ML-assisted Randomization Tests for Detecting Treatment Effects in A/B Experiments

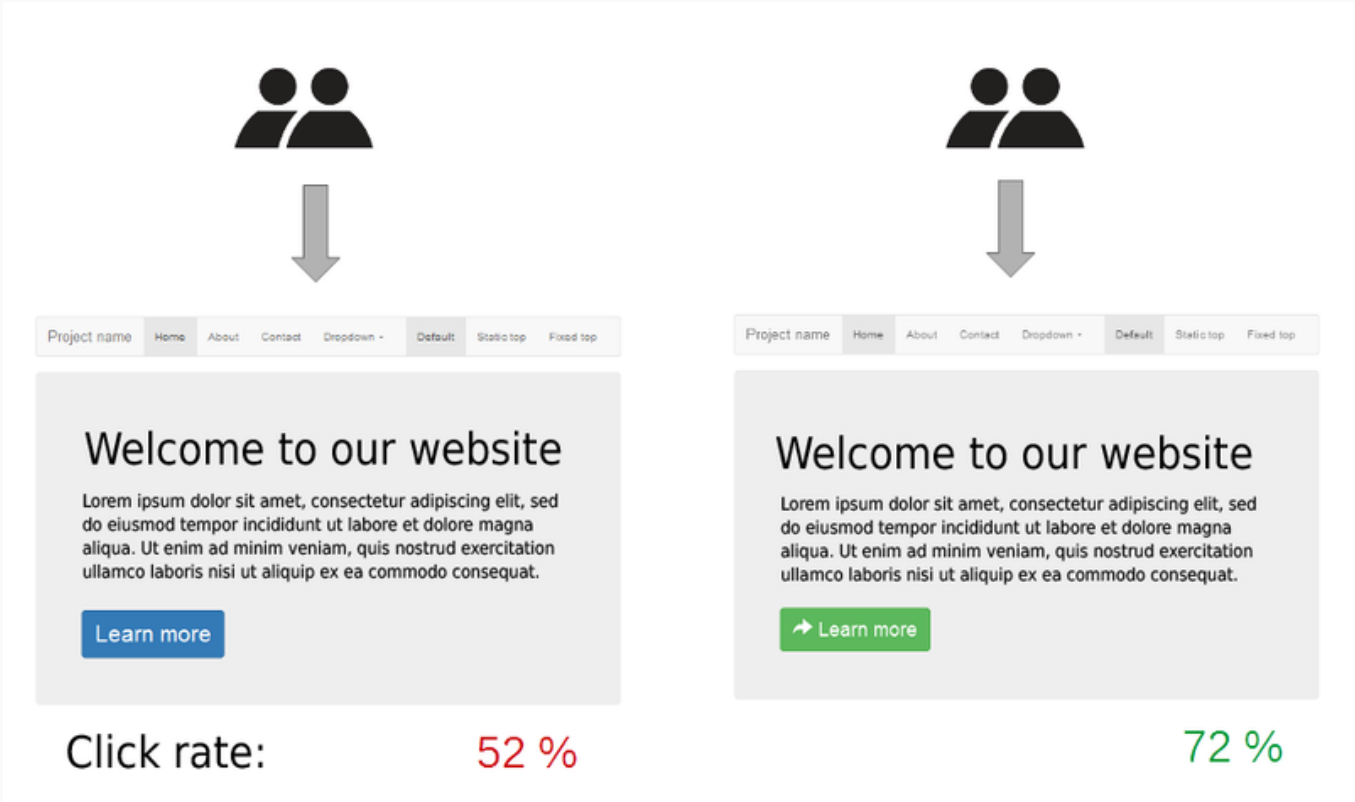
Wenxuan Guo, Fourth year Ph.D. student, University of Chicago
JungHo Lee, Second year Ph.D. student, Carnegie Mellon University
Panos Toulis, University of Chicago

American Causal Inference Conference, May 2025

Introduction

Randomized experiments lie at the heart of causal inference and data-driven decision making.

- In an A/B experiment, an online business randomizes two different treatments and aims to infer which is better.



Standard approaches

- A classical method to analyze A/B experiments is the t-test (Kohavi et al., 2020).
- Limited to average marginal effects and not finite-sample valid.
- Methods that use *Fisherian Randomization Tests* (FRTs) —e.g., permutation tests— tend to utilize standard t-statistics, producing results similar to t-tests.
- ANOVA-based methods can be more flexible but mainly used with linear models (Gerber and Green, 2012).

Contribution

We propose **Machine Learning (ML)-assisted randomization tests**.

The main idea is to:

- Utilize *ML-based* test statistics in the context of an FRT.
- Retain finite-sample validity of FRTs.
- Increased power compared to linear models thanks to ML.
  - New theoretical results on the test power.
- Flexible enough to test for global effects, heterogeneous treatment effects, and spillovers.

