Towards Optimal Variance Reduction in Online Experiments

Ying Jin

Department of Statistics, Stanford University

Joint work with Shan Ba at LinkedIn Applied Research

Conference on Digital Experimentation, November 5, 2021

Randomized experiments in tech companies

- A/B testing, randomized experiments, online controlled experiments...
  - Units are randomly assigned to treated / control groups
  - Measure outcomes after a period
  - Evaluate and compare the outcomes in two groups
  - If the effect is significantly positive, then adopt the new feature

Randomized experiments in tech companies

- A/B testing, randomized experiments, online controlled experiments...
  - Units are randomly assigned to treated / control groups
  - Measure outcomes after a period
  - Evaluate and compare the outcomes in two groups
  - If the effect is significantly positive, then adopt the new feature
- Powerful hypothesis testing is important
  - shorter experimental horizon, smaller sample, avoid potentially negative impacts

Randomized experiments in tech companies

- A/B testing, randomized experiments, online controlled experiments...
  - Units are randomly assigned to treated / control groups
  - Measure outcomes after a period
  - Evaluate and compare the outcomes in two groups
  - If the effect is significantly positive, then adopt the new feature
- Powerful hypothesis testing is important
  - shorter experimental horizon, smaller sample, avoid potentially negative impacts
- Desire estimator of treatment effects with smaller variance
  - What treatment effect?
  - How to reduce variance?
  - What is the best effort?

Overview of the work

- A rigorous statistical framework for variance reduction of count and ratio metrics
- Methodology of unbiased variance reduction with flexible machine learning tools and large numbers of covariates
- Optimality in the sense of semiparametric efficiency of all procedures
- Performance on (simulated and) real data

Potential outcome framework

- i.i.d. units $i = 1, \ldots, n$ from $\mathbb{P}$.
- Potential outcomes $Y_i(1)$, $Y_i(0)$.
- Treatment $T_i \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p)$.
- SUTVA: observe $Y_i(1)$ for $T_i = 1$ and $Y_i(0)$ for $T_i = 0$.
- $n_t = \sum_i T_i$ size of treated group, $n_c = n - n_t$ size of control group.

Randomization and metrics

- Case 1: Count metric $\tau = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$
- Default estimator: difference-in-mean

$$\widehat{\tau}_{\text{DIM}} = \frac{1}{n_t} \sum_{i \text{ treated}} Y_i - \frac{1}{n_c} \sum_{i \text{ control}} Y_i .$$

Randomization and metrics

- Case 2: ratio metric

$$\frac{\sum_{i \text{ treated}} Y_i}{\sum_{i \text{ treated}} Z_i} = \frac{\frac{1}{n_t}\sum_{i \text{ treated}} Y_i}{\frac{1}{n_t}\sum_{i \text{ treated}} Z_i}, \quad \frac{\sum_{i \text{ control}} Y_i}{\sum_{i \text{ control}} Z_i} = \frac{\frac{1}{n_c}\sum_{i \text{ control}} Y_i}{\frac{1}{n_c}\sum_{i \text{ control}} Z_i},$$

where i stands for a cluster, Y_i is the aggregated outcome, Z_i is the size of cluster.

Randomization and metrics

- Case 2: ratio metric

$$\frac{\sum_{i \text{ treated}} Y_i}{\sum_{i \text{ treated}} Z_i} = \frac{\frac{1}{n_t}\sum_{i \text{ treated}} Y_i}{\frac{1}{n_t}\sum_{i \text{ treated}} Z_i}, \quad \frac{\sum_{i \text{ control}} Y_i}{\sum_{i \text{ control}} Z_i} = \frac{\frac{1}{n_c}\sum_{i \text{ control}} Y_i}{\frac{1}{n_c}\sum_{i \text{ control}} Z_i},$$

  where i stands for a cluster, Y_i is the aggregated outcome, Z_i is the size of cluster.

- **Stable denominator assumption**: $Z_i = Z_i(1) = Z_i(0)$, not influenced by the treatment.

Randomization and metrics

- Case 2: ratio metric

$$\frac{\sum_{i \text{ treated}} Y_i}{\sum_{i \text{ treated}} Z_i} = \frac{\frac{1}{n_t}\sum_{i \text{ treated}} Y_i}{\frac{1}{n_t}\sum_{i \text{ treated}} Z_i}, \quad \frac{\sum_{i \text{ control}} Y_i}{\sum_{i \text{ control}} Z_i} = \frac{\frac{1}{n_c}\sum_{i \text{ control}} Y_i}{\frac{1}{n_c}\sum_{i \text{ control}} Z_i},$$

  where i stands for a cluster, Y_i is the aggregated outcome, Z_i is the size of cluster.

