TECHNOMETRICS
2023, VOL. 65, NO. 2, 231–242
https://doi.org/10.1080/00401706.2022.2142670

Taylor & Francis
Taylor & Francis Group

Toward Optimal Variance Reduction in Online Controlled Experiments

Ying Jin[a] and Shan Ba[b]

[a]Department of Statistics, Stanford University, Stanford, CA; [b]Data Science Applied Research, LinkedIn Corporation, Sunnyvale, CA

ABSTRACT

We study optimal variance reduction solutions for count and ratio metrics in online controlled experiments. Our methods apply flexible machine learning tools to incorporate covariates that are independent from the treatment but have predictive power for the outcomes, and employ the cross-fitting technique to remove the bias in complex machine learning models. We establish CLT-type asymptotic inference based on our estimators under mild convergence conditions. Our procedures are optimal (efficient) for the corresponding targets as long as the machine learning estimators are consistent, without any requirement for their convergence rates. In complement to the general optimal procedure, we also derive a linear adjustment method for ratio metrics as a special case that is computationally efficient and can flexibly incorporate any pretreatment covariates. We evaluate the proposed variance reduction procedures with comprehensive simulation studies and provide practical suggestions regarding commonly adopted assumptions in computing ratio metrics. When tested on real online experiment data from LinkedIn, the proposed optimal procedure for ratio metrics can reduce up to 80% of variance compared to the standard difference-in-mean estimator and also further reduce up to 30% of variance compared to the CUPED approach by going beyond linearity and incorporating a large number of extra covariates.

ARTICLE HISTORY
Received November 2021
Accepted October 2022

KEYWORDS
A/B test; Causal inference; Clustered randomized experiments; Covariate adjustment; Ratio metrics; Semiparametric efficiency

1. Introduction

Online controlled experiments, also known as A/B tests, are extensively used in tech companies to assess the impacts of product changes on business metrics. Although online experiments can involve millions of users, their sensitivity is still a major challenge because the treatment effect is often small compared to the noise. Failing to detect even small differences in key metrics can have significant business implications (Kohavi, Tang, and Xu 2020) and it is crucial to develop powerful statistical tools to quickly capture nonzero effects with fewer samples and shorter experimentation turn-around time.

To improve the sensitivity of online controlled experiments, variance reduction techniques are commonly used which leverage relevant covariates to remove explainable variance in the outcomes. For count metrics, many variance reduction solutions have been developed in the literature based on classical linear adjustment (Yang and Tsiatis 2001; Freedman 2008; Lin 2013; Deng et al. 2013) and machine learning tools (Hosseini and Najmi 2019; Guo et al. 2021). However, the optimality of variance reduction procedures has not been thoroughly studied and using suboptimal procedures may lead to unnecessary costs. In addition to count metrics, online experiments also often involve ratio metrics (Deng, Lu, and Litz 2017) whose variance reduction is more complex but much less studied. As we will discuss in Section 2.1, ratio metrics are similar to the setting of cluster randomized experiments. Existing variance reduction solutions for ratio metric are mostly extensions to those of count metrics (Deng et al. 2013) and a rigorous statistical framework for (optimal) variance reduction remains absent.

In this article, we target at the natural but unanswered question: *With access to a set of covariates that are independent of the treatment in the experiment, what is the optimal estimator for comparing the outcomes of treatment and control groups?* We study the optimal variance reduction procedures for both count and ratio metrics that are ubiquitous in online controlled experiments in the industry. The optimality we focus on is semiparametric efficiency (Bickel et al. 1993). Given an estimand that arises from the comparison of experiment outcomes, our goal is to develop an estimator with the smallest asymptotic variance among all asymptotically unbiased estimators. We propose procedures that reduce the variance of treatment effect estimation by incorporating flexible ML regressors with rigorous statistical guarantee. Based on classical semiparametric statistics theory, we establish the optimality of our procedures under mild conditions. For ratio metrics, in addition to the optimal (and perhaps nonlinear) approaches, we also propose a computationally efficient linear adjustment method which, to the best of our knowledge, is not available in the literature.

The rest of the article is organized as follows. In Section 2, we introduce the definition of count and ratio metrics in online experiments and provide an overview of the related literature. We develop the variance reduction procedures, asymptotic inference and optimality properties for count metrics in Section 3, and for ratio metrics in Sections 4 and 5. Section 6 is devoted to simulation studies and Section 7 illustrates the

CONTACT Shan Ba *shanbatr@gmail.com* Data Science Applied Research, LinkedIn Corporation, Sunnyvale, CA.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/TECH*.

performance of our methods using real online experiments from LinkedIn. Finally, we give concluding remarks in Section 8.

2. Problem Setting and Related Work

The *randomization unit* and the *analysis unit* are two important concepts for online experiments. The most common online experiments are randomized by users, while sometimes the experiments can also use alternative randomization units. For example, in online experiments for enterprise products, the randomization unit typically needs to be a cluster of users (such as an enterprise account or contract) because users in the same contract must have access to the same product feature. In another scenario where it is infeasible to identify/track users in a web service, the randomization unit is often chosen to be a service request or a pageview. The analysis unit of an experiment, on the other hand, may not necessarily be the same as the experiment's randomization unit. For example, in a cluster randomized experiment for enterprise products, we can choose the analysis unit at either the cluster level (e.g., revenue per cluster) or the individual user level (e.g., revenue per user). Depending on whether the randomization and the analysis units coincide, metrics in online experiments can be classified into different types, for which different analysis procedures are needed.

2.1. Count and Ratio Metrics

We adopt the potential outcomes framework and follow the standard Stable Unit Treatment Value Assumption (Imbens and Rubin 2015, SUTVA) so that there is no interference among the randomization units. We take a super-population perspective where the randomization units can be viewed as iid, but the analysis units may not.

Count metrics. The most common metrics are count metrics, whose analysis unit matches the randomization unit in the experiment. For example, in online experiments that are randomized by users, count metrics are those defined on the user level such as revenue per user, pageviews per user, number of clicks per user, etc. Because the analysis units are the same as the the iid randomization units, the variance of count metrics can be estimated directly by the sample variance formula.

Formally, suppose there are n units in the experiment, for which we have access to iid observations $\{(X_i, Y_i, T_i)\}_{i=1}^n$ from an unknown distribution, where X_i is the pretreatment covariates, Y_i is the measured metric and T_i is the treatment indicator. Following the standard practice of online controlled experiments, we assume the treatment indicators $T_i \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$ for $p \in (0,1)$ and are independent of all other information. The units in the treated group ($T_i = 1$) receive the treatment, and the units in the control group ($T_i = 0$) do not. The outcomes $\{Y_i\}_{i=1}^n$ are then measured after the experiment. Under the potential outcome framework, each unit has two potential outcomes $(Y_i(1), Y_i(0))$, where $Y_i(1)$ is the outcome that unit i exhibits under treatment, and $Y_i(0)$ is that under control. For each unit, we only observe one potential outcome $Y_i = Y_i(T_i)$ under SUTVA (Imbens and Rubin 2015). Typically, count metrics are aggregated by sample means such as $(\sum_{T_i=1} Y_i)/(\sum_i T_i)$ for the units in the treatment or control groups, and the difference between two groups shows the causal effect of the treatment. The corresponding estimand is

$$\tau = \mathbb{E}\big[Y_i(1)\big] - \mathbb{E}\big[Y_i(0)\big],$$

where the expectations are taken with respect to the distribution that the units are from.

Ratio metrics. When the analysis unit is at a lower level than the experiments' randomization unit, the metric of interest is a ratio metric (because it can be expressed as a ratio of two count metrics), whose unique structure requires distinct techniques for inference.

In user-randomized experiments, the click-through rate (average number of clicks per pageview, computed as number of clicks /number of pageviews) is an example of ratio metric whose analysis unit is at the pageview level. Because the experiment is randomized by users and different pageviews of the same user are correlated, only the outcomes aggregated at the user level (randomization unit level) can be viewed as iid. For analysis purposes, the click-through rate can be equivalently viewed as a ratio of two user-level count metrics (number of clicks per user /number of pageviews per user). Consider another experiment for enterprise products which needs to be randomized by contracts to ensure that all users within each contract receive the same treatment assignment. Revenue per contract is a standard count metric in this experiment, but in practice we are often more interested in analyzing revenue per user, which is a ratio metric. Because only the contracts are iid and users under each contract are not independent, variance of the revenue per user cannot be directly calculated by the sample variance formula. Instead, by viewing revenue per user as a "ratio" of two contract-level count metrics (revenue per contract/number of users per contract), we can estimate its variance based on the delta method (Deng, Lu, and Litz 2017). More broadly, this setting is similar to the cluster randomized experiments in causal inference (Green and Vavreck 2008; Middleton and Aronow 2015) as each randomization unit can be viewed as a cluster of analysis units.

Formally, let $i = 1, \ldots, n$ be iid randomization units in the experiment from some distribution $\mathbb{P}$. They are randomly allocated to treated or control groups, indicated by $T_i \in \{0,1\}$. Here we assume $T_i \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$ for some $p \in (0,1)$ and are independent from all other information. Suppose two count metrics Y_i and Z_i are measured for each randomization unit i (e.g., Y_i is the total number of clicks of user i and Z_i is the total number of pageviews of user i). The ratio metrics (e.g., click through rate) for the treated and control groups are defined as

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

where Y_i and Z_i are aggregated across all randomization units in the treatment/control groups first before taking the ratio. Because randomization units are iid, the law of large numbers implies that the population-level comparison target (the limit of difference in ratio metrics) is *the difference in ratios of expectations* $\mathbb{E}[Y_i|T_i = 1]/\mathbb{E}[Z_i|T_i = 1] - \mathbb{E}[Y_i|T_i =$

$0]/\mathbb{E}[Z_i|T_i = 0]$. In practice, the expectation of Z_i is always nonzero and hence both the population-level and sample-level ratios are well defined. Note that we do not define ratio metrics for the treatment and control groups as $\sum_{T_i=1}(Y_i/Z_i)/n_t$ and $\sum_{T_i=0}(Y_i/Z_i)/n_c$, because some individual Z_i may be zero whence Y_i/Z_i is not well-defined. The previous definition of ratio metrics can actually be viewed as a weighted average of Y_i/Z_i with the weights proportional to Z_i (e.g., giving more weights to more active users):

$$\frac{\frac{1}{n_c}\sum_{i \text{ control}} Y_i}{\frac{1}{n_c}\sum_{i \text{ control}} Z_i} = \sum_{i \text{ control}} \underbrace{\frac{Z_i}{\sum_{i \text{ control}} Z_i}}_{\text{"weights"}} \times \underbrace{\frac{Y_i}{Z_i}}_{\text{ratio for randomization unit } i}.$$

Ratio metrics can also be grouped into two types depending on their assumptions: (1) The first type allows both the metric Z in the denominator and the metric Y in the numerator to be changed by the treatment. The above click-through rate example belongs to this type. (2) The second type assumes that the metric Z in the denominator is a random variable associated with the experiment unit, but is *not* changed by the treatment. Deng, Lu, and Litz (2017) refers to (2) as the *stable denominator assumption* (SDA in the following). SDA is plausible when the ratio metric mainly uses Z in the denominator as a normalization factor to standardize the changes in the numerator metric Y. For example, when we are interested in a user-level metric "revenue per user" but the experiment needs to be randomized by clusters of users, we can choose Z_i as the number of active users in an cluster. This assumption always needs to be checked in practice using a separate test on Z_i. If the SDA is violated, the ratio metric itself is hard to interpret: we do not know whether an increase in the ratio is good (e.g., due to an increase in revenue) or bad (e.g., due to a decrease in the number of active users). For instance, a bad treatment which decreases the number of active users in the contracts can instead yield a higher revenue-per-user ratio as the remaining users most likely are the most active ones. When the SDA is violated, the analysis should also emphasize on studying the count metric change in the numerator Y and in the denominator Z separately before drawing conclusions.

