Improving the Sensitivity of Online Controlled Experiments by Utilizing Pre-Experiment Data

Alex Deng*
Microsoft
One Microsoft Way
Redmond, WA 98052
alexdeng@microsoft.com

Ya Xu*
Microsoft
1020 Enterprise Way
Sunnyvale, CA 94089
yaxu@microsoft.com

Ron Kohavi
Microsoft
One Microsoft Way
Redmond, WA 98052
ronnyk@microsoft.com

Toby Walker
Microsoft
One Microsoft Way
Redmond, WA 98052
towalker@microsoft.com

ABSTRACT

Online controlled experiments are at the heart of making data-driven decisions at a diverse set of companies, including Amazon, eBay, Facebook, Google, Microsoft, Yahoo, and Zynga. Small differences in key metrics, on the order of fractions of a percent, may have very significant business implications. At Bing it is not uncommon to see experiments that impact annual revenue by millions of dollars, even tens of millions of dollars, either positively or negatively. With thousands of experiments being run annually, improving the sensitivity of experiments allows for more precise assessment of value, or equivalently running the experiments on smaller populations (supporting more experiments) or for shorter durations (improving the feedback cycle and agility). We propose an approach (CUPED) that utilizes data from the pre-experiment period to reduce metric variability and hence achieve better sensitivity. This technique is applicable to a wide variety of key business metrics, and it is practical and easy to implement. The results on Bing's experimentation system are very successful: we can reduce variance by about 50%, effectively achieving the same statistical power with only half of the users, or half the duration.

Categories and Subject Descriptors

G.3 [**Probability and Statistics/Experiment Design**]: controlled experiments, randomized experiments, A/B testing

General Terms

Measurement, Variance, Experimentation

*Corresponding authors.

WSDM'13, February 4–8, 2013, Rome, Italy.

Keywords

Controlled experiment, Variance, A/B testing, Search quality evaluation, Pre-Experiment, Power, Sensitivity

1. INTRODUCTION

A controlled experiment is probably the oldest and the most widely accepted methodology to establish a causal relationship (Mason et al. 1989; Box et al. 2005; Keppel et al. 1992; Kohavi et al. 2009b; Manzi 2012). Now widely used and re-discovered by many companies, it is referred to as a randomized experiment, an A/B test (Wikipedia), a split test, a weblab (at Amazon), a live traffic experiment (at Google), a flight (at Microsoft), and a bucket test (at Yahoo!). This paper is focused on online controlled experiments, where experiments are conducted on live traffic to support data-driven decisions in online businesses, including e-business sites like Amazon and eBay, portal sites like Yahoo and MSN (Kohavi et al. 2009b), search engines like Microsoft Bing (Kohavi et al. 2012) and Google (Tang et al. 2010b).

Online controlled experiments are critical for businesses. Small differences in key metrics, on the order of fractions of a percent, can have very significant business implications. At Bing it is not uncommon to see experiments that impact annual revenue by millions of dollars, even tens of millions of dollars, either positively or negatively. We begin with a motivating example of an online controlled experiment run at MSN (Kohavi et al. 2009a). MSN Real Estate (`http://realestate.msn.com`) had six visual design candidates for the "Find a home" widget, as shown in Figure 1. During the experiment, users were randomly split between the 6 variants, where the control is the production version and treatment 1 through 5 are five new designs. The goal was to increase visits to the linked partner sites through the widget. Users' interactions with the widget were instrumented and key metrics such as the number of transfers to partner sites were computed. In this experiment, the winner, treatment 5, increased revenue from transfer fees by almost 10% compared to the control.

One challenge with any controlled experiment is the ability to detect the treatment effect when it indeed exists, usually referred to as "power" or "sensitivity." Improving sen-

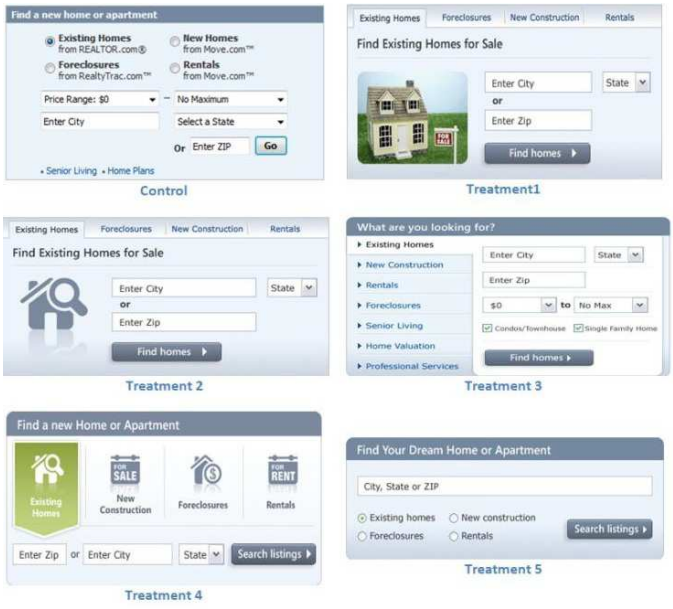

Figure 1: Widgets tested for MSN Real Estate.

sitivity is particularly important when running online experiments at large scale. A mature online experimentation platform runs thousands of experiments a year (Kohavi et al. 2012; Manzi 2012). The benefit of any increased sensitivity is therefore amplified by economies of scale. It might seem unnecessary to emphasize sensitivity for online experiments because they tend to have very large sample sizes already (e.g. millions of users (comScore 2012)) and increasing sample size is usually the most straightforward way to improve power. In reality, even with a large amount of traffic, online experiments cannot always reach enough statistical power. Google made it very clear that they are not satisfied with the amount of traffic they have (Tang et al. 2010a, Slide 6) even with over 10 billion searches per month (comScore 2012). There are several reasons for this. First, the treatment effects we would like to detect tend to be very small. The sensitivity of controlled experiments is inversely proportional to the number of users squared, so whereas a small site may need 10,000 users to detect a 5% delta, detecting a 0.5% delta requires 100 times (10 squared) more users, or one million users. Even a 0.5% change in revenue per user equates to millions of dollars for large online sites. Second, it is crucial to get results fast. One wants to launch good features early, and more importantly, if the treatment turns out to have a negative impact on users, we need to stop it as soon as possible. Third, there are many experiments that have low triggering rates; that is, only a small fraction of the experiment's users actually experience the treatment feature. For example, in an experiment affecting only recipe-related queries, the majority of the users will not see the target feature because they didn't search for recipes during the experiment. In these cases, the effective sample size can be small and statistical analysis can suffer from low statistical power. Finally, in a data-driven culture, there is always demand to run more experiments to keep up with the rate of innovation. A good online experimentation platform should allow many experiments to run together. This also requires that we make optimal use of large but still limited traffic.