- **Stable denominator assumption**: $Z_i = Z_i(1) = Z_i(0)$, not influenced by the treatment.
- If SDA holds, the population quantity is $\delta' = \frac{\mathbb{E}[Y_i(1)]}{\mathbb{E}[Z_i]} - \frac{\mathbb{E}[Y_i(0)]}{\mathbb{E}[Z_i]}$.

Randomization and metrics

- Case 2: ratio metric

$$\frac{\sum_{i \text{ treated}} Y_i}{\sum_{i \text{ treated}} Z_i} = \frac{\frac{1}{n_t}\sum_{i \text{ treated}} Y_i}{\frac{1}{n_t}\sum_{i \text{ treated}} Z_i}, \quad \frac{\sum_{i \text{ control}} Y_i}{\sum_{i \text{ control}} Z_i} = \frac{\frac{1}{n_c}\sum_{i \text{ control}} Y_i}{\frac{1}{n_c}\sum_{i \text{ control}} Z_i},$$

  where i stands for a cluster, Y_i is the aggregated outcome, Z_i is the size of cluster.

- **Stable denominator assumption**: $Z_i = Z_i(1) = Z_i(0)$, not influenced by the treatment.
- If SDA holds, the population quantity is $\delta' = \frac{\mathbb{E}[Y_i(1)]}{\mathbb{E}[Z_i]} - \frac{\mathbb{E}[Y_i(0)]}{\mathbb{E}[Z_i]}$.
- If SDA does not hold, the population quantity is $\delta = \frac{\mathbb{E}[Y_i(1)]}{\mathbb{E}[Z_i(1)]} - \frac{\mathbb{E}[Y_i(0)]}{\mathbb{E}[Z_i(0)]}$.

Methods in the literature

- For count metrics...
- Earlier: linear adjustment [Lin, 2013], CUPED [Deng et al., 2013]..
  - linearity is restrictive
  - cannot handle large numbers of covariates
- Recently: machine learning, large numbers of covariates [Guo et al., 2021]
  - optimality of procedures?

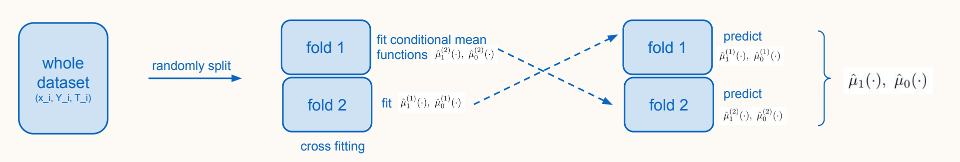
Methods in the literature

- For count metrics...
- Earlier: linear adjustment [Lin, 2013], CUPED [Deng et al., 2013]..
  - linearity is restrictive
  - cannot handle large numbers of covariates
- Recently: machine learning, large numbers of covariates [Guo et al., 2021]
  - optimality of procedures?
- For ratio metrics...
  - less studied, lack of rigorous statistical framework and guarantee
  - CUPED extension cannot use general covariates
  - optimality of procedures?

High-level idea: fit-then-debias

- Why is diff-in-mean estimator not efficient?
  - Decreased sample size (half of n units are treated / control)
  - Unpaired comparison (variance is the sum of treated and control)
  - Ideal estimator: $\frac{1}{n}\sum_{i=1}^{n}[Y_i(1) - Y_i(0)]$ (for count metrics)
- General idea:
  - with covariates X_i, use ML estimators to predict the missing outcome and plug in to perform pairwise comparison $\Rightarrow \frac{1}{n}\sum_{i=1}^{n}[\widehat{\mu}_1(X_i) - \widehat{\mu}_0(X_i)]$?
  - Bias can be larger than variance!
- De-biasing techniques
  - use cross-fitting to fit the estimators
  - add a de-biasing term to correct for the bias of $\widehat{\mu}_w$

Procedure for count metrics



- Step 1: sample splitting into $\mathcal{D}^{(k)}$, $k = 1, \ldots, K$.
- Step 2: use $\mathcal{D} \backslash \mathcal{D}^{(k)}$ to estimate $\widehat{\mu}_1^{(k)}$ and $\widehat{\mu}_0^{(k)}$.
- Step 3: Plug into $\mathcal{D}^{(k)}$ to obtain $\widehat{\mu}_1(X_i) = \widehat{\mu}_1^{(k)}(X_i)$, $\widehat{\mu}_0(X_i) = \widehat{\mu}_0^{(k)}(X_i)$ for all $i \in \mathcal{D}^{(k)}$.
- Step 4: Estimator