We now define the estimands for the two types of ratio metrics under the potential outcomes framework. For type (1) ratio metric, each unit has potential outcomes $(Y_i(1), Y_i(0), Z_i(1), Z_i(0))$, where we observe $(Y_i, Z_i) = (Y_i(T_i), Z_i(T_i))$. The estimand is

$$\delta = \frac{\mathbb{E}[Y_i(1)]}{\mathbb{E}[Z_i(1)]} - \frac{\mathbb{E}[Y_i(0)]}{\mathbb{E}[Z_i(0)]}, \tag{2.1}$$

where the expectation is with respect to the distribution the units are from. In the type (2) ratio metric, each unit has potential outcomes $(Y(1), Y(0))$ so that $Y_i = Y_i(T_i)$ while Z is a plain random variable, and the estimand is

$$\delta' = \frac{\mathbb{E}[Y_i(1)]}{\mathbb{E}[Z_i]} - \frac{\mathbb{E}[Y_i(0)]}{\mathbb{E}[Z_i]}.$$

In this work, we will consider both types of estimands for the ratio metrics and offer practical recommendations. As we will show in our numerical experiments, the assumption (2) should be made with caution.

2.2. Related Work

This work falls into the general randomized experiment setting in causal inference (Imbens and Rubin 2015), whereas we study the post-hoc variance reduction instead of the experimental design strategies. In addition, our method leverages machine learning models to learn the conditional relation between the potential outcomes and the covariates, which is similar to but with distinct goal from the investigation of treatment effect heterogeneity in causal inference (Imai and Ratkovic 2013; Athey and Imbens 2016; Chernozhukov et al. 2017; Wager and Athey 2018; Künzel et al. 2019; Nie and Wager 2020): we fit the conditional mean functions of the potential outcomes to remove explainable variations in the outcomes, rather than to investigate the heterogeneity of the conditional treatment effect.

This work adds to the rich literature of variance reduction with covariate adjustment in randomized experiments. Analysis of covariance (ANCOVA) has a long history of application in physical experiments (Wu and Hamada 2009). The classic linear regression adjustment (Yang and Tsiatis 2001; Freedman 2008; Lin 2013) is shown to work even when the linear model is misspecified. Deng et al. (2013) proposes to use pretreatment data as the regression covariates and the corresponding variance reduction method, called Controlled-experiment Using Pre-Experiment Data (CUPED), has been widely used in the industry. The blossom of ML research also inspires a line of recent work on using ML tools for variance reduction, including Hosseini and Najmi (2019), Cohen and Fogarty (2020), Guo et al. (2021) and the references therein, but few of them focus on the optimality of the variance reduction procedure. In particular, among the existing works for count metrics, both Guo et al. (2021) and Cohen and Fogarty (2020) uses predicted outcomes from ML estimators as covariates in a linear regression adjustment. Their intuitions are to create more relevant features with ML tools to improve upon CUPED, but not aimed at optimality. As would be discussed in details at the end of Section 3.3, careful considerations are needed for optimality and the method in Guo et al. (2021) can not be semiparametrically efficient in general. Hosseini and Najmi (2019) discusses ratio metrics and provides methods to obtain unbiased estimators with machine learning tools. However, to our knowledge, rigorous and explicit statistical inference guarantees and the optimality of the procedures are not provided.

Our approach for count metrics is asymptotically the same as the augmented inverse propensity weighting (AIPW) estimator (Robins, Rotnitzky, and Zhao 1994), whose well-established semiparametric efficiency (Hahn 1998) result forms the basis of our optimality guarantee. For count metrics, we develop valid inference procedures in randomized experiments (L_2 convergence in probability to any fixed function), which is much weaker than the pointwise convergence condition to true conditional mean functions as is often required in observational studies (Nichols 2007; Schuler and Rose 2017; Chernozhukov et al. 2018) or the investigation of heterogeneous treatment effects (Chernozhukov et al. 2017; Künzel et al. 2019; Athey and Wager 2019; Nie and Wager 2020; Kennedy 2020), and also differs from the traditional approach of Donsker conditions and empirical process theory (Andrews 1994; Van Der Vaart et al. 1996; Van der Vaart 2000) to control errors in estimating

nuisance components (the conditional mean functions in our setting).

Our estimators for ratio metrics as well as its inference and optimality results are new to the literature. Intuitively, they all have a fit-and-debias flavor related to the AIPW estimator (Robins, Rotnitzky, and Zhao 1994). The optimality theories we develop for ratio metrics finds roots in the classical semiparametric efficiency theory (Bickel et al. 1993; Hahn 1998).

3. Variance Reduction for Count Metrics

We begin our discussion with intuitions on how machine learning can assist variance reduction for count metrics. To estimate the treatment effect $\tau = \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)]$, an ideal "estimator" is $\widehat{\theta}_{\text{ideal}} = \frac{1}{n}\sum_{i=1}^{n}(Y_i(1) - Y_i(0))$ which directly compares pairs of potential outcomes, although it is not rigorously an estimator since $Y_i(1)$ and $Y_i(0)$ are not simultaneously observable. This "estimator" is ideal for two reasons: (a) large sample size: both the treated and the control group leverage n observations; (b) paired comparison: the variance of the individual treatment effect $Y_i(1) - Y_i(0)$ is typically smaller than the variances of $Y_i(1)$ and $Y_i(0)$ alone. A standard Difference-in-Mean (DiM) estimator compares the averages of the two groups: $\widehat{\theta}_{\text{DiM}} = \frac{1}{n_t}\sum_{i\text{ treated}} Y_i(1) - \frac{1}{n_c}\sum_{i\text{ control}} Y_i(0)$, which has larger variance than $\widehat{\theta}_{\text{ideal}}$ by Cauchy-Schwarz inequality. While $\widehat{\theta}_{\text{ideal}}$ is not computable, a natural idea is to impute the unobserved potential outcomes based on covariates X_i. Let $\widehat{\mu}_1(\cdot)$ and $\widehat{\mu}_0(\cdot)$ be some machine learning predictors for $Y(1)$ and $Y(0)$ based on X, respectively. By naively plugging in the predictors, we obtain $\widehat{\theta}_{\text{plug-in}} = \frac{1}{n}\sum_{i=1}^{n}\big(\widehat{Y}_i(1) - \widehat{Y}_i(0)\big)$, where

$$\widehat{Y}_i(1) = \begin{cases} Y_i(1), & T_i = 1 \\ \widehat{\mu}_1(X_i), & T_i = 0 \end{cases}, \qquad \widehat{Y}_i(0) = \begin{cases} Y_i(0), & T_i = 0 \\ \widehat{\mu}_0(X_i), & T_i = 1 \end{cases}.$$

The simplicity of $\widehat{\theta}_{\text{plug-in}}$ comes with two problems. First, without a well-posed parametric model, the convergence rate of $\widehat{\mu}_1(\cdot)$ and $\widehat{\mu}_0(\cdot)$ is often slower than $n^{-1/2}$, hence, $\widehat{\theta}_{\text{plug-in}}$ can introduce "regressor bias" that is even larger than the variance. Second, another "double-dipping bias" occurs as the same dataset is used both for model-fitting and for prediction, hence, $\{\widehat{\mu}_w(X_i)\}_{1\le i\le n}$ are no longer independent copies, posing challenges for statistical analysis. We will develop debiasing techniques for these issues.

We also note that the above intuition is related to Guo and Basse (2021) and Cohen and Fogarty (2020) where machine learning predictions are used to impute the counterfactuals. However, our work takes a specific approach that aims at optimality. The conditions for valid inference also differ from them.

3.1. Estimation Procedure

Our proposed variance reduction procedure improves upon the naive $\widehat{\theta}_{\text{plug-in}}$ in two aspects: it eliminates the "regressor bias" by adding a de-biasing term and employs the cross-fitting technique (Chernozhukov et al. 2018) to correct for the "double-dipping bias."

The first step is to randomly split the original dataset $\mathcal{D} = (Y_i, T_i, X_i)_{i=1}^n$ into K (roughly) equal-sized folds $\{\mathcal{D}^{(k)}\}_{1\le k\le K}$ with sizes $n_k = |\mathcal{D}^{(k)}|$, each fold containing $n_{k,t} = \sum_{i\in\mathcal{D}^{(k)}} T_i$ treated samples and $n_{k,c} = n_k - n_{k,t}$ control samples. Here the random splitting is conducted separately in the treated and control groups to achieve balanced numbers of treated samples across folds. In practice, $K = 2$ generally works well.

The second step is cross-fitting (Chernozhukov et al. 2018). Let $\mathcal{D}^{(-k)} = \mathcal{D}\backslash\mathcal{D}^{(k)}$ denote all data after holding out the kth fold. For each $k \in [K]$, fit a function $\widehat{\mu}_1^{(k)}(x)$ for $\mathbb{E}[Y(1)|X = x]$ using $\{(X_i, Y_i)\colon T_i = 1, i \in \mathcal{D}^{(-k)}\}$, and fit a function $\widehat{\mu}_0^{(k)}(x)$ for $\mathbb{E}[Y(0)|X = x]$ using $\{(X_i, Y_i)\colon T_i = 0, i \in \mathcal{D}^{(-k)}\}$. Then apply the fitted functions to the held-out samples in $\mathcal{D}^{(k)}$ to generate out-of-sample predictions $\widehat{\mu}_0(X_i) = \widehat{\mu}_0^{(k)}(X_i)$ and $\widehat{\mu}_1(X_i) = \widehat{\mu}_1^{(k)}(X_i)$ for $i \in \mathcal{D}^{(k)}$. These out-of-sample predictions are independent conditional on $\mathcal{D}^{(-k)}$, helping alleviate the "double-dipping bias."

