- (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.)

MC Control: Policy Improvement

- 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.