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\mu}_1(X_i) - \widehat{\mu}_0(X_i)\right) + \frac{1}{n_t}\sum_{i \text{ treated}}\left(Y_i(1) - \widehat{\mu}_1(X_i)\right) - \frac{1}{n_c}\sum_{i \text{ control}}\left(Y_i(0) - \widehat{\mu}_0(X_i)\right),$$

Procedure for count metrics

- Estimator

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n}\big(\widehat{\mu}_1(X_i) - \widehat{\mu}_0(X_i)\big) + \frac{1}{n_t}\sum_{i\ \text{treated}}\big(Y_i(1) - \widehat{\mu}_1(X_i)\big) - \frac{1}{n_c}\sum_{i\ \text{control}}\big(Y_i(0) - \widehat{\mu}_0(X_i)\big),$$

- Valid inference when $\|\widehat{\mu}_w^{(k)} - \mu_w^*\|_2 \xrightarrow{P} 0$ for deterministic functions μ_w^*, $w \in \{0,1\}$.
- Semiparametrically efficient when $\mu_w^*(x) = \mathbb{E}[Y(w) \,|\, X = x]$, $w \in \{0,1\}$.

Count metrics: implications on optimality

- When estimators converge in L_2 to true conditional mean functions,

$$\mathrm{Var}(\widehat{\theta}) \approx \underbrace{\frac{1}{n}\mathrm{Var}\left(\mu_1(X_i) - \mu_0(X_i)\right)}_{\text{(i) predictable part}} + \underbrace{\frac{1}{n_t}\mathrm{Var}\left(Y_i(1) - \mu_1(X_i)\right) + \frac{1}{n_c}\mathrm{Var}\left(Y_i(0) - \mu_0(X_i)\right)}_{\text{(ii) irreducible variance}}$$

  - (i) is the best efforts in predicting $Y(1) - Y(0)$ given X.
  - (ii) is the intrinsic uncertainty that cannot be eliminated.

Count metrics: implications on optimality

- When estimators converge in L_2 to true conditional mean functions,

$$\mathrm{Var}(\widehat{\theta}) \approx \underbrace{\frac{1}{n}\mathrm{Var}\left(\mu_1(X_i) - \mu_0(X_i)\right)}_{\text{(i) predictable part}} + \underbrace{\frac{1}{n_t}\mathrm{Var}\left(Y_i(1) - \mu_1(X_i)\right) + \frac{1}{n_c}\mathrm{Var}\left(Y_i(0) - \mu_0(X_i)\right)}_{\text{(ii) irreducible variance}}$$

  - (i) is the best efforts in predicting $Y(1) - Y(0)$ given X.
  - (ii) is the intrinsic uncertainty that cannot be eliminated.

- Take-away message on optimality:
  - Should target for conditional mean functions
    $\Rightarrow$ This method is better than CUPED when there is nonlinearity

Count metrics: implications on optimality

- When estimators converge in L_2 to true conditional mean functions,

$$\mathrm{Var}(\widehat{\theta}) \approx \underbrace{\frac{1}{n}\mathrm{Var}\left(\mu_1(X_i) - \mu_0(X_i)\right)}_{\text{(i) predictable part}} + \underbrace{\frac{1}{n_t}\mathrm{Var}\left(Y_i(1) - \mu_1(X_i)\right) + \frac{1}{n_c}\mathrm{Var}\left(Y_i(0) - \mu_0(X_i)\right)}_{\text{(ii) irreducible variance}}$$

  - (i) is the best efforts in predicting $Y(1) - Y(0)$ given X.
  - (ii) is the intrinsic uncertainty that cannot be eliminated.

- Take-away message on optimality:
  - Should target for conditional mean functions
    $\Rightarrow$ This method is better than CUPED when there is nonlinearity
  - Estimate for two groups $\mathbb{E}[Y(1) \mid X = x]$ and $\mathbb{E}[Y(0) \mid X = x]$ separately
    $\Rightarrow$ This method is better than one single estimator when there is treatment heterogeneity

Ratio metrics: optimal procedures

- We develop two procedures for δ and δ' under different conditions of denominators.
- For the target $\delta = \frac{\mathbb{E}[Y_i(1)]}{\mathbb{E}[Z_i(1)]} - \frac{\mathbb{E}[Y_i(0)]}{\mathbb{E}[Z_i(0)]}$ without SDA
  - fit-then-debias for $\mathbb{E}[Y(w) \mid X = x]$, $\mathbb{E}[Z(w) \mid X = x]$ respectively, $w \in \{0, 1\}$.
- For the target $\delta' = \frac{\mathbb{E}[Y_i(1)]}{\mathbb{E}[Z_i]} - \frac{\mathbb{E}[Y_i(0)]}{\mathbb{E}[Z_i]}$ under SDA,
  - pool all Z_i to estimate $\mathbb{E}[Z]$
  - fit-then-debias for $\mathbb{E}[Y(w) \mid X = x, Z = z]$, $w \in \{0, 1\}$.

