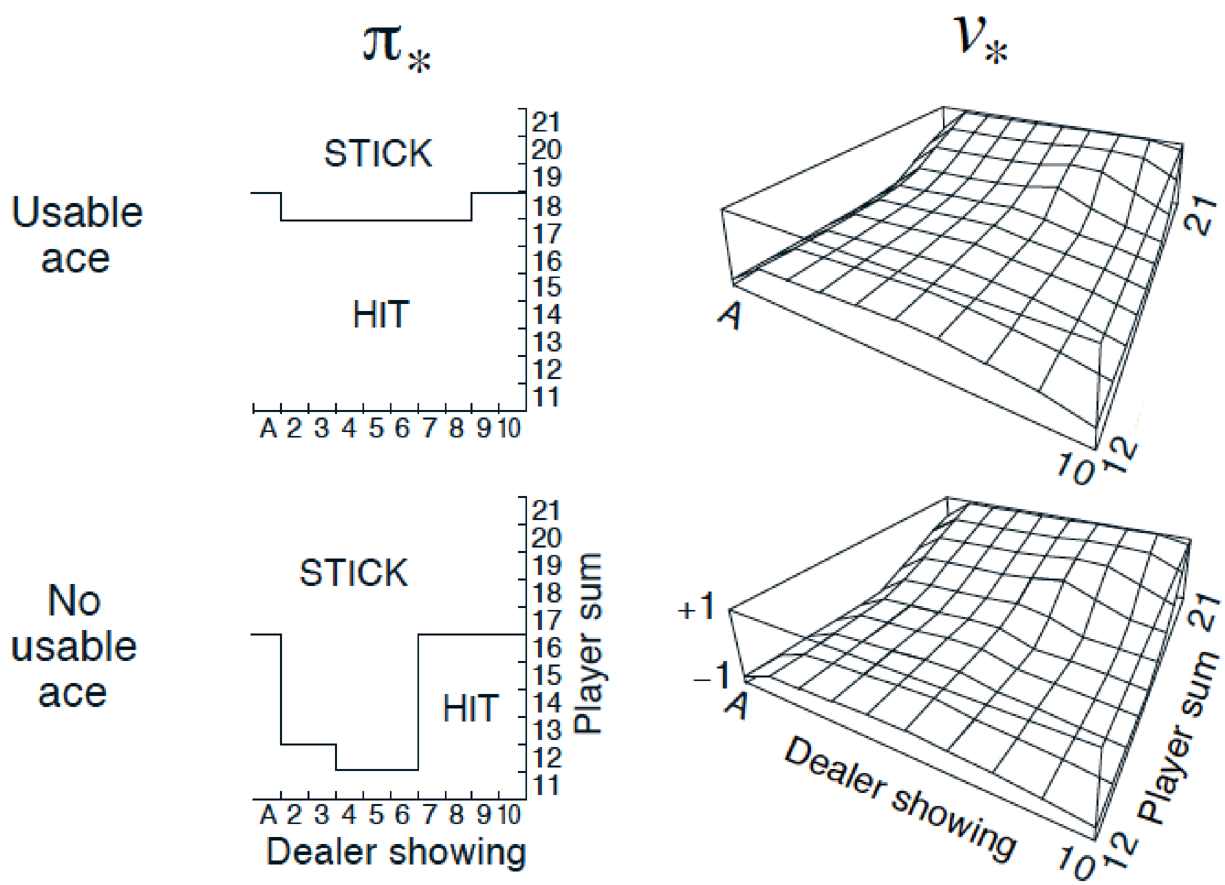
Summary



Optimal policy and state-value function in blackjack (Sutton and Barto, 2017)

MC Prediction: State Values

- Algorithms that solve the **prediction problem** determine the value function v_π (or q_π) corresponding to a policy π.
- Methods that evaluate a policy π from interaction with the environment fall under one of two categories:
  - **On-policy** methods have the agent interact with the environment by following the same policy π that it seeks to evaluate (or improve).
  - **Off-policy** methods have the agent interact with the environment by following a policy b (where $b \neq \pi$) that is different from the policy that it seeks to

- Each occurrence of state $s \in \mathcal{S}$ in an episode is called a **visit to** s.
- There are two types of Monte Carlo (MC) prediction methods (for estimating v_π):
  - **First-visit MC** estimates $v_\pi(s)$ as the average of the returns following *only first* visits to s (that is, it ignores returns that are associated to later visits).
  - **Every-visit MC** estimates $v_\pi(s)$ as the average of the returns following *all* visits to s.

First-Visit MC Prediction (for State Values)

Input: policy π, positive integer $num\_episodes$
Output: value function V ($\approx v_\pi$ if $num\_episodes$ is large enough)
Initialize $N(s) = 0$ for all $s \in \mathcal{S}$
Initialize $returns\_sum(s) = 0$ for all $s \in \mathcal{S}$
for $i \leftarrow 1$ ***to*** $num\_episodes$ **do**
  Generate an episode $S_0, A_0, R_1, \ldots, S_T$ using π
  for $t \leftarrow 0$ ***to*** $T - 1$ **do**
    if *S_t is a first visit (with return G_t)* **then**
      $N(S_t) \leftarrow N(S_t) + 1$
      $returns\_sum(S_t) \leftarrow returns\_sum(S_t) + G_t$
  end
end
$V(s) \leftarrow returns\_sum(s)/N(s)$ for all $s \in \mathcal{S}$
return V

MC Prediction: Action Values

- Each occurrence of the state-action pair s, a ($s \in \mathcal{S}, a \in \mathcal{A}$) in an episode is called a **visit to** s, a.
- There are two types of MC prediction methods (for estimating q_π):
  - **First-visit MC** estimates $q_\pi(s, a)$ as the average of the returns following *only first* visits to s, a (that is, it ignores returns that are associated to later visits).
  - **Every-visit MC** estimates $q_\pi(s, a)$ as the average of the returns following *all* visits to s, a.

