From Augmentation to Decomposition: A New Look at CUPED in 2023

Alex Deng[1] **Luke Hagar**[2] **Nathaniel Stevens** [2] **Tatiana Xifara**[1]
Lo-Hua Yuan[1] **Amit Gandhi**[1]
[1] Airbnb [2] University of Waterloo
{alex.deng, tatiana.xifara, lohua.yuan, amit.gandhi}@airbnb.com
{nstevens, lmhagar}@uwaterloo.ca

Abstract

Ten years ago, CUPED (Controlled Experiments Utilizing Pre-Experiment Data) (Deng et al., 2013) mainstreamed the idea of variance reduction leveraging pre-experiment covariates. Since its introduction, it has been implemented, extended, and modernized by major online experimentation platforms (Xie and Aurisset, 2016; Guo et al., 2021; Poyarkov et al., 2016; Jin and Ba, 2023; Cosgrove et al., 2022). Many researchers and practitioners often interpret CUPED as a regression adjustment Lin (2013); Tsiatis et al. (2008). In this article, we clarify its similarities and differences to regression adjustment and present CUPED as a more general augmentation framework which is closer to the spirit of the 2013 paper. We show that the augmentation view naturally leads to cleaner developments of variance reduction beyond simple average metrics, including ratio metrics and percentile metrics. Moreover, the augmentation view can go beyond using pre-experiment data and leverage in-experiment data, leading to significantly larger variance reduction. We further introduce metric decomposition using approximate null augmentation (ANA) as a mental model for in-experiment variance reduction and studied it under both a Bayesian framework and a frequentist optimal proxy metric framework. Metric decomposition also arises naturally in conversion funnels.

1 Introduction

The CUPED method proposed by Deng et al. (2013) was inspired by the method of control variates from stochastic simulation (Asmussen and Glynn, 2008; Owen, 2013). CUPED is a model-free method that relies only on the observation that any pre-experiment difference between two randomized groups is pure noise due to randomization and should be 0 in expectation. The model-free aspect is at the heart of CUPED; there is no assumption of a relationship of any form between the covariates and the target metric, as long as they have a nonzero correlation to exploit. Also, the authors developed the theory directly on estimators (of metrics and Δ's), instead of modeling individual subject-level data points like regression models do. However, many blogs and papers citing CUPED often interpret it as a regression adjustment method, focusing on defining it as averages of individual residuals.

2023 marks a decade since CUPED's initial publication. In this work, we present CUPED as an augmentation framework which is closer to the spirit of the in initial proposal in the 2013 paper. We show that the augmentation view naturally leads to variance reduction beyond simple average metrics; it easily handles ratio metrics and percentile metrics as well. Moreover, the augmentation view can go beyond using pre-experiment data and leverage in-experiment data, leading to significantly larger variance reduction. We further introduce metric decomposition as a mental model for in-experiment variance reduction and present a Bayesian variance reduction framework. Metric decomposition also arises naturally in conversion funnels

Notation Henceforth, we denote the observed metric value from the treatment and control groups as M_t and M_c, the unknown true average treatment effect (ATE) as δ, and we use $\widehat{\delta} = M_t - M_c$ to denote the naive estimate of δ based on the difference between two metric values. This is also denoted as Δ.

2 CUPED as Augmentation

2.1 Three Simple Insights that Underlie CUPED

First insight: Augmentation with Mean-Zero Term For any estimator $\widehat{\delta}$ of δ, define a new estimator

$$\widehat{\delta}^* = \widehat{\delta} + \widehat{\delta}_0 \tag{1}$$

Preprint. Under review.

such that $\mathbb{E}\left[\widehat{\delta}_0\right]=0$. Then $\widehat{\delta}^*$ is also an estimator of δ and is unbiased if $\widehat{\delta}$ is. Thus, the CUPED estimator is a mean-preserving augmentation of any existing estimator. The purpose of the augmentation is to find an augmentation such that $\mathrm{Var}\left[\widehat{\delta}^*\right]<\mathrm{Var}\left[\widehat{\delta}\right]$.

