# Off-policy estimation with adaptively collected data: the power of online learning.

Jeonghwan Lee and Cong Ma

Department of Statistics at the University of Chicago

Pre-recording for poster presentation at NeurIPS 2024

# Problem formulation

- We observe $\{(X_i, A_i, Y_i) \in \mathbb{O} := \mathbb{X} \times \mathbb{A} \times \mathbb{Y} : i \in [n]\}$ produced as:
  1. $\{X_i : i \in [n]\} \overset{\text{i.i.d.}}{\sim} \Xi^*$, where $\Xi^* \in \Delta(\mathbb{X})$ is a fixed *context distribution*;
  2. The $i$-th *behavioral policy* $\Pi_i^* : \mathbb{X} \times \mathbb{O}^{i-1} \to \Delta(\mathbb{A})$ selects the $i$-th action as $A_i \,|(X_i, \mathbf{O}_{i-1}) \sim \Pi_i^*(\cdot\,|X_i, \mathbf{O}_{i-1})$;
  3. For a Markov kernel $\Gamma^* : \mathbb{X} \times \mathbb{A} \to \Delta(\mathbb{Y})$, $Y_i \,|(X_i, A_i) \sim \Gamma^*(\cdot\,|X_i, A_i)$. The conditional mean of $Y_i$ is specified as $\mathbb{E}[Y_i\,|X_i, A_i] = \mu^*(X_i, A_i)$, where the function $\mu^* : \mathbb{X} \times \mathbb{A} \to \mathbb{R}$ is called the *treatment effect*.
- Let $\lambda_{\mathbb{A}}(\cdot)$ be a base measure over $\mathbb{A}$ such that $\Pi_i^*(\cdot\,|x_i, \mathbf{o}_{i-1}) \ll \lambda_{\mathbb{A}}$. Let $\pi_i^*(x, \mathbf{o}_{i-1}; \cdot) := \frac{\mathrm{d}\Pi_i^*(\cdot|x_i, \mathbf{o}_{i-1})}{\mathrm{d}\lambda_{\mathbb{A}}} : \mathbb{A} \to \mathbb{R}_+$ for each $i \in [n]$;
- **GOAL:** estimation of the *off-policy value* for an evaluation function $g : \mathbb{X} \times \mathbb{A} \to \mathbb{R}$ defined as $\tau(\mathcal{I}^*) := \mathbb{E}_{X \sim \Xi^*}\left[\langle g(X, \cdot), \mu^*(X, \cdot)\rangle_{\lambda_{\mathbb{A}}}\right]$, where $\mathcal{I}^* := (\Xi^*, \Gamma^*)$ defines our *problem instance*;
- The *propensity scores* $\{\pi_i^*(X_i, \mathbf{O}_{i-1}; A_i) : i \in [n]\}$ are revealed.

# Meta-algorithm: the class of AIPW estimators

**Algorithm 1** Meta-algorithm: augmented inverse propensity weighting (AIPW) estimator.

**Input:** the dataset $\mathcal{D} = \{(X_i, A_i, Y_i) \in \mathbb{O} : i \in [n]\}$ and an evaluation function $g : \mathbb{X} \times \mathbb{A} \to \mathbb{R}$.

1: For each step $i \in [n]$, we compute an estimate $\hat{\mu}_i(\mathbf{O}_{i-1}) \in (\mathbb{X} \times \mathbb{A} \to \mathbb{R})$ of the treatment effect based on the sample trajectory $\mathbf{O}_{i-1}$ up to the $(i-1)$-th step. `// Implement Algorithm 2 as a subroutine;`

2: Consider the AIPW estimator (a.k.a., the *doubly-robust* (DR) estimator) $\hat{\tau}_n^{\mathsf{AIPW}}(\cdot) : \mathbb{O}^n \to \mathbb{R}$:

$$\hat{\tau}_n^{\mathsf{AIPW}}(\mathbf{o}_n) := \frac{1}{n} \sum_{i=1}^{n} \hat{\Gamma}_i(\mathbf{o}_i), \tag{3.6}$$

where the objects being averaged are the AIPW scores $\hat{\Gamma}_i(\cdot) : \mathbb{O}^i \to \mathbb{R}$ is defined by

$$\hat{\Gamma}_i(\mathbf{o}_i) := \frac{g(x_i, a_i)}{\pi_i^*(x_i, \mathbf{o}_{i-1}; a_i)} \{y_i - \hat{\mu}_i(\mathbf{o}_{i-1})(x_i, a_i)\} + \langle g(x_i, \cdot), \hat{\mu}_i(\mathbf{o}_{i-1})(x_i, \cdot)\rangle_{\lambda_{\mathbb{A}}}. \tag{3.7}$$

3: **return** the AIPW estimate $\hat{\tau}_n^{\mathsf{AIPW}}(\mathbf{O}_n)$.

- AIPW estimation combines the direct method (DM) and the IPW estimation to leverage their complementary strengths;
- *Cross-fitted* DR estimator is $\sqrt{n}$-consistent and asymptotically efficient. $\rightarrow$ We aim at establishing its *non-asymptotic theory*!