One way to improve sensitivity is through variance reduction. Kohavi et al. (2009b) provides examples where we can achieve a lower variance using a different evaluation metric or through filtering out users who are not impacted by the change. Deng et al. (2011) shows how we can use page level randomization at the design stage to reduce variance of page level metrics (Chapelle et al. 2012). However, these methods are limited in their applicability to special cases and we want a technique that is applicable to any metric as, in practice, businesses are likely to have a set of Key Performance Indicators (KPIs) that cannot be changed easily. Moreover, the technique should preferably not be based on any parametric model because model assumptions tend to be unreliable and a model that works for one metric does not necessarily work for another.

In this paper, we propose a technique, called CUPED (Controlled-experiment Using Pre-Experiment Data), which adjusts metrics using pre-experiment data for users in the control and treatment to reduce metric variability.

The key contributions of our work include:

- A theoretically sound and practical method to reduce variances for online experiments using pre-experiment data, which greatly increases experiment sensitivity.
- Extensions of approach to non-user metrics and partially missing pre-experiment data.
- Criteria for selecting the best covariates, including the empirical result that using the same metric from the pre-experiment typically gives the greatest variance reduction.
- Validation of the results on real online experiments run at Bing, demonstrating a variance reduction of about 50%, equivalent to doubling our traffic or halving the time we need to run an experiment to get the same sensitivity.
- Practical guidance on choices important to successful application of CUPED to real-world online experimentation, including factors like the best length to use for the pre-experiment period and the use of multiple covariates.

2. BACKGROUND AND RELATED WORK

2.1 Analyzing Experiments

Because of its wide applicability, we focus on the case of the two-sample t-test (Student 1908; Wasserman 2003). This is the framework most commonly used in online experiment analysis. Suppose we are interested in some metric Y (e.g. Queries per user). To apply the t-test, we assume the observed values of the metric for users in the treatment and control are independent realizations of random variables $Y^{(t)}$ and $Y^{(c)}$. The null hypothesis is that $Y^{(t)}$ and $Y^{(c)}$ have the same mean and the alternative is that they do not. The t-test is based on the t-statistic:

$$\frac{\overline{Y}^{(t)} - \overline{Y}^{(c)}}{\sqrt{\text{var}\left(\overline{Y}^{(t)} - \overline{Y}^{(c)}\right)}}, \tag{1}$$

where $\Delta = \overline{Y}^{(t)} - \overline{Y}^{(c)}$ is an unbiased estimator for the shift of the mean and the t-statistic is a normalized version of that estimator. For online experiments, the sample sizes for both

control and treatment are at least in thousands, hence the normality assumption on Y is usually unnecessary because of the Central Limit Theorem.

Because the samples are independent,

$$\text{var}(\Delta) = \text{var}\left(\overline{Y}^{(t)} - \overline{Y}^{(c)}\right) = \text{var}\left(\overline{Y}^{(t)}\right) + \text{var}\left(\overline{Y}^{(c)}\right).$$

In this framework, the key to variance reduction for the difference in mean lies in reducing the variance of the means themselves. As we will see in Section 3, this connects our problem to techniques used in Monte Carlo sampling to improve estimation of a mean through variance reduction.

At a very high level, our proposal for variance reduction works as follows. We conduct the experiment as usual but when analyzing the data, we compute an adjusted or corrected estimate of the delta. That adjusted estimate, Δ^*, incorporates pre-experiment information, such that

- Δ^* is still an unbiased estimator for the shift in the means (same as Δ), and
- Δ^* has a smaller variance than Δ.

Note that because of the reduced variance, the corresponding t-statistic would be larger for the same expected effect size. We therefore achieved better sensitivity.

2.2 Linear Models

We begin with a short review of related work. Variance reduction has been a longstanding challenge for analyzing randomized experiments. The most popular parametric method is based on linear modeling (Gelman and Hill 2006). A linear model for an experiment assumes that the outcome is a linear combination of a treatment effect coupled with additional covariate terms. In particular, suppose Y_i is the outcome metric, Z_i is the treatment assignment indicator and $\mathbf{X}_i$ is a vector of covariates. The linear model assumes $\mathbb{E}(Y_i|Z_i, \mathbf{X}_i) = \theta_0 + \delta Z_i + \boldsymbol{\theta}^T \mathbf{X}_i$. Under the assumptions of the model, linear regression (also called ANCOVA when covariates are categorical variables) gives a consistent estimator for the average treatment effect and reduces variance. However, the linear model makes strong assumptions that are usually not satisfied in practice, i.e., the conditional expectation of the outcome metric is linear in the treatment assignment and covariates. In addition, it also requires all residuals to have a common variance.

2.3 Semi-Parametric Models

To overcome limitations of the linear model, researchers have developed less restrictive models called semi-parametric models (Tsiatis 2006), for which Generalized Estimating Equations (GEE) are used for fitting the model. Comparing standard linear models with semi-parametric models, Yang and Tsiatis (2001) showed that linear model (ANCOVA) and GEE are asymptotically equivalent under the less restrictive semi-parametric model and both give more efficient estimates for the average treatment effect than the unadjusted t-test. Leon et al. (2003), Davidian et al. (2005) and Tsiatis et al. (2008) further refined the work using semi-parametric statistical theory (Tsiatis 2006) and gave the analytical form of a class of estimators for the average treatment effect. This class is complete in the sense that all possible RAL (regular and asymptotically linear) estimators for the average treatment effect are asymptotically equivalent to one in the class. The problem left is to find the estimator in the class with the smallest variance and they provided general guidance.

In this paper, we look at the problem from a different perspective. By connecting the variance reduction problem in randomized experiments to a similar problem in Monte Carlo simulation, we are able to derive a very powerful result. Instead of diving into abstract Hilbert spaces and functional influence curves, our argument only involves elementary probability. In particular, we propose to use the data from the pre-experiment period to reduce metric variability, which turns out to be very effective and practically applicable.

3. VARIANCE REDUCTION

Variance reduction is a common topic in Monte Carlo sampling, where the goal is usually to estimate a parameter by repeatedly simulating possible values from the underlying distribution. In Monte Carlo sampling significant efficiency gains can be had if we use sampling schemes that reduce the variance by incorporating prior information. Unlike Monte Carlo simulations, in the world of online experiments, the population is dynamic and data arrive gradually as the experiment progresses. We cannot design a sampling scheme in advance and then collect data accordingly. However, we will show that because we have pre-experiment data, we can adapt Monte Carlo variance techniques by applying them "retrospectively."

The two Monte Carlo variance reduction techniques we consider here are stratification and control variates. For each technique, we review the basic concepts and then show how it can be adapted to the online experiment setting. We devote Section 3.3 to discussing the connections between these two approaches and the implications in practice.

3.1 Stratification

Stratification is a common technique used in Monte Carlo sampling to achieve variance reduction. In this section, we show how it can be adapted to achieve the same goal in the world of online experimentation.

