- **Every-visit MC** estimates $q_\pi(s, a)$ as the average of the returns following *all* visits to s, a.

First-Visit MC Prediction (for Action Values)

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 $S_0, A_0, R_1, \ldots, S_T$ using π
  for $t \leftarrow 0$ ***to*** $T-1$ **do**
    if (S_t, A_t) *is a first visit (with return* G_t*)* **then**
      $N(S_t, A_t) \leftarrow N(S_t, A_t) + 1$
      $returns\_sum(S_t, A_t) \leftarrow returns\_sum(S_t, A_t) + G_t$
  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