Finally, we estimate τ via

$$\begin{aligned}\widehat{\theta}_{\text{Debias}} &= \frac{1}{K}\sum_{k=1}^{K}\widehat{\theta}_{\text{Debias}}^{(k)}, \quad \text{where} \\ \widehat{\theta}_{\text{Debias}}^{(k)} &= \frac{1}{n_k}\sum_{i\in\mathcal{D}^{(k)}}\big(\widehat{\mu}_1(X_i) - \widehat{\mu}_0(X_i)\big) \\ &\quad + \frac{1}{n_{k,t}}\sum_{\substack{T_i=1,\\ i\in\mathcal{D}^{(k)}}}\big(Y_i - \widehat{\mu}_1(X_i)\big) - \frac{1}{n_{k,c}}\sum_{\substack{T_i=0,\\ i\in\mathcal{D}^{(k)}}}\big(Y_i - \widehat{\mu}_0(X_i)\big).\end{aligned} \tag{3.1}$$

As will be shown later, this estimator is finite-sample unbiased and one could also obtain similar performance by the following estimator

$$\begin{aligned}\widetilde{\theta}_{\text{Debias}} &= \frac{1}{n}\sum_{i=1}^{n}\big(\widehat{\mu}_1(X_i) - \widehat{\mu}_0(X_i)\big) + \frac{1}{n_t}\sum_{i\text{ treated}}\big(Y_i(1) - \widehat{\mu}_1(X_i)\big) \\ &\quad - \frac{1}{n_c}\sum_{i\text{ control}}\big(Y_i(0) - \widehat{\mu}_0(X_i)\big),\end{aligned} \tag{3.2}$$

which might have an $O(1/n)$ bias from unequal fold sizes. Note that $\widetilde{\theta}_{\text{Debias}}$ adds a de-biasing term to the naive plug-in estimator

$$\begin{aligned}\widetilde{\theta}_{\text{Debias}} \approx \widehat{\theta}_{\text{plug-in}} &+ \frac{n_c}{n\cdot n_t}\sum_{i\text{ treated}}\big(Y_i(1) - \widehat{\mu}_1(X_i)\big) \\ &- \frac{n_t}{n\cdot n_c}\sum_{i\text{ control}}\big(Y_i(0) - \widehat{\mu}_0(X_i)\big).\end{aligned}$$

In this equation, $\widehat{\theta}_{\text{plug-in}}$ uses machine learning tools to impute the unobserved potential outcomes, and the remaining two terms serve as corrections to the "regressor bias" in fitting the mean functions. As will be seen in Sections 4 and 5, the debiasing technique is also useful when developing estimators for ratio metrics with careful considerations on the imputation of the predicted values. The proposed procedure is summarized in Algorithm 1.

Algorithm 1 Debiased Variance Reduction

1: Input: Dataset $\mathcal{D} = \{(Y_i, X_i, T_i\}_{i=1}^n$, number of folds K.
2: Randomly split $\mathcal{D}$ into K folds $\mathcal{D}^{(k)}$, $k = 1, \ldots, K$.
3: **for** $k = 1, \ldots, K$ **do**
4: Use all (X_i, Y_i) with $T_i = 1$ and $i \notin \mathcal{D}^{(k)}$ to obtain $\widehat{\mu}_1^{(k)}(x)$ for $\mathbb{E}[Y(1)|X = x]$;
5: Use all (X_i, Y_i) with $T_i = 0$ and $i \notin \mathcal{D}^{(k)}$ to obtain $\widehat{\mu}_0^{(k)}(x)$ for $\mathbb{E}[Y(0)|X = x]$;
6: For all $i \in \mathcal{D}^{(k)}$, compute $\widehat{\mu}_1(X_i) = \widehat{\mu}_1^{(k)}(X_i)$ and $\widehat{\mu}_0(X_i) = \widehat{\mu}_0^{(k)}(X_i)$.
7: **end for**
8: Compute estimator $\widehat{\theta}_{\text{Debias}}$ or $\widetilde{\theta}_{\text{Debias}}$ on $\mathcal{D}$ according to (3.1) or (3.2).

Ignoring the nuisance in estimation and assuming $\widehat{\mu}_w(x) \to \mu_w(x) := \mathbb{E}[Y(w)|X = x]$ for $w \in \{0, 1\}$, one can show that the asymptotic variance of $\widetilde{\theta}_{\text{Debias}}$ and $\widehat{\theta}_{\text{Debias}}$ satisfies

$$\text{var}(\widetilde{\theta}_{\text{Debias}}) \approx \underbrace{\frac{1}{n}\text{var}\big(\mu_1(X) - \mu_0(X)\big)}_{\text{(i) predictable part}} + \underbrace{\frac{1}{n_t}\text{var}\big(Y(1) - \mu_1(X)\big) + \frac{1}{n_c}\text{var}\big(Y(0) - \mu_0(X)\big)}_{\text{(ii) irreducible variance}}$$

In the above decomposition, the (i) predictable part echoes our intuitions of imputing $Y_i(1) - Y_i(0)$ with predicted values, thereby inheriting the advantages of $\widehat{\theta}_{\text{ideal}}$ of increased sample size and decreased single-term variance when the treatment effects are typically small. In the meantime, (i) is the variance of the projection of $Y_i(1) - Y_i(0)$ on the X-space, which is the best effort to predict $Y_i(1) - Y_i(0)$ with X_i. The (ii) is the variance that cannot be eliminated by the information of X embodied in the n_t treated samples and n_c control samples. In other words, one might view $\widetilde{\theta}_{\text{Debias}}$ as the best efforts toward exploiting the information in X. As the variance in (i) is deflated by a larger factor n, predictors with smaller variance in (ii), that is, more accurate ML predictors $\mu_w(x)$, are desirable.

Remark 3.1. For the case of count metrics, our estimator is asymptotically equivalent to the AIPW estimator (Robins, Rotnitzky, and Zhao 1994), whose semiparametric efficiency (Hahn 1998) directly implies the optimality of our estimator. As will be shown in Sections 3.2 and 3.3, the differences here include (a) we establish valid inference as long as the ML estimator converges to deterministic functions without consistency, and (b) we establish the optimality under a weaker condition of L_2 convergence in probability. Our debiasing term is also related to the "prediction unbiasedness" condition in Guo and Basse (2021).

When we restrict $\widehat{\mu}_w$ in Algorithm 1 to be a linear function of x, our estimator is connected to the well-known CUPED estimator (Deng et al. 2013)

$$\widehat{\theta}_{\text{CUPED}} = \frac{1}{n_t}\sum_{i \text{ treated}}\big(Y_i(1) - \widehat{\theta}(X_i - \bar{X})\big) - \frac{1}{n_c}\sum_{i \text{ control}}\big(Y_i(0) - \widehat{\theta}(X_i - \bar{X})\big), \quad (3.3)$$

where $\widehat{\theta}$ is the OLS projection of Y_i (the pooled outcomes of control and treated groups) on (centered) pretreatment metric X_i. The connection can be easily established as follows. By replacing $\mu_w(x)$ with the linear predictor $\beta_w^\top x$ ($w = 0$ or 1) where β_w is the least-square linear coefficient of $Y(w)$ on X, the proposed optimal estimator $\widetilde{\theta}_{\text{Debias}}$ in (3.2) can be simplified as

$$\widetilde{\theta}_{\text{Debias}} = \frac{1}{n_t}\sum_{i \text{ treated}}\big(Y_i(1) - \beta_1^\top(X_i - \bar{X})\big) - \frac{1}{n_c}\sum_{i \text{ control}}\big(Y_i(0) - \beta_0^\top(X_i - \bar{X})\big).$$

This improves upon CUPED by running separate regressions in the treated and control groups. This estimator enjoys the agnostic property (Lin 2013): it leads to no larger variance than the diff-in-mean estimator without any assumptions, while the vanilla CUPED in (3.3) may not. In such low-dimensional case, due to the low complexity of linear function classes, the double-dipping bias is not a concern and cross-fitting is not necessary. But for high-dimensional linear regression such as the LASSO, it is important to use our proposed algorithm to cross-fit and debias with $\widehat{\mu}$ being the LASSO estimator, otherwise the bias induced by the high dimensional regression could be of the same order as the variance.

3.2. Unbiasedness and Asymptotic Inference

In this part, we establish the finite-sample unbiasedness and asymptotic inference for our procedures for count metrics. In the following theorem, we show that $\widehat{\theta}_{\text{Debias}}$ is unbiased and $\widetilde{\theta}_{\text{Debias}}$ has a negligible $O(1/n)$ bias due to potentially unequally-sized folds. Such finite-sample unbiasedness holds for any machine learning regressors.

Theorem 3.2 (Finite-sample unbiasedness). Suppose $(X_i, Y_i(0), Y_i(1)) \overset{\text{iid}}{\sim} \mathbb{P}$ are independent of the treatment assignments $\mathcal{T} = \{T_i\}_{i=1}^n$. Then $\mathbb{E}[\widehat{\theta}_{\text{Debias}}|\mathcal{T}] = \tau$ for any n. Furthermore, suppose $\big|\mathbb{E}[\widehat{\mu}_w^{(k)}(X_i)|\mathcal{D}^{(-k)}]\big|, \big|\mathbb{E}[Y_i(w)]\big| \le c_0$ for all $w \in \{0, 1\}$ for some absolute constant $c_0 > 0$. Then $\big|\mathbb{E}[\widetilde{\theta}_{\text{Debias}}|\mathcal{T}] - \tau\big| \le c/\min\{n_t, n_c\}$ for some absolute constant $c > 0$.

We now study the asymptotic inference for our estimators. We assume the convergence of $\widehat{\mu}_1, \widehat{\mu}_0$ to fixed functions (this is mild because we do not require it to converge to the true conditional mean functions) and the treatment assignment mechanism.

Assumption 3.3 (Convergence). There exists two fixed functions $\mu_1^*(\cdot)$ and $\mu_0^*(\cdot)$, so that $\|\widehat{\mu}_1^{(k)} - \mu_1^*\|_2 \xrightarrow{P} 0$ and $\|\widehat{\mu}_0^{(k)} - \mu_0^*\|_2 \xrightarrow{P} 0$ for all $k \in [K]$.

Assumption 3.4 (Treatment assignment mechanism). Assume $n_t/n \xrightarrow{P} p$ for some fixed $p \in (0, 1)$, so that $n_{k,t}/n_k \xrightarrow{P} p$ for all $k \in [K]$.

