SaVeR: Optimal Data Collection Strategy for Safe Policy Evaluation in Tabular MDP

Subhojyoti Mukherjee*, Josiah P. Hanna, Robert Nowak

Problem Setting and Contributions

In off-policy RL we evaluate a target policy on a pre-collected dataset before deploying it in real-world. This dataset is collected using a different behavior policy. Recent works in policy evaluation for RL assumes a fixed or adaptive behavior policy that generates data to evaluate a particular target policy. In our work, we study the optimal choice for such adaptive behavior policy under safety constraints and introduce a method that can learn the optimal behavior policy during data collection.

Contributions:

1) We show that there exists a class of intractable MDPs where no safe algorithm can efficiently collect data and satisfy the safety constraints

2) We define the tractability condition for an MDP such that a safe oracle algorithm can efficiently collect data and using that we prove the first lower bound for this setting

3) We introduce algorithm SaVeR that approximates a safe oracle algorithm and bound the finite-sample mean squared error of the algorithm while ensuring it satisfies the safety constraint.

Safety Constraints for Policy Evaluation

Our objective is to determine a sequence of behavior policies, $\{\mathbf{b}_1, .., \mathbf{b}_K\}$, that will produce a set of K episodes that lead to the most accurate estimate of $V^{\pi}(s_1)$ subject to the constraint that the cumulative expected constraint-value $V_c^{\mathbf{b}}(s_1)$ always exceeds a fixed percentage of $V_c^{\pi_0}(s_1)$. We consider the objective:

$$\min_{\mathbf{b}} \mathbb{E}_{\mathcal{D}}\left[\left(Y_n^{\pi}(s_1) - V^{\pi}(s_1)\right)^2\right]$$

$$\text{s.t.} \sum_{k'=1}^{k} V_c^{\mathbf{b}^{k'}}(s_1) \geq (1-\alpha)kV_c^{\pi_0}(s_1) \text{ for all } k \in [K]$$

where $Y_n(s_1)$ is our estimate of $V^{\pi}(s_1), \alpha \in (0,1]$ is the risk parameter, and the expectation is over the collected data set $\mathcal{D}$. We also make the following simplifying assumption.

Tractability condition for Policy Evaluation

Define $V_{\mathbf{b}^-}^c(s_1)$ as the value of worst case behavior policy $\mathbf{b}^-$ that suffers a cost value that is lower than any other behavior policy $\mathbf{b}$. So this policy $\mathbf{b}^-$ can be thought of as the worst possible behavior policy that can be followed by the agent during an episode. Then the tractability condition states that

MDP complexity dependent budget

$$\sqrt{n} \geq \frac{\frac{1}{\alpha}\left(1 - \frac{V_{\mathbf{b}^-}^c(s_1)}{V_{\pi_0}^c(s_1)}\right)}{\frac{C_\sigma}{\alpha}\left(1 - \frac{V_{\mathbf{b}^-}^c(s_1)}{V_{\pi_0}^c(s_1)}\right) - 1}$$

where $C_\sigma \in (0,1)$ is a MDP dependent parameter that depends on the reward variance of state-action pairs.

*Subhojyoti Mukherjee is looking for full time position in industry from Fall 2024

We lower the MSE of policy evaluation in MDPs with an adaptive behavior policy under safety constraints

Safe Variance Reduction (SaVeR)

1) In real world setting the variances are unknown.

