Simulating Discounted Costs

COMP 767 - Reinforcement Learning

Mathieu Nassif

McGill University

February 3, 2017

Discount in Rewards

γ is used to decrease the impact of future rewards on the present decision.

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots$$

Rewards far in the future will count a lot less than the next few rewards.

Discount in Rewards

γ is used to decrease the impact of future rewards on the present decision.

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots$$

Rewards far in the future will count a lot less than the next few rewards.

What if the discount was rather seen as a probability of survival to the next state?

Probability of Survival

If γ is the probability of survival, $(1-\gamma)$ is the probability of death.

$$\tilde{G_t} = R_{t+1} + S * \tilde{G}_{t+1}$$

where S is a Bernoulli variable that takes 1 with probability γ, 0 otherwise.

$$E(\tilde{G_t}) = R_{t+1} + \gamma \tilde{G}_{t+1} + (1-\gamma)0$$
$$E(\tilde{G_t}) = R_{t+1} + \gamma(R_{t+2} + \gamma \tilde{G}_{t+2})$$
$$E(\tilde{G_t}) = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots$$

Adapted Monte Carlo Approach

1. Instead of simulating a full episode, include at each step a probability of stopping.
2. Returns are *flat returns*

Then, we average the sampled returns (weighted by the importance sampling ratios).

$$V(s) = \frac{\sum_{t \in \mathcal{T}(s)} \rho_t^{T(t)} \tilde{G}_t}{|\mathcal{T}(s)|} \qquad \text{or} \qquad V(s) = \frac{\sum_{t \in \mathcal{T}(s)} \rho_t^{T(t)} \tilde{G}_t}{\sum_{t \in \mathcal{T}(s)} \rho_t^{T(t)}}$$

Why Use This Approach

1. More efficient
2. Simplify computations (but requires a random number generation)
3. **Can be used in non-episodic tasks**

Tested MDP

States: Integers from 0 to 9.
Actions: Choose to move left or right.
Next State: Move (left/right) by a random number of integers selected from a probability distribution: 1 (90%) or 2 (10%). If value overflow, wrap around.
Terminal State: 0
Initial State: 5
Reward: -1 for each action, until the terminal state is reached.

Evaluation Parameters

Discount (γ): 0.9
MC Limit: By the number of actions taken. The limit varies from 1000 to 100 000.
Repeat: 100 times for each method and limit.
Behavior policy: 50% left, 50% right
Target policy: 1-4: left, 5: 50%/50%, 6-9: right

Baseline

Recall:

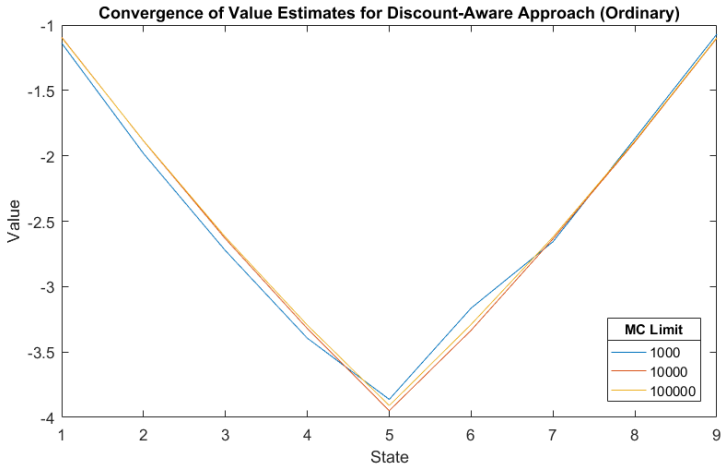
Discount-aware off-policy ordinary importance sampling