Theorem 3.5 (Asymptotic confidence intervals). Suppose Assumptions 3.3 and 3.4 hold. Then $\sqrt{n}(\widehat{\theta}_{\text{Debias}} - \tau) \xrightarrow{d} N(0, \sigma^2_{\text{Debias}})$

and $\sqrt{n}(\widetilde{\theta}_{\text{Debias}} - \tau) \xrightarrow{d} N(0, \sigma^2_{\text{Debias}})$, where $\sigma^2_{\text{Debias}} = \frac{1}{p}\text{var}\big(Y_i(1) - (1-p)\mu_1^*(X_i) - p\mu_0^*(X_i)\big) + \frac{1}{1-p}\text{var}\big(Y_i(0) - (1-p)\mu_1^*(X_i) - p\mu_0^*(X_i)\big)$. Furthermore, define the variance estimator $\widehat{\sigma}^2_{\text{Debias}} = \frac{n}{n_t^2}\sum_{i=1}^n T_i(A_i - \bar{A})^2 + \frac{n}{n_c^2}\sum_{i=1}^n (1-T_i)(B_i - \bar{B})^2$, where $\bar{A} = \frac{1}{n_t}\sum_{i=1}^n T_i A_i$, $\bar{B} = \frac{1}{n_c}\sum_{i=1}^n (1-T_i)B_i$, and $A_i = Y_i(1) - \frac{n_c}{n}\widehat{\mu}_1(X_i) - \frac{n_t}{n}\widehat{\mu}_0(X_i)$, $B_i = Y_i(0) - \frac{n_c}{n}\widehat{\mu}_1(X_i) - \frac{n_t}{n}\widehat{\mu}_0(X_i)$. Then $\widehat{\sigma}^2_{\text{Debias}} \xrightarrow{P} \sigma^2_{\text{Debias}}$, and $\widehat{\theta}_{\text{Debias}} \pm z_{1-\alpha/2}\widehat{\sigma}_{\text{Debias}}/\sqrt{n}$ and $\widetilde{\theta}_{\text{Debias}} \pm z_{1-\alpha/2}\widehat{\sigma}_{\text{Debias}}/\sqrt{n}$ are both asymptotically valid $(1-\alpha)$ confidence intervals for $\tau = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$.

3.3. Optimality: Semiparametric Efficiency

Returning to our motivating question, we establish the optimality of our procedure, building upon semiparametric statistics theory (Bickel et al. 1993; Robins, Rotnitzky, and Zhao 1994; Hahn 1998). We are to show that the asymptotic variance σ^2_{Debias} is no larger than any "regular" estimators (roughly speaking, those asymptotically linear ones with the form $\frac{1}{n}\sum_{i=1}^n \phi(X_i, Y_i, T_i) + o_P(1/\sqrt{n})$ for some function ϕ, including the ones obtained from linear regression as in Deng et al. 2013; Guo et al. 2021; Cohen and Fogarty 2020). Due to the limit of article length, we omit the formal backgrounds on semiparametric efficiency in the main text and defer the discussions to Section 4 of the supplementary materials, which includes notions of regular nonparametric space and efficient influence functions.

We impose the following condition on the consistency of the estimators $\widehat{\mu}_1(\cdot)$, $\widehat{\mu}_0(\cdot)$.

Assumption 3.6 (Consistency). Let $\mu_1(x) = \mathbb{E}[Y(1)|X = x]$ and $\mu_0(x) = \mathbb{E}[Y(0)|X = x]$ be the two true mean functions. Suppose $\|\widehat{\mu}_w^{(k)} - \mu_w\|_2 \xrightarrow{P} 0$ for $w \in \{0,1\}$ and all $k \in [K]$.

Assumption 3.6 is a mild condition without any requirement of the convergence rates. This is in contrast to the conditions on convergence rates (though doubly-robust properties help relax them to lower orders) for observational studies (Nichols 2007; Schuler and Rose 2017; Chernozhukov et al. 2018) where the propensity score function $e(x) = \mathbb{P}(T = 1|X = x)$ also needs to be estimated. Also, we do not need the convergence to be pointwise.

Theorem 3.7 (Semiparametric efficiency). Suppose Assumptions 3.4 and 3.6 hold. Then the asymptotic variance of $\widehat{\theta}_{\text{Debias}}$ and $\widetilde{\theta}_{\text{Debias}}$ in Theorem 3.5 is the semiparametric variance bound for $\tau = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$.

Theorem 3.7, together with Theorem 3.5, indicates that the covariates X_i should be chosen to be powerful predictors for the outcomes, such that $\text{Var}(Y(w) - \mu_w(X))$ is relatively small for $w \in \{0,1\}$. Since our estimator is finite-sample unbiased conditional on $\mathcal{T}$, if we view $\mathcal{T}$ as fixed, it has the smallest variance among all asymptotically unbiased regular estimators. This result might be of interest for two-sample mean test and completely randomized experiments as well; in the latter case, units in the treated and control groups still have iid potential outcomes if we assume the units are iid before treatment assignment.

We take a moment here to compare Theorem 3.7 to other machine-learning empowered methods in the literature (Hosseini and Najmi 2019; Cohen and Fogarty 2020; Guo et al. 2021). As indicated by (3.1) and (3.2), our estimator is essentially a linear combination of Y_i, $\widehat{\mu}_1(X_i)$ and $\widehat{\mu}_0(X_i)$ that achieves smallest variance. Thus, it can be obtained from a linear regression when Assumption 3.6 holds. Referring to (3.1) and (3.2), for optimality, one needs to include both $\widehat{\mu}_1(X_i)$ and $\widehat{\mu}_0(X_i)$ in the regression terms and run the regression separately on treated and control groups. Therefore, although machine learning helps to exploit nonlinearity, the regression in Guo et al. (2021) with one single predictor for Y_i (not separately for $Y_i(1)$ and $Y_i(0)$) would not achieve optimality in general, unless $\mu_1(x)$ and $\mu_0(x)$ are completely colinear. This issue is the same for Hosseini and Najmi (2019) where only one predictor is used. The method in Cohen and Fogarty (2020) may be optimal, which, however, requires a slightly stronger condition of L_4-distance convergence.

4. Variance Reduction for Ratio Metrics: Without SDA

We now study variance reduction procedures for ratio metrics introduced in Section 2.1 whose denominator $Z_i = Z_i(T_i)$ can be changed by the treatment (without SDA). The results are new to the literature while sharing similar ideas to our results for count metrics.

4.1. Estimation Procedure

The first step is the K-fold sample splitting for $\mathcal{D} = (Y_i, Z_i, T_i, X_i)_{i=1}^n$ as introduced in Section 3.1. The second step, cross-fitting, needs careful consideration: for each $k \in [K]$, we use the data $\big\{(X_i, Z_i, Y_i)\colon T_i = 1, i \in \mathcal{D}^{(-k)}\big\}$ to obtain estimators $\widehat{\mu}_1^{Y,(k)}(x)$ for $\mathbb{E}[Y(1)|X = x]$ and $\widehat{\mu}_1^{Z,(k)}(x)$ for $\mathbb{E}[Z(1)|X = x]$. Likewise, we use $\big\{(X_i, Z_i, Y_i)\colon T_i = 0, i \in \mathcal{D}^{(-k)}\big\}$ to obtain $\widehat{\mu}_0^{Y,(k)}(x)$ and $\widehat{\mu}_0^{Z,(k)}(x)$. Then, we calculate predictions $\widehat{\mu}_w^Y(X_i) = \widehat{\mu}_w^{Y,(k)}(X_i)$ and $\widehat{\mu}_w^Z(X_i) = \widehat{\mu}_w^{Z,(k)}(X_i)$ for all $i \in \mathcal{D}^{(k)}$, $w \in \{0,1\}$. Finally, we estimate δ in (2.1) by

$$\widehat{\delta} = \frac{\sum_{i=1}^n A_i}{\sum_{i=1}^n B_i} - \frac{\sum_{i=1}^n C_i}{\sum_{i=1}^n D_i}, \tag{4.1}$$

where $A_i = \widehat{\mu}_1^Y(X_i) + T_i\big(Y_i - \widehat{\mu}_1^Y(X_i)\big)/\widehat{p}$, $B_i = \widehat{\mu}_1^Z(X_i) + T_i\big(Z_i - \widehat{\mu}_1^Z(X_i)\big)/\widehat{p}$, $C_i = \widehat{\mu}_0^Y(X_i) + (1-T_i)\big(Y_i - \widehat{\mu}_0^Y(X_i)\big)/(1-\widehat{p})$, and $D_i = \widehat{\mu}_0^Z(X_i) + (1-T_i)\big(Z_i - \widehat{\mu}_0^Z(X_i)\big)/(1-\widehat{p})$ for $\widehat{p} = n_t/n$. The procedure is summarized in Algorithm 2. Compared to $\widehat{\delta}_{\text{DiM}} := \sum_{T_i=1} Y_i/\sum_{T_i=1} Z_i - \sum_{T_i=0} Y_i/\sum_{T_i=0} Z_i$, our estimator substitutes the sample means of the treated and control groups with the average fit-and-debias predictions, which is similar to our estimator for count metrics.

4.2. Asymptotic Inference

The analysis of ratio metrics is naturally asymptotic (Deng, Lu, and Litz 2017), and thus we focus more on the asymptotic properties of the proposed estimator.

We impose the following conditions on the treatment assignment mechanism and convergence of cross-fitted functions. In

Algorithm 2 Debiased Variance Reduction for Ratio Metric without SDA

1: Input: Dataset $\mathcal{D} = \{(Y_i, X_i, Z_i, T_i)\}_{i=1}^n$, number of folds K.
2: Randomly split $\mathcal{D}$ into K folds $\mathcal{D}^{(k)}$, $k = 1, \ldots, K$.
3: **for** $k = 1, \ldots, K$ **do**
4: Use all (X_i, Z_i, Y_i) with $T_i = 1$ and $i \notin \mathcal{D}^{(k)}$ to obtain $\widehat{\mu}_1^{Y,(k)}(x)$ and $\widehat{\mu}_1^{Z,(k)}(x)$;
5: Use all (X_i, Z_i, Y_i) with $T_i = 0$ and $i \notin \mathcal{D}^{(k)}$ to obtain $\widehat{\mu}_0^{Y,(k)}(x)$ and $\widehat{\mu}_0^{Z,(k)}(x)$;
6: Compute $\widehat{\mu}_w^{Y}(X_i) = \widehat{\mu}_w^{Y,(k)}(X_i)$ and $\widehat{\mu}_w^{Z}(X_i) = \widehat{\mu}_w^{Z,(k)}(X_i)$ for all $i \in \mathcal{D}^{(k)}$ and $w \in \{0, 1\}$.
7: **end for**
8: Compute estimator $\widehat{\delta}$ on $\mathcal{D}$ according to (4.1).

