Adaptive Experimental Design for Policy Learning: Contextual Best Arm Identification

Masahiro Kato, Mizuho-DL Financial Technology Co., Ltd.
Kyohei Okumura, Northwestern University
Takuya Ishihara, Tohoku University
Toru Kitagawa, Brown University

Mizuho-DL Financial Technology

1. Contribution

- ✓ New framework: **Contextual best arm identification (BAI).**
- ✓ Algorithm: **Adaptive experimental design for policy learning.**
- ✓ Results: **minimax optimality of the proposed algorithm.**

2. Problem Setting

➢ **Contextual BAI with a fixed budget.**

- K **treatment arms** $[K] = \{1, 2, \dots, K\}$:

  Treatment arms: alternatives of medicine, policy, and advertisements.

  - Each arm $a \in [K]$ has a potential outcome $Y^a \in \mathbb{R}$.
- **Context** $X \in \mathcal{X} \subseteq \mathbb{R}^d$.
- **Policy** π: $[K] \times \mathcal{X} \to [0,1]$ such that $\sum_{a \in [K]} \pi(a \mid x) = 1 \ \forall x \in \mathcal{X}$.
  - A policy is a function that returns a estimated best arm given a context x.
  - Policy class Π: a set of policies π. Ex. Neural networks.
- **Distribution** P of $(Y^1, \dots, Y^K, X)$.
  - $\mathcal{P}$: a set of P.
  - $\mathbb{E}_P$ and $\mathbb{P}_P$: expectation and probability laws under P.
  - $\mu^a(P) = \mathbb{E}_P[Y_t^a]$ and $\mu^a(P)(x)$
- **Policy value** $Q(P)(\pi) \coloneqq \mathbb{E}_P\left[\sum_{a \in [K]} \pi(a \mid X)\, \mu^a(P)(X)\right]$.
  - Expected outcome under a policy π and a distribution P.
- **Optimal policy** $\pi^*(P) \in \arg\max_{\pi \in \Pi} Q(P)(\pi)$.
- **Experimenter's interest**: training π^* using experimental data.
- **Adaptive experiment with T rounds:**
  - In each round $t \in [T] \coloneqq \{1, 2, \dots, T\}$, an experimenter
    1. observes d-dimensional context $X_t \in \mathcal{X} \subseteq \mathbb{R}^d$.
    2. assigns treatment $A_t \in [K]$, and
    3. observes outcome $Y_t \coloneqq \sum_{a \in [K]} 1[A_t = a] Y_t^a$.
  - After the final round T, an experimenter trains a policy π.
  - Denote the trained policy by $\hat{\pi} \in \Pi$.

3. Performance Measure

- **Performance measure of $\hat{\pi}$: Expected simple regret.**
  - Simple regret: $r_T(P)(\hat{\pi})(x) = \mu^{a^*(P)}(P)(x) - \mu^{\hat{a}_T}$ under P
  - **Expected simple regret**:
    $R_T(P)(\hat{\pi}) \coloneqq \mathbb{E}_P[r_t(P)(\hat{\pi})(X)] = Q(P)\left(\pi^*(P)\right) - Q(P)(\hat{\pi})$.
- **Our goal**:
  - Design of an algorithm (strategy) of an experimenter.
  - Under the strategy, the experimenter trains π while minimizing the expected simple regret $\mathbb{E}_P[r_t(P)]$.

→ We design the **Adaptive-Sampling (AS) and Policy-Learning (PL) strategy** and show its minimax optimality for $R_T(P)(\hat{\pi})$.

4. PLAS Strategy

- **AS**: adaptive sampling during an experiment.
- **PL**: policy learning at the end of the experiment.

➢ **Procedure of the PLAS strategy.**

- Define the target treatment-assignment ratio as follows:
  - $K = 2$: $w^*(a \mid x) = \frac{\sigma^a(x)}{(\sigma^1(x) + \sigma^2(x))^2}$ $\forall a \in [2]$ and $\forall x \in \mathcal{X}$.
  - $K \geq 3$: $w^*(a \mid x) = \frac{\left(\sigma^a(x)\right)^2}{\sum_{b \in [K]} \left(\sigma^b(x)\right)^2}$ $\forall a \in [K]$ and $\forall x \in \mathcal{X}$.