# Non-asymptotic guarantees of the AIPW estimator

## Theorem 1 (Non-asymptotic upper bound on the MSE)

*For any sequence of estimates* $\{\hat{\mu}_i(\mathbf{O}_{i-1}) \in (\mathbb{X} \times \mathbb{A} \to \mathbb{R}) : i \in [n]\}$ *for the treatment effect* $\mu^*$, *the AIPW estimator has the MSE bounded above by*

$$\mathbb{E}_{\mathcal{I}^*}\left[\left\{\hat{\tau}_n^{\mathrm{AIPW}}(\mathbf{O}_n) - \tau(\mathcal{I}^*)\right\}^2\right]$$
$$\leq \frac{1}{n}\left\{v_*^2 + \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\mathcal{I}^*}\left[\frac{g^2(X_i, A_i)\{\hat{\mu}_i(\mathbf{O}_{i-1})(X_i, A_i) - \mu^*(X_i, A_i)\}^2}{(\pi_i^*)^2(X_i, \mathbf{O}_{i-1}; A_i)}\right]\right\}.$$

- The first term $v_*^2$ is *unavoidable*;
- The second term measures the average estimation error of the sequence of estimates $\{\hat{\mu}_i(\mathbf{O}_{i-1}) \in (\mathbb{X} \times \mathbb{A} \to \mathbb{R}) : i \in [n]\}$.
  $\rightarrow$ We need to choose a sequence which minimizes this term!

# Reduction to online non-parametric regression

- To construct a desired sequence of estimates for $\mu^*$, we borrow the idea of **online non-parametric regression (NPR)**;

**Algorithm 2** Online non-parametric regression protocol for estimation of the treatment effect.

**Input:** the number of rounds $n \in \mathbb{N}$.

1: **for** $i = 1, 2, \cdots, n$, **do**

2: The learner selects a point $\hat{\mu}_i(\mathbf{O}_{i-1}) \in (\mathbb{X} \times \mathbb{A} \to \mathbb{R})$ based on the sample trajectory $\mathbf{O}_{i-1}$;

3: The environment then picks a loss function $l_i(\cdot) : (\mathbb{X} \times \mathbb{A} \to \mathbb{R}) \to \mathbb{R}$ defined as

$$l_i(\mu) := \frac{g^2(X_i, A_i)}{\left(\pi_i^*\right)^2 (X_i, \mathbf{O}_{i-1}; A_i)} \{Y_i - \mu(X_i, A_i)\}^2, \ \forall \mu(\cdot, \cdot) \in (\mathbb{X} \times \mathbb{A} \to \mathbb{R}). \tag{3.12}$$

4: **end for**

5: **return** the sequence of estimates $\{\hat{\mu}_i(\mathbf{O}_{i-1}) \in (\mathbb{X} \times \mathbb{A} \to \mathbb{R}) : i \in [n]\}$ of the treatment effect.

- **GOAL:** minimize the regret against the *best fixed action in hindsight* belonging to a pre-specified function class $\mathcal{F} \subseteq (\mathbb{X} \times \mathbb{A} \to \mathbb{R})$:

$$\text{Regret}(n, \mathcal{F}; \mathcal{A}) := \sum_{i=1}^{n} l_i \{\hat{\mu}_i(\mathbf{O}_{i-1})\} - \inf \left\{ \sum_{i=1}^{n} l_i(\mu) : \mu \in \mathcal{F} \right\}$$

# Reduction to online non-parametric regression

**Theorem 2 (Oracle inequality for the class of AIPW estimators)**

*The AIPW estimator using a sequence of estimates for $\mu^*$ produced by the online NPR algorithm $\mathcal{A}$ enjoys the following upper bound on the MSE:*

$$\begin{aligned} &\mathbb{E}_{\mathcal{I}^*}\left[\left\{\hat{\tau}_n^{\mathrm{AIPW}}\left(\mathbf{O}_n\right)-\tau\left(\mathcal{I}^*\right)\right\}^2\right] \\ &\leq \frac{1}{n}\left(v_*^2+\frac{1}{n}\mathbb{E}_{\mathcal{I}^*}\left[\mathrm{Regret}\left(n, \mathcal{F}; \mathcal{A}\right)\right]+\inf\left\{\|\mu-\mu^*\|_{(n)}^2 : \mu \in \mathcal{F}\right\}\right). \end{aligned} \tag{1}$$

# Reduction to online non-parametric regression

- If $\mathcal{A}$ exhibits a *no-regret learning dynamics*, i.e., $\mathbb{E}_{\mathcal{I}^*}[\text{Regret}(n, \mathcal{F}; \mathcal{A})] = o(n)$ as $n \to \infty$, the RHS of (1) is asymptotically the same as

$$\frac{1}{n}\left(v_*^2 + \inf\left\{\|\mu - \mu^*\|_{(n)}^2 : \mu \in \mathcal{F}\right\}\right).$$

- The AIPW estimator may suffer from an efficiency loss which depends on how well the unknown treatment effect $\mu^*$ can be approximated by a member of $\mathcal{F} \subseteq (\mathbb{X} \times \mathbb{A} \to \mathbb{R})$ under the $\|\cdot\|_{(n)}$-norm.

Thank you for listening!