2) SaVeR uses plug-in estimates of the variances for oracle policy b* defined by

$$\mathbf{b}_*^k = \begin{cases} \mathbf{b}_*, & \text{if } \widehat{Z}_L^{k-1} \geq 0, k > \sqrt{K} \\ \pi_0 & \text{if } \widehat{Z}_L^{k-1} < 0 \\ \pi_x, & \text{if } \widehat{Z}_L^{k-1} \geq 0, k \leq \sqrt{K} \end{cases}. \quad (6)$$

3)

$$\widehat{\mathbf{b}}^k = \begin{cases} \widehat{\mathbf{b}}_*^k & \text{if } \widehat{Z}^{k-1} \geq 0, k > \sqrt{K} \\ \pi_0 & \text{if } \widehat{Z}^{k-1} < 0 \\ \pi_x & \text{if } \widehat{Z}^{k-1} \geq 0, k \leq \sqrt{K} \end{cases} \quad (8)$$

4) Use UCB to introduce exploration

5) Keeps safety budget $\widehat{Z}^{k-1}$ and explores when safety budget is sufficiently high

Algorithm 1 Safe Variance Reduction (SaVeR) for $\mathcal{T}$

Input: Risk Parameter $\alpha > 0$, target policy π.
Output: Dataset $\mathcal{D}$.
Initialize $\mathcal{D} = \emptyset$, $\widehat{\mathbf{b}}_1(a|s)$ uniform over all actions.
for $k = 1, 2, \ldots, K$ **do**
  for $\ell = 1, 2, \ldots, L$ **do**
    Get $\mathcal{H}^k := \{S_\ell^k, A_\ell^k, R(S_\ell^k, A_\ell^k), C(S_\ell^k, A_\ell^k)\}_{\ell=1}^L$ by selecting $\mathbf{b}^k$ according to (8).
    $\mathcal{D} \leftarrow \mathcal{D} \cup \{(\mathcal{H}^k, \widehat{\mathbf{b}}^k)\}$
    Update model parameters and estimate $\widehat{\mathbf{b}}_1^{k+1}(a|s)$ for each s, a
  end for
end for
Return Dataset $\mathcal{D}$ to evaluate policy π.

$$\mathbf{b}_*(a|s_t^\ell) \propto \left(\pi^2(a|s_t^\ell)[\sigma^2(s_t^\ell, a) + \sum_{s_j^{\ell+1}} P(s_j^{\ell+1}|s_i^\ell, a)M^2(s_j^{\ell+1})]\right)^{\frac{1}{3}} \quad (4)$$

where, $M(s_t^\ell)$ is the normalization factor defined as follows:

$$M(s_t^\ell) = \sum_a \left(\pi^2(a|s_t^\ell)(\sigma^2(s_t^\ell, a) + \sum_{s_j^{\ell+1}} P(s_j^{\ell+1}|s_i^\ell, a)M^2(s_j^{\ell+1}))\right)^{\frac{1}{3}}. \quad (5)$$

π_0 = **Baseline Policy**
π_x = **Exploration Policy**
$\mathbf{b}_*$ = **Oracle policy**

Regret of SaVeR

Define the regret as $\overline{\mathcal{R}}_n = \mathcal{L}_n\left(\pi, \widehat{\mathbf{b}}^k\right) - \overline{\mathcal{L}}_n^*\left(\pi, \mathbf{b}_*^k\right)$ where $\overline{\mathcal{L}}_n^*\left(\pi, \mathbf{b}_*^k\right)$ is the upper bound to the safe oracle MSE.

Define the constraint regret as follows: $\overline{\mathcal{R}}_n^c = \mathcal{C}_n\left(\pi, \widehat{\mathbf{b}}^k\right) - \overline{\mathcal{C}}_n^*\left(\pi, \mathbf{b}_*^k\right)$ where $\overline{\mathcal{C}}_n^*\left(\pi, \mathbf{b}_*^k\right)$ is the upper bound to the oracle constraint viola-

Corollary 1 Under Tractability condition, the constraint regret of SaVeR is bounded by $\overline{\mathcal{R}}_n^c \leq O\left(\frac{\log(n)}{n^{1/2}}\right)$ and the regret is bounded by $\overline{\mathcal{R}}_n \leq O\left(\frac{\log(n)}{n^{3/2}}\right)$.

1) The regret of SaVeR matches the regret in the unconstrained MDP setting [1, 2] under tractability condition.

2) It depends on the minimum sampling proportion in each state.

