

MODEL

FREE

PREDICTION

MODEL-FREE PREDICTION

Problem: Finding Value function of unknown dynamics by acting and observing

- Monte Carlo (MC) methods can solve this problem
- Simple idea: Value = mean return
- Caveats: can only be applied to episodic environments
 - All episodes must end

MONTÉ-CARLO POLICY EVALUATION

Goal: learn V^π from episodes of experience under policy π :

$$x_1, u_1, l_1, x_2, u_2, l_2, \dots, x_n \sim \pi$$

Cost-to-go is total discounted cost:

$$J_t = l_t + \gamma l_{t+1} + \gamma^2 l_{t+2} + \dots + \gamma^{T-1} l_{t+T-1}$$

Value function is expected cost-to-go under policy π :

$$V^\pi(x) = \mathbb{E}_\pi [J_t \mid x_t = x] \leftarrow \text{implicitly assumes some stochasticity (policy, cost, MDP)}$$

MC policy evaluation estimates V^π as the average cost-to-go.

FIRST-VISIT MC POLICY EVALUATION

The first time that a state x is visited in an episode:

- Increment counter $N(x) := N(x) + 1$
- Increment total cost $C(x) := C(x) + J(x)$
- Estimate Value as $V(x) = C(x) / N(x)$

By the law of large numbers $V(x) \rightarrow V^\pi(x)$ as $N(x) \rightarrow \infty$

Another version exists where counter and total cost are incremented every time a state is visited (every visit)

INCREMENTAL MC UPDATES

IDEA: Compute mean incrementally:

$$V_N = \frac{1}{N} \sum_{j=1}^N T_j = \frac{T_N + \sum_{j=1}^{N-1} T_j}{N} = \frac{T_N + (N-1)V_{N-1}}{N} = V_{N-1} + \frac{1}{N} (T_N - V_{N-1})$$

In non-stationary problems it can be useful to track a running mean, i.e. forget old episodes:

$$V(x) := V(x) + \alpha (T - V(x))$$

TEMPORAL DIFFERENCE LEARNING - TD(0)

- Alternative to MC
- Estimate cost-to-go by 1-step lookahead:

$$V(x_t) := V(x_t) + d_t \left(\underbrace{d_t + \gamma V(x_{t+1})}_{\text{TD error}} - V(x_t) \right)$$

TD target = estimated return

- Compare to MC incremental updates:

$$V(x_t) := V(x_t) + d_t (J_t - V(x_t))$$

$$J_t = \sum_{i=0}^{\infty} \gamma^i l_i$$

- Policy evaluation is TD(0) with $d_t = 1$ and sampling all states uniformly

TD(0) PROPERTIES

- TD(0) is a stochastic approximation algorithm
- Since Bellman operator has a unique fixed point
⇒ if TD(0) converges (and all states are sampled infinitely many times) then it must converge to V^π
- Convergence is guaranteed if step-size sequence satisfies Robbins-Monro (RM) conditions

$$\sum_{t=0}^{\infty} d_t = \infty \quad \sum_{t=0}^{\infty} d_t^2 < \infty \quad \left[\text{e.g. } d_t = \frac{\bar{\alpha}}{t} \right]$$

- In practice people often use constant step size

MC vs TD

- TD can learn at every step
- MC must wait until the end of an episode to know J
- TD works in non-terminating environments
- MC works only in episodic environments
- TD target is a biased estimate of $V^\pi(x_t)$
- Cost-to-go J_t is unbiased estimate of $V^\pi(x_t)$
- TD target is lower variance than J_t
 - J_t depends on many random transitions
 - TD target depends only on one transition

- MC has high variance, zero bias
 - good convergence properties
 - (even with function approximation)
 - not very sensitive to initialization
 - simple to understand and use

- TD has low variance, some bias
 - usually more efficient than MC
 - $TD(0)$ converges to $V_{\pi}(x_t)$
 - (but not always with function approximation)
 - more sensitive to initialization

AB EXAMPLE

Two states (A, B), no discounting, 8 episodes
($\gamma = 1$)

B, 1

B, 1

B, 1

B, 1

B, 1

B, 1

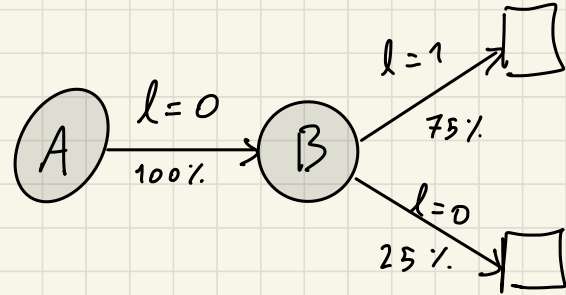
B, 0

A, 0, B, 0

$$V(B) = 0.75$$

$$V(A) = 0.75$$

MC would find $V(A) = 0$



What's $V(A)$, $V(B)$?

AB EXAMPLE

- 8) A, 0, B, 0
- 7) B, 1
- 6) B, 1
- 5) B, 1
- 4) B, 1
- 3) B, 1
- 2) B, 1
- 1) B, 0

$$1) V_1(B) = V_0(B) + \alpha (0 + 0 - V_0(B)) \\ = (1 - \alpha) V_0(B)$$

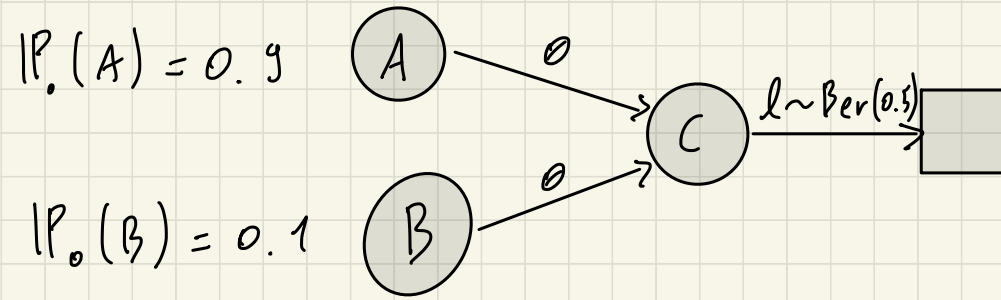
$$2) V_2(B) = V_1(B) + \alpha (1 + 0 - V_1(B)) \\ = \alpha + (1 - \alpha) V_1(B) \\ = \alpha + (1 - \alpha)^2 V_0(B)$$

$$3) V_3(B) = V_2(B) + \alpha (1 - V_2(B)) \\ = \alpha + (1 - \alpha) V_2(B) = \\ = \alpha + (1 - \alpha) \alpha + \cancel{(1 - \alpha)^3 V_0(B)}$$

....

$$8) V_8(A) = \cancel{V_0(A)} + \alpha (0 + V_7(B) - \cancel{V_0(A)}) = \\ = \alpha V_7(B)$$

ABC EXAMPLE



- After k visits of B , C has visited $\approx 10k$ times

- $\text{Var}[\hat{V}(C)] \approx \frac{1}{10k}$

- W, th TD(0) the variance of TD target for $V(B)$ decreases with the # of episodes

$$\hat{V}(B) := \hat{V}(B) + \alpha (0 + \hat{V}(C) - \hat{V}(B))$$

- W, th MC the variance of the target is always 0.25
 - MC is slower to converge