Second Insight: Optimal Variance Reduction from a Linear Family The second insight of CUPED is that variance reduction is almost guaranteed with any mean-zero augmentation $\widehat{\delta_0}$! This is because any mean-zero augmentation can be multiplied by a scalar θ yielding a whole family of mean-zero augmentations $\widehat{\delta}^*(\theta)=\widehat{\delta}+\theta\widehat{\delta_0}$, with variance $\mathrm{Var}\left[\widehat{\delta}^*(\theta)\right]=\mathrm{Var}\left[\widehat{\delta}\right]+\theta^2\mathrm{Var}\left[\widehat{\delta}_0\right]+2\theta\mathrm{Cov}\left[\widehat{\delta},\widehat{\delta}_0\right]$. This variance is minimized when

$$\theta:=-\frac{\mathrm{Cov}\left[\widehat{\delta}_0,\widehat{\delta}\right]}{\mathrm{Var}\left[\widehat{\delta}_0\right]}$$

with minimum variance

$$\mathrm{Var}\left[\widehat{\delta}^*(\theta)\right]=\mathrm{Var}\left[\widehat{\delta}\right]\times\left(1-\frac{\mathrm{Cov}\left[\widehat{\delta}_0,\widehat{\delta}\right]^2}{\mathrm{Var}\left[\widehat{\delta}\right]\mathrm{Var}\left[\widehat{\delta}_0\right]}\right)=\mathrm{Var}\left[\widehat{\delta}\right]\times\left(1-\mathrm{Corr}\left[\widehat{\delta},\widehat{\delta}_0\right]^2\right).$$

The amount of variance reduction from augmentation by $\widehat{\delta_0}$ is equal to the correlation between the augmentation term and the existing estimator $\mathrm{Corr}\left[\widehat{\delta},\widehat{\delta}_0\right]^2$. Any mean-zero augmentation is like a direction of gradient descent, and variance is optimally reduced with an appropriate choice of step size.

Third Insight: Existence of Mean-Zero Augmentation The final insight of CUPED is that mean-zero augmentations are abundant if we tap into pre-experiment period data. Let $M(\mathbf{X}_{pre})$ be any metric computed from a (multivariate) signal $\mathbf{X}$ on pre-experiment period data (e.g., $\overline{f(\mathbf{X})}$ or a percentile, etc.). Then $\Delta_{pre}(M)=M(\mathbf{X}_{pre})_t-M(\mathbf{X}_{pre})_c$ is a mean-zero augmentation.

2.2 Relation to Regression Adjustment

When the metric of interest is a simple average and the naive estimator of ATE is the difference in means $\widehat{\delta}=\Delta(Y)$, and the augmentation is also a difference in means $\Delta(Y)^*=\Delta(Y)-\theta\Delta(X)$, people often compare CUPED to regression adjustment of the form

$$Y_i=\alpha+\delta A_i+\beta X_i+\epsilon_i. \tag{ANCOVA1}$$

and

$$Y_i=\alpha+\delta A_i+\beta X_i+(\gamma X_i)A_i+\epsilon_i, \tag{ANCOVA2}$$

where A is the assignment indicator. Tsiatis et al. (2008) showed both ANCOVA1 and ANCOVA2 estimate δ asymptotically as

$$\overline{Y_t}-\overline{Y_c}-(\overline{f(X_t)}-\overline{f(X_c)}),$$

for some function f. Therefore, we can see both ANCOVA1 and ANCOVA2 asymptotically is a special form of CUPED. The differences between ANCOVA1 and ANCOVA2 are the choice of how to fit the function $f(X)$ using treatment and control data. The linear regression coefficient estimator is $\mathrm{Cov}(X,Y)/\mathrm{Var}[X]$. the denominator $\mathrm{Var}[X]$ is the same for treatment and control. $\mathrm{Cov}[X,Y]$ are different due to the treatment effect. ANCOVA1 pools the data together and fit a linear regression. This is to use