3.1.1 Stratification in Simulation

The basic idea of stratification is to divide the sampling region into strata, sample within each stratum separately and then combine results from individual strata together to give an overall estimate, which usually has a smaller variance than the estimate without stratification.

Mathematically, we want to estimate $\mathbb{E}(Y)$, the expected value of Y, where Y is the variable of interest. The standard Monte Carlo approach is to first simulate n independent samples $Y_i, i = 1, \ldots, n$, and then use the sample average $\overline{Y}$ as the estimator of $\mathbb{E}(Y)$. $\overline{Y}$ is unbiased and $\text{var}(\overline{Y}) = \text{var}(Y)/n$.

Let's consider a more strategic sampling scheme. Assume we can divide the sampling region of Y into K subregions (strata) with w_k the probability that Y falls into the kth stratum, $k = 1, \ldots, K$. If we fix the number of points sampled from the kth stratum to be $n_k = n \cdot w_k$, we can define a stratified average to be

$$\widehat{Y}_{strat} = \sum_{k=1}^{K} w_k \overline{Y}_k, \tag{2}$$

where $\overline{Y}_k$ is the average within the kth stratum.

The stratified average $\widehat{Y}_{strat}$ and the standard average $\overline{Y}$ have the same expected value but the former gives a smaller

variance when the means are different across the strata. The intuition is that the variance of $\overline{Y}$ can be decomposed into the within-strata variance and the between-strata variance, and the latter is removed through stratification. For example, the variance of children's heights in general is large. However, if we stratify them by their age, we can get a much smaller variance within each age group. More formally,

$$\begin{aligned}\text{var}(\overline{Y}) &= \sum_{k=1}^{K} \frac{w_k}{n}\sigma_k^2 + \sum_{k=1}^{K} \frac{w_k}{n}(\mu_k - \mu)^2 \\ &\geq \sum_{k=1}^{K} \frac{w_k}{n}\sigma_k^2 = \text{var}(\widehat{Y}_{strat})\end{aligned}$$

where (μ_k, σ_k^2) denote the mean and variance for users in the kth stratum. More detailed proof can be found in standard Monte Carlo books (e.g. Asmussen and Glynn (2008)). A good stratification is the one that aligns well with the underlying clusters in the data. By explicitly identifying these clusters as strata, we essentially remove the extra variance introduced by them.

3.1.2 Stratification in Online Experimentation

In the online world, because we collect data as they arrive over time, we are usually unable to sample from strata formed ahead of time. However, we can still utilize pre-experiment variables to construct strata after all the data are collected (for theoretical justification see Asmussen and Glynn (2008, Page 153)). For example, if Y_i is the number of queries from a user i, a covariate X_i could be the browser that the user used before the experiment started. The stratified average in (2) can then be computed by grouping Y according to the value of X,

$$\widehat{Y}_{strat} = \sum_{k=1}^{K} w_k \overline{Y}_k = \sum_{k=1}^{K} w_k \left(\frac{1}{n_k} \sum_{i:X_i=k} Y_i \right).$$

Using superscripts to denote treatment and control groups, the stratified delta

$$\Delta_{strat} = \widehat{Y}_{strat}^{(t)} - \widehat{Y}_{strat}^{(c)} = \sum_{k=1}^{K} w_k (\overline{Y}_k^{(t)} - \overline{Y}_k^{(c)})$$

enjoys the same variance reduction as the stratified average in Eq. (2). It is important to note that by using only the pre-experiment information, the stratification variable X is independent of the experiment effect. This ensures that the stratified delta is unbiased.

In practice, we don't always know the appropriate weights w_k to use. In the context of online experimentation, these can usually be computed from users not in the experiment. As we will see in Section 3.3, when we formulate the same problem in the form of control variates (Section 3.2), we no longer need to estimate the weights.

3.2 Control Variates

We showed how online experimentation can benefit from the stratification technique widely used in the simulation literature. In this section we show how another variance reduction technique used in simulation, called control variates, can also be adopted in online experimentation. In Section 3.2.1, we review control variates in its original form as a variance reduction technique for simulation. We then show how the same idea can be applied in the context of online experimentation in Section 3.2.2.

3.2.1 Control Variates in Simulation

The idea of variance reduction through control variates stems from the following observation. Assume we can simulate another random variable X in addition to Y with known expectation $\mathbb{E}(X)$. In other words, we have independent pairs of $(Y_i, X_i), i = 1, \ldots, n$. Define

$$\widehat{Y}_{cv} = \overline{Y} - \theta\overline{X} + \theta\mathbb{E}X, \tag{3}$$

where θ is any constant. $\widehat{Y}_{cv}$ is an unbiased estimator of $\mathbb{E}(Y)$ since $-\theta\mathbb{E}(\overline{X}) + \theta\mathbb{E}(X) = 0$. The variance of $\widehat{Y}_{cv}$ is

$$\begin{aligned}\text{var}(\widehat{Y}_{cv}) &= \text{var}(\overline{Y} - \theta\overline{X}) = \text{var}(Y - \theta X)/n \\ &= \frac{1}{n}(\text{var}(Y) + \theta^2\text{var}(X) - 2\theta\text{cov}(Y, X)).\end{aligned}$$

Note that $\text{var}(\widehat{Y}_{cv})$ is minimized when we choose

$$\theta = \text{cov}(Y, X)/\text{var}(X) \tag{4}$$

and with this optimal choice of θ, we have

$$\text{var}(\widehat{Y}_{cv}) = \text{var}(\overline{Y})(1 - \rho^2), \tag{5}$$

where $\rho = \text{cor}(Y, X)$ is the correlation between Y and X. Compare (5) to the variance of $\overline{Y}$, the variance is reduced by a factor of ρ^2. The larger ρ, the better the variance reduction. The single control variate case can be easily generalized to include multiple variables.

It is interesting to point out the connection with linear regression. The optimal $\boldsymbol{\theta}$ turns out to be the ordinary least square (OLS) solution of regressing (centered) Y on (centered) X, which gives variance

$$\text{var}(\widehat{Y}_{cv}) = \text{var}(\overline{Y})(1 - R^2),$$

with R^2 being the proportion of variance explained coefficient from the linear regression. It is also possible to use nonlinear adjustment. i.e., instead of allowing only linear adjustment as in (3), we can minimize variance in a more general functional space. Define

$$\widehat{Y}_{cv} = \overline{Y} - \overline{f(X)} + \mathbb{E}(f(X)), \tag{6}$$

and then try to minimize the variance of (6). It can be shown that the regression function $\mathbb{E}(Y|X)$ gives the optimal $f(X)$.

3.2.2 Control Variates in Online Experimentation

Utilizing control variates to reduce variance is a very common technique. The difficulty of applying it boils down to finding a control variate X that is highly correlated with Y and at the same time has known $\mathbb{E}(X)$.

