A new policy based RL algorithm with reduced bias and variance

Jayakumar Subramanian & Aditya Mahajan

ECE & CIM, McGill University and GERAD

Introduction

Policy iteration based methods for RL may be broadly classified as:

- Monte Carlo (MC) methods [low bias, high variance]
- Temporal Difference (TD) methods [high bias, low variance]

Our Contribution

- Monte Carlo (RMC) – a method based on MC – for infinite horizon models with a designated start state [bias comparable to MC and variance comparable to TD].
- RMC retains the advantages of the Monte Carlo approach including low bias, simplicity and ease of implementation while circumventing its key drawbacks of high variance and delayed (end of episode) updates.

Model

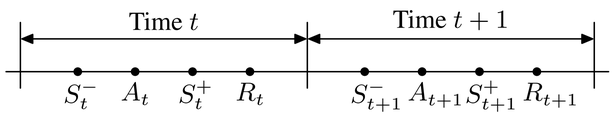

Fig. 1: A timing diagram showing the system variables.

We consider a generalized Markov decision process (MDP) $(\mathcal{S}^-, \mathcal{S}^+, \mathcal{A}, P^-, P^+, r)$ where

- $\mathcal{S}^-$ is the pre-decision state space;
- $\mathcal{S}^+$ is the post-decision state space;
- $\mathcal{A}$ is the action space;
- P^- is a Markov transition kernel from $\mathcal{S}^+$ to $\mathcal{S}^-$;
- P^+ is a controlled Markov transition kernel from $\mathcal{S}^- \times \mathcal{A}$ to $\mathcal{S}^+$;
- r: $\mathcal{S}^- \times \mathcal{A} \times \mathcal{S}^+ \to \mathbb{R}$ is the per-step reward function.

Renewal Theory based Terms

- **Stopping times**: $\tau_1 = \inf\{t > 1 : s_t^+ = s_*^+\}, \quad \tau_2 = \inf\{t > \tau_1 : s_t^+ = s_*^+\},$
- **Discounted reward & time**: $R_n = \gamma^{-\tau_{n-1}} \sum_{t=\tau_{n-1}+1}^{\tau_n} \gamma^{t-1} R_t \quad \text{and} \quad T_n = \gamma^{-\tau_{n-1}} \sum_{t=\tau_{n-1}+1}^{\tau_n} \gamma^{t-1}.$
- By the strong Markov property, $\{R_n\}_{n\geqslant 1}$ and $\{T_n\}_{n\geqslant 1}$ are i.i.d. sequences. Define: $\widehat{R}_N = \frac{1}{N}\sum_{n=1}^N R_n \quad \text{and} \quad \widehat{T}_N = \frac{1}{N}\sum_{n=1}^N T_n.$

Main Result

Renewal relationship: The post-decision value function at the reference state s_*^+ is given by:

$$V_\pi^+(s_*^+) = \frac{R_\pi}{(1-\gamma)T_\pi}.$$

RMC Algorithm

- Policies parameterized by a closed and convex subset Θ of the Euclidean space.
- Given a $\theta \in \Theta$, π_θ denotes the policy parametrized by θ and J_θ to denote $V_{\pi_\theta}^+(s_*^+)$
- Policy Improvement: $\theta_{m+1} = \left[\theta_m + \alpha_m \widehat{J}'_{\theta_m}\right]_\Theta$
- Using renewal relationship: $\nabla_\theta J_\theta = H_\theta / T_\theta^2, \quad$ where $H_\theta = T_\theta \nabla_\theta R_\theta - R_\theta \nabla_\theta T_\theta$.

