

Policy Evaluation

Monte Carlo Evaluation: Bias and Variance for MC

Marius Lindauer



Winter Term 2021

First-Visit Monte Carlo (MC) On Policy Evaluation

Initialize $N(s) = 0, G(s) = 0 \forall s \in \mathcal{S}$

Loop

- ▶ Sample episode $i = s_{i,1}, a_{i,1}, r_{i,1}, s_{i,2}, a_{i,2}, r_{i,2}, \dots$
- ▶ Define $G_{i,t} = r_{i,t} + \gamma r_{i,t+1} + \gamma^2 r_{i,t+2} + \dots$
- ▶ For each state s visited in episode i
 - ▶ for first time t that state s is visited in episode i
 - ▶ Increment counter of total first visits: $N(s) = N(s) + 1$
 - ▶ Increment total return $G(s) = G(s) + G_{i,t}$
 - ▶ Update estimate $V^\pi(s) = G(s)/N(s)$

Recap: Bias, Variance and MSE

- ▶ Consider a statistical model that is parameterized by θ and that determines a probability distribution over observed data $P(x|\theta)$
- ▶ Consider a statistic $\hat{\theta}$ that provides an estimate of θ and is a function of observed data x
 - ▶ E.g. for a Gaussian distribution with known variance, the average of a set of i.i.d data points is an estimate of the mean of the Gaussian
- ▶ Definition: the bias of an estimator $\hat{\theta}$ is:

$$Bias_{\theta}(\hat{\theta}) = \mathbb{E}_{x|\theta}[\hat{\theta}] - \theta$$

- ▶ Definition: the variance of an estimator $\hat{\theta}$ is:

$$Var(\hat{\theta}) = \mathbb{E}_{x|\theta}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2]$$

- ▶ Definition: mean squared error (MSE) of an estimator $\hat{\theta}$ is

$$MSE(\hat{\theta}) = Var(\hat{\theta}) + Bias_{\theta}(\hat{\theta})$$

First-Visit Monte Carlo (MC) On Policy Evaluation

Initialize $N(s) = 0, G(s) = 0 \forall s \in \mathcal{S}$

Loop

- ▶ Sample episode $i = s_{i,1}, a_{i,1}, r_{i,1}, s_{i,2}, a_{i,2}, r_{i,2}, \dots$
- ▶ Define $G_{i,t} = r_{i,t} + \gamma r_{i,t+1} + \gamma^2 r_{i,t+2} + \dots$
- ▶ For each state s visited in episode i
 - ▶ for first time t that state s is visited in episode i
 - ▶ Increment counter of total first visits: $N(s) = N(s) + 1$
 - ▶ Increment total return $G(s) = G(s) + G_{i,t}$
 - ▶ Update estimate $V^\pi(s) = G(s)/N(s)$

Properties:

- ▶ V^π estimator is an unbiased estimator of true $\mathbb{E}_\pi[G_t \mid s_t = s]$
- ▶ By law of large numbers, as $N(s) \rightarrow \infty, V^\pi(s) \rightarrow \mathbb{E}_\pi[G_t \mid s_t = s]$
- ▶ **every**-visit MC estimator:
 - ▶ is biased estimator of V^π (observations are correlated \leadsto not i.i.d)
 - ▶ often better RMSE, because more data per state

Monte Carlo (MC) Policy Evaluation Key Limitations

- ▶ Generally high variance estimator
 - ▶ Reducing variance can require a lot of data
 - ▶ In cases where data is very hard or expensive to acquire, or the stakes are high, MC may be impractical

Monte Carlo (MC) Policy Evaluation Key Limitations

- ▶ Generally high variance estimator
 - ▶ Reducing variance can require a lot of data
 - ▶ In cases where data is very hard or expensive to acquire, or the stakes are high, MC may be impractical
- ▶ Requires episodic settings
 - ▶ Episode must end before data from episode can be used to update V

Monte Carlo (MC) Policy Evaluation Summary

- ▶ Aim: estimate $V^\pi(s)$ given episodes generated under policy π
 - ▶ $s_{i,1}, a_{i,1}, r_{i,1}, s_{i,2}, a_{i,2}, r_{i,2}, \dots$ where the actions are sampled from π
 - ▶ $G_{i,t} = r_{i,t} + \gamma r_{i,t+1} + \gamma^2 r_{i,t+2} + \dots$ under policy π
 - ▶ $V^\pi(s) = \mathbb{E}[G_{t,s} \mid s_t = s]$
- ▶ Simple: Estimates expectation by empirical average (given episodes sampled from policy of interest)
- ▶ Updates V estimate using sample of return to approximate the expectation
- ▶ No bootstrapping
- ▶ Does not assume Markov process