Statistical Properties of Monte Carlo Estimates in Reinforcment Learning

Gautam Chandrasekaran   Saibal De

April 13, 2021

Table of Contents

1. Motivation
2. Reduction to Two-State Abstract MRP
3. Bias and Variance of Monte-Carlo Estimates
4. Finite Time Probability Bounds
5. Conclusion

Markov Reward Process and Monte Carlo Methods

- Markov reward process (MRP) has three elements
  - Markov chain $\{S_t : t \geq 0\}$ with state space $\{1, \ldots, n\}$ and transition probabilities

$$p_{ij} = \mathbb{P}(S_{t+1} = j \mid S_j = i)$$

  - Reward function $R : \{1, \ldots, n\} \to \mathbb{P}([0, 1])$
  - Discount factor $\gamma \in (0, 1]$
- Want to estimate value functions

$$V(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(S_t) \middle| S_0 = s\right]$$

- Can be estimated from trajectories $\{S_0, R_0, \ldots, S_{T-1}, R_{T-1}, S_T\}$ starting at $S_0 = s$ and terminating at $S_T = T$

$$G(S) = \sum_{t=0}^{T-1} \gamma^t R_t$$

Convergence of Monte Carlo Estimates

- Two algorithms exist; given trajectory $\{S_0, R_0, \ldots, S_{T-1}, R_{T-1}, S_T\}$
  - First-visit Monte Carlo computes gains $G(S_k)$ provided S_k does not repeat in $S_{k+1}, \ldots, S_T$
  - Every-visit Monte Carlo computes gains $G(S_k)$ for all $1 \leq k \leq T$ irrespective of repetitions
- Singh and Sutton (1996) derives bias and variance of these estimates in the undiscounted setting
  - First-visit Monte Carlo is unbiased
  - Every-visit Monte Carlo is asymptotically unbiased
  - After single episode, every-visit has lower variance than first-visit
  - In the long run, first-visit has lower MSE than every-visit
- Singh and Dayan (1998) analyzes the evolution of MSE curves

Our Goals

- Extend the analysis of Singh and Sutton (1996) to the discounted setting
- Derive finite time probability bounds for Monte Carlo estimates

Table of Contents

1. Motivation
2. Reduction to Two-State Abstract MRP
3. Bias and Variance of Monte-Carlo Estimates
4. Finite Time Probability Bounds
5. Conclusion

Partitioning the Transition Matrix

- Only focus on trajectories starting at state s to estimate $V(s)$
- Assume all trajectories terminate at a terminal state T
- Denote $\overline{\mathcal{S}} = \{1, \ldots, n\} \setminus \{s, T\}$
- Partition the Markov transition matrix

$$P = \begin{array}{c} \\ s \\ \overline{\mathcal{S}} \\ T \end{array} \begin{array}{c} \begin{array}{ccc} s & \overline{\mathcal{S}} & T \end{array} \\ \begin{pmatrix} p_s & u^\top & p_T \\ v & A & w \\ 0 & 0^\top & 1 \end{pmatrix} \end{array}$$

- Claim: Spectral radius $\rho(A) < 1$
- Consequence: We have for all $\gamma \in (0, 1]$

$$I + \gamma A + \gamma^2 A^2 + \cdots = (1 - \gamma A)^{-1}$$

Segmenting the Episode

- Episode $\{S_0, R_0, \ldots, S_{T-1}, R_{T-1}, S_T\}$ starting at s and terminating at T
- $T_0, T_1, \ldots, T_k$ the sorted time indices when $S_{T_k} = s$
- Trajectory segments $\{S_{T_{j-1}}, R_{T_{j-1}}, \ldots, S_{T_j-1}, R_{T_j-1}\}$ for $\leq j \leq k$ and $\{S_{T_k}, R_{T_k}, \ldots, S_{T-1}, R_{T-1}\}$ are independent
- Rewards accumulated over each segment