Setup

- $Z = (Z_1, \ldots, Z_n) \in \{0,1\}^n$: binary *treatments*.
- Treatment assignment is known: $Z \sim \mathbb{P}_n(Z)$.
- $Y = (Y_1, \ldots, Y_n) \in \mathbb{R}^n$: *outcomes*.
- $X_1, \ldots, X_n$: *covariates*, $X_i \in \mathbb{R}^p$. $\mathbf{X} \in \mathbb{R}^{n \times p}$ for entire matrix.

We posit the *outcome model*:

$$Y_i = \mu + \underbrace{b(X_i)}_{\text{baseline}} + Z_i \underbrace{h(X_i)}_{\text{direct effect}} + \underbrace{g(\mathbf{X}, Z_{-i})}_{\text{spillover}} + \varepsilon_i, \tag{1}$$

where ε_i is an independent noise with $\mathbb{E}(\varepsilon_i \mid \mathbf{X}) = 0$, $\varepsilon \perp\!\!\!\perp Z \mid \mathbf{X}$.

(1) follows Causal ML literature (Hill, 2011; Chernozhukov et al., 2018; Künzel et al., 2019).

Null hypothesis of no treatment effect

Outcome model:

$$Y_i = \mu + \underbrace{b(X_i)}_{\text{baseline}} + Z_i \underbrace{h(X_i)}_{\text{direct effect}} + \underbrace{g(\mathbf{X}, Z_{-i})}_{\text{spillover}} + \varepsilon_i.$$

As a starting point, consider the null

$$H_0^{\text{glob}} : h = 0, g = 0 \quad \text{v.s.} \quad H_1^{\text{glob}} : h \neq 0, g = 0.$$

- Under the potential outcomes framework, $H_0^{\text{glob}} \equiv Y_i(0) \stackrel{d}{=} Y_i(1)$, which is weaker than *Fisher's sharp null*.

ML-based test statistic

To test H_0^{glob}, we propose constructing two models using ML:

$$\mathcal{M}_0^{\text{glob}} : Y_i \sim X_i, \quad \mathcal{M}_1^{\text{glob}} : Y_i \sim Z_i + X_i.$$

Define the test statistic as

$$t_n(Y, Z, \mathbf{X}) := \text{CV}_{n,k}(\mathcal{M}_0^{\text{glob}}) - \text{CV}_{n,k}(\mathcal{M}_1^{\text{glob}}), \qquad (2)$$

where $\text{CV}_{n,k}(\mathcal{M})$: k-fold cross-validated squared loss of model $\mathcal{M}$.

- Intuitively, $t_n(Y, Z, \mathbf{X})$ measures whether Z is predictive of Y.
- An ANOVA-type statistic (Gerber and Green, 2012; Breiman, 2001; Strobl et al., 2008; Williamson et al., 2021; Bénard et al., 2022).
- Omnibus test: only detects effect; does not quantify an ATE.
- Captures non-linear treatment effects through $\mathcal{M}_1^{\text{glob}}$.

Finite-sample valid testing procedure

Procedure 1 (ML-assisted Randomization Test)

1. Obtain observed value of $T_n = t_n(Y, Z, \mathbf{X})$ as defined in (2).
2. Compute $t^{(r)} = t_n(Y, Z^{(r)}, \mathbf{X})$, $Z^{(r)} \overset{iid}{\sim} \mathbb{P}_n$, for $r = 1, \ldots, R$.
3. Calculate p-value:

$$\text{pval} = \frac{1}{1+R}\left[\sum_{r=1}^{R} \mathbb{1}\{t^{(r)} > T_n\} + 1\right]. \tag{3}$$

- The test is finite-sample valid (Lehmann and Romano, 2005, e.g.):

$$\mathbb{P}\big(\text{pval} \leq \alpha \mid \mathbf{X}, H_0^{\text{glob}}\big) \leq \alpha, \text{ for any } \alpha \in [0,1] \text{ and any } n > 0.$$

- What about power?

Type II error bound

Assumption

- *Bernoulli design with probability* $\pi \in (0, 1)$,
- $(X_i, \varepsilon_i)_{i \in [n]}$ *are i.i.d. with* $\mathbb{E}(\varepsilon_i | X_i) = 0$ *and* $\mathbb{E}(\varepsilon_i^2) < \infty$,
- $|Y_i| \leq M$ *with probability one.*

Define

- $\mathcal{F}$ = function class of ML models in $\mathcal{M}_1$ (full model) with domain $\mathcal{X} \times \{0, 1\}$.
- $\mathcal{F}_0$ = function class of ML models in $\mathcal{M}_0$ (reduced model) with domain $\mathcal{X}$.
- Alternative hypothesis H_1^{glob}: $h \neq 0, g = 0 \Rightarrow$ nonzero direct effect.