Input: policy π, positive integer $num\_episodes$
Output: value function Q ($\approx q_\pi$ if $num\_episodes$ is large enough)
Initialize $N(s,a) = 0$ for all $s \in \mathcal{S}, a \in \mathcal{A}(s)$
Initialize $returns\_sum(s,a) = 0$ for all $s \in \mathcal{S}, a \in \mathcal{A}(s)$
for $i \leftarrow 1$ ***to*** $num\_episodes$ **do**
  Generate an episode $S_0, A_0, R_1, \dots, S_T$ using π
  for $t \leftarrow 0$ ***to*** $T-1$ **do**
    if (S_t, A_t) *is a first visit (with return* G_t*)* **then**
      $N(S_t, A_t) \leftarrow N(S_t, A_t) + 1$
      $returns\_sum(S_t, A_t) \leftarrow returns\_sum(S_t, A_t) + G_t$
  end
end
$Q(s,a) \leftarrow returns\_sum(s,a)/N(s,a)$ for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$
return Q

Generalized Policy Iteration

- Algorithms designed to solve the **control problem** determine the optimal policy π_* from interaction with the environment.
- **Generalized policy iteration (GPI)** refers to the general method of using alternating rounds of policy evaluation and improvement in the search for an optimal policy, All of the reinforcement learning algorithms we examine in this course can be classified as GPI.

MC Control: Incremental Mean

- (In this concept, we derived an algorithm that keeps a running average of a sequence of numbers.)

MC Control: Policy Evaluation

- (In this concept, we amended the policy evaluation step to update the value function after every episode of interaction.)

- A policy is **greedy** with respect to an action-value function estimate Q if for every state $s \in \mathcal{S}$, it is guaranteed to select an action $a \in \mathcal{A}(s)$ such that $a = \arg\max_{a \in \mathcal{A}(s)} Q(s, a)$. (It is common to refer to the selected action as the **greedy action**.)
- A policy is **ϵ-greedy** with respect to an action-value function estimate Q if for every state $s \in \mathcal{S}$,
  - with probability $1 - \epsilon$, the agent selects the greedy action, and
  - with probability ϵ, the agent selects an action (uniformly) at random.

Exploration vs. Exploitation

- All reinforcement learning agents face the **Exploration-Exploitation Dilemma**, where they must find a way to balance the drive to behave optimally based on their current knowledge (**exploitation**) and the need to acquire knowledge to attain better judgment (**exploration**).
- In order for MC control to converge to the optimal policy, the **Greedy in the Limit with Infinite Exploration (GLIE)** conditions must be met:
  - every state-action pair s, a (for all $s \in \mathcal{S}$ and $a \in \mathcal{A}(s)$) is visited infinitely many times, and
  - the policy converges to a policy that is greedy with respect to the action-value function estimate Q.

Input: positive integer *num_episodes*
Output: policy π ($\approx \pi_*$ if *num_episodes* is large enough)
Initialize $Q(s,a) = 0$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}(s)$
Initialize $N(s,a) = 0$ for all $s \in \mathcal{S}, a \in \mathcal{A}(s)$
for $i \leftarrow 1$ ***to*** *num_episodes* **do**
  $\epsilon \leftarrow \frac{1}{i}$
  $\pi \leftarrow \epsilon\text{-greedy}(Q)$
  Generate an episode $S_0, A_0, R_1, \ldots, S_T$ using π
  for $t \leftarrow 0$ ***to*** $T - 1$ **do**
    if *(S_t, A_t) is a first visit (with return G_t)* **then**
      $N(S_t, A_t) \leftarrow N(S_t, A_t) + 1$
      $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{1}{N(S_t, A_t)}(G_t - Q(S_t, A_t))$
  end
end
return π

MC Control: Constant-alpha

- (In this concept, we derived the algorithm for **constant-α MC control**, which uses a constant step-size parameter α.)
- The step-size parameter α must satisfy $0 < \alpha \leq 1$. Higher values of α will result in faster learning, but values of α that are too high can prevent MC control from converging to π_*.

Summary

Input: positive integer $num\_episodes$, small positive fraction α
Output: policy π ($\approx \pi_*$ if $num\_episodes$ is large enough)
Initialize Q arbitrarily (e.g., $Q(s,a) = 0$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}(s)$)
for $i \leftarrow 1$ ***to*** $num\_episodes$ **do**
  $\epsilon \leftarrow \frac{1}{i}$
  $\pi \leftarrow \epsilon$-greedy(Q)
  Generate an episode $S_0, A_0, R_1, \ldots, S_T$ using π
  for $t \leftarrow 0$ ***to*** $T-1$ **do**
    if (S_t, A_t) *is a first visit (with return* G_t*)* **then**
      $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(G_t - Q(S_t, A_t))$
  end
end
return π

NEXT