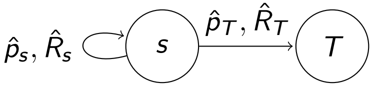
$$\hat{R}_{s_j} = \sum_{t=0}^{T_j - T_{j-1} - 1} \gamma^t R_{T_{j-1}+t}, \; 1 \leq j \leq k, \qquad \hat{R}_T = \sum_{t=0}^{T - T_k - 1} \gamma^t R_{T_k+t}$$

  are independent
- Gain over episode is given by

$$G = \sum_{j=1}^{k} \gamma^{T_{j-1}} \hat{R}_{s_j} + \gamma^{T_k} \hat{R}_T$$

Summary

- Informally, we integrate out all states except the starting and terminal states



- Gain is given by

$$G = \sum_{j=1}^{k} \gamma^{T_{j-1}} \hat{R}_{s_j} + \gamma^{T_k} \hat{R}_T$$

- How does $\hat{p}_s$, $\hat{p}_T$, $\hat{R}_s$ and $\hat{R}_T$ look like?
- What are the expressions for $\mathbb{E}[G \mid k \text{ revisits}]$ and $\mathbb{V}\text{ar}(G \mid k \text{ revisits})$?

Probabilities

- Self-loop in reduced MRP can happen in two ways:
  - A self loop in full MRP
  - Transition to $s_1, \ldots, s_{k+1} \in \overline{\mathcal{S}}$ followed by transition back to s, $k \geq 0$
- Combining all possible paths

$$
\begin{aligned}
\hat{p}_s &= p_{ss} + \sum_{k=0}^{\infty} \sum_{s_1,\ldots,s_{k+1} \in \overline{\mathcal{S}}} p_{ss_1} p_{s_1 s_2} \cdots p_{s_k s_{k+1}} p_{s_{k+1} s} \\
&= p_s + \sum_{k=0}^{\infty} u^\top A^k v \\
&= p_s + u^\top (I - A)^{-1} v
\end{aligned}
$$

- Similarly

$$
\hat{p}_T = p_T + u^\top (I - A)^{-1} w
$$

Reward Distributions

- Expectations are computed by law of total expectation:

$$\mathbb{E}[\hat{R}_s] = p_s r_s + r_s u^\top (I - A)^{-1} v + \gamma u^\top (I - \gamma A)^{-1} \mathrm{diag}(r_{\overline{S}})(I - A)^{-1} v$$

  and

$$\mathbb{E}[\hat{R}_T] = p_T r_s + r_s u^\top (I - A)^{-1} w + \gamma u^\top (I - \gamma A)^{-1} \mathrm{diag}(r_{\overline{S}})(I - A)^{-1} w$$

- The variances $\mathbb{V}\mathrm{ar}(\hat{R}_s)$ and $\mathbb{V}\mathrm{ar}(\hat{R}_T)$ can also be computed using law of total variance

Expected Gain

- Define $\hat{T}_j = T_j - T_{j-1}$ for $1 \leq j \leq k$; then

$$\mathbb{E}[\gamma^{T_j}] = \mathbb{E}[\gamma^{\hat{T}_1 + \cdots + \hat{T}_j}] = \mathbb{E}[\gamma^{\hat{T}_1}] \cdots \mathbb{E}[\gamma^{\hat{T}_j}] = \hat{\gamma}^j$$

  where we denote $\hat{\gamma} = \mathbb{E}[\gamma^{\hat{T}_1}]$
- It follows that

$$\mathbb{E}[G \mid k \text{ revisits}] = \mathbb{E}\left[\sum_{j=1}^{k} \gamma^{T_{j-1}} \hat{R}_{s_j} + \gamma^{T_k} \hat{R}_T\right]$$

$$= \sum_{j=1}^{k} \hat{\gamma}^{j-1} \hat{r}_s + \hat{\gamma}^k \hat{r}_T$$

- We can derive similar expression for variance; it has an extra term to account for the randomness in $\gamma^{\hat{T}_j}$
- The expressions become particularly simple when $\gamma = 1$