  $(\sigma^a(x))^2$: Conditional variance of Y_t^a given x.

  → Our designed policy adaptively estimates the ratio and assigns arms following the probability.
- **(AS) In each round $t \in [T]$**, an experimenter
  1. observes X_t,
  2. estimates $w^*(a \mid X_t)$ by replacing $(\sigma^a(X_t))^2$ with its estimator $(\hat{\sigma}^a(X_t))^2$, and
  3. assigns an arm a following the probability $\hat{w}_t(a \mid X_t)$, an estimator of $w^*(a \mid X_t)$.
- **(PL)** At the end, the experimenter trains π as

$$\hat{\pi}^{\mathrm{PLAS}} \coloneqq \arg\max_{\pi \in \Pi} \sum_{a \in [K]} \frac{1}{T} \sum_{t \in [T]} \pi(a \mid X_t) \left(\frac{1[A_t = a](Y_t - \hat{\mu}_t(a \mid X_t))}{\hat{w}_t(a \mid X_t)} + \hat{\mu}_t(a \mid X_t) \right).$$

- $\hat{w}_t(a \mid X_t)$ and $\hat{\mu}_t(a \mid X_t)$ are estimators of w^* and $\mu(P)(x)$ using samples until t.

5. Minimax Optimality

- Evaluate the strategy by using the worst-case analysis.
- Lower and upper bounds depend on the policy class complexity.
  - Use the Vapnik-Chervonenkis dimension when $K = 2$ and the Natarajan dimension when $K \geq 3$ for the complexity.
  - Denote the complexity of Π by M.
- Make several restrictions on the policy class and distribution.

➢ **Worst-case regret lower bounds.**

- If $K = 2$, any strategy with a trained policy $\hat{\pi}$ satisfies

$$\sup_{P \in \mathcal{P}} \sqrt{T}\, \mathbb{E}_P[R_T(P)(\hat{\pi})] \geq \frac{1}{8} \mathbb{E}_X\left[\sqrt{M(\sigma^1(X) + \sigma^2(X))^2}\right] + o(1) \quad as \ T \to \infty.$$

- If $K \geq 3$, any strategy with a trained policy $\hat{\pi}$ satisfies

$$\sup_{P \in \mathcal{P}} \sqrt{T}\, \mathbb{E}_P[R_T(P)(\hat{\pi})] \geq \frac{1}{8} \mathbb{E}_X\left[\sqrt{M \sum_{a \in [K]} (\sigma^a(X))^2}\right] + o(1) \quad as \ T \to \infty.$$

- Note that $\mathbb{E}_X\left[\sqrt{M \sum_{a \in [K]} (\sigma^a(X))^2}\right] \geq \mathbb{E}_X\left[\sqrt{M(\sigma^1(X) + \sigma^2(X))^2}\right]$.

➢ **Worst-case regret upper bounds of the PLAS strategy.**

- If $K = 2$, the PLAS strategy satisfies

$$\sqrt{T}\, \mathbb{E}_P\left[R_T(P)\left(\hat{\pi}^{\mathrm{PLAS}}\right)\right] \leq \left(272\sqrt{M} + 870.4\right) \mathbb{E}_X\left[\sqrt{(\sigma^1(X) + \sigma^2(X))^2}\right] + o(1) \quad as \ T \to \infty.$$

- If $K \geq 3$, the PLAS strategy satisfies

$$\sqrt{T}\, \mathbb{E}_P\left[R_T(P)\left(\hat{\pi}^{\mathrm{PLAS}}\right)\right] \leq \left(C\sqrt{\log(d) M} + 870.4\right) \mathbb{E}_X\left[\sqrt{M \sum_{a \in [K]} (\sigma^a(X))^2}\right] + o(1) \ as \ T \to \infty,$$

where $C > 0$ is universal constant.

➢ The leading factor depending on M, T, and $\left(\sigma^a(x)\right)^2$ of the upper bounds aligns with that of the lower bounds.

- $\mathbb{E}_X\left[\sqrt{M(\sigma^1(X) + \sigma^2(X))^2}\right]$ and $\mathbb{E}_X\left[\sqrt{M \sum_{a \in [K]} (\sigma^a(X))^2}\right]$.

→ **Minimax (rate-)optimal for the expected simple regret.**