Assumption 4.2, we only require the convergence of estimated functions to deterministic functions, not the true conditional mean functions. This is a mild condition that holds for general machine learning regression methods.

Assumption 4.1 (Data Generating Process). $(X_i, Z_i(0), Z_i(1), Y_i(0), Y_i(1)) \overset{\text{iid}}{\sim} \mathbb{P}$ and the treatment assignments $T_i \overset{\text{iid}}{\sim}$ Bernoulli(p) are independent of all other random variables.

Assumption 4.2 (Convergence). There exists some fixed functions $\mu_{1,Y}^*(\cdot), \mu_{0,Y}^*(\cdot), \mu_{1,Z}^*(\cdot), \mu_{0,Z}^*(\cdot)$, so that both $\|\widehat{\mu}_w^{Y,(k)} - \mu_{w,Y}^*\|_2, \|\widehat{\mu}_w^{Z,(k)} - \mu_{w,Z}^*\|_2 \xrightarrow{P} 0$ for $w \in \{0, 1\}$.

In preparation for inferential guarantees, we define the influence function

$$\phi_\delta(Y_i, Z_i, X_i, T_i) = \frac{A_i^*}{\mathbb{E}[Z_i(1)]} - \frac{\mathbb{E}[Y_i(1)]}{\mathbb{E}[Z_i(1)]^2} B_i^* - \frac{C_i^*}{\mathbb{E}[Z_i(0)]} + \frac{\mathbb{E}[Y_i(0)]}{\mathbb{E}[Z_i(0)]^2} D_i^*, \quad (4.2)$$

where $A_i^* = \widetilde{\mu}_{1,Y}^*(X_i) + \frac{T_i}{p}\big(Y_i - \mathbb{E}[Y_i(1)] - \widetilde{\mu}_{1,Y}^*(X_i)\big)$, $B_i^* = \widetilde{\mu}_{1,Z}^*(X_i) + \frac{T_i}{p}\big(Z_i - \mathbb{E}[Z_i(1)] - \widetilde{\mu}_{1,Z}^*(X_i)\big)$, $C_i^* = \widetilde{\mu}_{0,Y}^*(X_i) + \frac{1-T_i}{1-p}\big(Y_i - \mathbb{E}[Y_i(0)] - \widetilde{\mu}_{0,Y}^*(X_i)\big)$, and $D_i^* = \widetilde{\mu}_{0,Z}^*(X_i) + \frac{1-T_i}{1-p}\big(Z_i - \mathbb{E}[Z_i(0)] - \widetilde{\mu}_{0,Z}^*(X_i)\big)$. Here $\widetilde{\mu}_{w,Y}^*(X_i) = \mu_{w,Y}^*(X_i) - \mathbb{E}[\mu_{w,Y}^*(X_i)]$ and $\widetilde{\mu}_{w,Z}^*(X_i) = \mu_{w,Z}^*(X_i) - \mathbb{E}[\mu_{w,Z}^*(X_i)]$, $w \in \{0, 1\}$ are the centered limiting functions. The following theorem establishes the asymptotic confidence intervals, whose proof is in the supplementary materials.

Theorem 4.3. Suppose Assumptions 4.1 and 4.2 hold, and let $\widehat{\delta}$ be the output of Algorithm 2. Then $\sqrt{n}(\widehat{\delta} - \delta) \xrightarrow{d} N(0, \sigma_\delta^2)$, where $\sigma_\delta^2 = \text{var}(\phi_\delta(Y_i, Z_i, X_i, T_i))$. Moreover, define the variance estimator $\widehat{\sigma}_\delta^2 = \frac{1}{n}\sum_{T_i=1}\big(d_{1,i} - \bar{d}_1\big)^2 + \frac{1}{n}\sum_{T_i=0}\big(d_{0,i} - \bar{d}_0\big)^2$, where

$$d_{1,i} = -\frac{n_c \widehat{\mu}_1^Y(X_i)}{n_t \bar{Z}(1)} + \frac{nY_i}{n_t \bar{Z}(1)} + \frac{n_c}{n_t}\frac{\bar{Y}(1)}{\bar{Z}(1)^2}\widehat{\mu}_1^Z(X_i) - \frac{n}{n_t}\frac{\bar{Y}(1)}{\bar{Z}(1)^2}Z_i - \frac{\widehat{\mu}_0^Y(X_i)}{\bar{Z}(0)} + \frac{\bar{Y}(0)}{\bar{Z}(0)^2}\widehat{\mu}_0^Z(X_i),$$

$$d_{0,i} = \frac{\widehat{\mu}_1^Y(X_i)}{\bar{Z}(1)} - \frac{\bar{Y}(1)}{\bar{Z}(1)^2}\widehat{\mu}_1^Z(X_i) - \frac{nY_i}{n_c \bar{Z}(0)} + \frac{n_t \widehat{\mu}_0^Y(X_i)}{n_c \bar{Z}(0)} + \frac{n}{n_c}\frac{\bar{Y}(0)}{\bar{Z}(0)^2}Z_i - \frac{n_t}{n_c}\frac{\bar{Y}(0)}{\bar{Z}(0)^2}\widehat{\mu}_0^Z(X_i).$$

Then $\widehat{\delta} \pm \widehat{\sigma}_\delta \cdot z_{1-\alpha/2}/\sqrt{n}$ is an asymptotically valid $(1-\alpha)$ confidence interval for δ.

4.3. Optimality

Our estimator is optimal (semiparametric efficient) when the estimated functions are consistent. We begin with mild convergence assumptions.

Assumption 4.4. $\|\widehat{\mu}_w^{Y,(k)} - \mu_{w,Y}\|_2 \xrightarrow{P} 0$ and $\|\widehat{\mu}_0^{Z,(k)} - \mu_{w,Z}\|_2 \xrightarrow{P} 0$ for all $k \in [K]$ and $w \in \{0, 1\}$, where $\mu_{w,Y}(x) = \mathbb{E}[Y(w)|X = x]$ and $\mu_{w,Z}(x) = \mathbb{E}[Z(w)|X = x]$.

We are to show that the efficient influence function for the estimation of δ is given by

$$\phi_\delta^\dagger(Y_i, Z_i, X_i, T_i) = \frac{A_i^\dagger}{\mathbb{E}[Z_i(1)]} - \frac{\mathbb{E}[Y_i(1)]}{\mathbb{E}[Z_i(1)]^2} B_i^\dagger - \frac{C_i^\dagger}{\mathbb{E}[Z_i(0)]} + \frac{\mathbb{E}[Y_i(0)]}{\mathbb{E}[Z_i(0)]^2} D_i^\dagger,$$

where $A_i^\dagger = \mu_{1,Y}(X_i) + \frac{T_i}{p}\big(Y_i - \mu_{1,Y}(X_i)\big) - \mathbb{E}[Y_i(1)]$, $B_i^\dagger = \mu_{1,Z}(X_i) + \frac{T_i}{p}\big(Z_i - \mu_{1,Z}(X_i)\big) - \mathbb{E}[Z_i(1)]$, $C_i^\dagger = \mu_{0,Y}^*(X_i) + \frac{1-T_i}{1-p}\big(Y_i - \mu_{0,Y}(X_i)\big) - \mathbb{E}[Y_i(0)]$, $D_i^\dagger = \mu_{0,Z}^*(X_i) + \frac{1-T_i}{1-p}\big(Z_i - \mu_{0,Z}(X_i)\big) - \mathbb{E}[Z_i(0)]$. The following theorem shows that under consistency, the asymptotic variance of $\widehat{\delta}$ coincides with the variance of $\phi_{\delta,\dagger}$, which is also the efficient variance bound. Thus, the optimality of the proposed procedure is established under appropriate conditions.

Theorem 4.5. Suppose Assumptions 4.1 and 4.4 hold. Then it holds that $\sqrt{n}(\widehat{\delta} - \delta) \xrightarrow{d} N(0, \sigma_{\delta,\dagger}^2)$, where $\sigma_{\delta,\dagger}^2 = \text{Var}\big(\phi_{\delta,\dagger}(Y_i, Z_i, X_i, T_i)\big)$ (see, (4.2)) is the semiparametric asymptotic variance bound for $\delta = \mathbb{E}[Y(1)]/\mathbb{E}[Z(1)] - \mathbb{E}[Y(0)]/\mathbb{E}[Z(0)]$.

Theorem 4.5 is based on a more general result (Theorem 4.7 in the supplementary materials) in situations where the treatment assignment might depend on X. It may be of independent interest for efficient estimation of ratio metrics in stratified experiments and observational studies.

4.4. Special Case: Optimal Linear Adjustment for Ratio Metrics

As a special case of Algorithm 2, we derive an optimal *linear adjustment* method for ratio metrics, which is computationally efficient and has several advantages over the existing approaches in the literature.

We suppose $X_i \in \mathbb{R}^p$ for some fixed p (with intercept). In Algorithm 2, let $\widehat{\mu}_w^{Y,(k)}(x) = \widehat{\beta}_{Y,w,(k)}^\top x$ for all $w \in \{0, 1\}$ and $k = 1, \ldots, K$, where $\widehat{\beta}_{Y,w,(k)}$ is the OLS coefficient of $\{Y_i\colon T_i = w, i \in$

