Continuous Treatment: Methods Comparison

Yifei Ding @ UC Riverside, Meng Xu @ Snap Inc.

September 2023

1 Introduction

In many business questions at tech companies like Snap, the treatment is not binary but continuous. In the realm of monetization, our focus lies in understanding the impact of ad frequency on user conversions and the influence of ad load on user engagement, with precise estimations guiding our advertising strategy optimization efforts. Within the field of engineering, our concern pertains to how app performance factors such as latency affect user engagement, determining the appropriate level of engineering investment required to enhance latency metrics. In the sphere of marketing, our attention is directed toward assessing how geolocation events and campaigns influence local engagement concerning proximity, such as the diminishing effect of billboards or local community events as distance from the target location increases. In the domain of sales, our primary interest revolves around examining how sales respond to variations in coupon quantities.

To address these business inquiries, beyond computing the Average Marginal Effect (AME), which quantifies the average change in outcome variables when the continuous treatment variable increases by 1 unit, we are also keen on (1) determining the Dose Response Curve (DRC) to understand how the outcome variable responds to the continuous treatment variable in a non-linear fashion, and (2) estimating the marginal effect at each specific point along the continuous treatment variable.

There are many methods to estimate the the Dose Response Curve and marginal effect at each dose. Imbens and Hirano (2004) propose a method to regress outcome variable on the continuous treatment and include Generalized Propensity Score (GPS) as a covariate. Imai and van Dyk (2004) propose a method about matching or subclassifying on the propensity function. Robins et al. (2000) propose inverse probability weighting (IPW) method to estimate the average marginal effect of a continuous treatment. This method is further considered and developed by Flores et al. (2012) and Galvao and Wang (2015). Instead of using the IPW-based method to make the continuous treatment and observed confounders orthogonal, there is a series of newly developed balancing methods for continuous treatment directly minimize the correlation between the continuous treatment and observed confounders Tübbicke (2020); Vegetabile et al. (2021); Bahadori et al. (2022); Greifer et al. (2020) or estimate the generalized propensity score (GPS) and achieve balance simultaneously Fong et al. (2018). We can also estimate the dose-response curve by doubly robust approaches. Kennedy et al. (2017) propose a kernel smooth approach while Colangelo and Lee (2022); Klosin (2021) focus on using double/debiased machine learning techniques.

Within the context of tech companies like Snap, the consideration of scalability plays a pivotal role in selecting appropriate methods, particularly when dealing with datasets comprising millions of observations. This study places its primary focus on two emerging methodologies with the potential to efficiently handle the estimation of continuous treatment effects for vast numbers of analysis units: entropy balancing for continuous treatment (Tübbicke (2020); Vegetabile et al. (2021)) and double/debiased machine learning for continuous treatment (Colangelo and Lee (2022)). To evaluate their performance, we employ semi-synthetic data, generated based on actual Snapchat user data, to replicate intricate, nonlinear relationships among outcome variables, continuous treatment, and observed confounding factors, subsequently employing this simulated data to compare the performance of these two methodologies. In the context of the double/debiased machine learning method, we additionally assess the performance of various machine learning algorithms.

2 Methods for Continuous Treatment

2.1 Identification Assumptions of Continuous Treatment

The identification of continuous treatment relies on some conditions. For any value (d) of the continuous treatment D, we have (Imbens (2000)):

- Weak unconfoundedness: $Y(d)|X$.
- Common support: $f(D = d|X) > 0$.
- Balancing condition: $X(d) = X$.

For continuous treatment, we usually focus on two estimates:

- Average treatment effect at any treatment value d: $ATE(d) = E[Y(d)] - E[Y(0]$.
- Marginal treatment effect at any treatment value d: $MTE(d) = \frac{\partial E[Y(d)]}{\partial d}$

2.2 Balancing Approach

The balancing approach encompasses a set of weighting techniques designed to minimize the correlation between treatment variables and observed confounding factors. A notable method within this approach is entropy balancing, pioneered by Hainmueller (2012). Initially devised for estimating treatment effects in binary variables, entropy balancing has demonstrated doubly robust properties when applied to linear data generation, as validated by Zhao and Percival (2017). Tübbicke (2020); Vegetabile et al. (2021); Bahadori et al. (2022) further extend entropy balancing to handle continuous treatment. Essentially, in the context of continuous treatment, entropy balancing seeks to determine a set of weights denoted as w that minimize the covariance between the continuous treatment variable and each observed confounder. This optimization is achieved while preserving the distributions of both the continuous treatment variable and all observed confounders. It is formalized below Vegetabile et al. (2021):

