Introduction to Monte Carlo Reinforcement Learning

- What is Monte Carlo RL?
- The MC RL agent
- Review of Monte Carlo sampling
- Monte Carlo state value estimation
- Monte Carol policy improvement

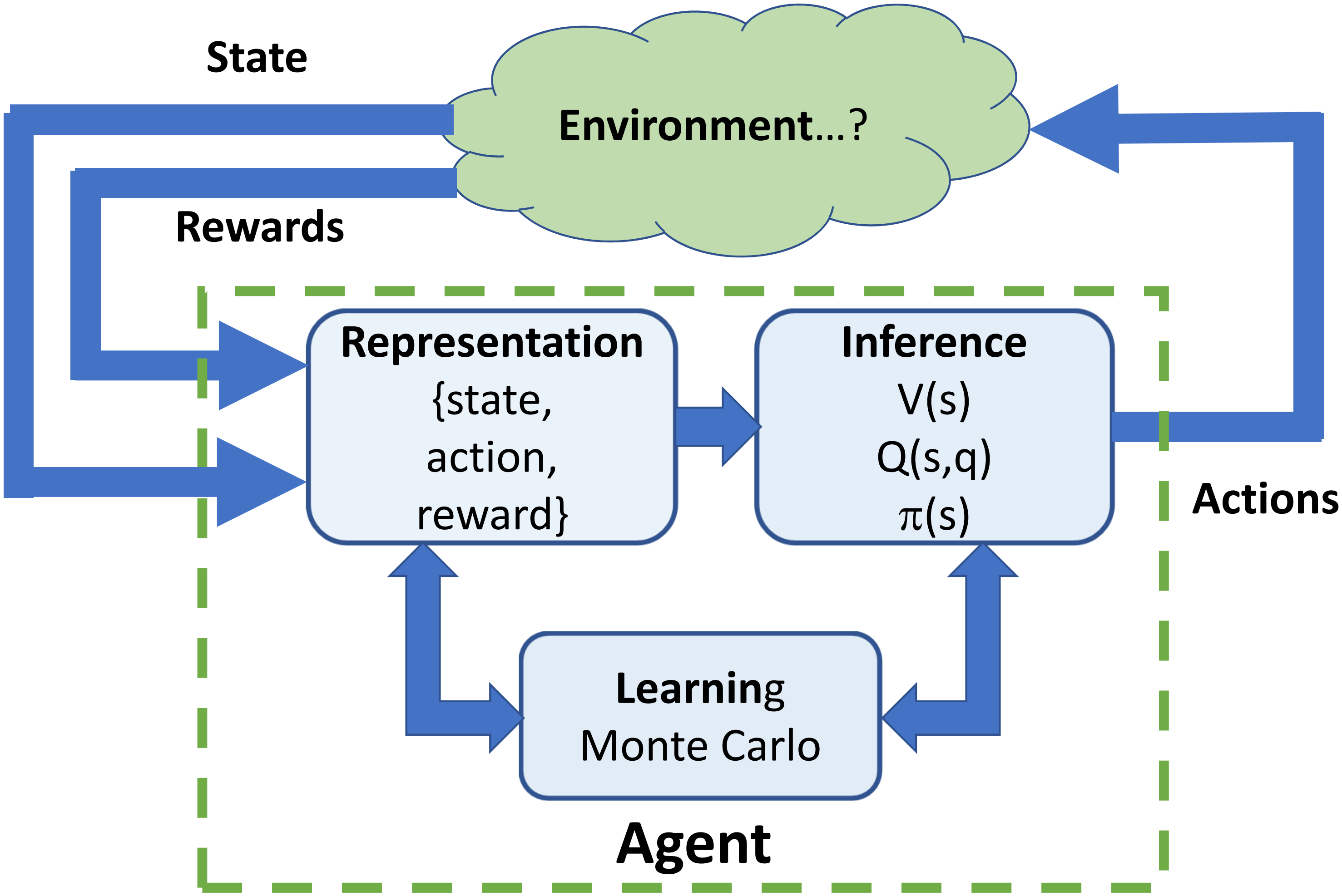
What is Monte Carlo Reinforcement Learning

- RL is **model free**
  - No specified model
  - Learn the value of state and action
- Monte Carlo agents take random samples of the values
  - Update average values with new samples
- Monte Carlo agents **must complete episodes**
  - Can only update values once episode terminates
- Monte Carlo RL is often used as a reference for performance of other algorithms

Introduction to Monte Carlo Reinforcement Learning

Model Type	Model?	State	Labeled Data	Loss Function
Supervised Learning	Yes	No	Yes	Error Metric
Unsupervised Learning	Yes	No	No	Error Metric
Bandit Agent	No	No	No	Reward
Dynamic Programming	Yes	Yes	No	Reward
Reinforcement Learning	No	Yes	No	Reward

The Reinforcement Learning Agent



Review of Monte Carlo Sampling

- Monte Carlo methods **randomly sample**
- Repetitive sampling creates a **Markov chain**
- Sample values are averaged
- Convergence of sample estimates converges by the **weak law of large numbers**

Review of Monte Carlo Sampling

- Sample estimates converge by the weak law of large numbers
- For **expected value** of underlying distribuend, μ, use sample estimate of the mean

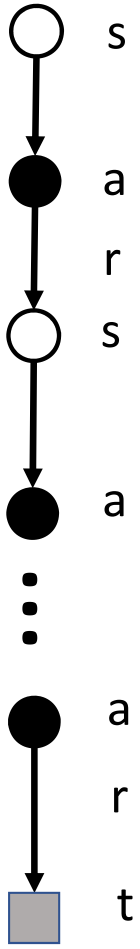
$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Then by the weak law of large numbers

$$\bar{X} \rightarrow E(X) = \mu$$

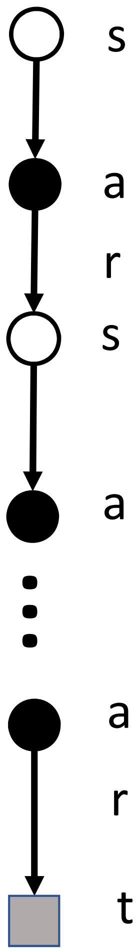
as, $n \rightarrow \infty$

Monte Carlo State Value Estimation



- The backup diagram aids understand the **MC RL state value estimation** algorithm
- MC sampling algorithm:
  - Start in state, s
  - Take action, a, based on policy, π
  - Record reward, r
  - Repeat above
  - Until terminal state, t
- MC algorithms **do not bootstrap**
  - **Complete backup**
  - **Strong convergence properties**
  - **High variance**
  - **Cannot work online**

Monte Carlo State Value Estimation



- Upon termination of the Markov chain, compute return

$$G_t = R_{t+1} + R_{t+2} + \ldots = R_T = \sum_{k=0}^{T} R_{t+k+1}$$

- Process is episodic so do not need to discount
- Two possible sampling methods:
    - **First visit Monte Carlo** estimates returns from rewards of the first visit to a state in an episode
    - **Every visit Monte Carlo** accumulates the rewards for any visit to a state in an episode
- Use first-visit MC in this course

Monte Carlo Policy Improvement

- Monte Carlo **policy improvement, or control, samples action values**, q(s,a)
- Rewards are accumulated for each action, a, from each state, s, following policy, π(s,a)
- At end of episode return for each action, a, from each state, s, are computed
- After a specified number of episodes, the policy is updated
  - Greedy improvement
  - ε-greedy improvement
- Above steps may be repeated