$\mathcal{D}^{(-k)}\}$ on $\{X_i\colon T_i = w, i \in \mathcal{D}^{(-k)}\}$. The fitted function $\widehat{\mu}_w^{Z,(k)}(x)$ can be similarly obtained. As we have discussed at the end of Section 3.1, in this fixed-p setting, because of the low complexity of linear function classes, the cross-fitting step may not be needed and we can further simplify the approach by letting $\widehat{\mu}_w^Y(x) = \widehat{\beta}_{Y,w}^\top x$ in Algorithm 2, where $\widehat{\beta}_{Y,w}$ is the empirical OLS coefficient of $\{Y_i\colon T_i = w\}$ on $\{X_i\colon T_i = w\}$ for $w \in \{0,1\}$, that is, running linear regressions separately on treated and control groups without sample splitting. The imputations $\widehat{\mu}_w^Z(x)$ can be similarly obtained. One can show that the two estimators with or without sample splitting are asymptotically equivalent due to the convergence property of linear regression coefficients. Both estimators admit the asymptotic linear expansion (up to additive constants)

$$\frac{1}{n}\sum_{i=1}^{n}\Big(\alpha_{1,*}^\top X_i - \alpha_{0,*}^\top X_i + \frac{T_i}{p}\big(\Gamma_i - \alpha_{1,*}^\top X_i\big) - \frac{1-T_i}{1-p}\big(\Gamma_i - \alpha_{0,*}^\top X_i\big)\Big) + o_P(1/\sqrt{n}),$$

where $\Gamma_i = T_iY_i/\mathbb{E}[Z(1)] - T_iZ_i\mathbb{E}[Y(1)]/\mathbb{E}[Z(1)]^2 - (1-T_i)Y_i/\mathbb{E}[Z(0)] + (1-T_i)Z_i\mathbb{E}[Y(0)]/\mathbb{E}[Z(0)]^2$, and $\alpha_{1,*}$, $\alpha_{0,*}$ are the population OLS coefficients of Γ_i on X_i in treated and control groups. Because of the "agnostic" property of linear adjustment (Lin 2013), the variance of our estimator is always no larger than that of the diff-in-mean estimator. Moreover, our estimator achieves semiparametric efficiency if the actual conditional means are linear. Our approach also improves the ratio-metric extention of CUPED (Deng et al. 2013):

$$\frac{\sum_{T_i=1} Y_i}{\sum_{T_i=1} Z_i} - \frac{\sum_{T_i=0} Y_i}{\sum_{T_i=0} Z_i} - \widehat{\theta}\cdot\left(\frac{\sum_{T_i=1}\widetilde{Y}_i}{\sum_{T_i=1}\widetilde{Z}_i} - \frac{\sum_{T_i=0}\widetilde{Y}_i}{\sum_{T_i=0}\widetilde{Z}_i}\right) \tag{4.3}$$

for some $\widehat{\theta} \in \mathbb{R}$, where $\widetilde{Y}_i$ and $\widetilde{Z}_i$ are pretreatment versions of Y_i and Z_i.

(i) First, the estimator (4.3) only uses pretreatment metrics as covariates, while our method incorporates arbitrary covariates, hence, more flexible and powerful.

(ii) Second, our method always reduces variance compared to the diff-in-mean estimator because it runs separate regressions in the treated and control groups. A single $\widehat{\theta}$ in (4.3) is not guaranteed to reduce variance in some adversarial settings (Lin 2013).

(iii) Third, the linear expansion of the estimator in (4.3) (assuming $\widehat{\theta} \xrightarrow{P} \theta$) is $\frac{1}{n}\sum_{i=1}^n\big[\frac{T_i}{p}\big(\Gamma_i - \theta\widetilde{\Gamma}_i\big) - \frac{1-T_i}{1-p}\big(\Gamma_i - \theta\widetilde{\Gamma}_i\big)\big] + o_P(1/\sqrt{n})$, where $\widetilde{\Gamma}_i$ is similarly defined as Γ_i with Y_i, Z_i replaced by $\widetilde{Y}_i, \widetilde{Z}_i$, and θ is the OLS coefficient of Γ_i on $\widetilde{\Gamma}_i$. Our estimator achieves the OLS projection of Γ_i on the whole linear space of X_i, while (4.3) only projects on the linear space of $\widetilde{\Gamma}_i$, a subspace of X_i when X_i contains $\widetilde{Y}_i, \widetilde{Z}_i$. Therefore, our estimator achieves more reduction of variance.

5. Variance Reduction for Ratio Metrics: with SDA

In this section, we consider the second type of ratio metrics introduced in Section 2.1 whose denominator Z is assumed to be stable (SDA). Due to the page limit, we only describe the procedure here in Algorithm 3 for the reference of practitioners. Inference and optimality guarantees can be found in the supplementary materials. We adopt the similar plug-in-and-debias idea as in Sections 3 and 4, while here we pool all samples to estimate $\mathbb{E}[Z]$, and fit the conditional mean functions for Y based on (X, Z) separately in two groups.

Algorithm 3 Debiased Variance Reduction for Ratio Metric with Stable Z

1: Input: Dataset $\mathcal{D} = \{(Y_i, X_i, Z_i, T_i\}_{i=1}^n$, number of folds K.
2: Randomly split $\mathcal{D}$ into K folds $\mathcal{D}^{(k)}$, $k = 1, \ldots, K$.
3: **for** $k = 1, \ldots, K$ **do**
4: Use $\{(X_i, Z_i, Y_i)\colon T_i = 1, i \notin \mathcal{D}^{(k)}\}$ to obtain $\widehat{\mu}_1^{(k)}(x, z)$ for $\mathbb{E}[Y(1)|X = x, Z = z]$;
5: Use $\{(X_i, Z_i, Y_i)\colon T_i = 0, i \notin \mathcal{D}^{(k)}\}$ to obtain $\widehat{\mu}_0^{(k)}(x, z)$ for $\mathbb{E}[Y(0)|X = x, Z = z]$;
6: Compute $\widehat{\mu}_1(X_i, Z_i) = \widehat{\mu}_1^{(k)}(X_i, Z_i)$ and $\widehat{\mu}_0(X_i) = \widehat{\mu}_0^{(k)}(X_i, Z_i)$ for all $i \in \mathcal{D}^{(k)}$.
7: **end for**
8: Compute $\widehat{p} = n_t/n$ and $\Gamma_i = \widehat{\mu}_1(X_i, Z_i) - \widehat{\mu}_0(X_i, Z_i) + \frac{T_i}{\widehat{p}}\big(Y_i - \widehat{\mu}_1(X_i, Z_i)\big) - \frac{1-T_i}{1-\widehat{p}}\big(Y_i - \widehat{\mu}_0(X_i, Z_i)\big)$ for all $i \in \mathcal{D}$.
9: Compute estimator $\widehat{\delta}' = \frac{\sum_{i=1}^n \Gamma_i}{\sum_{i=1}^n Z_i}$.

6. Simulation Studies

We conduct simulations to demonstrate the performance of all proposed optimal variance reduction procedures, based on which we offer practical suggestions.

6.1. Count Metrics

In this section, we use simulations to validate whether Algorithm 1 can outperform state-of-the-art methods when Y and X have a nonlinear relationship, while achieving comparable performance when the relationship is indeed linear (in which case linear adjustment is optimal).

We design one nonlinear and one linear data-generating processes where $X \in \mathbb{R}^d$ for $d \in \{10, 100\}$. The sample size is fixed at $n = 10000$ and we generate $X_i \overset{\text{iid}}{\sim} N(0, I_d)$ for $d \in \{10, 100\}$ except for a categorical variable $X_6 \sim \text{Unif}\{1, 2, \ldots, 10\}$ to represent both continuous and categorical covariates. Let treatments $T_i \overset{\text{iid}}{\sim} \text{Bernoulli}(0.5)$ and the outcomes are generated by $Y_i = b(X_i) + T_i \cdot \tau(X_i) + \epsilon_i$, where $\epsilon_i \overset{\text{iid}}{\sim} N(0, 1)$ is the random noise. In the nonlinear setting inspired by Guo et al. (2021); Friedman (1991), we define the conditional treatment effect $\tau(x) = 10x_1 + 5\log(1 + \exp(x_2)) + \mathbb{1}\{x_6 \in \{1, 5, 9\}\}$ and baseline $b(x) = 10\sin(\pi \cdot x_1x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5\,\mathbb{1}\{x_6 \in \{1, 5, 9\}\}$. In the linear setting, we specify $b(x) = \beta^\top x$ and $\tau(x) = 1 + \delta^\top x + \mathbb{1}\{x_6 \in \{1, 5, 9\}\}$, where $\beta = (5.31, 1.26, 3.12, -0.85, 0, \ldots, 0)^\top$, $\delta = (1.26, -3.14, 0, \ldots, 0)^\top \in \mathbb{R}^d$. The ground truth is $\tau = 4.34$ for the nonlinear setting and $\tau = 0.303$ for the linear setting. In $N = 1000$ independent runs, we evaluate the estimated standard deviation and the empirical coverage of the

Table 1. Variance reduction % compared to DiM (Var. red%.) and the empirical coverage of the 0.95-C.I.s (Emp. cov.) for all methods in the linear (Lin) and nonliner (Nonlin) settings.

Setting	Algorithm 1						CUPED		*DiM*
	Var.Red%.			Emp.Cov.			Var.Red%.	Emp.Cov.	Emp.Cov.
	RF	GB	NN	RF	GB	NN			
Lin, $d = 10$	90.51	91.91	91.80	0.950	0.958	0.947	92.12	0.941	0.951
Lin, $d = 100$	89.92	91.48	88.31	0.951	0.944	0.951	92.12	0.945	0.949
Nonlin, $d = 10$	88.51	89.59	67.57	0.953	0.943	0.949	34.11	0.944	0.945
Nonlin, $d = 100$	88.00	90.42	73.48	0.956	0.953	0.960	37.62	0.936	0.941

0.95-confidence interval. Valid empirical coverage would certify the validity of the inference procedure, under which smaller estimated standard deviation indicates shorter confidence intervals and higher efficiency.

We compare our method to the diff-in-mean (DiM) estimator and the popular CUPED estimator. To demonstrate the performance with different types of machine learning algorithms, we use the random forest regressor (RF), gradient boosting (GB) and neural networks (NN, two layers) from `scikit-learn` python library, all without manual model tuning. The CUPED method we implement is equivalent to Lin (2013) which uses multivariate covariates and runs separate regressions in the treated and control groups. It has better performance than the vanilla version of Deng et al. (2013), and is optimal in the linear setting. Results averaged over 1000 replicates are summarized in Table 1.

From Table 1, the results for linear setting show that our proposed Algorithm 1 is valid and achieves comparable efficiency as the optimal linear method. The results for nonlinear setting show that Algorithm 1 has much higher efficiency than the state-of-the-art linear method, as machine learning regressors are capable to capture nonlinear dependencies. The variance reduction of NN is less than that of RF and GB, which is probably because NN has lower prediction accuracy due to the lack of model tuning.

6.2. Ratio Metrics

Variance reduction for ratio metrics is rarely studied in the literature and existing benchmarks are scarce. In this section, we conduct simulations to evaluate the performance of our methods and some alternative solutions.

We design two data generating processes, one satisfying the SDA introduced in Section 2.1 and one does not. In both settings, the marginal distributions of X_i and Z_i are the same, and $Y_i = b(X_i, Z_i) + T_i \cdot \tau(X_i, Z_i) + \epsilon_i$ for $\epsilon_i \overset{\text{iid}}{\sim} N(0,1)$, where we define $b(x,z) = \big(1.5 + \sin(\pi x_1 x_2)\big) \cdot z + 0.5x_4^2$ and $\tau(x,z) = 0.5z \cdot \big(x_1 + \log(1 + e^{x_3})\big) + 0.2\,\mathbb{1}\{x_6 \in \{1,5,9\}\}$. The only difference is whether Z_i is influenced by the treatment. In both settings, we generate $X_i \overset{\text{iid}}{\sim} N(0, I_d)$ for $d \in \{10, 100\}$ except for a categorical variable $X_6 \sim \text{Unif}\{1, 2, \ldots, 10\}$. To illustrate the impact of imposing the SDA on Z, we test Algorithms 2 and 3 in both settings.