$$min_w \sum_{i=1}^{N} w_i log(\frac{w_i}{N}),\ subject\ to$$

$$\sum_{i=1}^{N} w_i(X_{ij}^p - \mu_j^p)(D_i - \mu_D) = 0, \sum_{i=1}^{N} w_i = 1,$$

$$\sum_{i=1}^{N} w_i(D_i^q - \mu_D^q) = 0,\ \sum_{i=1}^{N} w_i(X_{ij}^p - \mu_{X_{ij}}^q) = 0.$$

In which j refers to the dimension of observable and p, q are related moments.

After calculating the balancing weights, a nonlinear model (e.g. local linear regression, generalized additive model, etc) can be applied to estimate the dose response curve and marginal effect at each point.

Overall, the balancing approach offers scalability advantages. As outlined in a recent discussion by Lin et al. (2023), it can be readily implemented within distributed computing frameworks like Spark and Hive. However, it does have a limitation in its ability to handle intricate interactions between variables, which can be overcome by complex machine learning algorithms.

2.3 Machine Learning Method

In contrast to the balancing approach, the machine learning method offers distinct advantages by integrating variable selection and exhibiting flexibility in capturing nonlinearities and interactions. Chernozhukov et al. (2018) present a double/debiased machine learning approach for calculating treatment effects in binary treatments and average marginal effects in continuous treatments. Building upon this foundation, Colangelo and Lee (2022) further extend the framework to estimate dose-response curves and marginal effects at various dosage levels. In the following section, we provide a concise overview of this methodology and explore considerations for selecting appropriate machine learning models.

2.3.1 Double Debiased Machine Learning for Continuous Treatment

The estimation procedure of the double/debiased machine learning method for continuous treatment in Colangelo and Lee (2022) is divided into the following steps:

- Step 1. (Cross-fitting) Data sample is randomly partitioned into L distinct groups $I_\ell, \ell = 1, \ldots, L$. For each $\ell = 1, \ldots, L$, the estimators of outcome model, $\hat{\gamma}_\ell(d, x)$ for $\gamma(d, x) \equiv \mathbb{E}[Y \mid D = d, X = x]$, and selection model, $\hat{f}_\ell(d \mid x)$ for $f_{D|X}(d \mid x)$, are estimated by the observations not in I_ℓ.
- Step 2. (Dose response curve) The double debiased ML (DML) estimator is defined as

  $$\hat{\beta}_d \equiv \frac{1}{n} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} \left\{ \hat{\gamma}_\ell(d, X_i) + \frac{K_h(D_i - d)}{\hat{f}_\ell(d \mid X_i)} (Y_i - \hat{\gamma}_\ell(d, X_i)) \right\},$$

  where the generalized propensity score, $\hat{f}_\ell(d \mid X_i)$, is estimated by a kernel-based estimation method by following the reciprocal of the generalized propensity score (GPS).
- Step 3. (Marginal effect) Let $d^+ \equiv (d_1 + \eta/2, t_2, \ldots, d_{k_d})'$ and $d^- \equiv (d_1 - \eta/2, d_2, \ldots, d_{k_d})'$, where η is a positive sequence converging to zero as $n \to \infty$. We estimate the partial effect of the first component of the continuous treatment $\theta_d \equiv \partial\beta_d/\partial d_1$ by $\hat{\theta}_d \equiv \left(\hat{\beta}_{d^+} - \hat{\beta}_{d^-}\right)/\eta$.

2.3.2 Machine Learning Model Selection

In the double/debiased machine learning framework, the question of what kind of machine learning algorithm can deliver good potential for estimating dose response curve and marginal effects deserves further investigation. When we estimate marginal effect, conventional tree-based model like XGBOOST could suffer....

Therefore, the majority of the work in this paragraph focuses on discussing

A novel tree-based machine learning model, boosting smooth transition regression trees(BooST), is introduced for estimating partial derivative (?). Boosting is an ensemble method to strengthen the stability of learning mapping for weaker learners. A long-tested and useful weaker learner applied in boosting is regression/decision tree. Boosting regression/decision tree has many virtues including natural handling of data of *mixed* category, computational scalability and less sensitive to dimensional problems while not differentiable. In addition, deep learning has achieved unprecedented success in a great deal of prediction problems, which is largely explained by its considerable capacity of learning the unknown structures in prediction tasks. The various universal approximation theorems of deep learning have justified its effectiveness in approximating functions (????). As a result of the expressive power of DNN for approximating functions, we propose deep neural networks as a nonparametric model to recover partial derivatives and participate in estimation competitions.

