Deep Learning Based Casual Inference with Combinatorial A/B Tests on Large-Scale Platforms

1 Introduction

Internet-based online platforms have substantially impacted people's lives and the global economy. They have penetrated billions of people's daily lives in various areas such as social media (Facebook and WeChat), online shopping (Amazon and Alibaba), and urban transportation (Uber and Didi), to name a few. Because of the tremendous values created by these platforms, it is also estimated by the Committee on Judiciary of the USA that the total market value of platform-based tech firms will reach more than 30% of the annual global GDP within the next 10 years.[1] The prosperity of these platforms relies heavily on the enormous data they own, and the data analytics methodologies that drive their strategic and operations decisions. Randomized experiments (a.k.a. A/B tests) have now become a ubiquitous and critical data-driven decision tool to efficiently evaluate and optimize their strategies. In practice, a large-scale online platform like Facebook usually launches thousands of experiments every day to fast iterate their business operations such as product designs and recommendation algorithms. Consequently, each user of the platform is independently treated by thousands of A/B tests simultaneously. This triggers interesting and important research questions for platforms to best leverage the power of A/B tests:

- How to estimate and infer the overall treatment effect of multiple experiments on a platform?
- Without observing the outcomes of all experiment combinations, how to identify the optimal experiment combination?

Platform managers usually assume the treatment effects of different experiments are linearly additive, so the optimal combination is that of all the experiments with a positive average treatment effect. However, we observe from the real data of a large-scale online video-sharing platform (Platform O hereafter) that linear additivity does not hold, and that the overall treatment effect of multiple experiments varies for different users. The goal of this paper is to address the aforementioned research questions taking into account the non-additivity and heterogeneous treatment effect (HTE) of multiple experiments.

2 Statistical Framework

Based on the Double/De-biased Machine Learning (DML) approach (see, e.g., Chernozhukov et al., 2018; Farrell et al., 2020), we develop a novel statistical framework to capture the HTE and non-additive treatment effects observed from the data of Platform O. In this abstract, we use capital letters to denote random variables/vectors and small letters to denote their realizations. We denote $X \in \mathbb{R}^d$ as the feature vector capturing user heterogeneity on the platform, $Y \in \mathbb{R}$ as the outcome variable (e.g., average App time per day), and $T \in \{0,1\}^k$ as the treatment vector capturing whether the user is in the treatment or control condition of each experiment. Hence, for experiment $j \in \{1,2,...,k\}$, $T_j = 1$ (resp. $T_j = 0$) means the user is in the treatment (resp. control) condition. We assume that, for some unknown functions $G(\cdot,\cdot) \in \mathbb{R}$ and $\theta(\cdot) \in \mathbb{R}^{d_\theta}$, the conditional expected outcome satisfies $\mathbb{E}[Y|X = x, T = t] = G(\theta(x), t)$, where $x \in \mathbb{R}^d$ is a given user feature vector and $t \in \{0,1\}^k$ is a given treatment vector. It is clear from the model formulation that the function $\theta(\cdot)$ captures the HTE of any experiment and the non-linearity of $G(\cdot,\cdot)$ captures the non-additivity of multiple treatment effects. The (link) function $G(\cdot,\cdot)$ can be parameterized in a fairly general fashion and we have used the sigmoid functions in our actual implementation. The function $\theta(\cdot)$ is non-parametric. For a given real-valued function $H(\cdot)$, we are interested in the estimation and inference of a generic quantity $\tau := \mathbb{E}\left[H(X, \theta(X); t^*)\right]$, where t^* is any given experiment combination of interest. For example, if we are interested in the average treatment effect (ATE) of the combined treatment of all k experiments, then we

[1] See https://judiciary.house.gov/uploadedfiles/competition_in_digital_markets.pdf?utm_campaign=4493-519.

Table 1: Ground-Truth ATE of 8 Treatment Combinations

Treatment Combination	Re-scaled ATE	Observed or Not	Number of Users
[0, 0, 0]	0.000	Observable	258,249
[0, 0, 1]	4.747**	Observable	258,340
[0, 1, 0]	-1.161	Observable	258,367
[1, 0, 0]	3.297*	Observable	258,321
[1, 1, 1]	9.223****	Observable	258,375
[1, 1, 0]	2.995	Unobservable	258,480
[1, 0, 1]	10.000****	Unobservable	258,305
[0, 1, 1]	6.031***	Unobservable	258,172

*p<0.05; **p<0.01; ***p<0.001; ****p<0.0001.

assign $H := G(\theta(x), t^*) - G(\theta(x), t_0)$, where $t^* = (1, 1, ..., 1)$ and $t_0 = (0, 0, ..., 0)$.

We leverage the flexibility and scalability of deep neural networks (DNNs) to estimate the unknown parameters $\theta(\cdot)$. Specifically, given a training data set $\{w_i = (y_i, x_i, t_i) \in \mathbb{R} \times \mathbb{R}^d \times \{0,1\}^k : i = 1, 2, ...n\}$ and a relevant class of DNNs $\mathcal{F}_{DNN}$ (e.g., multi-layer perceptrons, MLPs), we follow the standard empirical loss minimization procedure to obtain the estimates of $\theta(\cdot)$, which we denote as $\hat{\theta}(\cdot)$ (see, e.g., Chernozhukov et al., 2018; Farrell et al., 2020, 2021).[2]