$$\mathrm{Cov}[X,Y]=p\mathrm{Cov}_T[X,Y]+(1-p)\mathrm{Cov}_C[X,Y],$$

where $\mathrm{Cov}_T(X,Y)$ is the covariance in the treatment, $\mathrm{Cov}_C(X,Y)$ for control and p is the proportion of treatment sample size. On the contrary, ANCOVA2 uses

$$\mathrm{Cov}[X,Y]=(1-p)\mathrm{Cov}_T[X,Y]+p\mathrm{Cov}_C[X,Y].$$

For CUPED, the optimal θ is

$$\theta^*=\frac{Cov[\overline{Y_t},\overline{X_t}]+Cov[\overline{Y_c},\overline{X_c}]}{Var[\overline{X_t}]+Var[\overline{X_c}]}. \tag{2}$$

This θ asymptotically converge to $(1-p)\mathrm{Cov}_T[X,Y]+p\mathrm{Cov}_C[X,Y]$. Hence CUPED with this arrangement of θ is asymptotically equivalent to ANCOVA2. Because $\mathrm{Cov}[\overline{Y_t},\overline{X_t}]=\mathrm{Cov}_T[X,Y]/n_t$ and $\mathrm{Cov}[\overline{Y_c},\overline{X_c}]=\mathrm{Cov}_C[X,Y]/n_c$, we can see covariances are weighted inversely proportional to the sample sizes. This explains why it is better to weight $\mathrm{Cov}_T[X,Y]$ by $1-p$ and $\mathrm{Cov}_C[X,Y]$ by p, instead of using a more straightforward choice of p for $\mathrm{Cov}_T[X,Y]$ and $1-p$ for $\mathrm{Cov}_C[X,Y]$ as in ANCOVA1. From here we also see ANCOVA2 is theoretically better than ANCOVA1, as advocated by (Lin, 2013). In practice this difference is small unless p is away from 0.5 and $\mathrm{Cov}_T[X,Y]$ is very different from $\mathrm{Cov}_C[X,Y]$. When $p=0.5$, ANCOVA1 and ANCOVA2 are equivalent.

3 Advantages of the Augmentation View

CUPED is asymptotically equivalent to ANCOVA2 when applied to simple average metrics, and also ANCOVA1 when in addition treatment and control are equal size. But the advantage of CUPED is its augmentation view can naturally lead to variance reduction beyond simple average metrics. The augmentation term does not even need to be related to difference of metric values observed in treatment and control groups.

Flexible Metric Form As an augmentation to any estimator of interest, it is clear that the theory of CUPED doesn't depend on the metric being an average; CUPED can also be applied to percentile metrics and ratio metrics straightforwardly. These are common challenges when practitioners try to implement CUPED with the regression residual interpretation.

Flexible Augmentation Form The augmentation $\widehat{\delta_0}$ does not have to be in the form of a difference of two metric values. One recent development of this idea is illustrated in Deng et al. (2023b), where the augmentation term $\widehat{\delta_0}$ is constructed from matching and balancing methods in observational causal inference.

4 Metric Decomposition with Approximately Null Augmentation (ANA)

We can take the augmentation view further into the realm where the augmentation is not guaranteed to have mean zero. This is motivated by the idea of metric decomposition, where we decompose a metric into two components where treatment effects are believed to be mostly captured in one of the two components.

To be more specific, let $M = M_1 + M_2$, $\Delta(M) = \Delta(M_1) + \Delta(M_2)$, and the true effect also has the decomposition $\delta = \delta_1 + \delta_2$. If δ_1 is close to 0 in most cases, and M_1 accounts for a significant portion of variation in M, treating $\Delta(M_1)$ as an augmentation and using $\Delta(M_2) = \Delta(M) - \Delta(M_1)$ can yield significant variance reduction.

Compared to CUPED, we no longer have the guarantee that the augmentation has mean zero. We call this scenario approximate null augmentation (**ANA**). Let $\underset{\sim}{\Delta} = (\Delta_1, \Delta_2)$ be the decomposed vector of $\Delta(M_1)$ and $\Delta(M_2)$, $\underset{\sim}{\delta} = (\delta_1, \delta_2)$ be the vector of true effect for the two components. Then

$$\underset{\sim}{\Delta} \sim \underset{\sim}{\delta} + \underset{\sim}{\varepsilon},$$

where $\underset{\sim}{\varepsilon}$ is has known covariance matrix $\Sigma(n)$ and is assumed to be approximately normally distributed from central limit theorem. The effect $\underset{\sim}{\delta}$ follows a bivariate distribution with covariance Λ. Our inference target is $\delta = \delta_1 + \delta_2$.