Main Theorem

Theorem (G., Lee, Toulis)

Suppose the previous assumption holds with additional regularity conditions and $k = O(1)$. Then, under the alternative H_1^{glob}, for some small constant $C > 0$,

$$\mathbb{P}(\text{pval} > \alpha) = O\left(k \exp\left(-\frac{Cn\Delta^2}{kM^4}\right)\right),$$

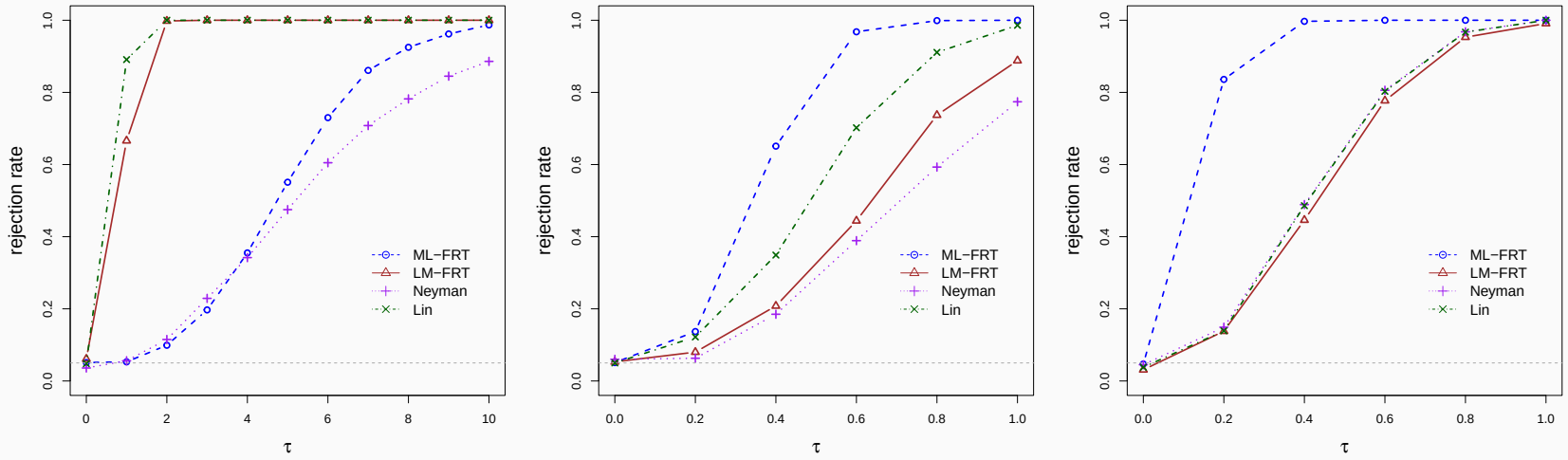
- Quantity Δ measures the *variable importance* of treatment:

$$\Delta := \underbrace{\inf_{f_0 \in \mathcal{F}_0} \mathbb{E}(Y - f_0(X))^2}_{\text{prediction error in the reduced model}} - \underbrace{\inf_{f \in \mathcal{F}} \mathbb{E}(Y - f(X, Z))^2}_{\text{prediction error in the full model}}.$$

- e.g., $\Delta = \pi(1-\pi)\tau^2$, in a linear model $y = a + bx + \tau z$.

Takeaway: better prediction $\Rightarrow$ larger Δ $\Rightarrow$ higher power!

Simulations



(a) Linear **(b)** Piecewise-constant **(c)** Cosine

Figure 1: Rejection rates for constant treatment effects.

- We implement random forests and linear model in our test (ML-FRT, LM-FRT).
- Compared to Neyman's difference-in-means estimator and Lin's estimator (interacted regression).
- Benefits from our procedure in more complex outcome models.

Extensions

$$Y_i = \mu + \underbrace{b(X_i)}_{\text{baseline}} + Z_i \underbrace{h(X_i)}_{\text{direct effect}} + \underbrace{g(\mathbf{X}, Z_{-i})}_{\text{spillover}} + \varepsilon_i.$$

Treatment heterogeneity. $H_0^{\text{het}} : h(x) = \tau, g = 0$ vs. $h(x) \neq \tau, g = 0$.

- Repeat Procedure 1 for $Y - \tau_0 Z$ to get $\mathrm{pval}(\tau_0)$ and "sup" over τ_0.

Spillover. $H_0^{\text{sp}} : g = 0$ vs. $g \neq 0$. Modify Procedure 1 as:

$$\mathcal{M}_0^{\text{sp}} : Y_i \sim Z_i + X_i, \quad \mathcal{M}_1^{\text{sp}} : Y_i \sim Z_i + \mathbf{A}_{i\cdot}^\top Z + \mathbf{X}.$$

- $\mathbf{A} \in \{0,1\}^{n \times n}$: adjacency matrix between units.
- Spillover effects can be captured by $\mathcal{M}_1^{\text{sp}}$ through $\mathbf{A}_{i\cdot}^\top Z$ and $\mathbf{X}$.
- Conditional randomization test: fixing individual treatments Z_i but varying $\mathbf{A}_i^\top Z$. (Athey et al., 2018; Basse et al., 2019, 2024)

Thank you!

- Wenxuan Guo, JungHo Lee, and Panos Toulis, "ML-assisted Randomization Tests for Detecting Treatment Effects in A/B Experiments ," `https://arxiv.org/abs/2501.07722`, 2025.