AIL 722: Reinforcement Learning

Lecture 16: Temporal-Difference Prediction (Part 2)

Raunak Bhattacharyya

YARDI SCHOOL OF ARTIFICIAL INTELLIGENCE
INDIAN INSTITUTE OF TECHNOLOGY DELHI

Outline

- TD prediction
- Vs. ground truth: model-based policy evaluation
- Q-Learning

Incremental Model-Free Policy Evaluation

mth sample

$$\hat{V}^{\pi}_{m}(s^j) \longleftarrow \hat{V}^{\pi}_{m-1}(s^j) + \alpha\big[G^{(m)} - \hat{V}^{\pi}_{m-1}(s^j)\big]$$

Estimate at m^th iteration | **Estimate at m-1^th iteration** | **Estimate at m-1^th iteration**

$$\underbrace{\hat{V}^{\pi}_{m}(s^j)}_{\text{New estimate}} \longleftarrow \underbrace{\hat{V}^{\pi}_{m-1}(s^j)}_{\text{Old estimate}} + \alpha\big[\underbrace{G^{(m)}}_{\text{Target}} - \hat{V}^{\pi}_{m-1}(s^j)\big]$$

Temporal Difference Policy Evaluation

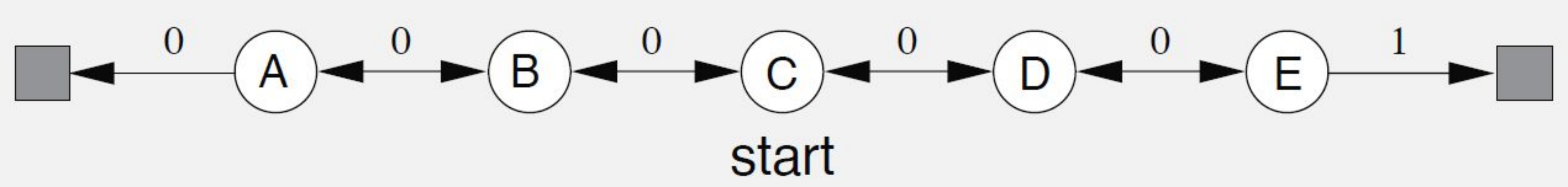
$$\hat{V}_m^\pi(s^j) \longleftarrow \hat{V}_{m-1}^\pi(s^j) + \alpha\left[\left(r_{t+1} + \gamma \cdot \hat{V}_{m-1}^\pi(s_{t+1})\right)^{(m)} - \hat{V}_{m-1}^\pi(s^j)\right]$$

Tabular TD(0) for estimating v_π

Input: the policy π to be evaluated
Algorithm parameter: step size $\alpha \in (0, 1]$
Initialize $V(s)$, for all $s \in \mathcal{S}^+$, arbitrarily except that $V(terminal) = 0$

Loop for each episode:
    Initialize S
    Loop for each step of episode:
        $A \leftarrow$ action given by π for S
        Take action A, observe R, S'
        $V(S) \leftarrow V(S) + \alpha\left[R + \gamma V(S') - V(S)\right]$
        $S \leftarrow S'$
    until S is terminal

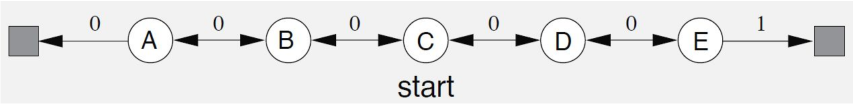
Example: Random Walk

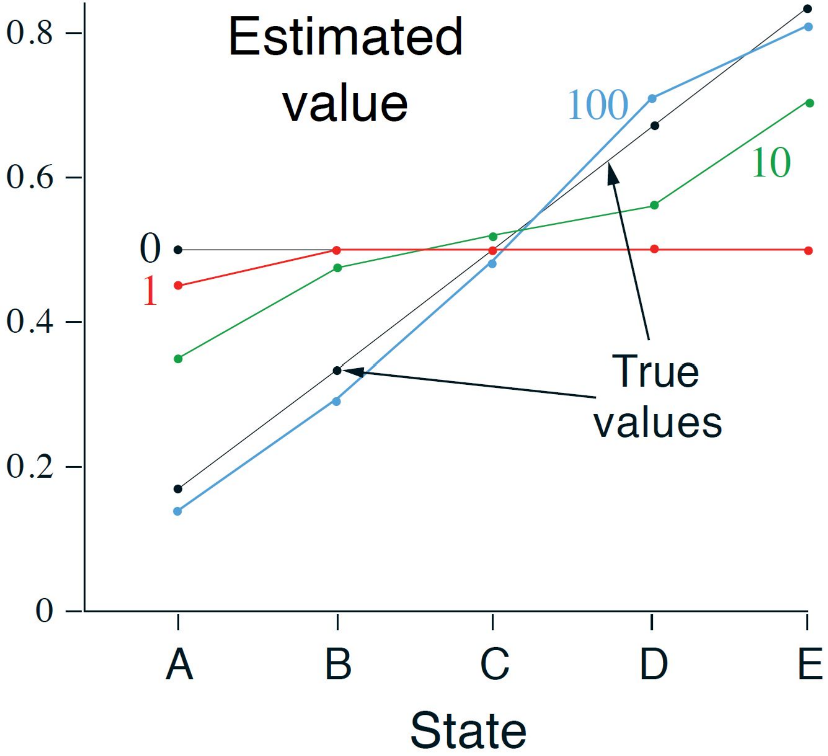


What is the value function?

