MODEL-FREE PREDICTION WITH MONTE CARLO

Kevin Chen and Zack Khan

University of Maryland

CMSC389F: Reinforcement Learning, Spring 2018

LEARNING GOALS

- Understand benefit of MC algorithms over DP algorithms
- Figure out how to use MC to evaluate policies
- Understand on-policy first-visit MC algorithm
- Understand on-policy every-visit MC algorithm
- Understand incremental mean (online algorithm)

WHY MODEL-FREE?

Dynamic programming (value/policy iteration) requires **complete knowledge** about the environment (MDP)

In the real-world, we don't have complete knowledge about the environment

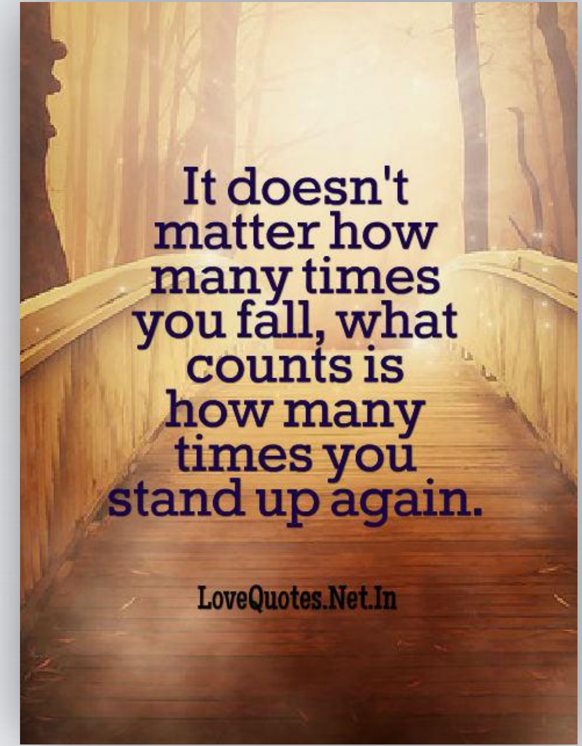
MODEL-FREE RL

Previously: Solve MDP with model-based planning using DP

Now: Estimate value function without a model of the environment using MC

ESSENCE OF MONTE-CARLO

MC methods learn from episodes of experience, which are collected by interacting with the environment



EPISODE OF EXPERIENCE

(State, Action, Reward, Next State)
(State, Action, Reward, Next State)
(State, Action, Reward, Next State)
...

MONTE-CARLO

1. MC methods learn from episodes of experience
2. MC is model-free: no knowledge of transitions / rewards
3. MC learns from complete episodes (no stopping midway)
4. MC uses a simple idea: treating value = empirical mean return
5. Can only apply MC to episodic MDPs (all episodes must end)

OTHER MODEL-FREE METHODS

Monte Carlo

TD Learning

TD-Lambda

MONTE-CARLO ALGORITHM

1. Have some initial estimates
2. Act and gain experience
3. Sample episodes of experience
4. Improve estimates
5. Repeat (new episode)

MONTE CARLO POLICY EVALUATION

Goal: Learn the value function for a policy from episodes of experience

Recall that return is the total discounted reward

$G_t = R_{t+1} + \gamma R_{t+2} + \ldots + \gamma^{(T-1)} R_T$

MC policy evaluation uses empirical mean return to estimate the value function

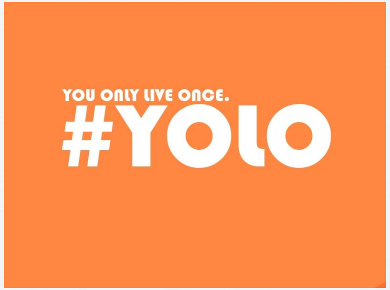
FIRST-VISIT MC POLICY EVALUATION

The first time-step t that state s is visited in an episode

- Increment counter $N(s) \leftarrow N(s) + 1$
- Increment total return $S(s) \leftarrow S(s) + G_t$

Value is estimated by mean return $V(s) = S(s)/N(s)$

By law of large numbers, $V(s) \rightarrow v_\pi(s)$ as $N(s) \rightarrow \infty$



EVERY-VISIT MC POLICY EVALUATION

Every time-step t that state s is visited in an episode

- Increment counter $N(s) \leftarrow N(s) + 1$
- Increment total return $S(s) \leftarrow S(s) + G_t$