Taking an empirical Bayes approach, assuming $\underset{\sim}{\delta}$ has mean $(0,0)$, we can estimate Λ from a set of historical experiments with many realization of $\underset{\sim}{\Delta}$.

4.1 Bayesian Variance Reduction

The first research question we studied is how decomposing a metric into two more more components changes the Bayesian posterior distribution. In the simple normal-normal model, assuming $\underset{\sim}{\delta}$ has mean 0 and covariance Λ, we know

$$\mathrm{E}\left[\underset{\sim}{\delta} | \underset{\sim}{\Delta}\right] = S\underset{\sim}{\Delta}\ , \quad \mathrm{Var}\left[\underset{\sim}{\delta} | \underset{\sim}{\Delta}\right] = (I - S)\Lambda\ ,$$

where $S = \Lambda(\Lambda + \Sigma)^{-1}$. Since $\delta = (1,1) \cdot \underset{\sim}{\delta}$, we derive the posterior mean to be

$$\begin{aligned} \mathrm{E}[\delta | \Delta_1, \Delta_2] = {} & C^{-1}[(\Lambda_{11} + \Lambda_{12})(\Lambda_{22} + \sigma_{22}) - (\Lambda_{12} + \sigma_{12})(\Lambda_{12} + \Lambda_{22})]\Delta_1 \\ & + C^{-1}[(\Lambda_{12} + \Lambda_{22})(\Lambda_{11} + \sigma_{11}) - (\Lambda_{12} + \sigma_{12})(\Lambda_{11} + \Lambda_{12})]\Delta_2, \end{aligned} \tag{3}$$

where $C = \Lambda_{11}\Lambda_{22} + \Lambda_{11}\Sigma_{22} + \Lambda_{22}\Sigma_{11} + \Sigma_{11}\Sigma_{22} - \Lambda_{12}^2 - 2\Lambda_{12}\Sigma_{12} - \Sigma_{12}^2$.

Alternatively, without the ANA metric decomposition,

$$\mathrm{E}[\delta | \Delta] = A\Delta\ , \quad \mathrm{Var}[\delta | \Delta] = A\sigma^2\ ,$$

where $A = \frac{\lambda^2}{\lambda^2 + \sigma^2}$, $\lambda^2 = \Lambda_{11} + \Lambda_{12} + \Lambda_{21} + \Lambda_{22}$ and $\sigma^2 = \Sigma_{11} + \Sigma_{12} + \Sigma_{21} + \Sigma_{22}$.

ANA metric decomposition naturally lead to variance reduction under the Bayesian framework. We proved that

$$1^T \cdot (I - S)\Lambda \cdot 1 < A\sigma^2.$$

The proof is omitted here and we will demonstrate using empirical results.

4.2 Frequentist Optimal Proxy Metric with Variance Reduction

As an extension of CUPED, ANA can be used as a frequentest estimator and analyzed as a proxy metric of the form $\Delta^* := \Delta_2 + \theta\Delta_1$. Comparing to Bayesian posterior mean, the main difference is that we do not put any shrinkage factor on the signal component Δ_2, and only shrink the ANA component Δ_1.