3 Simulation

3.1 Data Generating Process

Two data generating processes are performed to investigate the estimation preciousness of dose response curve and its corresponding partial derivatives among the balancing approach and machine learning methods.

3.1.1 Friedman Model

The first synthetic data generation process to be considered is the classical data generation process in machine learning first proposed by Friedman (1991). The outcome model is

$$\begin{aligned} y &= f(\boldsymbol{x}) + u \\ &= 20sin(\pi x_1 x_2 + 5) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 + u \end{aligned}$$

and selection model is specifically designed by

$$x_1 = sin(x_2 x_3) - cos(x_3 x_4) + sin(x_5^2) + sin(x_4 - x_5^2),$$

where the Gaussian noise u follows $N(0, \sigma^2)$ with $\sigma = 5$ and $x_i \sim N(0,1)$ for all $i \in \{1,2,3,4\}$. The role of x_1 in Friedman outcome model represents a high interactive component with other features, which means its first order derivative is varying and influenced by other variables.

3.1.2 Regression in Tensor Product Spaces by the Method of Sieves (RTPS)

The second semi-synthetic data generating process is performed to evaluate the performance of various machine learning models and balancing approaches. The steps to generate semi-synthetic data are following:

- Step 1. We randomly select 2N (N = 10 Million) users from whole population and divide it into a training dataset and testing dataset.
- Step 2. We train a regression model $\hat{g}_j(X_i)$ by a multivariate linear sieve model in tensor product space developed by Zhang and Simon (2022).

3.2 Results

Within this section, we undertake a performance evaluation, contrasting the effectiveness of the balancing approach against various machine learning methodologies for the estimation of both dose-response curves and marginal effects at varying doses. Furthermore, we engage in a discussion regarding the indispensability of the double machine learning process, enhanced by cross-fitting, as elaborated in Colangelo and Lee (2022).

3.2.1 Balancing approach moment and model selection

In the context of the balancing approach, we examine various combinations of balancing model moments (m = 1, 2, 3) and two dose-response curve models (LOESS and GAM[1]). When working with data generated using the Friedman model (characterized by low dimensions and a sample size of 10,000), we assess the performance of all six combinations of balancing moments and dose-response curve models. In the case of data generated using the RTPS model, particularly when dealing with high-dimensional data and a small sample size (10,000), we limit our evaluation to the moment = 1 setting, as balancing all second and third moments becomes infeasible in high-dimensional scenarios. Lastly, for data generated using the RTPS model with a large sample size (1 million), our assessment focuses solely on the GAM model, as the LOESS model lacks scalability for such extensive datasets.

In Figure 1, the following observations come to light: (1) In the context of the Friedman model, the balancing approach, while introducing some bias, effectively reconstructs the shape of the dose-response curve. However, when dealing with the RTPS model, the balancing approach exhibits substantial bias and fails to accurately restore the true dose-response curve. (2) Contrary to the findings in Vegetabile et al. (2021), we note that increasing the balancing moments does not lead to a reduction in RMSE (Root Mean Square Error). (3) In a comparative analysis between the LOESS and GAM models, no clear superiority emerges in the context of the Friedman model. However, in the case of the RTPS model, the GAM model outperforms the LOESS model.

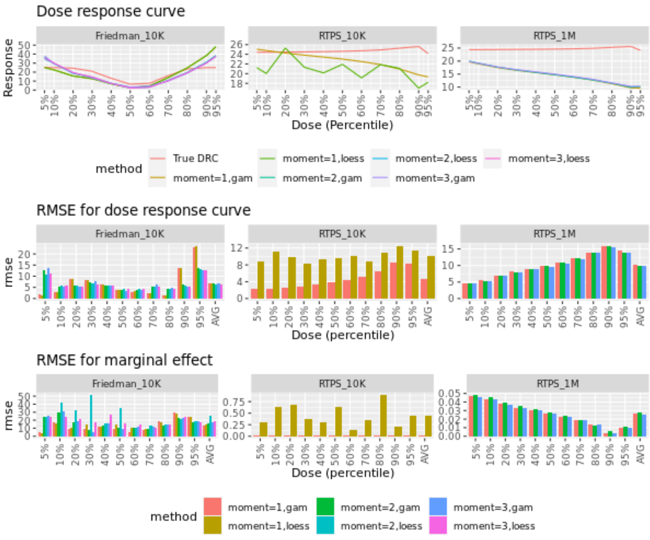

Figure 1: Balancing approach moment and model comparison

[1] We employ cross-validation to determine the optimal span for the LOESS model and the appropriate value of k for the GAM model.