Table of Contents

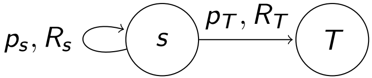
1. Motivation
2. Reduction to Two-State Abstract MRP
3. Bias and Variance of Monte-Carlo Estimates
4. Finite Time Probability Bounds
5. Conclusion

Change of Notation

- From now on, we only consider the two-state abstract MRP
- To simplify notation, we drop the "hat" from $\hat{p}_s, \hat{R}_s, \hat{r}_s, \hat{\gamma}, \ldots$

Value Function for the Two-State MRP

- With the change of notation, we have



- Bellman optimality condition

$$V(s) = p_s(r_s + \gamma V(s)) + p_T r_T \implies V(s) = \frac{p_s r_s + p_T r_T}{1 - \gamma p_s}$$

First-Visit Monte-Carlo is Unbiased

- Given reward sequence $\{R_{s_1}, \ldots, R_{s_k}, R_T\}$ the first visit estimate after single episode is

$$V_1^F(s) = R_{s_1} + \gamma R_{s_2} + \cdots + \gamma^{k-1} R_{s_k} + \gamma^k R_T$$

- Expected value of the first-visit gain is

$$\begin{aligned}
\mathbb{E}[V_1^F(s)] &= \sum_{k=0}^{\infty} \mathbb{P}(k \text{ revisits})\, \mathbb{E}[V_1^F(s) \mid k \text{ revisits}] \\
&= \sum_{k=0}^{\infty} p_s^k p_T \left(\sum_{j=1}^{k} \gamma^{j-1} r_s + \gamma^k r_T \right) \\
&= \cdots \\
&= V(s)
\end{aligned}$$

Every-Visit Monte-Carlo After Single Episode is Biased

- Given reward sequence $\{R_{s_1}, \ldots, R_{s_k}, R_T\}$ the every visit estimate after single episode is

$$V_1^E(s) = \frac{\left(\sum_{j=1}^{k} \gamma^{j-1} R_{s_k} + \gamma^k R_T\right) + \left(\sum_{j=2}^{k} \gamma^{j-2} R_{s_k} + \gamma^{k-1} R_T\right) + \cdots + R_T}{k+1}$$

- Expected value of the every-visit gain after a single episode is

$$\begin{aligned} \mathbb{E}[V_1^E(s)] &= \sum_{k=0}^{\infty} \mathbb{P}(k \text{ revisits}) \, \mathbb{E}[V_1^E(s) \mid k \text{ revisits}] \\ &= \cdots \\ &= \frac{1}{1-\gamma}\left[r_s + \frac{p_T}{p_s}\left(r_T - \frac{r_s}{1-\gamma}\right) \ln\left(1 + \frac{p_s}{p_T}(1-\gamma)\right)\right] \end{aligned}$$

Table of Contents

1 Motivation

2 Reduction to Two-State Abstract MRP

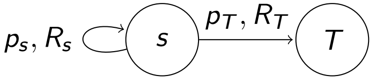
3 Bias and Variance of Monte-Carlo Estimates

4 Finite Time Probability Bounds

5 Conclusion

Overview

- Finite time bounds for two state undiscounted MRP with deterministic rewards
- Recap of notation:



Cumulant Generating Function and Concentration Bounds

- The Cumulant Generating Function of a random variable X is defined as

$$\psi_X(t) = \log \mathbb{E}[e^{tx}]$$

- Using this, we can derive concentration bounds as follows:

$$\begin{aligned}\mathbb{P}[X \geq \lambda\mu] &= \mathbb{P}[e^{tX} \geq e^{t\lambda\mu}] \\ &\leq \frac{\mathbb{E}[e^{tX}]}{e^{t\lambda\mu}} \\ &= \exp(-\lambda\mu t + \psi_X(t))\end{aligned}$$