Although in general it is not easy to find control variate X with known $\mathbb{E}(X^{(t)})$ and $\mathbb{E}(X^{(c)})$, a key observation is that $\mathbb{E}(X^{(t)}) - \mathbb{E}(X^{(c)}) = 0$ in the pre-experiment period because we have not yet introduced any treatment effect. By using only information from before the launch of the experiment to construct the control variate, the randomization between treatment and control ensures that we have $\mathbb{E}X^{(t)} = \mathbb{E}X^{(c)}$.

Given $\mathbb{E}X^{(t)} - \mathbb{E}X^{(c)} = 0$, it is easy to see the delta

$$\Delta_{cv} = \widehat{Y}_{cv}^{(t)} - \widehat{Y}_{cv}^{(c)} \tag{7}$$

is an unbiased estimator of $\delta = \mathbb{E}(\Delta)$. Notice how Δ_{cv} does not depend on the unknown $\mathbb{E}(X^{(t)})$ and $\mathbb{E}(X^{(c)})$ at all as

they cancel each other. With the optimal choice of θ from Eq (4), we have that Δ_{cv} reduces variance by a factor of ρ^2 compared to Δ, i.e.

$$\text{var}(\Delta_{cv}) = \text{var}(\Delta)(1 - \rho^2).$$

To achieve a large correlation and hence better variance reduction, an obvious approach is to choose X to be the same as Y, which naturally leads to using the same variable during pre-experiment observation window as the control variate. As we will see in the empirical results in Section 5, this indeed turns out to be the most effective choice we found for control variates.

There is a slight subtlety that's worth pointing out. The pair (Y, X) may have different distributions in treatment and control when there is an experiment effect. For Δ_{cv} to be unbiased, the same θ has to be used for both control and treatment. The simplest way to estimate it is from the pooled population of control and treatment. The impact on variance reduction will likely be negligible. In the general nonlinear control covariates case, we should use the same functional form in both $\widehat{Y}_{cv}^{(t)}$ and $\widehat{Y}_{cv}^{(c)}$.

3.3 Connection between Stratification and Control Variates

We have discussed two techniques that both utilize covariates to achieve variance reduction. The stratification approach uses the covariates to construct strata while the control variates approach uses them as regression variables. The former uses discrete (or discretized) covariates, whereas control variates seem more naturally to be continuous variables. It is, however, not surprising that these two approaches are closely related. In fact, we can show that when the covariate X is categorical (say, with discrete values $1, \ldots, K$) the two approaches produce identical estimates. Details are included in Appendix A. The basic idea is to construct an indicator variable $1_{X=k}$ for each stratum and use it as a control variate with mean being the stratum weight w_k. To this end, the control variates technique is an extension of stratification, where both continuous and discrete covariates are applicable.

While these two techniques are well connected mathematically, they provide different insights into understanding why and how to achieve variance reduction. The stratification formulation has a nice analogy with mixture models, where each stratum is one component of the mixture model. Stratification is equivalent to separating samples according to their component memberships, effectively removing the between-component variance and achieving a reduced variance. A better covariate is hence the one that can better classify the samples and align with their underlying structure. On the other hand, the control variates formulation quantifies the amount of variance reduction as a function of the correlation between the covariates and the variable itself. It is mathematically simpler and more elegant. Clearly, a better covariate should be the one with larger (absolute) correlation.

4. CUPED IN PRACTICE

A simple yet effective way to implement CUPED is to use the same variable from the pre-experiment period as the covariate. Indeed, this is essentially what we have implemented in practice for Bing's experimentation system. However, there are situations when this is not possible or practical. For example, if we want to measure user retention rate or conduct an experiment on new users, there are no pre-experiment data to work with. In fact, in most online experiments, we may not have pre-experiment information on *all* users. An additional challenge is how to use pre-experiment data for metrics whose analysis unit is not a user. This section is devoted to address practical challenges like these using Bing's experimentation system as a case study.

4.1 Selecting Covariates

The choice of covariates is critical, as it directly determines the effectiveness of variance reduction. With the choice of the right variables we can halve the variance but with the wrong choice there is little reduction in variance. To understand which pre-experiment variables worked best we evaluated a large number of possible pre-experiment variables. Across a large class of metrics, our results consistently showed that using the same variable from the pre-experiment period as the covariate tends to give the best variance reduction. In addition, the lengths of the pre-experiment and the experiment periods also play a role. Given the same pre-experiment period, extending the length of the experiment does not necessarily improve the variance reduction rate. On the other hand, a longer pre-period tends to give a higher reduction for the same experiment period. We discuss more details in the context of an empirical example in Section 5.

4.2 Handling Missing Pre-Experiment Data

In online sites, we might not have pre-experiment data on all users in the experiment. This can occur because some users are visiting for the first time, or users simply do not visit the site frequently enough to appear during the pre-experiment period. In addition, users are identified by cookies, which are unreliable and can "churn" (i.e. change due to users clearing their cookies).

This poses a challenge for using the pre-experiment information to construct covariates. For users who are in the experiment but not in the pre-experiment period, the corresponding covariates are not well-defined. One way to address this is to define another covariate that indicates whether or not a user appeared in the pre-experiment period. With this additional binary covariate, we can set the missing covariate values to be any constant we like. Intuitively, this is equivalent to first splitting users into two strata: those that appeared in the pre-experiment period and those that did not. Note that for the stratum of users with pre-experiment data, their pre-experiment covariates are well-defined so further variance reduction based on these covariates is possible. In addition, the stratification by presence in the pre-period is a further source of variance reduction.

4.3 Beyond Pre-Experiment Data

So far we only considered reducing metrics variability based on covariates constructed using the pre-experiment data. This is not only because using the same variable from the pre-period tends to give the best variance reduction, but also because the pre-experiment information is guaranteed to be independent of the experiment's effect, which is crucial to avoid biased results.

It is probably easier to demonstrate this mathematically with the control variates formulation. In Eq. (7), the delta

Δ_{cv} is computed assuming $\mathbb{E}(X^{(t)}) = \mathbb{E}(X^{(c)})$. If there is truly a difference between control and treatment in terms of X and this equality does not hold, Δ_{cv} will be biased. For example, we know a faster page load-time usually leads to more clicks on a search page (Kohavi et al. 2009b). If we use the number of clicks as the covariate in an experiment that improves page load-time, we will end up underestimating the experiment impact on the load-time because part of the improvement is "adjusted away" by the covariate. See Section 5 for a real example of biased results caused by using a covariate that violates the requirements described here.

However, this does not mean that covariates based on pre-experiment data are the only choice. All that is required is that the covariate X is not affected by the experiment's treatment. A natural extension is the class of covariates constructed using information gathered at the first time a user appears in the experiment. For instance, the day-of-week a user is first observed in the experiment is independent of the experiment itself. Such covariates can serve as an additional powerful source for variance reduction. To further extend this idea, covariates based on any information established before a user actually *triggers* the experiment feature are also valid. This can be particularly helpful if the feature to be evaluated has a low triggering rate.