Modified Policy Improvement

$\theta_{m+1} = \left[\theta_m + \alpha_m \widehat{H}_m\right]_\Theta$; $\quad \widehat{H}_m$ is estimated as $\widehat{H}_N = \widehat{T}_N \widehat{R}'_N - \widehat{R}_N \widehat{T}'_N$

Policy Differentiable w.r.t θ

- Likelihood Ratio: $L_t = \nabla_\theta \log[\pi_\theta(A_t \mid S_t^-)]$, $L_n = \sum_{t=\tau_{n-1}+1}^{\tau_n} L_t$
- Estimators for $\nabla_\theta R_\theta$ and $\nabla_\theta T_\theta$: $\widehat{R}'_N = \frac{1}{N}\sum_{n=1}^N R_n L_n \quad \text{and} \quad \widehat{T}'_N = \frac{1}{N}\sum_{n=1}^N T_n L_n.$

Algorithm 1: RMC Algorithm for RL: Likelihood ratio (or score function) based variant

input : θ_0 : Initial policy
γ : Discount factor
s_*^+ : Reference state (same as initial state)
N : Number of regenerative cycles
M : Number of iterations
output : θ_M: Estimate for best policy
initialize: $t = 1$
for *iteration* $m = 0$ *up to* $M-1$ **do**
  Set $\overline{R} = 0$; $\overline{R}' = 0$; $\overline{T} = 0$; $\overline{T}' = 0$
  for *regenerative cycle* $n = 1$ *up to* N **do**
    Set $\widehat{R}_n = 0$; $\widehat{T}_n = 0$; $L_n = 0$; $\texttt{scale} = 1$
    do
      Observe (S_t^-, A_t, S_t^+, R_t) where $A_t \sim \theta_m(S_t^-)$
      $\widehat{R}_n \leftarrow \widehat{R}_n + \texttt{scale} * R_t$; $\quad \widehat{T}_n \leftarrow \widehat{T}_n + \texttt{scale}$; $\quad L_n \leftarrow L_n + \nabla_\theta \log\left[\pi_\theta(A_t \mid S_t^-)\right]$
      $\texttt{scale} \leftarrow \texttt{scale} * \gamma$;
      $t \leftarrow t+1$
    while $S_t^+ \neq s_*^+$
    $\overline{R} \leftarrow \overline{R} + \frac{1}{n}(\widehat{R}_n - \overline{R})$; $\quad \overline{R}' \leftarrow \overline{R}' + \frac{1}{n}(\widehat{R}_n L_n - \overline{R}')$
    $\overline{T} \leftarrow \overline{T} + \frac{1}{n}(\widehat{T}_n - \overline{T})$; $\quad \overline{T}' \leftarrow \overline{T}' + \frac{1}{n}(\widehat{T}_n L_n - \overline{T}')$
  $\widehat{H}'_m = \overline{T}\,\overline{R}' - \overline{R}\,\overline{T}'$
  $\theta_{m+1} = \left[\theta_m + \alpha_m \widehat{H}'_m\right]_\Theta$
return θ_M

Policy Not Differentiable w.r.t θ – Simultaneous Perturbation FD Methods

- SPSA: $\widehat{R}'_\theta = (R_{\theta+\beta\delta} - R_\theta)/\beta\delta$, where δ: a θ-dim Bernoulli/Binomial random variable.
- SF: $\widehat{R}'_\theta = \eta(R_{\theta+\beta\eta} - R_\theta)/\beta$, where η: a θ-dim Gaussian ($\mathcal{N}(0,1)$) random variable.
- β is a small positive value and expressions for $\widehat{T}'_N$ are similar.