Example 6.2, Reinforcement Learning: An Introduction, Sutton & Barto

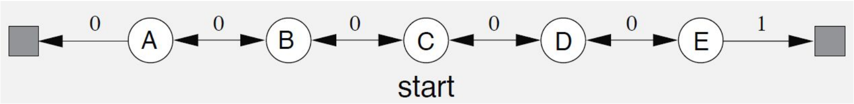
Example: Random Walk

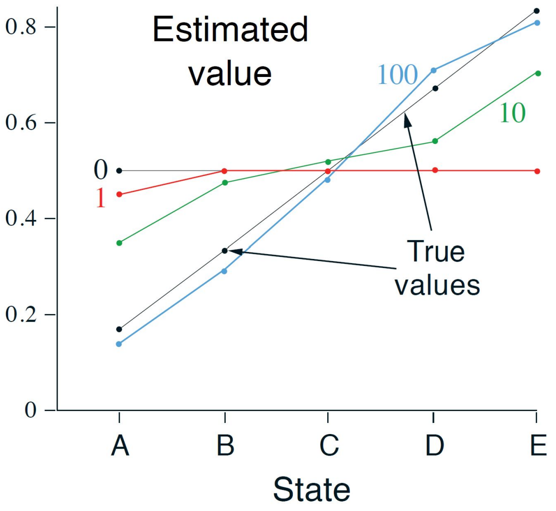


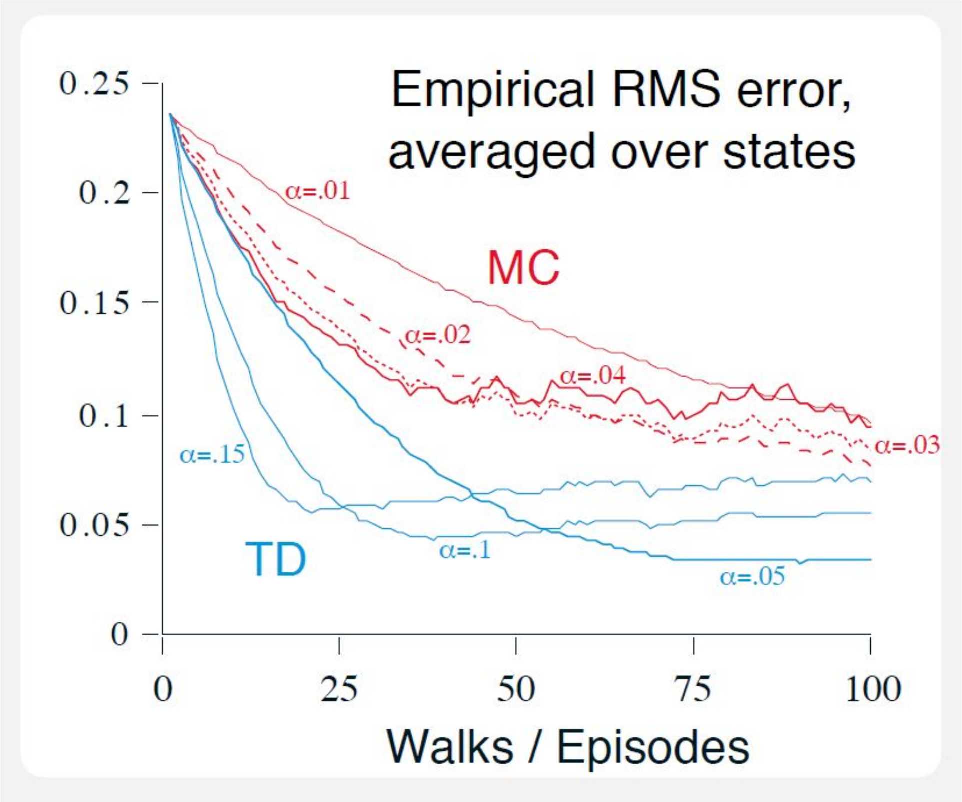


Example 6.2, [Reinforcement Learning: An Introduction](), Sutton & Barto

Example: Random Walk

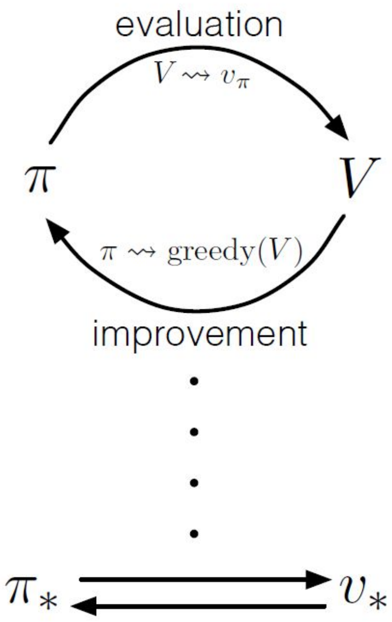






Example 6.2, [Reinforcement Learning: An Introduction](), Sutton & Barto

Generalised Policy Iteration



- Two simultaneous, interacting processes
  - Make value fun consistent with current policy
  - Make policy greedy w.r.t. current value function
- In PI, these processes alternate, each completing before other begins
- In VI, single iteration of policy evaluation between each policy improvement

GPI: Evaluation and improvement processes interact, independent of granularity

Model-free evaluation in GPI?

Section 4.6, [Reinforcement Learning: An Introduction](), Sutton & Barto