4.4 Handling Non-User Metrics

As we mentioned in Section 1, in online experiments, users (cookies) are the common randomization unit. In the discussion thus far, we have assumed that the analysis unit is also "user." However, this is not always the case. For example, we may want to compute click-through-rate (CTR) as the total number of clicks divided by the total number of pages. The analysis unit here is a "page" instead of a "user." Variance estimation itself is harder when there is a mismatch between the analysis unit and the experiment unit. The most common solution is to use the delta method to produce correct estimate of variance that takes into account the correlation of pages from the same user (Deng et al. 2011; Tang et al. 2010b; Kohavi et al. 2009b). To achieve variance reduction for these non-user level metrics, we need to combine the delta method and our variance reduction techniques together. The details are provided in Appendix B. In fact, not only can the metric of interest (e.g. CTR) be at page level, we can have page-level covariates as well. This opens the door to a larger class of covariates that are based on features specific to a page that may be not specific to a user, e.g. time stamp on a page-view.

5. EMPIRICAL RESULTS

In this section, we share empirical results that show the effectiveness of CUPED for Microsoft Bing's experimentation system. First we show how CUPED can greatly improve the sensitivity for a real experiment run at Bing. Next we look deeper into a 3-week A/A test, which is a controlled experiment where treatment is identical to control and hence the treatment effect is known to be 0. Using an A/A experiment we can examine important decisions that can have a large impact of the success of variance reduction for real-world online experimentation. Finally, we show how biased estimates arise if we go past the user triggering into the experiment and choose a covariate inappropriately.

5.1 Slowdown Experiment in Bing

To show the impact of CUPED in a real experiment we examine an experiment that tested the relationship between page load-time and user engagement on Bing. Delays, on the order of hundreds of milliseconds, are known to hurt user engagement (Kohavi et al. 2009b, Section 6.1.2). In this experiment, we deliberately delayed the server response to Bing queries by 250 milliseconds. The experiment first ran for two weeks on a small fraction of Bing users, and we observed an impact to click-through-rate (CTR) that was borderline statistically significant, i.e., the p-value was just slightly below our threshold of 0.05. To confirm that the treatment effect on this metric is real and not a false positive, a much larger experiment was run, which showed that this was indeed a real effect with a p-value of 2e-13.

We applied CUPED using CTR from the 2-week pre-period as the covariate. The result is impressive: the delta was statistically significant from day 1! The top plot of Figure 2 shows the p-values over time in log scale. The black horizontal line is the 0.05 significance bar. The vanilla t-test trends slowly down and by the time the experiment was stopped in 2 weeks, it barely reached the threshold. When CUPED is applied, the entire p-value curve is below the bar. The bottom plot of Figure 2 compares the p-value curves when CUPED runs on only half the users. Even with half the users exposed to the experiment, CUPED results in a more sensitive test, allowing for more non-overlapping experiments to be run. While most experiments are not known to be negative to the user experience a-priori, it has been well documented that most experiments are flat or negative (Kohavi et al. 2009a; Manzi 2012), so being able to run

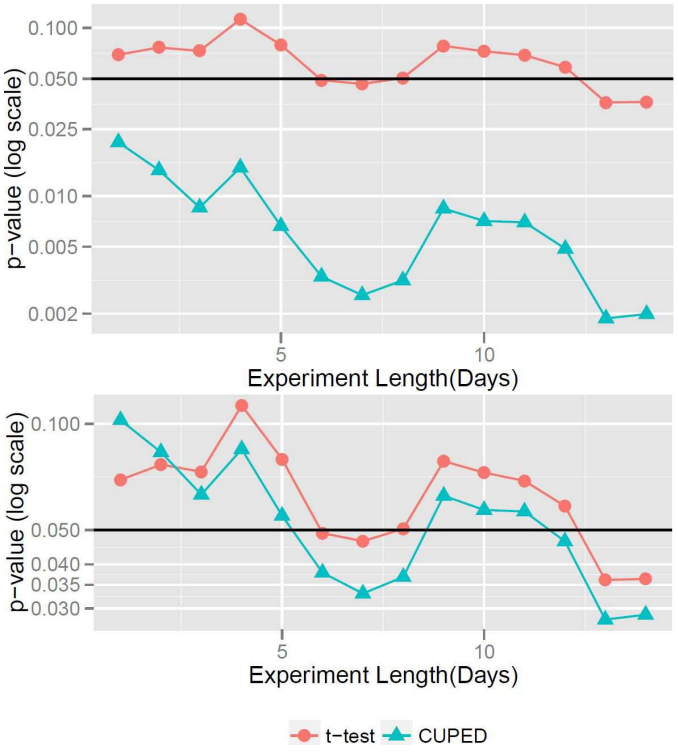

Figure 2: Slowdown experiment. Top: p-value. Bottom: p-value when using only half the users for CUPED.

experiments with the same statistical power on smaller populations of users is very important.

Applying CUPED to other experiments at Bing has resulted in similar increases to statistical power.

5.2 Factors Affecting CUPED Effectiveness

In this section we look at factors that can have a large impact on the success of using CUPED in practice. The data we use here are from a 3-week A/A experiment.

5.2.1 Covariates

Figure 3 shows CUPED's variance reduction rate for the metric queries-per-user. Two covariates are considered: (1) entry-day, which is a categorical variable indicating the first day a user appears in the experiment and (2) queries-per-user in the 1-week pre-experiment period. Note that the entry-day is not pre-experiment data, but it satisfies the condition required in Section 4.3 because the treatment has no effect on when a user will come for the first time during the experiment.

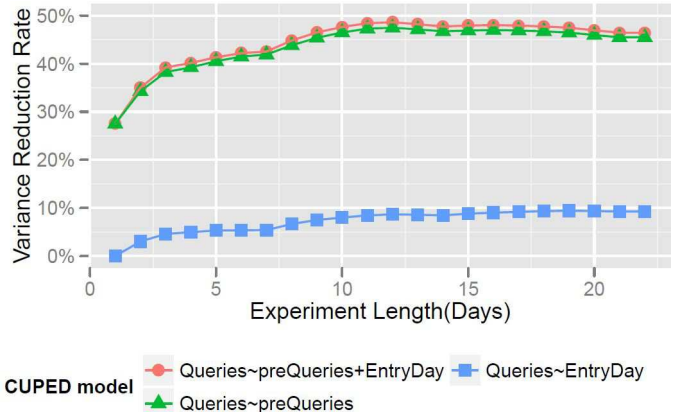

Figure 3: Variance reduction for Queries/UU using different covariates.