3.2.2 Machine learning

Machine learning model selection

In Figure 2, we have implemented the double/debiased machine learning method as outlined in Colangelo and Lee (2022). We then proceeded to compare the performance of various machine learning algorithms, including the best balancing approach model. Our analysis yielded the following key findings: (1) When estimating dose-response curves within the context of the Friedman model, the tree-based model consistently outperforms both the balancing approach and the DNN model. Impressively, Boostsmooth exhibits even better performance than XGBOOST. For the RTPS model, as Boostsmooth struggles to scalably handle high-dimensional data, and the DNN model fails to accurately estimate the generalized propensity score. It is worth noting that XGBOOST displays subpar performance with a 10K sample size but significantly improves its performance compared to the balancing approach when dealing with 1 million samples. (2) In the domain of marginal effect estimation, the balancing approach consistently demonstrates the best performance across the board.

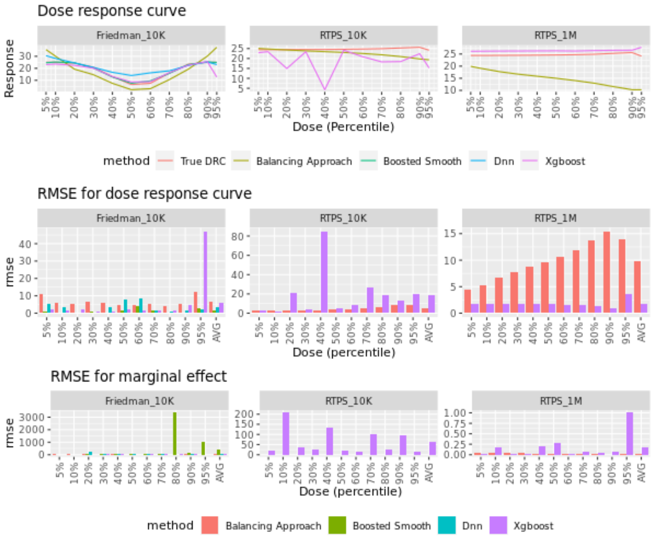

Figure 2: DML comparisons

The necessity of double machine learning with cross fitting

The double/debiased machine learning process, augmented by cross-fitting, offers a theoretical solution to eliminate regularization bias and achieve a doubly robust outcome. We were intrigued by the potential performance improvements it might bring to our simulated data. In our analysis, we focus on the machine learning model that exhibited the best performance for each data generation process, specifically Boostsmooth for the Friedman model and XGBOOST for the RTPS model. We then proceeded to compare the performance of the following specifications: (1) DML with Cross-Fitting (Baseline), (2) DML with Full Data, (3) Naive ML (Outcome Model Only) with Cross-Fitting, (4) Naive ML with Full Data. Our findings reveal that, overall, there is minimal disparity between estimates derived from full data and cross-fitting techniques. The primary distinction arises in the comparison between Naive ML and DML. Surprisingly, in all data generation processes, the Naive model outperforms DML in both dose-response curve estimation and marginal effects estimation. This leads us to reconsider the utility of kernel-based methods in estimating the selection model.