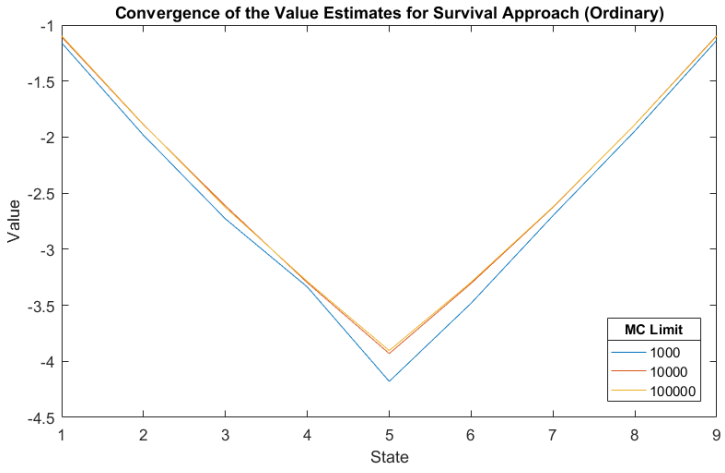
$$V(s) = \frac{\sum_{t \in \mathcal{T}(s)} \left((1-\gamma) \sum_{h=t+1}^{T(t)-1} \gamma^{h-t-1} \rho_t^h \bar{G}_t^h + \gamma^{T(t)-t-1} \rho_t^{T(t)} \bar{G}_t^{T(t)} \right)}{|\mathcal{T}(s)|}$$

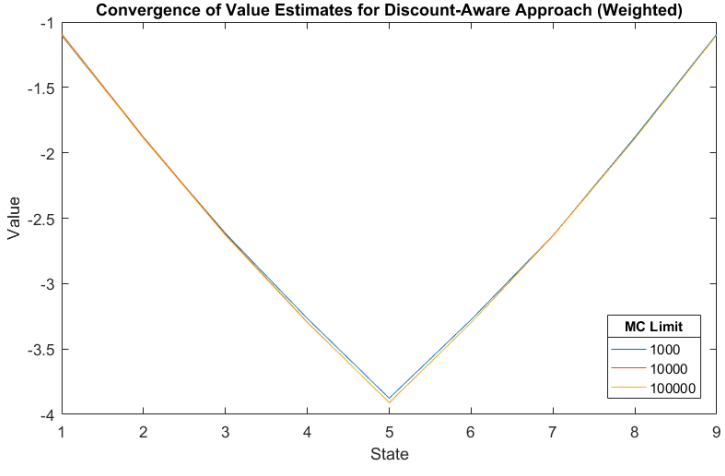
Discount-aware off-policy weighted importance sampling

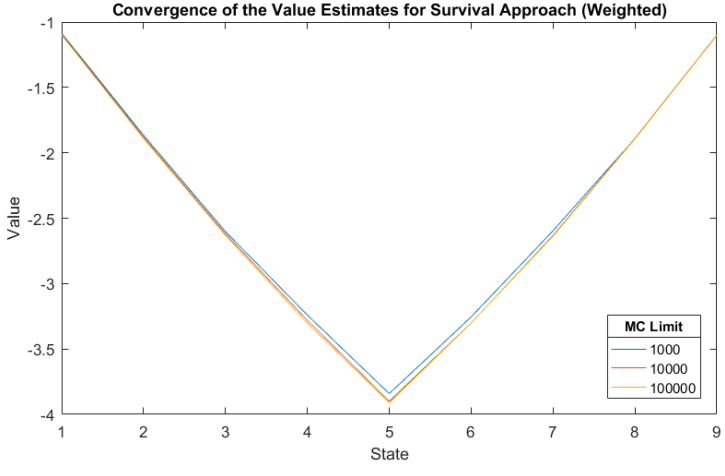
$$V(s) = \frac{\sum_{t \in \mathcal{T}(s)} \left((1-\gamma) \sum_{h=t+1}^{T(t)-1} \gamma^{h-t-1} \rho_t^h \bar{G}_t^h + \gamma^{T(t)-t-1} \rho_t^{T(t)} \bar{G}_t^{T(t)} \right)}{\sum_{t \in \mathcal{T}(s)} \left((1-\gamma) \sum_{h=t+1}^{T(t)-1} \gamma^{h-t-1} \rho_t^h + \gamma^{T(t)-t-1} \rho_t^{T(t)} \right)}$$

Results: Convergence









Results: Variance