Note that the full experiment ran for 3 weeks, but the reduction rate is evaluated and plotted as the experiment accumulates data up until to the full 3 weeks. From the plot we can see that when only entry-day is used as the covariate, the variance reduction rate increases as the experiment runs longer and stays at about 9% to 10% after 2 weeks. On the other hand, using only the pre-experiment period queries-per-user, variance reduction rate can reach more than 45%. When we combine the two covariates together, we only gain an extra 2% to 3% more reduction compared to the pre-experiment queries-per-user alone. This suggests that the same metric computed in the pre-experiment period is a better single covariate. That is intuitive since the same metric in the experiment and pre-experiment periods should naturally have high correlation. The fact that when combining two covariates together the marginal variance reduction is small means most correlation between entry-day and queries-per-user can be "explained away" by the pre-experiment queries-per-user. More precisely, the partial correlation between entry-day and queries-per-user given pre-experiment queries-per-user is low.

5.2.2 Lengths of Experiment and Pre-Experiment

In addition to the effect of covariates, Figure 3 also provides insights on the impact of the experiment length. An interesting observation is that the variance reduction rate for the pre-experiment covariate (green/triangle) is not monotonically increasing as the experiment duration increases. It reaches the maximum at about 2 weeks and then starts to slowly decrease. To help understand the underlying reasons for this trend, we plot 4 variations of this curve with pre-period length varying from 3 days to 2 weeks. As shown in Figure 4, we see that a longer pre-period gives a higher variance reduction rate.

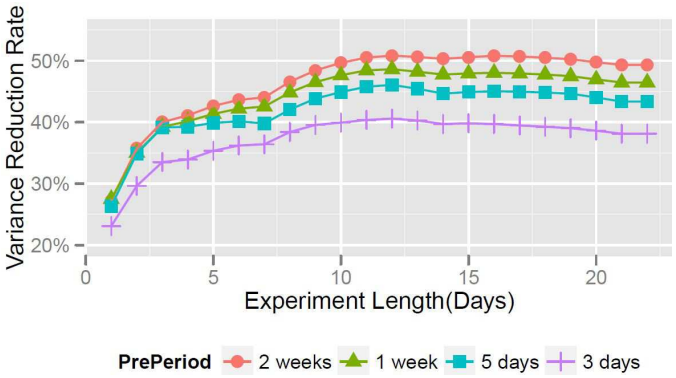

Figure 4: Impact of pre-experiment period length.

The trends in Figure 3 and 4 together reveal two conflicting factors that impact the effectiveness of CUPED.

- *Correlation.* The higher the correlation, the better the variance reduction. When we increase the pre-experiment period length or increase the experiment duration, the correlation increases for "cumulative" metrics such as queries-per-user. This is because the longer the period, the higher the signal-to-noise ratio.
- *Coverage.* Coverage is the percentage of users in the experiment that also appeared in the pre-experiment period. Coverage is determined by:
  - *Experiment Duration.* As the experiment duration increases, coverage decreases. The reason is that frequent visitors are seen early in the experiment and users seen later in the experiment are often new or "churned" users. The result is that as coverage decreases, the rate of variance reduction goes down.
  - *Pre-period Duration.* Increasing the pre-period length increases coverage because we have a better chance of matching an experiment user in the pre-period.

When pre-experiment period was chosen to be 2 weeks, the variance reduction rate is about 50% for a large range of experiment durations.

5.2.3 Metric of Interest

Besides the choice of covariates and the lengths of experiment and pre-experiment, CUPED effectiveness varies from metric to metric, depending on the correlation between the metric and its pre-experiment counterpart. We applied CUPED on a few metrics, such as clicks-per-user and visits-per-user. CUPED performed well on all these metrics with similar variance reduction curves as in Figure 3. One notable exception is revenue-per-user, where CUPED reduced the variance by less than 5% due to the low correlation of revenue-per-user between the pre-experiment and the experiment periods. We also applied CUPED on a few page level

metrics (see the discussion in Section 4.4). Figure 5 shows the reduction rate for click-through-rate (CTR), using 2-week pre-experiment CTR as the CUPED covariate. Figure 6 plots the correlation curve between the two periods. Both the variance reduction rate and the correlation are similar to those of queries-per-user.

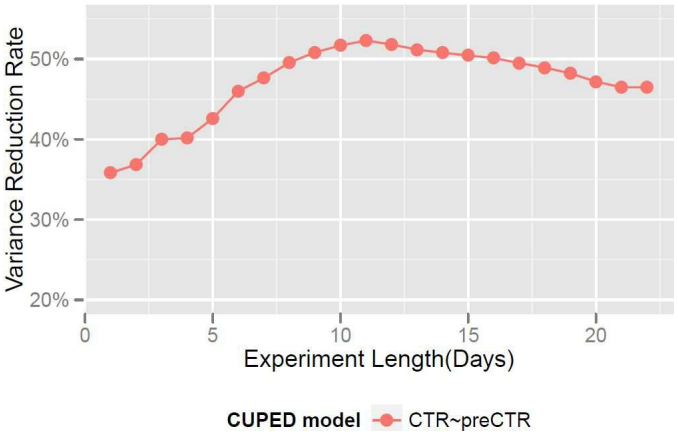

Figure 5: Variance reduction rate for CTR.

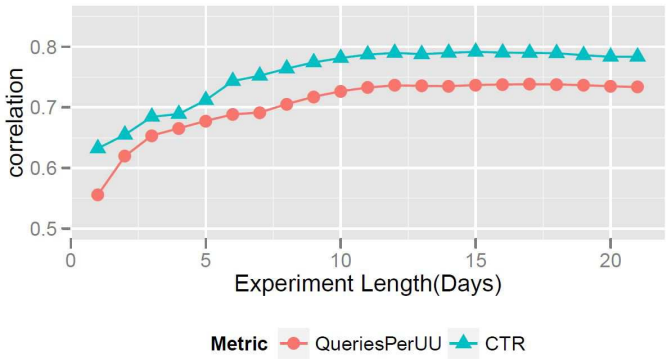

Figure 6: Correlation between the metrics of interest and their covariates within the matched users.

5.3 Warning on Using Post-Triggering Data

In Section 4, we mentioned that to guarantee unbiased results, any covariate X has to satisfy the condition $\mathbb{E}(X^{(t)}) = \mathbb{E}(X^{(c)})$. We illustrate this point in Figure 7. The data is from another Bing experiment, where queries-per-user increased statistically significantly. If we were to look for a metric with high correlation to queries-per-user, there is actually a much better candidate than the pre-experiment queries-per-user: the "in-experiment" Distinct Queries-per-user (DQ-per-user). DQ is defined as query counts for a user after consecutive duplicated queries are removed. As a result, DQ is extremely highly correlated with queries-per-user. Seemingly, DQ-per-user as covariate sounds like a good idea. However, Figure 7 shows that the CUPED estimates Δ_{CUPED} are negative and the confidence intervals are almost always below 0 (Note how narrow the confidence intervals are. DQ-per-user indeed reduced variance a lot!). This suggests the queries-per-user difference between the treatment and control is negative with 95% confidence, a result that is directionally opposite of the known effect. The contradiction is only apparent because DQ-per-user does not satisfy $\mathbb{E}(X^{(t)}) = \mathbb{E}(X^{(c)})$. In fact, since we know treatment has larger queries-per-user, it has larger DQ-per-user too. By (7), CUPED estimate Δ_{cv} can be interpreted as the in-experiment delta for the metric of interest "corrected" by the delta for the covariate. In this case the covariate delta is also positive, driving down the CUPED estimate below 0. This example illustrates the pitfall when extending CUPED beyond using pre-experiment (or pre-triggering) data.