Setting 1 satisfies the SDA, and $\delta = \delta' = \mathbb{E}[Y(1) - Y(0)]/\mathbb{E}[Z]$. We generate $Z_i = \log(1 + \exp(1 + X_{i,1})) + D_i(0.2X_{i,3}^2 + 0.1\,\mathbb{1}\{x_6 \in \{1,5,9\}\})$ for $D_i \overset{\text{iid}}{\sim} \text{Bernoulli}(0.5)$ to ensure $Z_i > 0$. We compare our Algorithms 2 and 3 to the difference-in-mean estimator $\widehat{\delta}'_{\text{DiM}} = \big(\frac{1}{n_t}\sum_{T_i=1} Y_i - \frac{1}{n_c}\sum_{T_i=0} Y_i\big)/\big(\frac{1}{n}\sum_{i=1}^n Z_i\big)$ as well as $\widehat{\delta}'_{\text{CUPED}} = \widehat{\theta}_{\text{CUPED}}/\big(\frac{1}{n}\sum_{i=1}^n Z_i\big)$, where $\widehat{\theta}_{\text{CUPED}}$ is the linear adjustment estimator for count metrics (multivariate and separate regression version). Both of them pool all Z_i to estimate the denominator and apply methods for count metrics to estimate the numerator, hence, comparable to Algorithm 3. The true estimands are $\delta = \delta' = 0.641$.

Setting 2 does not satisfy the SDA, and the two estimands δ and δ' are distinct. The only difference to setting 1 is that we generate $Z_i = \log(1+\exp(X_{i,1})) + T_i(0.2X_{i,3}^2 + 0.1\,\mathbb{1}\{x_6 \in \{1,5,9\}\})$, where we replace D_i with T_i so that Z_i is a realized potential outcome. As the two algorithms are not directly comparable, our evaluation focuses on Algorithm 2, while we still evaluate Algorithm 3 to show the consequences of erroneously assuming SDA. We compare these two nonlinear algorithms to the difference-in-mean estimator $\widehat{\theta}_{\text{DiM}} = \sum_{T_i=1} Y_i / \sum_{T_i=1} Z_i - \sum_{T_i=0} Y_i / \sum_{T_i=0} Z_i$ as well as our linear adjustment method proposed in Section 4.4. (Note that we do not include the ratio metric extension of the CUPED method outlined in the appendix of Deng et al. (2013) because it cannot incorporate covariates other than pretreatment metrics.) The true estimands are $\delta = 0.463$ and $\delta' = 0.740$.

In both settings, our procedures are implemented with random forest, gradient boosting, and neural network from `scikit-learn` python library. We evaluate the estimated variance and the empirical coverage of the 0.95-confidence intervals over $N = 1000$ independent runs with sample size $n = 10{,}000$, and we also assess the proportion of reduced variance compared to the diff-in-mean estimator, as well as the coverage of confidence intervals for estimands (i.e., δ for $\widehat{\delta}$, while δ' for $\widehat{\delta}'$). Due to the page limit, Table 2 summarizes the results with random forest, with the rests are deferred to the supplementary material.

In setting 1 with SDA, the first two lines in Table 2 confirms the optimality of Algorithm 3: its coverage is above the nominal level 0.95, and it achieves the best variance reduction performance. As $\delta = \delta'$, Algorithm 2 is also valid for δ with comparable variance reduction to Algorithm 3; this is due to the nature of the simulation design, not necessarily true in general. For linear methods, CUPED (linear regression without cross-fitting) for the SDA, $d = 100$ setting (line 2) does not provide valid coverage, showing problems with linear regression asymptotics under relatively high dimensionality. On the contrary, our cross-fitting based linear method from Section 4.4 reliably handles high dimensionality ($d = 100$) and achieves the desired coverage (line 4).

Table 2. Variance reduction (relative to the corresponding diff-in-mean estimator) and empirical coverage for ratio metrics.

Setting	Algorithm 3		Algorithm 2		Linear	
	Var.Red%	Emp.Cov.	Var.Red%	Emp.Cov.	Var.Red%	Emp.Cov.
SDA, $d = 10$	71.39%	0.977	70.73%	0.947	46.71%	0.959
SDA, $d = 100$	71.05%	0.991	70.74%	0.939	42.43%	0.817
non-SDA, $d = 10$	42.23%	0.000	41.30%	0.942	1.25%	0.954
non-SDA, $d = 100$	48.01%	0.000	39.64%	0.941	2.63%	0.954

NOTE: "Linear" is $\widehat{\delta}'_{\text{CUPED}}$ for SDA settings and the linear adjustment method in Section 4.4 for non-SDA settings.

The last two lines in Table 2 illustrate the performance of different methods in absence of SDA (setting 2) and Algorithm 2 is clearly the best. Linear adjustment does not reduce much variance due to the nonlinearity of the data. Although Algorithm 3 achieves more variance reduction due to the nature of the data-generating process, the inference based on Algorithm 3 for δ' is not valid (coverage is zero). This happens because the asymptotic unbiasedness of $\widehat{\delta}'$ relies crucially on $\mathbb{P}_{(X_i,Z_i)|T_i=1} = \mathbb{P}_{(X_i,Z_i)|T_i=0}$. However, without SDA, the debiasing term $\frac{1}{n_c}\sum_{T_i=0}(Y_i(0) - \widehat{\mu}_0(X_i, Z_i(0)))$ cannot correct for the bias of $\frac{1}{n_t}\sum_{T_i=1}\widehat{\mu}_0(X_i, Z_i(1))$. In fact, the confidence intervals derived from Algorithm 3 covers another quantity that is neither δ nor δ' and whose practical interpretation is unclear (the quantity equals $\mathbb{E}[\Gamma_i]/\mathbb{E}[Z_i]$ for Γ_i defined in Section 1 of the supplementary material). Thus, the smaller variance of Algorithm 3 does not make it more attractive than Algorithm 2.

Some takeaway messages and practical suggestions are summarized below.

(i) When the SDA does *not* hold, the estimand is $\delta = \mathbb{E}[Y(1)]/\mathbb{E}[Z(1)] - \mathbb{E}[Y(0)]/\mathbb{E}[Z(0)]$ and $\widehat{\delta}$ from Algorithm 2 is optimal.

(ii) When the SDA holds, the target is $\delta' = \mathbb{E}[Y(1) - Y(0)]/\mathbb{E}[Z]$ and Algorithm 3 is optimal.

(iii) In practice, the SDA for variance reduction needs to be made with caution, because if it is violated, $\widehat{\delta}'$ from Algorithm 3 may be invalid and its actual target lacks clear interpretation. There should always be a separate test (such as applying Algorithm 1 for count metrics) on whether $\mathbb{E}[Z(1)] = \mathbb{E}[Z(0)]$. Without strong evidence for SDA, we recommend dropping the SDA and using Algorithm 2 for optimal variance reduction. Even if SDA *actually* holds, Algorithm 2 is still valid and in some cases only slightly inferior to Algorithm 3.

7. Real Examples

In this section, we provide two real examples at LinkedIn: one applying Algorithm 1 to analyze the count metrics in a LinkedIn feed experiment, and the other one applying Algorithm 2 to analyze ratio metrics in an enterprise experiment for LinkedIn learning.

7.1. Count Metrics in a LinkedIn Feed Experiment

The LinkedIn feed is an online system exposing members to contents posted in their network, including career news, ideas, questions, and jobs in the form of short text, articles, images, and videos. ML algorithms are used to rank tens of thousands of candidate updates for each member to help them discover the most relevant contents. We consider an experiment on LinkedIn's feed homepage where members are randomly assigned into a treatment group and a control group. The treatment group is assigned a new version of feed relevance algorithm, which would be compared to the baseline algorithm in the control group. The goal of the experiment is to understand how the new ranking algorithm would impact the revenue from feed. As discussed in Deng et al. (2013), the effectiveness of variance reduction using CUPED depends on the linear correlation of the experiment outcome with the pre-experiment metric. It is challenging to reduce the variance for revenue because member's revenue is a volatile metric whose autocorrelation across different time periods is weak. We hope to improve the performance of variance reduction by incorporating more covariates and exploiting nonlinearity with ML methods.

The experiment takes a random sample of $n = 400{,}000$ members from the LinkedIn online feed traffic and remove outliers whose revenues are above the 99.5% quantile. We implement Algorithm 1 with gradient boosting in the `scikit-learn` Python library and compare it to CUPED (with separate linear regressions on treated and control outcomes) as well as the diff-in-mean estimator. To illustrate the advantage of incorporating side information, the CUPED method only uses pretreatment revenue metric, whilst two sets of covariates are used in Algorithm 1: (a) pretreatment revenue metric; (b) pretreatment revenue metric and other member attributes including country code, industry, membership status, job seeker class, profile viewer count, connection count, network density, etc., where categorical features are transformed into binary variables with one-hot encoding.

With only pretreatment revenue metrics, CUPED reduces 15.91% of variance compared to the diff-in-mean estimator, whereas Algorithm 1 reduces 19.77% by exploiting nonlinear relations with gradient boosting models. By further incorporating member attribute covariates, Algorithm 1 reduces 22.22% of variance compared to the diff-in-mean estimator, showing the efficiency gained from more side information.

7.2. Ratio Metrics in a LinkedIn Learning Experiment

LinkedIn Learning platform is launching a new notification center from which learners can receive customized course recommendations based on what they watched, saved or the trending courses. It also ensures that the learners would not miss assignments from their managers, active conversations with instructors and learners on their favorite contents.

