リグレット vs 平均二乗誤差
オフ方策学習と条件付き平均処置効果推定の橋渡し

Bridging the Gap between Empirical Welfare Maximization and Conditional Average Treatment Effect Estimation in Policy Learning
arXiv Preprint. URL:
https://arxiv.org/abs/2510.26723

Masahiro Kato
Mizuho-DL Financial Technology

- Consider **binary treatments** 1 and 0, e.g., AB testing.
- Two approaches in policy learning in causal inference:
- **Counterfactual Risk Minimization (CRM)**.
- **Plugin approach with Conditional Average Treatment Effect (CATE) estimation**.

- Existing results: we should use CRM since it directly solves the decision-making problem directly.
  → Really?
- We reveal that at the optimization level, these problems are closely

Policy learning

1. Potential Outcomes and Observations

- **Potential outcome** for each treatment 1 and 0: Y_1 and $Y_0 \in \mathcal{Y} \subset \mathbb{R}$.
- k-dimensional **Covariates** of a unit: $X \in \mathcal{X} \subset \mathbb{R}^k$.
- **Treatment** indicator: $D \in \{1,0\}$.
- **Observations** $\{(X_i, D_i, Y_i)\}_{i=1}^n$, where (X_i, D_i, Y_i) is an i.i.d. copy of (X, D, Y).

2. Policy

- A decision-maker recommends a treatment d given a unit's covariates.
- For the recommendation, the decision-maker uses a policy

$$\pi: \mathcal{X} \to [0,1].$$

  - $\pi(X)$ denotes a probability of recommending treatment 1 for the unit.
  - A deterministic (0-1) policy: $\pi = 1\{X \in G\}$ for a region $G \subset \mathcal{X}$.
- From the observations, we aim to train a policy.
- We can only observe either of Y_1 or Y_0 due to the counterfactual property.
- We need to estimate the welfare function $W(\pi)$.

3. Welfare

- Good policy = a policy maximizing the following welfare:

$$W(\pi) \coloneqq E\big[\pi(X)Y_1 + \big(1-\pi(X)\big)Y_0\big].$$

- Given a policy class Π, the optimal policy is defined as

$$\pi^* \coloneqq \arg\max_{\pi\in\Pi} W(\pi).$$

4. Counterfactual Risk Minimization

- Assume that the propensity score $e_0(X) = P(D = 1 \mid X)$ is known.
- For example, an inverse probability weighting (IPW) estimator is

$$\widehat{W}(\pi) \coloneqq \frac{1}{n}\sum_{i=1}^{n}\left(\pi(X_i)\frac{D_i}{e_0(X_i)}Y_i + \big(1-\pi(X_i)\big)\frac{1-D_i}{1-e_0(X_i)}Y_i\right).$$

- Then, the decision-maker can train a policy as

$$\hat{\pi}^{\mathrm{CRM}} \coloneqq \arg\max_{\pi\in\Pi}\{\widehat{W}(\pi)\}.$$

  - Such a method is called **empirical welfare maximization** or CRM.

✓ We can also use the doubly robust estimator or other estimators for $W(\pi)$

Plugin Policy with CATE Estimation

5. CATE and its Estimation

- The CATE is defined as $\tau_0(X) \coloneqq E[\,Y_1 - Y_0 \mid X\,]$.
- The CATE can be estimated as $g^* \coloneqq \arg\min_{g\in\mathcal{G}} E\left[\big(Y_1 - Y_0 - g(X)\big)^2\right]$,
  - $\mathcal{G}$ is a set of regression functions g.
- Empirically, we can estimate the CATE as

$$\hat{g} \coloneqq \arg\min_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{D_i}{e_0(X_i)}Y_i - \frac{1-D_i}{1-e_0(X_i)}Y_i - g(X_i)\right)^2.$$

6. Plugin Policy

- We define a plugin policy as

$$\hat{\pi}^{\mathrm{plugin}}(X_i) \coloneqq \begin{cases} 1 & if\ \hat{g}(X_i) \geq 0 \\ 0 & if\ \hat{g}(X_i) < 0 \end{cases}.$$

- **Existing studies claim that a plugin policy with CATE estimation underperforms compared with the CRM approach.**
  - The latter solves the decision-making problem more directly.
- Below, we show that under restricted $\mathcal{G}$, CATE estimation coincides with CRM.

Equivalence

- Let $\mathcal{G}_\Pi \coloneqq \{g = 2\pi - 1 : \pi \in \Pi\}$, where Π is a policy class, and we use $\mathcal{G}_\Pi$ as a model of the CATE function.

7. 0-1 Policy

- Consider the case where $\pi: \mathcal{X} \to \{1,0\}$ and Π is its set.
- Recall that the optimal policy is given as $\pi^* \coloneqq \arg\max_{\pi\in\Pi} W(\pi)$.
- The optimal CATE predictor is given as $g^* \coloneqq \arg\min_{g\in\mathcal{G}_\Pi} E\left[\big(Y_1 - Y_0 - g(X)\big)^2\right]$.
- We can show that π^* is equivalent to g^*.

Theorem: It holds that $g^* = 2\pi^* - 1$.

[Proof] Given $\mathcal{G}_\Pi$, we have

$$\begin{aligned}
&\arg\min_{g\in\mathcal{G}_\Pi} E\left[\big(Y_1 - Y_0 - g(X)\big)^2\right] \\
&= \arg\min_{g\in\mathcal{G}_\Pi} E[(Y_1 - Y_0)^2 - 2g(X)(Y_1 - Y_0) + g(X)^2] \\
&= \arg\min_{g\in\mathcal{G}_\Pi} E[-2g(X)(Y_1 - Y_0) + g(X)^2] = \arg\min_{g\in\mathcal{G}} E[-2g(X)(Y_1 - Y_0) + 1] \\
&= \arg\min_{g\in\mathcal{G}_\Pi} E[-2g(X)(Y_1 - Y_0)]
\end{aligned}$$

We also have $\pi^* = \arg\min_{\pi\in\{\pi:\ (g+1)./2\ \ g\in\mathcal{G}_\Pi\}} E\big[-\pi(X)Y_1 - \big(1-\pi(X)\big)Y_0\big]$.

Thus, the proof completes.

- We can derive an empirical version of this result using the IPW estimator.

8. [0, 1] Policy

- Consider the case where $\pi: \mathcal{X} \to [0,1]$ and Π is its set.
- Define the regularized CRM as

$$\pi^*(\lambda) = \arg\max_{\pi\in\Pi}\{W(\pi) + \lambda E[(2\pi - 1)^2]\}.$$

  - $\lambda > 0$ is a regularization parameter.
- The optimal policy $\pi^*(\lambda)$ is equivalent to the predictor trained via

$$g^* \coloneqq \arg\min_{g\in\mathcal{G}_\Pi} E\left[\left(\frac{1}{\sqrt{\lambda}}(Y_1 - Y_0) - \sqrt{\lambda}g(X)\right)^2\right]$$

Theorem: $g^*(\lambda) = 2\pi^*(4\lambda) - 1$.

Conclusion

- Showed the equivalence between CRM and CATE estimation (plugin policy).
- This finding suggests the following points:
  - **Optimization**: New regularization terms in policy learning. ($\lambda E[(2\pi - 1)^2]$).
  - **Theory**: Possibility of developing new theoretical results.