- Similarly, we get can bounds for $Pr[X \leq \lambda\mu]$

$$\begin{aligned}\mathbb{P}[X \leq \lambda\mu] &= \mathbb{P}[e^{-tX} \geq e^{-t\lambda\mu}] \\ &\leq \frac{\mathbb{E}[e^{-tX}]}{e^{-t\lambda\mu}} \\ &= \exp(-\lambda\mu \cdot (-t) + \psi_X(-t))\end{aligned}$$

First Visit MC

- The first visit estimate after one trial is

$$V_1^F(s) = kR_s + R_t$$

  where k is the number of revisits to s.
- We know that $\mu = \mathbb{E}[V_1^F(s)] = V(s)$.
- We now compute $\mathbb{E}[e^{tV_1^F(s)}]$

$$\mathbb{E}\left[e^{tV_1^F(s)}\right] = \sum_{k=0}^{\infty} p_s^k \cdot p_T \cdot e^{t(kR_s+R_T)}$$

$$= p_T \cdot e^{tR_T} \sum_{k=0}^{\infty} p_s^k \left(e^{tR_s}\right)^k$$

$$= \frac{p_T . e^{tR_T}}{1 - p_s e^{tR_s}} \quad \text{when } p_s e^{tR_s} < 1$$

First Visit MC

- Thus, we get

$$\psi_{V_1^F(s)}(t) = \log\left(\frac{p_T.e^{tR_T}}{1 - p_s e^{tR_s}}\right)$$

  when $t < \frac{-\log p_s}{R_s}$

- Using the fact that $e^{-x} \geq 1 - x$, we get

$$\begin{aligned}
\psi_{V_1^F(s)}(t) &= \log\left(\frac{p_T \cdot e^{t(R_T - R_s)}}{e^{-tR_s} - p_s}\right) \\
&\leq t(R_T - R_s) + \log\left(\frac{1 - p_s}{1 - tR_s - p_s}\right) \\
&\leq t(R_T - R_s) - \log\left(1 - \frac{tR_s}{1 - p_s}\right), \quad t < \frac{1 - p_s}{R_s}
\end{aligned}$$

First Visit MC

- The first visit estimate after n episodes is $V_n^F(s) = \frac{1}{n}\sum_{i=1}^{n} V_1^F(s)_i$.
- Since the trials are independant, we have

$$\psi_{V_n^F(s)}(t) = \sum_{i=1}^{n} \psi_{V_1^F(s)}(t/n) \leq t(R_T - R_s) - n \cdot \log\left(1 - \frac{tR_s/n}{1-p_s}\right)$$

- We need the upper bound

$$\mathbb{P}[V_n^F(s) \geq \lambda\mu] \leq \exp\Big(-\lambda\mu t + \psi_{V_n^F(s)}(t)\Big)$$

  to be as tight as possible. So, on optimizing with respect to t, we get that

$$t = \frac{n(1-p_s)}{R_s} + \frac{n}{R_T - R_s - \lambda\mu}$$

- We also have the constraint that $t < \frac{n(1-p_s)}{R_s}$. This is satisfied when

$$\lambda > \frac{(1-p_s)(R_T - R_s)}{(1-p_s)R_T + R_s}$$

First Visit MC

- Assuming $R_S > 0$, we have the conditions satisfied for $\lambda > 1$. Substituting for t in the concentration bound, we get

$$\mathbb{P}[V_n^F(s) \geq \lambda \cdot V(s)] \leq \exp(-n(\alpha - 1 - \log(\alpha)))$$

where $\alpha = \frac{(1-p_s)(\lambda V(S)+R_s-R_T)}{R_s} = \frac{\lambda-1}{R_s}[p_s \cdot R_s + (1-p_s) \cdot R_T] + 1.$

- Note that $\alpha - 1 - \log(\alpha) > 0$ for all positive $\alpha \neq 1$
- Similarly, we get that

$$\mathbb{P}[V_n^F(s) \leq \lambda \cdot V(s)] \leq \exp(-n(\alpha - 1 - \log(\alpha)))$$

for $\frac{(1-p_s)(R_T-R_s)}{(1-p_s)R_T+R_s} < \lambda < 1$

Every Visit MC

- The every visit estimate after n trials is

$$V_n^E(s) = \frac{\left[\sum_{i=1}^{n} \frac{k_i(k_i+1)}{2} R_s + (k_i + 1) R_t\right]}{\left(\sum_{i=1}^{n} k_i\right) + n}$$

