

Policy Evaluation

Summary: Policy Evaluation

Marius Lindauer



Leibniz
Universität
Hannover



Winter Term 2021

Bias/Variance of Model-free Policy Evaluation Algorithms

- ▶ Return G_t is an unbiased estimate of $V^\pi(s_t)$
- ▶ TD target $[r_t + \gamma V^\pi(s_{t+1})]$ is biased estimate of $V^\pi(s)$
- ▶ But often TD much lower variance than a single return G_t
 - ▶ MC: Return function of multi-step sequence of random actions, states & rewards
 - ▶ TD target only has one random action, reward and next state

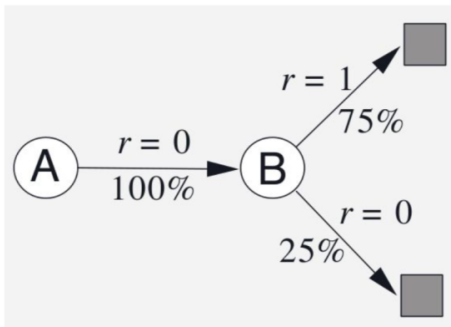
Bias/Variance of Model-free Policy Evaluation Algorithms

- ▶ Return G_t is an unbiased estimate of $V^\pi(s_t)$
- ▶ TD target $[r_t + \gamma V^\pi(s_{t+1})]$ is biased estimate of $V^\pi(s)$
- ▶ But often TD much lower variance than a single return G_t
 - ▶ MC: Return function of multi-step sequence of random actions, states & rewards
 - ▶ TD target only has one random action, reward and next state
- ▶ MC:
 - ▶ Unbiased (for first visit MC)
 - ▶ High variance
 - ▶ Consistent (converges to true) even with function approximation

Bias/Variance of Model-free Policy Evaluation Algorithms

- ▶ Return G_t is an unbiased estimate of $V^\pi(s_t)$
- ▶ TD target $[r_t + \gamma V^\pi(s_{t+1})]$ is biased estimate of $V^\pi(s)$
- ▶ But often TD much lower variance than a single return G_t
 - ▶ MC: Return function of multi-step sequence of random actions, states & rewards
 - ▶ TD target only has one random action, reward and next state
- ▶ MC:
 - ▶ Unbiased (for first visit MC)
 - ▶ High variance
 - ▶ Consistent (converges to true) even with function approximation
- ▶ TD
 - ▶ Some bias
 - ▶ Lower variance
 - ▶ TD(0) converges to true value with tabular representation
 - ▶ TD(0) does not always converge with function approximation

AB Example [Sutton & Barto, 2018]



- ▶ Two states A, B with $\gamma = 1$
- ▶ Given 8 episodes of experience:
 - ▶ $A, 0, B, 0$
 - ▶ $B, 1$ (observed 6 times)
 - ▶ $B, 0$
- ▶ Under batch (offline) solution for this finite set of observations, what do MC and TD(0)

AB Example [Sutton & Barto, 2018]

- ▶ Given 8 episodes of experience:

- ▶ $A, 0, B, 0$
- ▶ $B, 1$ (observed 6 times)
- ▶ $B, 0$

- ▶ For B :

- ▶ MC: $V(B) = \frac{6}{8} = 0.75$
- ▶ TD: $V(B) = \frac{6}{8} = 0.75$

AB Example [Sutton & Barto, 2018]

- ▶ Given 8 episodes of experience:
 - ▶ $A, 0, B, 0$
 - ▶ $B, 1$ (observed 6 times)
 - ▶ $B, 0$
- ▶ For B :
 - ▶ MC: $V(B) = \frac{6}{8} = 0.75$
 - ▶ TD: $V(B) = \frac{6}{8} = 0.75$
- ▶ For A :
 - ▶ MC: only one episode with $A \rightsquigarrow V(A) = 0$
 - ▶ TD: bootstraps from $V(B) \rightsquigarrow V(A) = 0.75$

$$V^\pi(s) = V^\pi(s) + \alpha(\underbrace{[r_t + \gamma V^\pi(s_{t+1})]}_{\text{TD target}} - V^\pi(s))$$

↪ Monte Carlo in batch setting converges to minimal
MSE (mean squared error)

↪ TD(0) converges to DP policy V^π for the MDP with

Efficiency

- ▶ Data efficiency & Computational efficiency
- ▶ In simplest TD, use (s, a, r, s') once to update $V(s)$
 - ▶ $O(1)$ operation per update
 - ▶ In an episode of length L , $O(L)$
- ▶ In MC have to wait till episode finishes, then also $O(L)$
- ▶ MC can be more data efficient than simple TD in non-Markov domains
- ▶ TD can exploit Markov structure \leadsto leveraging this is helpful

Summary: Policy Evaluation

Estimating the expected return of a particular policy if don't have access to true MDP models.
Example: evaluating average purchases per session of new product recommendation system

- ▶ Dynamic Programming
- ▶ Monte Carlo policy evaluation
 - ▶ Policy evaluation when we don't have a model of how the world works
- ▶ Temporal Difference (TD)
- ▶ Metrics to evaluate and compare algorithms
 - ▶ Robustness to Markov assumption
 - ▶ Bias/variance characteristics
 - ▶ Data efficiency
 - ▶ Computational efficiency