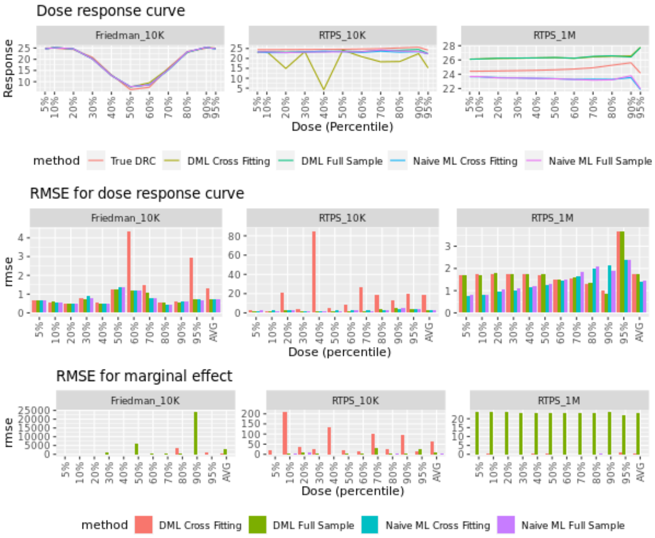

Figure 3: DML comparisons

4 Discussion

In this research, we conduct a comparative analysis between two state-of-the-art methods: the balancing approach and double/debiased machine learning, specifically for continuous treatment scenarios involving the estimation of dose-response curves and marginal effects at various dosage levels. Our findings indicate that machine learning techniques, particularly tree-based methods like XGBOOST and Boostsmooth, consistently outperform the balancing approach and DNN. Boostsmooth exhibits superior performance even when compared to XGBOOST. It prompts us to contemplate potential strategies for surmounting the method's scalability limitations.

Additionally, we delve into an examination of the necessity of the double machine learning process, coupled with cross-fitting, as outlined in Colangelo and Lee (2022). Surprisingly, our investigation reveals that Naive machine learning, which exclusively models outcomes, demonstrates superior performance when compared to the double machine learning process. Despite the strong theoretical underpinnings of the double machine learning process, we are prompted to reconsider the viability of the kernel-based selection model detailed in Colangelo and Lee (2022).

References

Bahadori, T., E. T. Tchetgen, and D. Heckerman (2022, 17–23 Jul). End-to-end balancing for causal continuous treatment-effect estimation. In K. Chaudhuri, S. Jegelka,

L. Song, C. Szepesvari, G. Niu, and S. Sabato (Eds.), *Proceedings of the 39th International Conference on Machine Learning*, Volume 162 of *Proceedings of Machine Learning Research*, pp. 1313–1326. PMLR.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018, 01). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal 21*(1), C1–C68.

Colangelo, K. and Y.-Y. Lee (2022). Double debiased machine learning nonparametric inference with continuous treatments.

Flores, C. A., A. Flores-Lagunes, A. Gonzalez, and T. C. Neumann (2012, 02). Estimating the Effects of Length of Exposure to Instruction in a Training Program: The Case of Job Corps. *The Review of Economics and Statistics 94*(1), 153–171.

Fong, C., C. Hazlett, and K. Imai (2018). Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements. *The Annals of Applied Statistics 12*(1), 156 – 177.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The annals of statistics 19*(1), 1–67.

Galvao, A. F. and L. Wang (2015). Uniformly semiparametric efficient estimation of treatment effects with a continuous treatment. *Journal of the American Statistical Association 110*(512), 1528–1542.

Greifer, N., K. Bollen, D. Bauer, M. Prinstein, and M. Hudgens (2020). *Estimating Balancing Weights for Continuous Treatments Using Constrained Optimization*. Ph. D. thesis. AAI28003097.

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis 20*(1), 25–46.

Imai, K. and D. A. van Dyk (2004). Causal inference with general treatment regimes. *Journal of the American Statistical Association 99*(467), 854–866.

Imbens, G. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika 87*(3), 706–710.

Imbens, G. and K. Hirano (2004). The propensity score with continuous treatments.

Kennedy, E. H., Z. Ma, M. D. McHugh, and D. S. Small (2017). Non-parametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society. Series B (Statistical Methodology) 79*(4), 1229–1245.

Klosin, S. (2021). Automatic double machine learning for continuous treatment effects.

Lin, S., M. Xu, X. Zhang, S.-K. Chao, Y.-K. Huang, and X. Shi (2023). Balancing approach for causal inference at scale.

Robins, J., M. Hernán, and B. Brumback (2000, September). Marginal structural models and causal inference in epidemiology. *Epidemiology (Cambridge, Mass.) 11*(5), 550—560.

Tübbicke, S. (2020, October). Entropy Balancing for Continuous Treatments. CEPA Discussion Papers 21, Center for Economic Policy Analysis.

Vegetabile, B., B. A. Griffin, D. Coffman, M. Cefalu, M. Robbins, and D. Mccaffrey (2021, 03). Nonparametric estimation of population average dose-response curves using entropy balancing weights for continuous exposures. *Health Services and Outcomes Research Methodology 21*.

Zhang, T. and N. Simon (2022). Regression in tensor product spaces by the method of sieves.

Zhao, Q. and D. Percival (2017). Entropy balancing is doubly robust. *Journal of Causal Inference 5*(1), 20160010.