The core of our proposed method is adopting the DML framework to correct the bias of a plug-in estimator from the perturbations of $\hat{\theta}(\cdot)$ as a result of the variations in feature x (see, also, Chernozhukov et al., 2018; Farrell et al., 2020). Specifically, we derive the Neyman orthogonal score function for parameter τ (also called an influence function) as $\psi(w, \theta, \Lambda; t^*) := H(x, \theta(x); t^*) - \partial_\theta H(x, \theta(x); t^*)' \Lambda(x)^{-1} \partial_\theta \ell(y, t, \theta(x))$, where $\ell(\cdot)$ is the loss function and $\Lambda(x) := \mathbb{E}[\partial^2_\theta \ell(Y, T, \theta(X))|X = x]$. It is clear that the first-term of the score function $\psi(\cdot)$ (i.e., $H(x, \theta(x); t^*)$) is the plug-in estimator, whereas the second-term (i.e., $-\partial_\theta H(x, \theta(x); t^*)' \Lambda(x)^{-1} \partial_\theta \ell(y, t, \theta(x))$) is the bias-correction for more accurate and efficient estimation of τ. We theoretically show that, under some standard technical assumptions, our debiased deep learning (DeDL) estimator, $\hat{\tau}_{\mathsf{DeDL}} := \psi(w, \hat{\theta}, \hat{\Lambda}; t^*)$, is consistent and asymptotically normal, thus naturally giving rise a valid estimation and inference procedure for the treatment effect of interest τ.

3 Online Implementation and Empirical Results

We collaborate with Platform O who has more than 300 million daily active users (DAU) and 600 million monthly active users (MAU) worldwide, and runs hundreds of A/B tests simultaneously everyday. We leverage 3 A/B tests, each of which examines the treatment effect of a major adjustment to the video recommendation algorithm on one of three main pages of Platform O. The outcome of interest is the App time duration per day for each user. Consistent with our statistical framework, we use a 3-dimensional binary vector $t \in \{0,1\}^3$ to represent the treatment combination applied to each user. Based on these 3 A/B tests, Table 1 reports the ground-truth re-scaled ATE of each treatment combination with the maximum ATE (obtained when $t = (1, 0, 1)$) set as 10 to protect the sensitive data of Platform O. The platform usually launches each individual experiment independently (most likely in an asynchronous and uncoordinated fashion) and conducts a back-test for the treatment effect of the totally combined experiment (i.e., $t = (1, 1, 1)$) at the end. Hence, we could only observe the outcomes of the first 5 treatment combinations in Table 1.

In our online implementation of the DeDL framework, we have also collected pre-experiment user feature data, including 16 discrete variables such as gender, region, age range and users' activeness level, and 10 continuous variables such as the video watching duration on each page per day in the 10 days right before the start of the experiments. The link function is set as the *sigmoid function*, $G(\theta(x), t) =$

[2] In most applications, the squared loss is selected as the loss function to train the DNN.

Table 2: Comparison of Different Estimators

Method	Unobserved Treatment Combinations				All Treatment Combinations			
	MAPE	MSE	MAE	Significance	MAPE	MSE	MAE	Significance
PA	29.60%	3.516	1.754	N/A	12.68%	1.507	0.752	N/A
LR	21.34%	0.991	0.795	2/3	33.85%	1.234	1.017	7/8
DL	7.85%	0.144	0.316	2/3	21.61%	0.693	0.628	6/8
DeDL	5.77%	0.102	0.314	3/3	8.13%	0.136	0.330	8/8

$\frac{\theta_4(x)}{1+\exp(-(\theta_0(x)+\theta_1(x)t_1+\theta_2(x)t_2+\theta_3(x)t_3))}$, to capture the non-linear effects of treatment combinations, whereas the loss function is set as the squared loss $\ell(y,t,\theta) = (y - G(\theta(x),t))^2$. We approximate the parameter $\theta(\cdot)$ with 3-layer DNNs and train our model on TensorFlow.

Next, we describe the benchmark estimators to validate our DeDL framework. We consider 4 natural benchmarks. The first is the purely additive (PA) estimator which directly uses the sum of the treatment effect for each individual experiment as that for the experiment combination. The PA estimator is indeed the one actually adopted by Platform O's managers. The second is the linear regression LR estimator which predicts the outcome of unobserved treatment combinations using linear regression, and conducts a standard t-test to estimate the ATE afterwards. The third is the deep learning (DL) estimator which uses the plug-in term of the DeDL estimator without de-biasing.

Table 2 compares the 4 estimators with respect to different metrics including mean absolute percentage error (MAPE), mean squared error (MSE), mean absolute error (MAE) and the accuracy for correctly identifying whether the treatment effect is significant (the Significance columns in Table 2)[3]. Our empirical results clearly demonstrate the superiority of our DeDL estimator over the benchmarks in all performance metrics of interest. Not only does DeDL correctly identifies which experiment combination yields the highest ATE (i.e., best-arm identification) in a statistically reliable fashion, but it also more accurately estimates the ATE of any experiment combination. Notably, comparing DeDL with DL reveals additional insights. Even if DNNs could accurately recover the underlying ground-truth data generating process, it may still suffer from the substantial bias caused from the aforementioned data perturbation issues when estimating the function $\theta(\cdot)$. However, the debiased correction based on the Neyman orthogonal score helps sharpen the estimation and, eventually, significantly improve the estimation and inference of the treatment effects of interest.

To sum up, we develop a novel deep learning based statistical framework to estimate and infer the treatment effect of any experiment combination for large-scale online platforms. Our framework is simple, scalable, and theoretically sound. More importantly, we implement it on a large-scale video-sharing platform and demonstrate its significant superiority in generating more accurate and statistically reliable estimations than the commonly used benchmarks such as the pure deep learning estimator with de-biasing.

References

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters.

Farrell, M. H., Liang, T., and Misra, S. (2020). Deep learning for individual heterogeneity. *arXiv preprint arXiv:2010.14694*.

Farrell, M. H., Liang, T., and Misra, S. (2021). Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213.

[3]These metrics are quantifiable because the experimental architecture enables us to obtain the ground-truth outcomes and treatment effects.