Numerical Examples

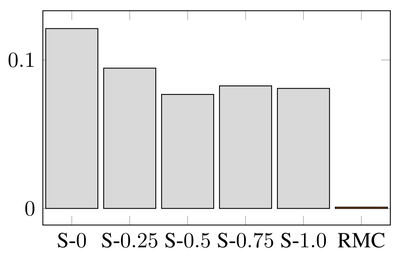

Fig. 2: Bias for Howard Taxi

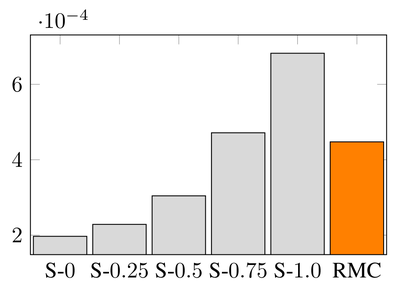

Fig. 3: Variance for Howard Taxi

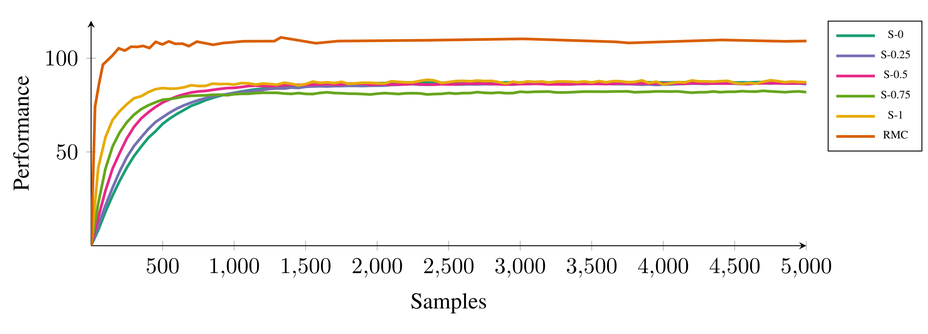

Fig. 4: Comparison of policy improvement for Howard Taxi using SARSA-λ

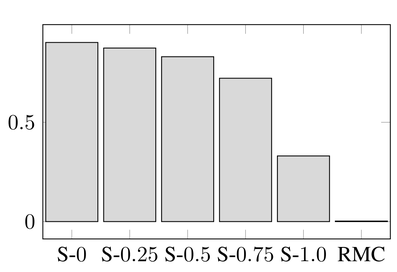

Fig. 5: Bias for Random MDP

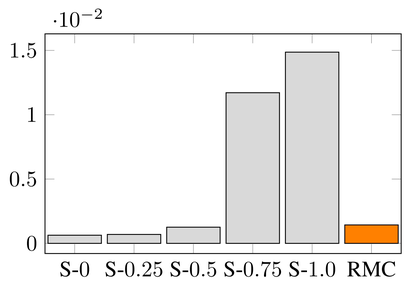

Fig. 6: Variance for Random MDP

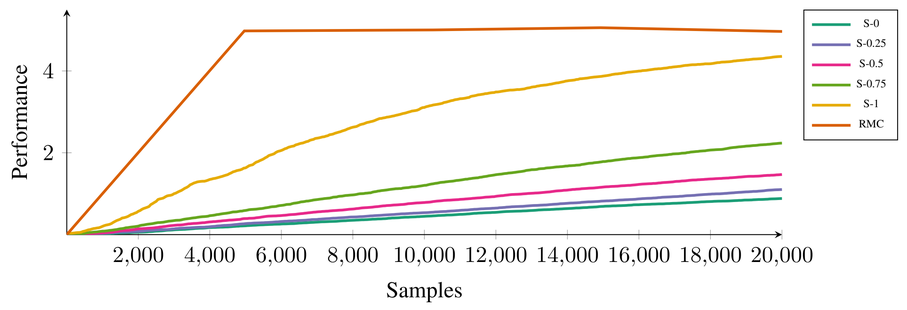

Fig. 7: Comparison of policy improvement for Random MDP using SARSA-λ

Conclusions

- A new RL algorithm, called Renewal Monte Carlo (RMC), for infinite-horizon discounted reward problems with a designated state state
- Experimental study suggests that RMC has significantly low bias with variance that is similar to SARSA-λ
- RMC is an Actor only method and works for models with continuous state and action spaces without the need to approximate the value function and for average rewards
- Possible to obtain an "every step" variant of RMC that can be used to to estimate the entire value function (or its approximation)