AML, 24 Oct 2019

Given an MDP $p(s', r | s, a)$

compute an optimal policy π_*

Policy evaluation $\pi \rightarrow V_\pi$

Policy iteration $\pi_0 \xrightarrow{P.E.} V_{\pi_0} \xrightarrow[\text{Policy}]{\text{Greedy}} \pi_1 \longrightarrow V_{\pi_1} \rightarrow \cdots$

Value iteration $\pi \rightarrow V_\pi$, do only one round

Collapse single step of updates

Value iteration $\quad \pi_0 \to V_{\pi_0} \to V_{\pi_1} \to V_{\pi_2} \to \cdots$

Similarly for $\quad q(s,a)$

Generalize this update procedure quite radically

$$\begin{tikzcd} \pi \arrow[r, bend left] & V_\pi \arrow[l, bend left] \end{tikzcd}$$

update one more

update one state

Generalized Policy Iteration

Dynamic Programming

- Assumes that MDP structure is known in advance

In Reinforcement learning $p(s', r | s, a)$ is not known

- Experimentation to discover MDP structure, and solve for π_*

Monte Carlo Methods

- Sampling strategies to learn MDP structure & compute π_*

Each sample should produce some well defined value

Assume episode based MDP

Generate random trajectories in MDP

Need partial information about model $P(s'|s,a)$

$$G_t = R_{t+1} + \lambda G_{t+1}$$

$$S_0 \ A_0 \ R_1 \ S_1 \ A_1 \ R_2 \ \cdots \ S_{T-1} \ A_{T-1} \ R_T$$

G_t

Estimate $v_\pi(s)$ for a fixed π

↓

Estimate by counting and averaging

In each run of simulation, s occurs 0 or more times

Multiple visits to s — multiple samples $v_\pi(s)$

First visit

Any visit

First-visit MC prediction, for estimating $V \approx v_\pi$

Input: a policy π to be evaluated

Initialize:

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

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

Assumed to terminate

Loop forever (for each episode):

system

  Generate an episode following π: $S_0, A_0 \xrightarrow{\pi} R_1, S_1, A_1, R_2, \ldots, S_{T-1}, A_{T-1}, R_T$

  π p

  $G \leftarrow 0$

  R_T

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

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

    Unless S_t appears in $S_0, S_1, \ldots, S_{t-1}$:   First Visit

      Append G to $Returns(S_t)$

      $V(S_t) \leftarrow \text{average}(Returns(S_t))$

$$s \rightarrow Returns(s) = [v_1, v_2, v_3 \ldots, v_n]$$

$$\left(\sum v_i\right) / n$$

Blackjack [21]

Cards have values

2 - 10
J, Q, K — 10
A — 1 or 11

Aim: Get as close to 21 as poss without crossing

Dealer gives you 2 cards (hidden)

Dealer has 2 cards, 1 visible

Action: Take another card (Hit)
Stop (Stick)

Cross 21 - lose

State :