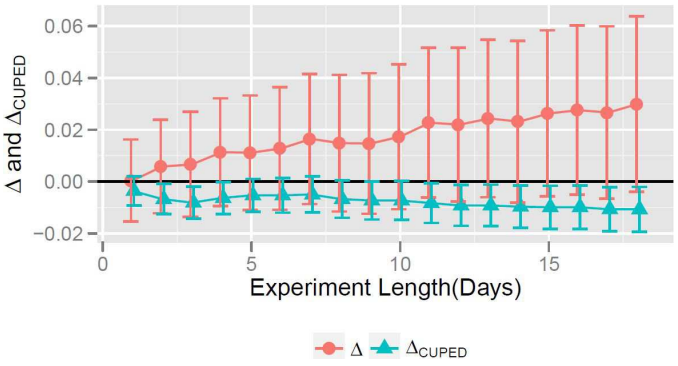

Figure 7: Example where results are directionally incorrect when covariates violate the pre-triggering requirement.

6. CONCLUSIONS

Increasing the sensitivity of online experiments allows for more precise assessment of value, or equivalently running the experiments on smaller populations (supporting more non-overlapping experiments) or for shorter durations (improving the feedback cycle and agility). We introduced CUPED, a new technique for increasing the sensitivity of controlled experiments by utilizing pre-experiment data. CUPED is currently live in Bing's online experimentation system. Three important recent experiments showed variance reductions of 45%, 52% and 49% with one week of experiment and one week of pre-experiment data. This reassures that CUPED can indeed help us effectively achieve the same statistical power with about only half the users, or half the duration.

CUPED is widely applicable to organizations running online experimentation systems because of its simplicity, ability to be added easily to existing systems, and its support for metrics commonly used in online businesses. Based on our experience applying CUPED at Bing, we can make the following recommendations for others interested in applying CUPED to their online experiments:

- Variance reduction works best for metrics where the distribution varies significantly across the user population. One common class of such metrics where the value is very different for light and heavy users. Queries-per-user is a paradigmatic example of such a metric.
- Using the metric measured in the pre-period as the covariate typically provides the best variance reduction.
- Using a pre-experiment period of 1-2 weeks works well for variance reduction. Too short a period will lead to poor matching, whereas too long a period will reduce correlation with the outcome metric during the experiment period.
- Never use covariates that could be affected by the treatment, as this could bias the results. We have

shown an example where directionally opposite conclusions could results if this requirement is violated.

While CUPED significantly improves the sensitivity of online experiments, we would like to explore improvements:

- *Optimized Covariate Selection*. Extend CUPED to optimize the selection of covariates, both for particular metrics and for particular types of experiments (e.g., a backend experiment might use data center as a covariate). We also plan to study the theory and practice to optimize selection of multiple covariates from a large library of potential covariate variables.
- *Incorporating Covariate Information into Assignment*. Rather than adjusting the data after the experiment completes, if we can make randomization aware of covariates we can potentially improve the sensitivity of our experiments even more as well as allocating traffic more efficiently to experiments.

7. ACKNOWLEDGMENTS

We wish to thank Sam Burkman, Leon Bottou, Thomas Crook, Kaustav Das, Brian Frasca, Xin Fu, Mario Garzia, Li-wei He, Greg Linden, Roger Longbotham, Carlos Gomez Uribe, Zijian Zheng and many members of the Bing Data Mining team.

References

Soren Asmussen and Peter Glynn. *Stochastic Simulation*. Springer-Verlag, 2008.

George E. P. Box, J. Stuart Hunter, and William G. Hunter. *Statistics for experimenters: design, innovation, and discovery*. Wiley, 2005.

Olivier Chapelle, Thorsten Joachims, Filip Radlinski, and Yisong Yue. Large-scale validation and analysis of interleaved search evaluation. *ACM Trans. Inf. Syst.*, 30(1): 6:1–6:41, March 2012. ISSN 1046-8188.

comScore. comscore releases june 2012 u.s. search engine rankings, June 2012. `http://www.comscore.com/Press_Events/Press_Releases/2012/6/comScore_Releases_June_2012_U.S._Search_Engine_Rankings`.

M. Davidian, A.A. Tsiatis, and S. Leon. Semiparametric estimation of treatment effect in a pretest-posttest study with missing data. *Statistical Science*, 20, 2005.

Shaojie Deng, Roger Longbotham, Toby Walker, and Ya Xu. Choice of the randomization unit in online controlled experiment. *JSM proceedings*, 2011.

Andrew Gelman and Jennifer Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, 2006.

Geoffrey Keppel, William H. Saufley, and Howard Tokunaga. *Introduction to design and analysis: A Student's Handbook*. Worth Publishers, 1992.

Ron Kohavi, Thomas Crook, and Roger Longbotham. Online experimentation at Microsoft. *Third Workshop on Data Mining Case Studies and Practice Prize.*, 2009a. `http://www.exp-platform.com/Pages/expMicrosoft.aspx`.

Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M. Henne. Controlled experiments on the web: survey and practical guide. *Data Mining Knowledge Discovery*, 18:140–181, 2009b. `http://www.exp-platform.com/Pages/hippo_long.aspx`.

Ron Kohavi, Alex Deng, Brian Frasca, Roger Longbotham, Toby Walker, and Ya Xu. Trustworthy online controlled experiments: Five puzzling outcomes explained. *Proceedings of the 18th Conference on Knowledge Discovery and Data Mining*, 2012. `http://www.exp-platform.com/Pages/PuzzingOutcomesExplained.aspx`.

S. Leon, A.A. Tsiatis, and M. Davidian. Semiparametric estimation of treatment effect in a pretest-posttest study. *Biometrika*, 59, 2003.

Jim Manzi. *Uncontrolled: The Surprising Payoff of Trial-and-Error for Business, Politics, and Society*. Basic Books, 2012.

Robert L. Mason, Richard F. Gunst, and James L. Hess. *Statistical design and analysis of experiments with applications to engineering and science*. Wiley, 1989.

Student. The probable error of a mean. *Biometrika*, 6:1–25, 1908.

Diane Tang, Ashish Agarwal, Deirdre O'Brien, and Mike Meyer. Overlapping experiment infrastructure: More, better, faster experimentation (presentation). 2010a. URL `http://research.google.com/archive/papers/Overlapping_Experiment_Infrastructure_More_Be.pdf`.

