A General Framework for Off-Policy Learning with Partially-Observed Reward (ICLR 2025)

Rikiya Takehi[1], Masahiro Asami[2], Kosuke Kawakami[2], Yuta Saito[3]

[1] Waseda University, [2] Hakuhodo Technologies Inc., [3] Cornell University

Off-Policy Learning in Contextual Bandits

Context $x \sim p(x)$ (e.g., user info)

Action $a \sim \underline{\pi(a|x)}$ (e.g., product, movies)

decision-making policy

Target Reward $r \sim p(r|x, a)$ (e.g., ratings)

We aim to learn a policy π_θ using only logged bandit data:

$$\mathcal{D} := \{(x_i, a_i, r_i)\}_{i=1}^{n} \sim \underbrace{\pi_0}_{\text{(old) logging policy}}$$

Goal of Off-Policy Learning

We take **Policy Gradient (PG)** iterations to maximize target reward r

$$\theta_{t+1} \leftarrow \theta_t + \underbrace{\nabla_\theta V(\pi_\theta)}_{\text{Policy Gradient (PG)}}$$

$$\underline{\nabla_\theta V(\pi_\theta)} := \mathbb{E}_{p(x)\pi_\theta(a|x)}\left[\underbrace{q(x,a)}_{= \mathbb{E}[r|x,a]}\nabla_\theta \log \pi_\theta(a|x)\right]$$

We aim to accurately estimate the policy gradient using dataset $\mathcal{D}$

Problem: rewards are often *only* ***partially*** observed

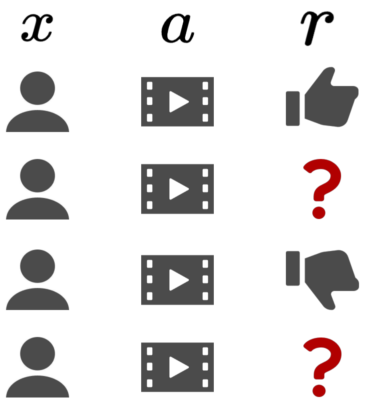
On various platforms, (target) rewards are often partially observed due to missing data, delayed observations, data fusion, multi-stage rewards,....

> e.g., ***explicit ratings*** in streaming platforms, ***conversion signals*** in e-commerce platforms

Problem: rewards are often *only* *partially* observed

On various platforms, (target) rewards are often partially observed due to missing data, delayed observations, data fusion, multi-stage rewards,....

> e.g., ***explicit ratings*** in streaming platforms, ***conversion signals*** in e-commerce platforms



Dataset with partial rewards:

$$\mathcal{D} := \{x_i, a_i, o_i, r_i\}$$

Reward observation indicator

$$o_i = \begin{cases} 1, \text{ if } r_i \text{ is captured} \\ 0, \text{ if } r_i \text{ is not captured} \end{cases}$$

Existing Solution 1: Only use Partial Target Rewards

$$\nabla_\theta \hat{V}_{\mathrm{r-IPS}}(\pi_\theta; \mathcal{D}) = \frac{1}{n}\sum_{i=1}^{n} \underbrace{\frac{o_i}{p(o_i|x_i)}}_{\text{address observation bias}} \frac{\pi_\theta(a_i|x_i)}{\pi_0(a_i|x_i)} r_i \nabla_\theta \log \pi_\theta(a_i|x_i)$$

Unbiased but **high variance**

due to small useable data size

Existing Solution 1: Only use Partial Target Rewards

$$\nabla_\theta \hat{V}_{\mathrm{r-IPS}}(\pi_\theta;\mathcal{D}) = \frac{1}{n}\sum_{i=1}^{n} \underbrace{\frac{o_i}{p(o_i|x_i)}}_{\text{address observation bias}} \frac{\pi_\theta(a_i|x_i)}{\pi_0(a_i|x_i)} r_i \nabla_\theta \log \pi_\theta(a_i|x_i)$$

Unbiased but **high variance**

due to small useable data size

Main motivation of our study

How can we perform effective OPL given sparsity in reward observation?

Use of secondary rewards?

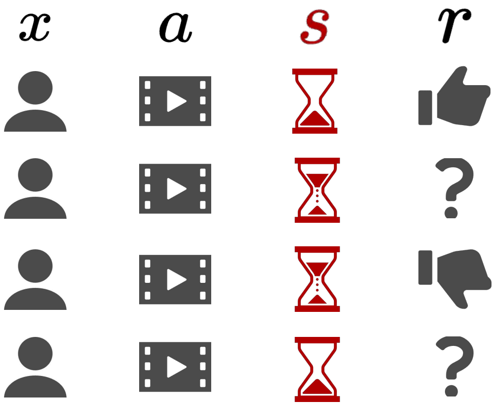
In many real-life scenarios, we have **densely-observed secondary rewards** $\mathcal{S}$.

> e.g., ***watch time, dwell time, clicks, …***

Use of secondary rewards?

In many real-life scenarios, we have **densely-observed secondary rewards** s .