- Note that $\mu = \mathbb{E}[V_n^E(s)] = \frac{np_s}{(n+1)(1-p_s)} R_s + R_T = V(s) - \frac{p_s}{(n+1)(1-p_s)} R_s$

Every Visit MC

- We now compute $\mathbb{E}[e^{tV_n^E(s)}]$

$$\mathbb{E}\left[\exp\left(tV_n^E(s)\right)\right] = \sum_{k_1=0}^{\infty}\dots\sum_{k_n=0}^{\infty} p_s^{\sum_{i=1}^n k_i} p_T^n \exp\left[t\left(\frac{\sum_{i=1}^n \frac{k_i(k_i+1)}{2}R_s + (k_i+1)R_t}{(\sum_{i=1}^n k_i) + n}\right)\right]$$

$$= \sum_{k=0}^{\infty} p_s^k p_T^n \sum_{\sum_{i=1}^n k_i = k} \exp\left[t\left(\frac{\sum_{i=1}^n \frac{k_i(k_i+1)}{2}R_s + (k_i+1)R_t}{k+n}\right)\right]$$

$$= \sum_{k=0}^{\infty} p_s^k p_T^n \sum_{\sum_{i=1}^n k_i = k} \exp\left[t\left(\frac{\sum_{i=1}^k \frac{1}{2}k_i^2 R_s + \frac{k}{2}R_s + (k+n)R_t}{k+n}\right)\right]$$

$$\leq \sum_{k=0}^{\infty} p_s^k p_T^n \sum_{\sum_{i=1}^n k_i = k} \exp\left[t\left(\frac{\frac{1}{2}k^2 R_s + \frac{k}{2}R_s + (k+n)R_t}{k+n}\right)\right]$$

$$= \sum_{k=0}^{\infty} p_s^k p_T^n \binom{n+k-1}{n-1} \exp\left[t\left(\frac{\frac{1}{2}k^2 R_s + \frac{k}{2}R_s + (k+n)R_t}{k+n}\right)\right]$$

Every Visit MC

$$\mathbb{E}\left[\exp\left(tV_n^E(s)\right)\right] \leq \sum_{k=0}^{\infty} p_s^k p_T^n \binom{n+k-1}{n-1} \exp\left[t\left(\frac{\frac{1}{2}k(k+1)}{k+n}R_s + R_t\right)\right]$$

$$\leq \sum_{k=0}^{\infty} p_s^k p_T^n \binom{n+k-1}{n-1} \exp\left[t\left(\frac{kR_s}{2} + R_t\right)\right]$$

$$= p_T^n \exp\left(tR_t\right) \sum_{k=0}^{\infty} \binom{n+k-1}{n-1} \left[\exp\left(\frac{tR_s}{2}\right) p_s\right]^k$$

$$= p_T^n \exp\left(tR_t\right) \sum_{k=0}^{\infty} \binom{n+k-1}{k} (-1)^k \left[-\exp\left(\frac{tR_s}{2}\right) p_s\right]^k$$

Negative Binomial Series

- We have the following

$$(x+a)^{-n} = \sum_{k=0}^{\infty} \binom{-n}{k} x^k a^{-n-k}$$

for $|x| < a$ and

$$\binom{-n}{k} = \frac{(-n)(-n-1)\cdots(-n-k+1)}{k!}$$

- We can rewrite

$$\binom{-n}{k} = (-1)^k \binom{n+k-1}{k}$$

Every Visit MC

- Using the Negative Binomial series, we get

$$\mathbb{E}\left[\exp\left(tV_n^E(s)\right)\right] \leq p_T^n \exp\left(tR_t\right) \sum_{k=0}^{\infty} \binom{-n}{k} \left[-\exp\left(\frac{tR_s}{2}\right) p_s\right]^k$$

$$= p_T^n \exp\left(tR_t\right) \left(1 - \exp\left(\frac{tR_s}{2}\right) p_s\right)^{-n}$$

$$= \exp(tR_t) \left[\frac{p_T}{1 - \exp\left(\frac{tR_s}{2}\right) p_s}\right]^n$$

- We need $\exp\left(\frac{tR_s}{2}\right)p_s < 1$ to use the expansion. Thus, $t < \frac{-2\log(p_s)}{R_s}$.

