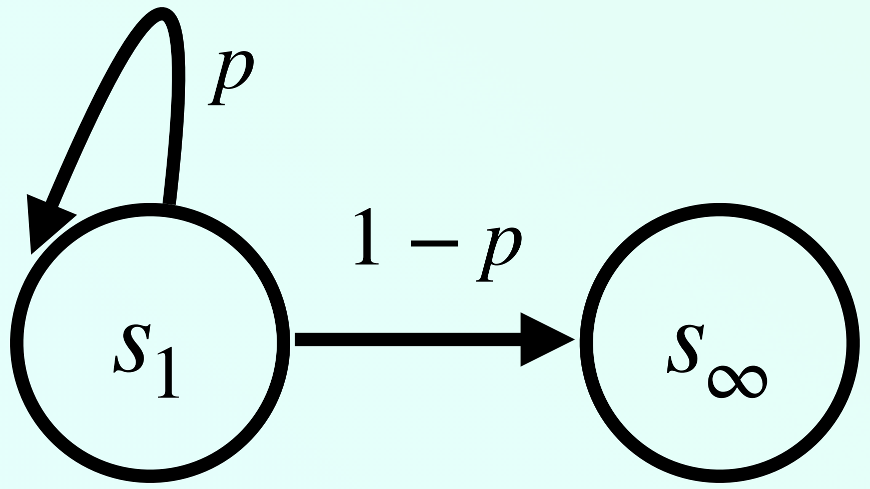
$$Q(s_3, a_1) = \frac{1}{2}(G_1 + G_3)$$

$$Q(s_1, a_1) = G_2$$

EXERCISE

FIRST VISIT VS EVERY VISIT

Exercise 5.5 Consider an MDP with a single nonterminal state and a single action that transitions back to the nonterminal state with probability p and transitions to the terminal state with probability $1-p$. Let the reward be +1 on all transitions, and let $\gamma=1$. Suppose you observe one episode that lasts 10 steps, with a return of 10. What are the first-visit and every-visit estimators of the value of the nonterminal state? □



STRONG LAW OF LARGE NUMBERS

KOLMOGOROV STRONG LAW OF LARGE NUMBERS

Let $\{X_i\}_{i=1}^{\infty}$ be independent random variables. If all X_i have the same mean and bounded variance, then $\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)_{n=1}^{\infty}$ is a sequence of random variables that converges almost surely to $\mathbb{E}[X_1]$

STRONG LAW OF LARGE NUMBERS

KOLMOGOROV STRONG LAW OF LARGE NUMBERS

Let $\{X_i\}_{i=1}^{\infty}$ be independent random variables. If all X_i have the same mean and bounded variance, then $\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)_{n=1}^{\infty}$ is a sequence of random variables that converges almost surely to $\mathbb{E}[X_1]$

If $V_n(s) = \frac{1}{n}\sum_{i=1}^{n} G_{t,i}$ has bounded variance then $V_n(s) \to v_\pi(s)$

VARIANCE OF VALUE ESTIMATE

DERIVATION

$\text{Var}(V_n(s)) =$

VARIANCE OF VALUE ESTIMATE

DERIVATION

$$
\begin{aligned}
\text{Var}(V_n(s)) &= \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} G_{t,i}\right) \\
&= \frac{1}{n^2}\text{Var}\left(\sum_{i=1}^{n} G_{t,i}\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \text{Var}(G_{t,i}) \\
&= \frac{1}{n^2} n \text{Var}(G_t) \\
&= \frac{1}{n}\text{Var}(G_t)
\end{aligned}
$$

First visit converges to v_π

EVERY-VISIT

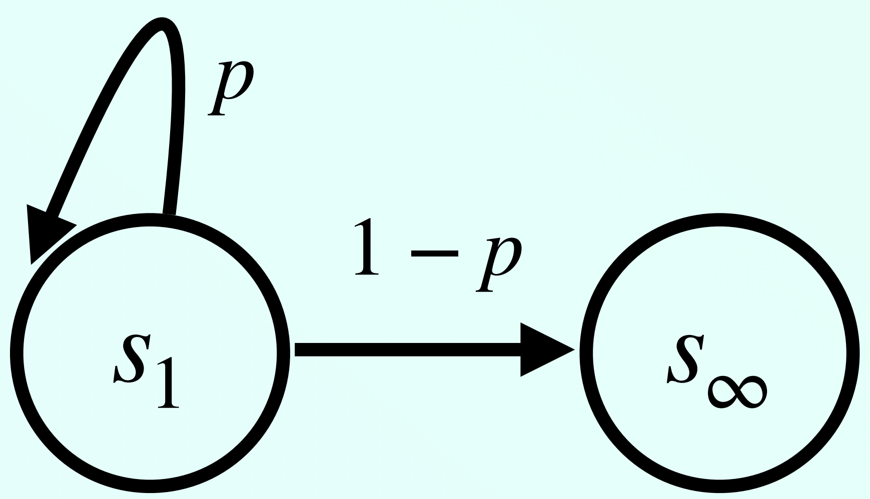
WHY IS IT BAD

Need samples to be independent and have the same mean to use the strong law of large numbers.

Returns from the same episode are not independent

Do they have the same mean?

$\mathbb{E}[G_0] = \mathbb{E}[G_1] = \ldots = \mathbb{E}[G_{T-1}]$?



NEXT CLASS

WHAT YOU SHOULD DO

1. Quiz Due tonight
2. Blackjack assignment due tonight (not graded and takes no effort)

Monday: Monte Carlo and Off-Policy