Ratio metrics: optimal procedures

- We develop two procedures for δ and δ' under different conditions of denominators.
- For the target $\delta = \frac{\mathbb{E}[Y_i(1)]}{\mathbb{E}[Z_i(1)]} - \frac{\mathbb{E}[Y_i(0)]}{\mathbb{E}[Z_i(0)]}$ without SDA
  - fit-then-debias for $\mathbb{E}[Y(w) \mid X = x]$, $\mathbb{E}[Z(w) \mid X = x]$ respectively, $w \in \{0, 1\}$.
- For the target $\delta' = \frac{\mathbb{E}[Y_i(1)]}{\mathbb{E}[Z_i]} - \frac{\mathbb{E}[Y_i(0)]}{\mathbb{E}[Z_i]}$ under SDA,
  - pool all Z_i to estimate $\mathbb{E}[Z]$
  - fit-then-debias for $\mathbb{E}[Y(w) \mid X = x, Z = z]$, $w \in \{0, 1\}$.
- Valid inference when estimators converge to deterministic functions (L_2 distance in probability)
- Optimality when they converge to true conditional mean functions

Ratio metrics: optimal procedures

- Ratio metric: target quantity $\delta = \frac{\mathbb{E}[Y_i(1)]}{\mathbb{E}[Z_i(1)]} - \frac{\mathbb{E}[Y_i(0)]}{\mathbb{E}[Z_i(0)]}$

$$\widehat{\delta} = \frac{\sum_{i=1}^n A_i}{\sum_{i=1}^n B_i} - \frac{\sum_{i=1}^n C_i}{\sum_{i=1}^n D_i},$$

$$\text{where } A_i = \widehat{\mu}_1^Y(X_i) + \frac{T_i}{\widehat{p}}(Y_i - \widehat{\mu}_1^Y(X_i)), \quad B_i = \widehat{\mu}_1^Z(X_i) + \frac{T_i}{\widehat{p}}(Z_i - \widehat{\mu}_1^Z(X_i)),$$

$$C_i = \underbrace{\widehat{\mu}_0^Y(X_i)}_{\text{regressor + plug-in}} + \underbrace{\frac{1-T_i}{1-\widehat{p}}(Y_i - \widehat{\mu}_0^Y(X_i))}_{\text{de-bias}}, \quad D_i = \widehat{\mu}_0^Z(X_i) + \frac{1-T_i}{1-\widehat{p}}(Z_i - \widehat{\mu}_0^Z(X_i)).$$

- Ratio metric: target quantity $\delta' = \frac{\mathbb{E}[Y_i(1)]}{\mathbb{E}[Z_i]} - \frac{\mathbb{E}[Y_i(0)]}{\mathbb{E}[Z_i]}$

$$\widehat{\delta}' = \frac{\sum_{i=1}^n \Gamma_i}{\sum_{i=1}^n Z_i}, \quad \text{where } \Gamma_i = \underbrace{\widehat{\mu}_1(X_i, Z_i) - \widehat{\mu}_0(X_i, Z_i)}_{\text{regressor + plug-in}} + \underbrace{\frac{T_i}{\widehat{p}}(Y_i - \widehat{\mu}_1(X_i, Z_i)) - \frac{1-T_i}{1-\widehat{p}}(Y_i - \widehat{\mu}_0(X_i, Z_i))}_{\text{de-bias}},$$

Real data performance

- Count metric: LinkedIn Feed experiment, revenue metric
  - $n = 400,000$ subsample of users
  - Our estimator with random forest from `scikit-learn` python library
  - Incorporate user covariates
  - Reduces 22.22% of variance compared to diff-in-mean, while CUPED reduces 15.91%

Real data performance

- Ratio metric: enterprise experiment in LinkedIn Learning, 'learning engagement' metric
  - $n = 10,299$ enterprise accounts
  - Our estimator with `XGBoost` python library
  - Incorporate enterprise covariates
  - For δ, reduces 12.35% of variance compared to diff-in-mean, while CUPED reduces 1.76%
  - For δ', reduces 83.6% of variance compared to diff-in-mean, while CUPED reduces 76.62%

Thanks!

Our paper: https://arxiv.org/abs/2110.13406