Value is estimated by mean return $V(s) = S(s)/N(s)$

By law of large numbers, $V(s) \rightarrow v_\pi(s)$ as $N(s) \rightarrow \infty$

BLACKJACK

(Example borrowed from Silver 2016)

States (200 of them):
Current sum (12-21)
Dealer's showing card (ace-10)
Do I have a "useable" ace? (yes-no)

Actions:
Stop: Stop receiving cards (and terminate)
Hit: Take another card

Rewards for Stop:
+1 if sum of cards > sum of dealer cards
0 if sum of cards = sum of dealer cards
-1 if sum of cards < sum of dealer cards

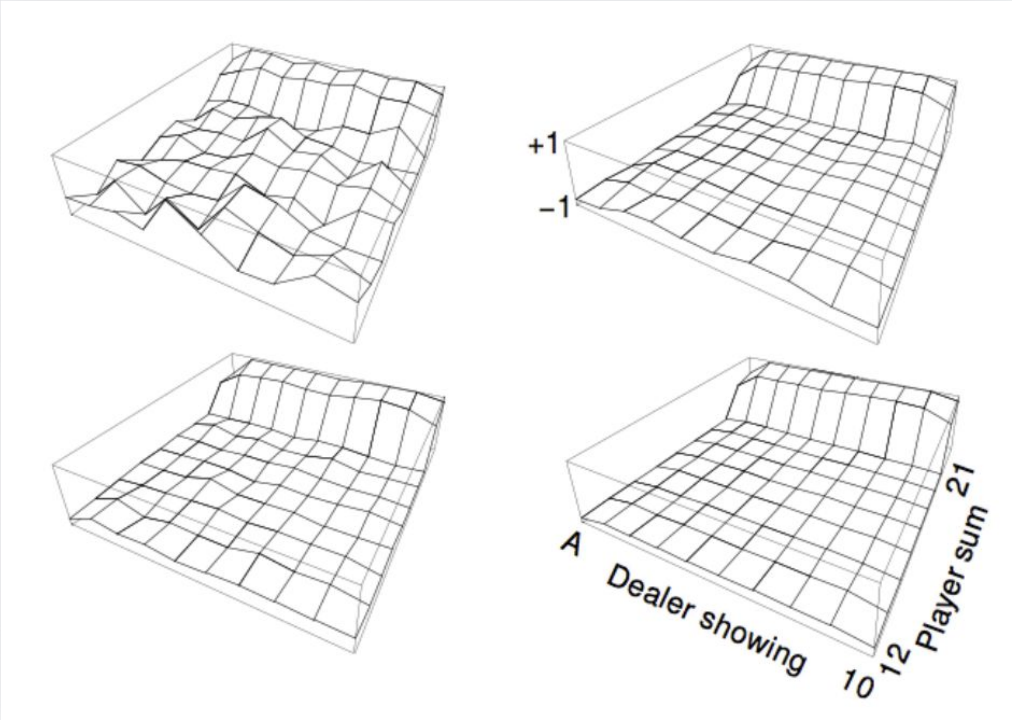
Reward for Hit:
-1 if sum of cards > 21 (and terminate)
0 otherwise

Transitions:
Automatically hit if sum of cards < 12

MC POLICY EVALUATION FOR BLACKJACK

Usable Ace

No Ace



10,000 episodes

500,000 episodes

Policy: **stop** if sum of cards **≥ 20**, otherwise **hit**

INCREMENTAL MEAN

The mean of an incoming sequence can be computed incrementally

$$
\begin{aligned}
\mu_k &= \frac{1}{k}\sum_{j=1}^{k} x_j \\
&= \frac{1}{k}\left(x_k + \sum_{j=1}^{k-1} x_j\right) \\
&= \frac{1}{k}\left(x_k + (k-1)\mu_{k-1}\right) \\
&= \mu_{k-1} + \frac{1}{k}\left(x_k - \mu_{k-1}\right)
\end{aligned}
$$

INCREMENTAL MONTE CARLO

Update V(s) incrementally after episode

For each state S_t with return G_t

$$N(S_t) \leftarrow N(S_t) + 1$$

$$V(S_t) \leftarrow V(S_t) + \frac{1}{N(S_t)}\left(G_t - V(S_t)\right)$$

Weight recent episodes higher (forget old episodes):

$V(St) \leftarrow V(St) + \alpha\ (Gt - V(St))$

ENJOY YOUR SPRING BREAK!