Diane Tang, Ashish Agarwal, Deirdre O'Brien, and Mike Meyer. Overlapping experiment infrastructure: More, better, faster experimentation. *Proceedings of the 16th Conference on Knowledge Discovery and Data Mining*, 2010b.

Anastasios Tsiatis. *Semiparametric Theory and Missing Data*. Springer-Verlag, 2006.

Anastasios A. Tsiatis, Marie Davidian, Min Zhang, and Xiaomin Lu. Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach. *Statistics in Medicine*, 27, 2008.

Larry Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer, 2003.

Wikipedia. A/b testing. `http://en.wikipedia.org/wiki/A/B_testing`.

Li Yang and Anastasios A. Tsiatis. Efficiency study of estimators for a treatment effect in a pretest-posttest trial. *The American Statistician*, 55, 2001.

APPENDIX

A. CONTROL VARIATES AS AN EXTENSION OF STRATIFICATION

Here we show that when the covariates are categorical, stratification and control variates produce identical results.

For clarity and simplicity, we assume X is binary with values 1 and 0. Let $w = \mathbb{E}(X)$. The two estimates are

$$\widehat{Y}_{strat} = w\overline{Y}_1 + (1-w)\overline{Y}_0,$$
$$\widehat{Y}_{cv} = \overline{Y} - \hat{\theta}\overline{X} + \hat{\theta}w,$$

where $\overline{Y}_1$ denotes the average of Y in the $\{X=1\}$ stratum and $\hat{\theta} = \widehat{\text{cov}}(Y,X)/\widehat{\text{var}}(X) = \overline{Y}_1 - \overline{Y}_0$. Plugging in the expression for $\hat{\theta}$, we have

$$\begin{aligned}\widehat{Y}_{cv} &= \overline{Y} - (\overline{Y}_1 - \overline{Y}_0)\overline{X} + (\overline{Y}_1 - \overline{Y}_0)w \\ &= (1-\overline{X})\overline{Y}_0 + \overline{Y}_0\overline{X} + (\overline{Y}_1 - \overline{Y}_0)w \\ &= w\overline{Y}_1 + (1-w)\overline{Y}_0 = \widehat{Y}_{strat},\end{aligned}$$

where the second equality follows from the fact that $\overline{Y} = \overline{XY}_1 + (1-\overline{X})\overline{Y}_0$.

To prove for the case with $K > 2$, we construct $K-1$ indicator variables as control variates. With the observation that the coefficients $\hat{\theta}_k = \overline{Y}_k - \overline{Y}_0$, the proof follows the same steps as the binary case outlined above.

B. GENERALIZATION TO OTHER ANALYSIS UNIT

As we mentioned in Section 4, to achieve variance reduction for non-user level metrics, we need to incorporate delta method. The formulation we lay out in Section 3 makes it easy to achieve this, as we will see below.

We use CTR as an example and derive for the control variates formulation since it's more general.

Let n be the number of users (non-random). Denote $Y_{i,j}$ the number of clicks on user i's jth page-view during the experiment and $X_{i,k}$ the number of clicks on user i's kth page-view during the pre-experiment period. Let N_i and M_i be the numbers of page-views from user i during the experiment and pre-experiment respectively. The estimate for CTR in Eq. (3) using $X_{i,j}$ as the control variate becomes

$$\begin{aligned}\widehat{Y}_{cv} &= \frac{\sum_{i,j} Y_{i,j}}{\sum_{i,j} 1} - \theta\frac{\sum_{i,k} X_{i,k}}{\sum_{i,k} 1} + \theta\mathbb{E}(X_{i,k}) \\ &= \frac{\sum_i Y_{i,+}}{\sum_i N_i} - \theta\frac{\sum_i X_{i,+}}{\sum_i M_i} + \theta\mathbb{E}(X_{i,j}),\end{aligned}$$

where $Y_{i,+} = \sum_j Y_{i,j}$ is the total number of clicks from user i. Similar notation applies to $X_{i,+}$.

Following the same derivation as in Section 3.2.1, we know $\text{var}(\widehat{Y}_{cv})$ is minimized at

$$\begin{aligned}\theta &= \text{cov}\left(\frac{\sum_i Y_{i,+}}{\sum_i N_i}, \frac{\sum_i X_{i,+}}{\sum_i M_i}\right) / \text{var}\left(\frac{\sum_i X_{i,+}}{\sum_i M_i}\right) \\ &\doteq \text{cov}\left(\frac{\overline{Y}}{\mu_N} - \frac{\mu_Y\overline{N}}{\mu_N^2} - \frac{\mu_Y}{\mu_N}, \frac{\overline{X}}{\mu_M} - \frac{\mu_X\overline{M}}{\mu_M^2} - \frac{\mu_X}{\mu_M}\right) \\ &\qquad / \text{var}\left(\frac{\overline{X}}{\mu_M} - \frac{\mu_X\overline{M}}{\mu_M^2} - \frac{\mu_X}{\mu_M}\right) \\ &= \text{cov}\left(\frac{\overline{Y}}{\mu_N} - \frac{\mu_Y\overline{N}}{\mu_N^2}, \frac{\overline{X}}{\mu_M} - \frac{\mu_X\overline{M}}{\mu_M^2}\right) / \text{var}\left(\frac{\overline{X}}{\mu_M} - \frac{\mu_X\overline{M}}{\mu_M^2}\right)\end{aligned} \tag{8}$$

where the second equality follows from using Taylor expansion to linearize the ratios and $\overline{Y} = \frac{1}{n}\sum_i Y_{i,+}$ with $\mu_Y = \mathbb{E}(\overline{Y})$ (similarly for μ_X, μ_N and μ_M).

Because the user is the randomization unit and user level observations are i.i.d., we have

$$\sqrt{n}\left(\overline{Y}, \overline{N}, \overline{X}, \overline{M}\right) \Rightarrow N(\mu, \Sigma),$$

following a multivariate normal distribution with mean vector μ and covariance matrix Σ easily estimated from the i.i.d. samples.

It is now straight forward to estimate θ in Eq. 8 using

$$\theta = (\beta_1^T\Sigma\beta_2)/(\beta_2^T\Sigma\beta_2),$$

where $\beta_1 = (1/\mu_N, -\mu_Y/\mu_N^2, 0, 0)^T$ and $\beta_2 = (0, 0, 1/\mu_M, -\mu_X/\mu_M^2)^T$ are the coefficients in Eq. 8.

Note that in the example above, both the metric of interest (CTR) and the covariate metric are at page-view level. We can easily see that the derivation works generally for various combinations. The metric can be at user level while the covariate can be at page-view level, etc. This opens door to a whole new class of covariates which are based on features specific to a page not to a user. Finally, it is easy to see that the case with multiple control variates follow similarly.