Every Visit MC

- We now compute $\psi_{V_n^E(s)}$.

$$
\begin{aligned}
\psi_{V_n^E}(t) &\leq tR_t + n\log\left(\frac{p_T}{1-\exp\left(\frac{tR_s}{2}\right)p_s}\right)\\
&= tR_t - n\frac{tR_s}{2} + n\log\left(\frac{p_T}{\exp\left(-\frac{tR_s}{2}\right)-p_s}\right)\\
&\leq t\left(R_t - \frac{nR_s}{2}\right) + n\log\left(\frac{1-p_s}{1-\frac{tR_s}{2}-p_s}\right)\\
&\leq t\left(R_t - \frac{nR_s}{2}\right) - n\log\left(1-\frac{\frac{tR_s}{2}}{1-p_s}\right)
\end{aligned}
$$

with $t < \frac{2(1-p_s)}{R_s}$

Every Visit MC

- We now optimize t to make the upper bound

$$\mathbb{P}[V_n^E(s) \geq \lambda\mu] \leq \exp\left(-\lambda\mu t + \psi_{V_n^E(s)}(t)\right)$$

as tight as possible. We get the best value of t as

$$t = \frac{2(1-p_s)}{R_s} + \frac{n}{R_T - \frac{nR_s}{2} - \lambda\mu}$$

- To satify $t < \frac{2(1-p_s)}{R_s}$, we have that

$$\lambda > \frac{(1-p)\left[R_T - \frac{R_s n}{2}\right]}{R_T(1-p) + \frac{n}{n+1} p \cdot R_s}$$

Every Visit MC

- Assuming positive rewards, the previous condition is satisfied for all $\lambda > 1$. Thus, we get

$$Pr[V_n^E(s) \geq \lambda\mu] \leq \exp\left(-n\left(\alpha - 1 - \log(\alpha)\right)\right)$$

with $\alpha = \frac{2(1-p_s)}{n \cdot R_s} \cdot (\lambda\mu + \frac{nR_s}{2} - R_t)$.

- Similarly,

$$Pr[V_n^E(s) \leq \lambda\mu] \leq \exp\left(-n(\alpha - 1 - \log(\alpha))\right)$$

for $\frac{(1-p_s)\left[R_T - \frac{R_s n}{2}\right]}{R_T(1-p_s) + \frac{n}{n+1} p_s \cdot R_s} < \lambda < 1$

Table of Contents

1. Motivation
2. Reduction to Two-State Abstract MRP
3. Bias and Variance of Monte-Carlo Estimates
4. Finite Time Probability Bounds
5. Conclusion

Summary

- Explicit expressions in two-state MRP reduction
- Bias of first- and every-visit MC in discounted settings
- Finite time bounds with deterministic rewards in undiscounted setting

Future Work

- Variance of first- and every-visit MC in discounted setting
- Finite time bounds with random rewards in discounted setting
- Potential new estimates?

References

Satinder Singh and Peter Dayan. Analytical mean squared error curves for temporal difference learning. *Machine Learning*, 32(1):5–40, 1998.

Satinder P Singh and Richard S Sutton. Reinforcement learning with replacing eligibility traces. *Machine Learning*, 22(1), 1996.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. Adaptive Computation and Machine Learning. MIT Press, 2018. ISBN 978-0-262-19398-6.

Mathukumalli Vidyasagar. *Hidden markov processes: Theory and applications to biology*, volume 44. Princeton University Press, 2014.