Data-Efficient Off-Policy Policy Evaluation for Reinforcement Learning Review

- Abstract:
  - Historical data 를 기존 학습된 Policy 에 적용하는데, 잘못된 Policy 만들어 지는 문제가 생길 수 도 있다. 따라서 이런 문제점을 해결하기 위해서 변화된 Policy 를 평가할 수 있는 new estimator 를 만들었다.
    - New estimator :
      - Doubly robust estimator 를 확장
      - Model based estimator 와 importance sampling based estimator 를 혼합하기 위한 방법을 제안

- 1. Introduce
  - Contribution
    - 기존 Doubly robust OPE 알고리즘을 확장 -> Weighted Double Robust (WDR) estimated
      - Horizon 이 유한하고, known 하다는 가정을 제거하고, 자신들의 Condition 을 넣음
      - 그래서 작은 양의 Bias 를 넣어서, Variance 를 줄이고, 효과적으로 mean squared error 를 줄여 나감
    - Model and Guided Importance Sampling Combining(MAGIC) estimator
      - **Combining two estimator(model base WDR, Model Free AM importance sampling)
- 정리
  - Weighted Doubly Robust (WDR) Estimator
    - 기존 DR Variance 를 많이 줄였으나, 많은 데이터가 필요
    - 작은 bias 가 필요
    - Guided important sampling 을 사용해서 bias 와 variance 의 trade off 를 잡는다
    - Weighted importance sampling 을 한다

        - Sampling 에 가중치를 준다는 말
    - Trajectory 증가시 v(π_b) towards v(π_e) 가된다
    - 식
    - $$\mathrm{WDR}(D) := \underbrace{\frac{1}{n}\sum_{i=1}^{n}\hat{v}^{\pi_e}(S_0^{H_i})}_{\text{(a)}} + \sum_{i=1}^{n}\sum_{t=0}^{\infty}\gamma^t w_t^i\left[\underbrace{R_t^{H_i} - \hat{q}^{\pi_e}\left(S_t^{H_i}, A_t^{H_i}\right) + \gamma\hat{v}^{\pi_e}\left(S_{t+1}^{H_i}\right)}_{\text{(b)}}\right]. \tag{3}$$
        - (a) 만약 Rt 와 St+1 이 S, A 의 deterministic function 이면 (b)가 zero 일 것이다 .
        - 만약 stochastic 이라면 (b)가 0 가 되지 않을 것이다.
        - 만약 importance weight w_t 가 high variance 라면 (b)는 0 가 아니며, High variance 를 가진다
        - AM(Approximate Model) 이 Lower MSE 를 가진다
        - 따라서 WDR 과 AM 사이에서 자동으로 스위칭을 잘하면 학습을 잘할 수 있다
- Blending IS and Model (BIM) Estimator
    - IS 와 BIM 을 잘 정의해서 특정 조건일 때 스위칭을 해보자
        - AM 은 High bias(일관성이 떨어짐) 및 MDP 환경에 수렴이 안될 수 있다
        - IS 는 강력한 일관성을 가지지만, High variance 문제가 있음
    - Partial importance sampling estimator(blending IS, BIM)
        - G(j)(D) -> off policy, J-step return
        - 식:
        - $$g^{(j)}(D) := \mathrm{IS}^{(j)}(D) + \mathrm{AM}^{(j+1)}(D)$$
          $$g^{(\infty)}(D) := \lim_{j\to\infty} g^{(j)}(D).$$
            - $g^{(-1)}$(D): **Model Base**,
                - J 의 작다: j-step important sampling from AM predict reward
            - $g^{(\infty)}$(D) , **Important sampling**
                - J 가 크다: important sampling predict reward only few reward from trajectory

- Model and Guided Importance Sampling Combining (MAGIC) Estimator

$$g^{(j)}(D) := \underbrace{\sum_{i=1}^{n}\sum_{t=0}^{j}\gamma^t w_t^i R_t^{H_i}}_{\text{(a)}} + \underbrace{\sum_{i=1}^{n}\gamma^{j+1} w_j^i \hat{v}^{\pi_e}(S_{j+1}^{H_i})}_{\text{(b)}}$$
$$-\underbrace{\sum_{i=1}^{n}\sum_{t=0}^{j}\gamma^t \left(w_t^i \hat{q}^{\pi_e}\left(S_t^{H_i}, A_t^{H_i}\right) - w_{t-1}^i \hat{v}^{\pi_e}\left(S_t^{H_i}\right)\right)}_{\text{(c)}}.$$

  - 최종 텀
    - (a) model base term
    - (b) WDR definition Wj
    - (c) combined control variate, important sampling term
  - 최종 알고리즘

Algorithm 1 MAGIC(D)

Input:
- $\mathcal{D}$: Historical data.
- π_e: Evaluation policy.
- Approximate model that allows for computation of $\hat{r}^{\pi_e}(s, a, t)$.
- $\mathcal{J}$: The set of return lengths to consider. The first element, $\mathcal{J}_1$, should be -1 and the last, $\mathcal{J}_{|\mathcal{J}|}$, should be ∞.
- κ: The number of bootstrap resamplings.

Compute $\widehat{\Omega}_n$ according to (25).
Allocate $D_{(\cdot)}$ so that for all $i \in \{1, \dots, \kappa\}$, D_i can hold n trajectories.
for $i = 1$ **to** κ **do**
    Load D_i with n uniform random samples drawn from D with replacement.
end for
$\mathbf{v} = \text{sort}\left(g^{(\infty)}(D_{(\cdot)})\right)$
$l \leftarrow \min\{\text{WDR}(D), \mathbf{v}(\lfloor 0.05n \rfloor)\}$
$u \leftarrow \max\{\text{WDR}(D), \mathbf{v}(\lceil 0.95n \rceil)\}$
for $j = 1$ **to** $|\mathcal{J}|$ **do**

$$\widehat{\mathbf{b}}_n(j) \leftarrow \begin{cases} g^{(\mathcal{J}_j)}(D) - u & \text{if } g^{(\mathcal{J}_j)}(D) > u \\ g^{(\mathcal{J}_j)}(D) - l & \text{if } g^{(\mathcal{J}_j)}(D) < l \\ 0 & \text{otherwise.} \end{cases}$$

end for
$\mathbf{x} \leftarrow \arg\min_{\mathbf{x} \in \Delta^{|\mathcal{J}|}} \mathbf{x}^\intercal [\widehat{\Omega}_n + \widehat{\mathbf{b}}_n \widehat{\mathbf{b}}_n^\intercal]\mathbf{x}$
return $\mathbf{x}^\intercal \mathbf{g}_{\mathcal{J}}(D)$