> e.g., ***watch time, dwell time, clicks, …***



$$\mathcal{D} := \{x_i, a_i, s_i, o_i, r_i\}$$

Secondary rewards are:

- Densely observed
- Little correlation with target reward

Existing Solution 2: Only use Secondary Rewards

$$\nabla_\theta \hat{V}_{\text{s-IPS}}(\pi_\theta; \mathcal{D}) = \frac{1}{n} \sum_{i=1}^{n} \frac{\pi_\theta(a_i | x_i)}{\pi_0(a_i | x_i)} F(s_i) \nabla_\theta \log \pi_\theta(a_i | x_i)$$

$F(s_i)$: some aggregation function to imitate the target reward

Lower variance but **high bias**

due to imperfect correlation with objective

A new PG estimator that uses both types of rewards

Unbiased & lower-variance estimation of target reward PG

$$\nabla_\theta \hat{V}_r(\pi_\theta; \mathcal{D}) := \frac{1}{n} \sum_{i=1}^{n} \Bigg\{ \mathbb{E}_{a \sim \pi_\theta(a|x_i)} \left[\hat{q}(x_i, a) \nabla_\theta \log \pi_\theta(a|x_i) \right]$$

$$+ \frac{\pi_\theta(a_i|x_i)}{\pi_0(a_i|x_i)} \left(\hat{q}(x_i, a_i, s_i) - \hat{q}(x_i, a_i) \right) \nabla_\theta \log \pi_\theta(a_i|x_i)$$

$$+ \frac{o_i}{p(o_i|x_i)} \frac{\pi_\theta(a_i|x_i)}{\pi_0(a_i|x_i)} \left(r_i - \hat{q}(x_i, a_i, s_i) \right) \nabla_\theta \log \pi_\theta(a_i|x_i) \Bigg\}$$

$\hat{q}(\cdot)$: regression model

No bias & low variance by the additional use of secondary rewards.

Our Idea: Lower variance even further!

HyPeR

$$\nabla_\theta \hat{V}_{\mathrm{HyPeR}}(\pi_\theta; \mathcal{D}, \gamma) = (1-\gamma)\underbrace{\nabla_\theta \hat{V}_r(\pi_\theta; \mathcal{D})}_{\text{target reward maximization}} + \overset{\text{weight optimized from data}}{\gamma}\underbrace{\nabla_\theta \hat{V}_s(\pi_\theta; \mathcal{D})}_{\text{secondary reward maximization}}$$

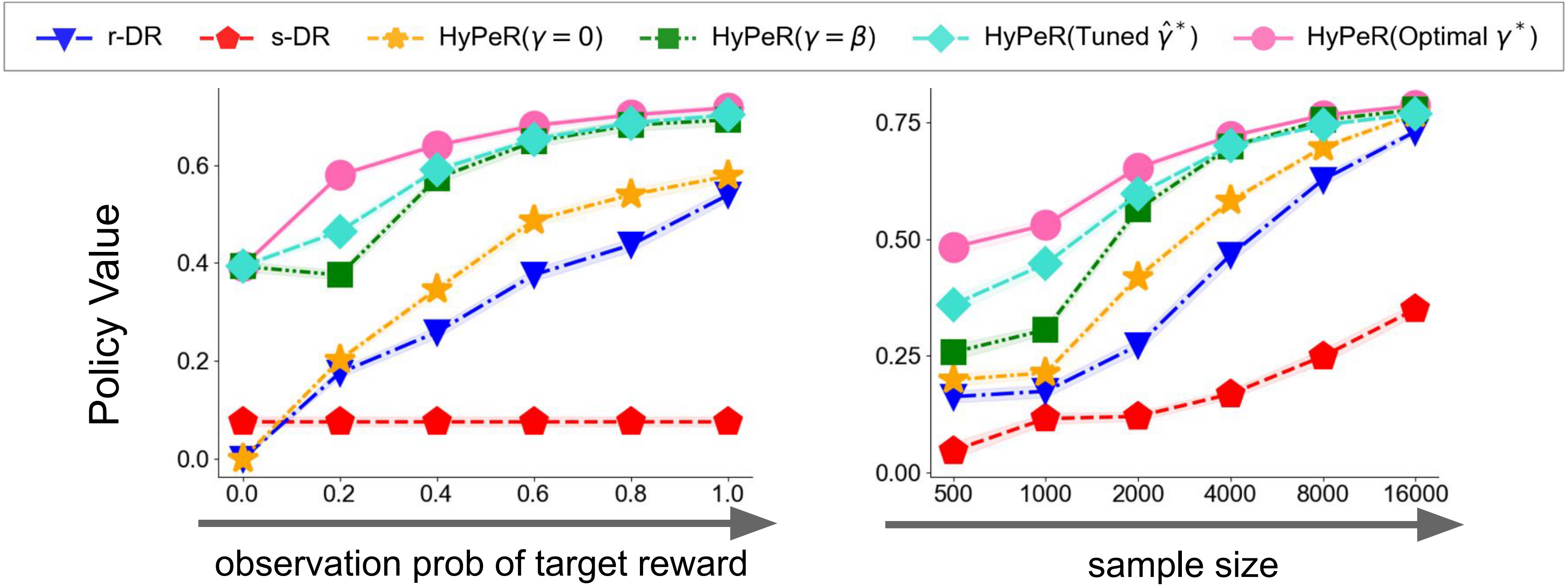
unbiased & high variance — high bias & low variance

We lower variance by leveraging $\nabla_\theta \hat{V}_s(\pi_\theta; \mathcal{D})$, *at the cost of introducing bias*.

Summary of Theoretical Findings

- HyPeR can lower variance while being unbiased.
- We find a general concept that it is often better to strategically introduce bias when in a multi-reward setting.
- We provide a data-driven method for optimizing the weight for a bias-variance balance.

Strong Empirical Performance



HyPeR is effective even when target rewards are sparse

Summary

- We introduce & formulate the problem of OPL with partial observations of rewards.
- Our *HyPeR* leverages secondary rewards to enable effective OPL.
- Other interesting theoretical & empirical results are in the paper!