Theorem 1. *Among ANA estimators* $\Delta^* := \Delta_2 + \theta\Delta_1$*, the mean squared error* $\mathrm{E}[(\delta - \Delta^*)^2]$ *is minimized when*

$$\theta = \frac{\Lambda_{11} - \Sigma_{12}}{\Lambda_{11} + \Sigma_{11}}. \tag{4}$$

The correlation between δ *and* Δ^* *is maximized when*

$$\theta = \frac{(\Lambda_{12} + \Sigma_{12})(\Lambda_{12} + \Lambda_{22}) - (\Lambda_{11} + \Lambda_{12})(\Lambda_{22} + \Sigma_{22})}{(\Lambda_{12} + \Sigma_{12})(\Lambda_{11} + \Lambda_{12}) - (\Lambda_{12} + \Lambda_{22})(\Lambda_{11} + \Sigma_{11})}. \tag{5}$$

Minimizing Effect Prediction Error ANA estimator with (4) minimizes the effect prediction error. It is a generalization of CUPED in the sense that when $\Lambda_{11} = 0$, $\theta = -\Sigma_{12}/\Sigma_{11}$ reduces to CUPED.

Maximizing Correlation The objective to maximize correlation between the true effect and the estimate is proposed in Tripuraneni et al. (2023). Comparing (5) to (3), we see the ANA estimator $\Delta_2 + \theta\Delta_1$ maximizing the correlation between δ and Δ^* is simply a rescaled Bayesian posterior mean estimator such that Δ_2 receives no shrinkage.

5 Empirical Results

We applied Approximate Null Augmentation to 25 early stage ranking experiments at Airbnb. These early stage experiments all run for about 1 week taking a small percentage of total traffic. The main target metric of interest is booking per guest. To construct the ANA, we leverage counterfactual ranking results. That is, for each search, we compare the ranked results produced by the treatment and control ranker. If a click on a booked listing is ranked in close proximity according to both rankers, or if a click is from map and both rankers would show the listing on the map, then the attributed booking from this click would be approximately the same regardless of the treatment assignment. In addition, we use a utility model to attribute a user's booking to searches (Deng et al., 2023a). The end result is for each booking, we can construct an approximate null augmentation component representing a fraction of the booking that both rankers would have contributed almost equally.

The covariance of effect Λ and the average covariance of the noise $\underset{\sim}{\varepsilon}$ were estimated to be (after scaled by the same constant)

$$\Lambda = \begin{pmatrix} 0.576 & -0.896 \\ -0.896 & 4.329 \end{pmatrix} \quad \text{and} \quad E[\Sigma] = \begin{pmatrix} 4.020 & 0.169 \\ 0.169 & 0.811 \end{pmatrix}.$$

We found the first ANA component Δ_1 displays a variance 5 times of the second component Δ_2; while the variance of the effect for the ANA component is less less than 1/7 of Δ_2. If the ANA component has theoretical mean 0, then the variance of the effect should also be 0. A close to 0 effect variance and a large noise variance (both relatively, comparing to Δ_2) mean CUPED using ANA can lead to significant variance reduction with a small bias trade-off.

Δ	Δ_1	Δ_2	ANA_{corr}	ANA_{err}
4/25	2/25	8/25	8/25	8/25

Table 1: Number of statistically significant results out of 25 experiments. CUPED with ANA maximizing correlation and minimizing error has the same number of significant results as using Component 2 only ($\theta = 0$).

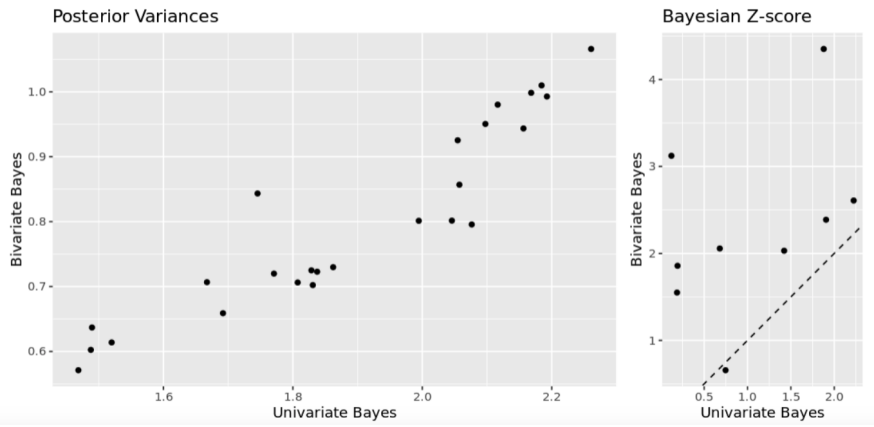

Figure 1: Left: Bayesian posterior using bivariate decomposed model is smaller than without decomposition. Right: Bivariate model produces larger absolute Bayesian Z-score.