3) Regret decreases at the rate of $n^{-3/2}$ which is optimal in the bandit setting.

Regret Lower Bound

We use an alternate definition of regret than the standard pseudo-regret definition in bandits. The regret of the learning algorithm is defined as $\mathcal{R}_n = \mathcal{L}_n - \mathcal{L}_n^*$

where, $\mathcal{L}_n$ is the loss of the algorithm and $\mathcal{L}_n^*$ is the loss of the oracle.

Theorem 1. (Lower Bounds) Under Tractability condition the regret $\mathcal{R}_n = \mathcal{L}_n(\pi, \mathbf{b}) - \mathcal{L}_n^*\left(\pi, \mathbf{b}_*^k\right)$ is lower bounded by

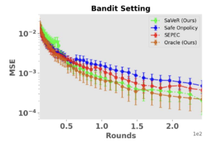
$$\mathbb{E}[\mathcal{R}_n] \geq \begin{cases} \Omega\left(\max\left\{\frac{A^{1/3}}{n^{3/2}}, \left(\frac{H_{*,(1)}^2 A^{2/3}}{n^{3/2}}\right)\right\}\right), (\boldsymbol{MAB}) \\ \Omega\left(\max\left\{\frac{\sqrt{SAL^2}}{n^{3/2}}, \left(\frac{H_{*,(1)}^2 SAL^2}{n^{3/2}}\right)\right\}\right) (\boldsymbol{MDP}) \end{cases}$$

where, $\Delta_0 = V_c^{\mathbf{b}_*}(s_1^1) - V_c^{\pi_0}(s_1^1)$ and $H_{*,(1)} = \frac{1}{\alpha V_c^{\pi_0}(s_1^1)}\left(\alpha V_c^{\pi_0}(s_1^1) + \Delta_0\right)$ is the hardness parameter.

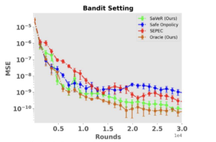
Experiments

1) SaVeR achieves the lowest regret among all baselines

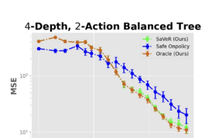
2) SaVeR balances exploration and exploitation by collecting sufficient safety budget and exploring new state-action pairs with high variance to reduce MSE



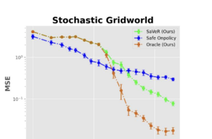
(a) Bandit setting



(b) Movielens setting

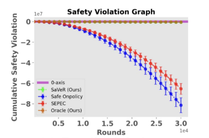


(c) Tree MDP

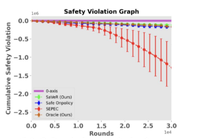


(d) Grid MDP

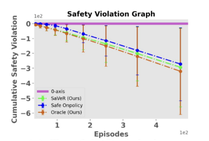
gure 1. MSE in different settings. The vertical axis (log-scaled) ves MSE and the horizontal axis is the number of episodes (or unds for bandits). Confidence bars show one standard error.



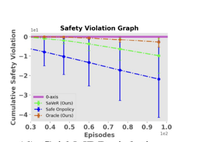
(a) Bandit violation



(b) Movielens violation



(c) Tree MDP violation



(d) Grid MDP violation

Figure 2. The vertical axis gives cumulative constraint violation and the horizontal axis is the number of episodes/rounds. The 0-axis is shown in pink. A safe algorithm has its plot below the 0-axis with the plot showing the cumulative unsafe budget.

References

[1] Alexandra Carpentier, Remi Munos, and András Antos. Adaptive strategy for stratified monte carlo sampling. J. Mach. Learn. Res., 16:2231–2271, 2015.

[2] Subhojyoti Mukherjee, Josiah P Hanna, and Robert D Nowak. Revar: Strengthening policy evaluation via reduced variance sampling. In Uncertainty in Artificial Intelligence, pages 1413–1422. PMLR, 2022a.