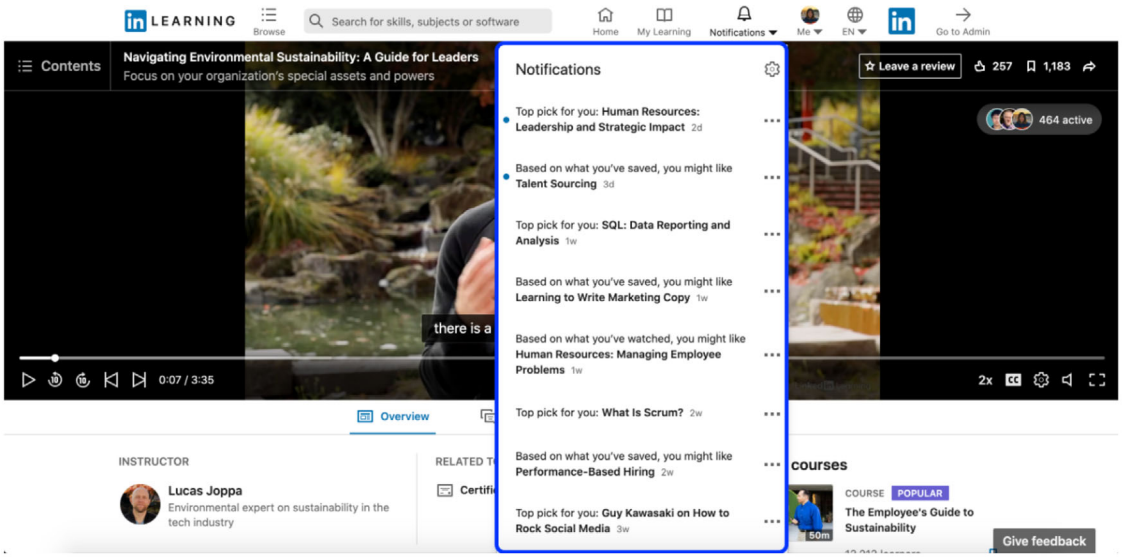

Figure 1. The new LinkedIn Learning notification center.

An online experiment is conducted to assess the impact of this new notification center on the enterprise learners. Because the notification center is an explicit new feature with UI changes (as shown in Figure 1), the experiment requires that learners from the same enterprise account will either be all in control or all in treatment to avoid jeopardizing customer trust. In order to ensure "same account same experience," the experiment is randomized by enterprise accounts (instead of by individual learners). We focus on the "learning engagement per learner" metric, which measures the contribution of one day's engagement to future engagement (video time watched). As introduced in Section 2.1, since the analysis unit (learner) is at a lower level than the randomization unit (enterprise account), our metric of interest is a ratio metric "learning engagement per account/number of active learners per account." This experiment can also be viewed as a cluster randomized experiment.

For target δ, we compare Algorithm 2 to an extension (separate regressions on each group) of the CUPED outlined in the appendix of Deng et al. (2013). When the target is δ', CUPED means $\widehat{\delta}'_{\text{CUPED}} = \widehat{\theta}_{\text{CUPED}}/\left(\frac{1}{n}\sum_{i=1}^{n} Z_i\right)$, where $\widehat{\theta}_{\text{CUPED}}$ is the CUPED estimator for count metrics "learning engagement per account." The corresponding diff-in-mean estimators are the same as introduced in Section 6.2. To illustrate the advantage of incorporating extra covariates, the CUPED method only uses pretreatment versions of Y, Z, whereas for each target, two sets of covariates are used in Algorithms 2 and 3: (a) only pretreatment versions of Y, Z; (b) pretreatment versions of Y, Z and features of contracts, including industry segment, SSO information, country code, account size category, etc. In both settings, we implement our procedures with the `XGBoost` python library without special tuning. The results are summarized in Table 3.

For both targets, we see the efficacy of using machine learning tools and incorporating large numbers of covariates. For target δ when CUPED does not reduce much variance, flexible machine learning tools improves CUPED by going beyond linearity as in Line 1 and incorporating a large number of extra covariates as in Line 2. For δ', we see the substantial improvement (about 70% to 80% of variance reduced) compared to the diff-in-mean estimator and our optimal estimator can further achieve 30% lower variance than CUPED by exploiting the information in the many covariates.

Table 3. Proportion of reduced variance compared to DiM (compared to CUPED).

Covariates	Target: δ		Target: δ'	
	Algorithm 2	CUPED	Algorithm 3	CUPED
(1)	3.13%(1.40%)	1.76%	74.1%(−11.13%)	76.62%
(2)	12.35%(10.78%)		83.6%(29.58%)	

8. Conclusions

We establish a rigorous statistical framework for variance reduction of count and ratio metrics that are popular in online controlled experiments. We propose variance reduction methods that use flexible ML tools to incorporate large numbers of covariates, yielding unbiased estimators for treatment effects under mild conditions. Based on semiparametric efficiency theory, we establish the optimality of the proposed procedures, sheding light on ideal solutions for variance reduction. Simulation studies illustrate the performance of our methods and also give practical suggestions for ratio metrics under different assumptions. Finally, two real experiments from LinkedIn illustrate the applicability of our methods.

Supplementary Materials

Deferred results and technical proofs Deferred theoretical results (inference and optimality) for Algorithm 3, as well as proofs of all theoretical results in the article. (pdf)

Reproduction codes Reproduction codes for the simulation studies in the article. (python codes and readme file)

Acknowledgments

This work is done during the first author's internship at LinkedIn Applied Research team. The authors would like to thank Dominik Rothenhäusler for helpful discussions and thank Weitao Duan, Rina Friedberg, Reza Hosseini, Juanyan Li, Min Liu, Jackie Zhao, Sishi Tang and Parvez Ahammad for their suggestions and feedbacks. The authors would also like to thank the Editor, AE and the anonymous referee for their valuable comments.

ORCID

Ying Jin http://orcid.org/0000-0002-5211-5083
Shan Ba http://orcid.org/0000-0002-4060-949X

References

Andrews, D. W. (1994), "Empirical Process Methods in Econometrics," *Handbook of Econometrics*, 4, 2247–2294. [233]

Athey, S., and Imbens, G. W. (2016), "Recursive Partitioning for Heterogeneous Causal Effects," *Proceedings of the National Academy of Sciences*, 113, 7353–7360. [233]

Athey, S., and Wager, S. (2019), "Estimating Treatment Effects with Causal Forests: An Application," *Observational Studies*, 5, 36–51. [233]

Bickel, P. J., Klaassen, C. A., Ritov, Y., and Wellner, J. A. (1993), *Efficient and Adaptive Estimation for Semiparametric Models*, Baltimore: Johns Hopkins University Press. [231,234,236]

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018), "Double/Debiased Machine Learning for Treatment and Structural Parameters," *The Econometrics Journal*, 21, 1–68. [233,234,236]

Chernozhukov, V., Demirer, M., Duflo, E., and Fernández-Val, I. (2017), "Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments," arXiv preprint arXiv:1712.04802. [233]

Cohen, P. L., and Fogarty, C. B. (2020), "No-harm Calibration for Generalized Oaxaca-Blinder Estimators," arXiv preprint arXiv:2012.09246. [233,234,236]

Deng, A., Lu, J., and Litz, J. (2017), "Trustworthy Analysis of Online a/b Tests: Pitfalls, Challenges and Solutions," in *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pp. 641–649. [231,232,233,236]

Deng, A., Xu, Y., Kohavi, R., and Walker, T. (2013), "Improving the Sensitivity of Online Controlled Experiments by using Pre-experiment Data," in *Proceedings of the sixth ACM international conference on Web search and data mining*, pp. 123–132. [231,233,235,236,238,239,240,241]

Freedman, D. A. (2008), "On Regression Adjustments to Experimental Data," *Advances in Applied Mathematics*, 40, 180–193. [231,233]

Friedman, J. H. (1991), "Multivariate Adaptive Regression Splines," *The Annals of Statistics*, 19, 1–67. [238]

Green, D. P., and Vavreck, L. (2008), "Analysis of Cluster-Randomized Experiments: A Comparison of Alternative Estimation Approaches," *Political Analysis*, 16, 138–152. [232]

Guo, K., and Basse, G. (2021), "The Generalized Oaxaca-Blinder Estimator," *Journal of the American Statistical Association*, 1–35. [234,235]

Guo, Y., Coey, D., Konutgan, M., Li, W., Schoener, C., and Goldman, M. (2021), "Machine Learning for Variance Reduction in Online Experiments," arXiv preprint arXiv:2106.07263. [231,233,236,238]

Hahn, J. (1998), "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 66, 315–331. [233,234,235,236]

Hosseini, R., and Najmi, A. (2019), "Unbiased Variance Reduction in Randomized Experiments," arXiv preprint arXiv:1904.03817. [231,233,236]

Imai, K., and Ratkovic, M. (2013), "Estimating Treatment Effect Heterogeneity in Randomized Program Evaluation," *The Annals of Applied Statistics*, 7, 443–470. [233]

Imbens, G. W., and Rubin, D. B. (2015), *Causal Inference in Statistics, Social, and Biomedical Sciences*, Cambridge: Cambridge University Press. [232,233]

Kennedy, E. H. (2020), "Optimal Doubly Robust Estimation of Heterogeneous Causal Effects," arXiv preprint arXiv:2004.14497. [233]

Kohavi, R., Tang, D., and Xu, Y. (2020), *Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing*, Cambridge: Cambridge University Press. [231]

Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. (2019), "Metalearners for Estimating Heterogeneous Treatment Effects using Machine Learning," *Proceedings of the National Academy of Sciences*, 116, 4156–4165. [233]

Lin, W. (2013), "Agnostic Notes on Regression Adjustments to Experimental Data: Reexamining Freedman's Critique," *The Annals of Applied Statistics*, 7, 295–318. [231,233,235,238,239]

Middleton, J. A., and Aronow, P. M. (2015), "Unbiased Estimation of the Average Treatment Effect in Cluster-Randomized Experiments," *Statistics, Politics and Policy*, 6, 39–75. [232]

Nichols, A. (2007), "Causal Inference with Observational Data," *The Stata Journal*, 7, 507–541. [233,236]

Nie, X., and Wager, S. (2020), "Quasi-Oracle Estimation of Heterogeneous Treatment Effects," *Biometrika*, forthcoming. [233]

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994), "Estimation of Regression Coefficients When Some Regressors are not Always Observed," *Journal of the American Statistical Association*, 89, 846–866. [233,234,235,236]

Schuler, M. S., and Rose, S. (2017), "Targeted Maximum Likelihood Estimation for Causal Inference in Observational Studies," *American Journal of Epidemiology*, 185, 65–73. [233,236]

Van der Vaart, A. W. (2000), *Asymptotic Statistics* (Vol. 3), Cambridge: Cambridge University Press. [233]

Van Der Vaart, A. W., van der Vaart, A., van der Vaart, A. W., and Wellner, J. (1996), *Weak Convergence and Empirical Processes: With Applications to Statistics*, New York: Springer. [233]

Wager, S., and Athey, S. (2018), "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests," *Journal of the American Statistical Association*, 113, 1228–1242. [233]

Wu, C. F. J., and Hamada, M. S. (2009), *Experiments: Planning, Analysis, and Optimization*, Hoboken, NJ: Wiley. [233]

Yang, L., and Tsiatis, A. A. (2001), "Efficiency Study of Estimators for a Treatment Effect in a Pretest–Posttest Trial," *The American Statistician*, 55, 314–321. [231,233]