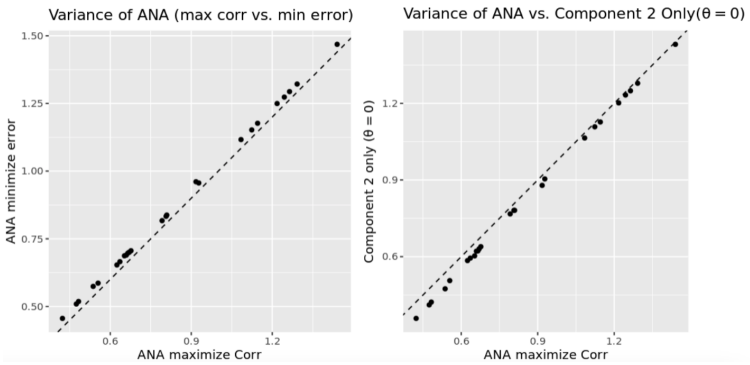

Figure 2: ANA maximizing correlation has a slightly smaller variance than ANA minimizing error (Left); but slightly larger variance than Δ_2 (Right). Difference in variances in these 25 experiments are all very small.

Table 1 shows using ANA with $\theta = 0$ (Δ_2), or maximizing correlation, or minimizing error all lead to more stat. sig. results, comparing to the original Δ before decomposition. Figure 1 shows posterior variance is greatly reduced with the bivariate model. Bayesian Z-score (posterior mean over posterior standard deviation) has greater absolute value under the bivariate model. Figure 2 compares variances of ANA with $\theta = 0$ (Δ_2), ANA maximizing correlation, and ANA minimizing error. All three produces similar variances.

References

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

Laura Cosgrove, Jen Townsend, and Jonathan Litz. 2022. Deep Dive Into Variance Reduction. `https://www.microsoft.com/en-us/research/group/experimentation-platform-exp/articles/deep-dive-into-variance-reduction/`.

Alex Deng, Michelle Du, Anna Matlin, and Qing Zhang. 2023a. Variance Reduction Using In-Experiment Data: Efficient and Targeted Online Measurement for Sparse and Delayed Outcomes. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 3937–3946.

Alex Deng, Ya Xu, Ron Kohavi, and Toby Walker. 2013. Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. In *Proceedings of the 6th ACM WSDM Conference*. 123–132.

Alex Deng, Lo-Hua Yuan, Naoya Kanai, and Alexandre Salama-Manteau. 2023b. Zero to hero: Exploiting null effects to achieve variance reduction in experiments with one-sided triggering. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*. 823–831.

Yongyi Guo, Dominic Coey, Mikael Konutgan, Wenting Li, Chris Schoener, and Matt Goldman. 2021. Machine Learning for Variance Reduction in Online Experiments. *arXiv preprint arXiv:2106.07263* (2021).

Ying Jin and Shan Ba. 2023. Toward Optimal Variance Reduction in Online Controlled Experiments. *Technometrics* 65, 2 (2023), 231–242.

Winston Lin. 2013. Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *The Annals of Applied Statistics* 7, 1 (2013), 295–318.

Art B Owen. 2013. Monte Carlo theory, methods and examples. (2013).

Alexey Poyarkov, Alexey Drutsa, Andrey Khalyavin, Gleb Gusev, and Pavel Serdyukov. 2016. Boosted decision tree regression adjustment for variance reduction in online controlled experiments. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 235–244.

Nilesh Tripuraneni, Lee Richardson, Alexander D'Amour, Jacopo Soriano, and Steve Yadlowsky. 2023. Choosing a Proxy Metric from Past Experiments. *arXiv preprint arXiv:2309.07893* (2023).

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

Huizhi Xie and Juliette Aurisset. 2016. Improving the sensitivity of online controlled experiments: Case studies at netflix. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 645–654.