1. Your current total*
2. Dealer's visible card

*Depends on Ace

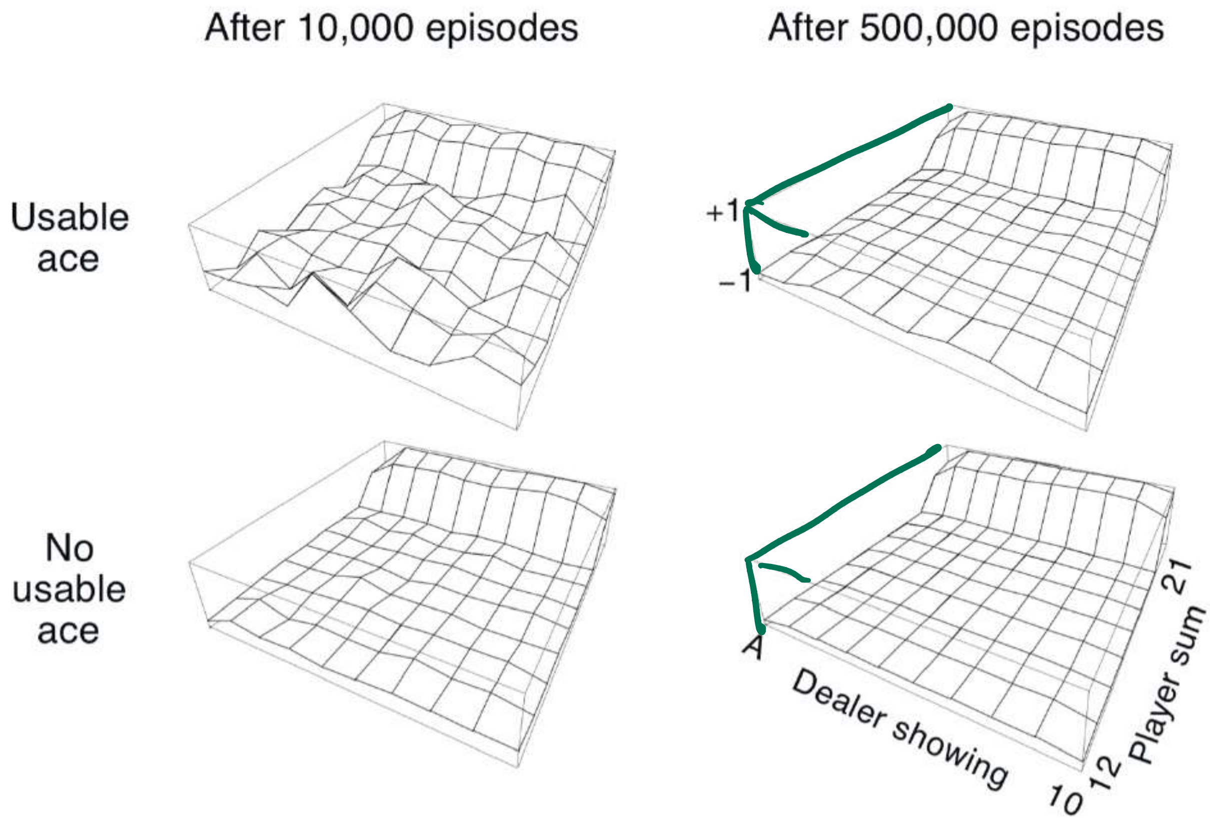
3. "Has a usable Ace"

→

Action

Hit
Stick

Our policy : Stick at 20/21



After 10,000 episodes
After 500,000 episodes
Usable ace
No usable ace
+1
−1
A
Dealer showing
10
12
Player sum
21

Similar strategy for $q_\pi(s,a)$

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, $t = T-1, T-2, \ldots, 0$:

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

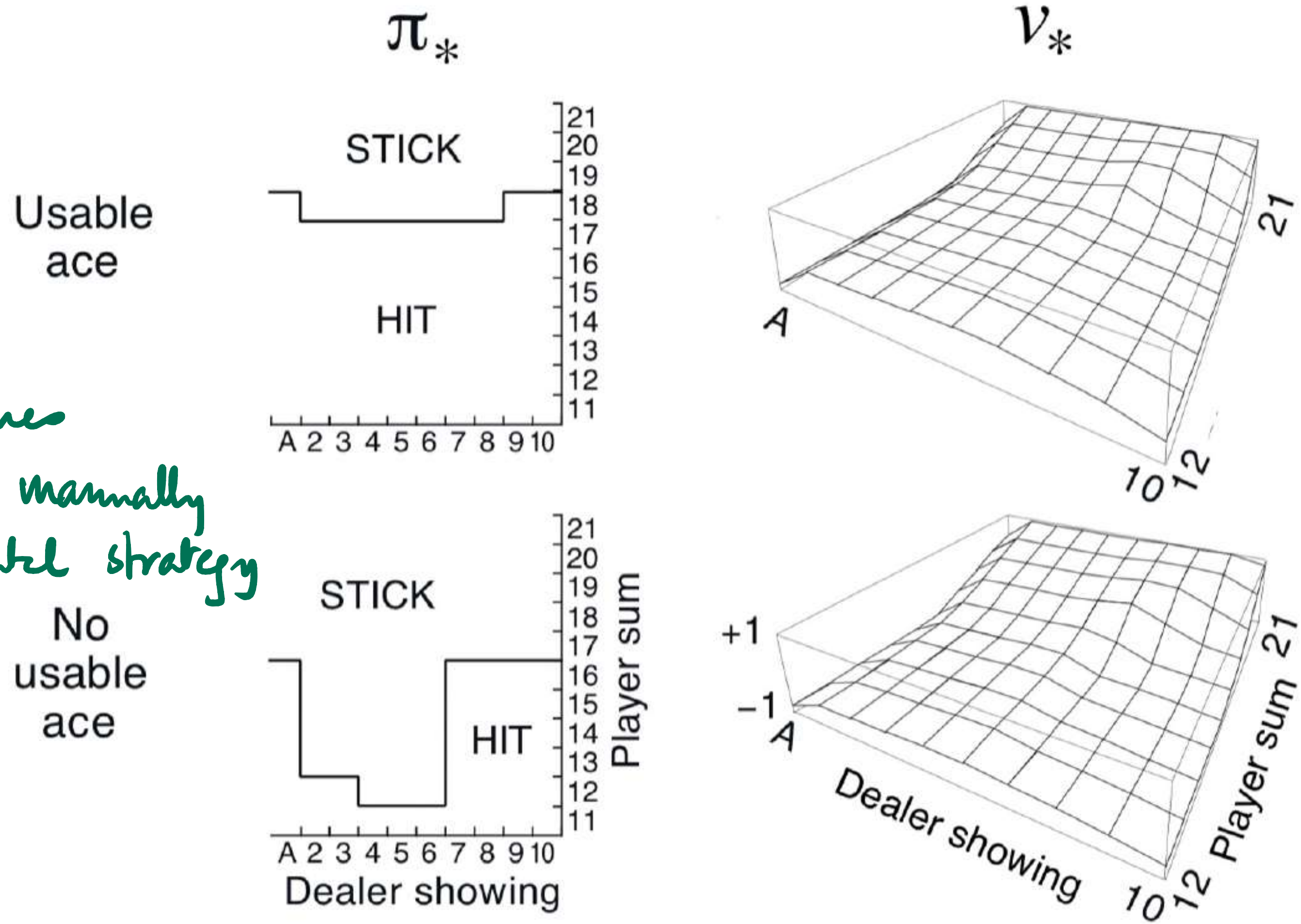
Unless the pair S_t, A_t appears in $S_0, A_0, S_1, A_1 \ldots, S_{t-1}, A_{t-1}$:

Append G to $Returns(S_t, A_t)$

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

$\pi(S_t) \leftarrow \text{argmax}_a\, Q(S_t, a)$ Policy Update

Matches best manually computed strategy



Avoid exploring starts?

Go back to Multi-Arm Bandits

ε - Greedy strategy

ε - randomly choose non greedy action

$1-\varepsilon$ - choose greedy action

ε-soft strategy: every action has probability at least $\frac{\varepsilon}{|A(s)|}$

# of actions at s

On-policy first-visit MC control (for ε-soft policies), estimates $\pi \approx \pi_*$

Algorithm parameter: small $\varepsilon > 0$

Initialize:

$\pi \leftarrow$ an arbitrary ε-soft policy

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

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

Repeat forever (for each episode):

Generate an episode 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 $Returns(S_t, A_t)$

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

$A^* \leftarrow \arg\max_a Q(S_t, a)$ (with ties broken arbitrarily)

For all $a \in \mathcal{A}(S_t)$:

$$\pi(a|S_t) \leftarrow \begin{cases} 1 - \varepsilon + \varepsilon/|\mathcal{A}(S_t)| & \text{if } a = A^* \\ \varepsilon/|\mathcal{A}(S_t)| & \text{if } a \neq A^* \end{cases}$$

On policy methods

Start with π

Sample & improve π

Off policy methods

Candidate π to improve iteratively

Separate sampling policy b