Deep-Learning-Based Causal Inference for Large-Scale Combinatorial Experiments: Theory and Empirical Evidence

Zikun Ye[1], Zhiqi Zhang[2], Dennis J. Zhang[2], Heng Zhang[3], Renyu Zhang[4]
[1] University of Washington, Seattle, WA
[2] Washington University in St. Louis, St. Louis, MO
[3] Arizona State University, Tempe, AZ
[4] The Chinese University of Hong Kong, Hong Kong, China
zikunye@uw.edu, z.zhiqi@wustl.edu, denniszhang@wustl.edu, hengzhang24@asu.edu, philipzhang@cuhk.edu.hk

Large-scale online platforms launch hundreds of randomized experiments (a.k.a. A/B tests) every day to iterate their operations and marketing strategies. The combinations of these treatments are typically not exhaustively tested, which triggers an important question of both academic and practical interest: Without observing the outcomes of all treatment combinations, how does one estimate the causal effect of any treatment combination and identify the optimal treatment combination? We develop a novel framework combining deep learning and doubly robust estimation to estimate the causal effect of any treatment combination for each user on the platform when observing only a small subset of treatment combinations. Our proposed framework (called debiased deep learning, DeDL) exploits Neyman orthogonality and combines interpretable and flexible structural layers in deep learning. We show theoretically that this framework yields efficient, consistent, and asymptotically normal estimators under mild assumptions, thus allowing for identifying the best treatment combination when observing only a few combinations. To empirically validate our method, we collaborated with a large-scale video-sharing platform and implemented our framework for three experiments involving three treatments where each combination of treatments is tested. When observing only a subset of treatment combinations, our DeDL approach significantly outperforms other benchmarks to accurately estimate and infer the average treatment effect of any treatment combination, and to identify the optimal treatment combination.

Key words: Deep Learning, Double Machine Learning, Causal Inference, Field Experiments, Experimentation on Online Platforms

1. Introduction

Large-scale online platforms have penetrated the daily lives of billions of people. As of January 2021, more than 53% of the world population (about 4.2 billion people) are active social media users.[1] People connect on social network platforms such as Facebook and TikTok, shop online on e-commerce platforms such as Amazon and Alibaba, and hail a ride on ride-sharing platforms such as Uber and Lyft. These platforms create tremendous value—the firms that develop and own such businesses are

[1] See https://datareportal.com/reports/digital-2021-global-overview-report.

now worth more than USD 2 trillion. For example, in November 2024, the market value of Amazon was USD 2.1 trillion, that of Alphabet was USD 2.0 trillion, and that of Microsoft was USD 3.1 trillion. McKinsey & Company estimates that the total market value of platform-based tech firms will reach more than 30% of the annual global gross economic outcome within the next ten years.[2]

Equipped with mountains of user data and advanced information technology, online platforms base their critical business decisions on data analytics techniques. Of central importance are randomized experiments (a.k.a. A/B tests or field experiments; we use A/B tests and experiments interchangeably hereafter), which are widely considered the gold standard for causal inference and policy evaluation. Under an A/B test, a platform randomly assigns its users to different groups and applies a different treatment to users in each group. The controlled randomization enables the platform to credibly attribute the outcome differences of different user groups to the treatment effect of the strategies.

Because of the online nature of their business and their vast user traffic, platforms can conveniently run A/B tests to evaluate and optimize their product design, pricing, and recommendation strategies (Kohavi et al. 2020). Usually, the analyst casts a policy change in these aspects as a treatment and compares it with the existing policy through an A/B test. Leading online platforms such as Facebook, Amazon, Google, and TikTok each run more than 10,000 online experiments annually, many of which engage millions of their users (Kohavi and Thomke 2017).

To quickly iterate its business operations, a large-scale online platform typically runs hundreds of A/B tests concurrently (see, e.g., Xiong et al. 2020). The sheer number of tests makes it difficult to test the joint effect of different treatments. In particular, due to limited user traffic, a standard online experimentation method for the platform is the orthogonal traffic-assignment design (Tang et al. 2010, Xiong et al. 2020): The treatment assignments of different individual A/B tests are independent. As a consequence, each user of the platform may be treated by numerous A/B tests simultaneously. On the one hand, the orthogonal experiment design utilizes the user traffic of the platform more efficiently. Orthogonality ensures noninterference among experiments, so the platform gets a credible causal estimate for each treatment in each experiment. On the other hand, it largely ignores the joint effects caused by the combination of treatments in practice. It does not allow the platform managers to find the best combination of treatments for each user.

In practice, platform managers typically assume that the treatment effects of different A/B tests are linearly additive. Hence, the decision on whether and how to expand the traffic of one treatment, for example, to all platform users, is irrespective of other concurrent experiments. Such a decision paradigm is particularly prevalent due to organizational reasons. For example, different stakeholders in a firm (e.g., the machine learning (ML) engineers and product managers) often manage their own

[2] See https://www.mckinsey.com/capabilities/growth-marketing-and-sales/our-insights/prime-day-and-the-broad-reach-of-amazons-ecosystem.

set of A/B tests, and often there is little coordination. The combined effects of multiple experiments are largely treated in a simple manner with the linear additivity assumption made.

Combining different treatments can create synergistic or antagonistic effects, depending on how different treatments interact with each other. It is usually difficult, without a formal test, to predict which one is in effect. In the worst case, two treatments that both benefit the platform can in fact hurt the platform once combined.

To empirically illustrate that the interactions between different treatments should not be ignored, we collaborate with a large-scale online video-sharing platform (referred to as Platform O hereafter). We plot in Figure 1 the relationships between the treatment effects of two concurrent treatments (more institutional details will be provided in Section 4). We observed the causal effects of all four treatment combinations (2×2) in this example, because the experiments are run under the full factorial design. To investigate the heterogeneous responses to the treatments with respect to user covariates, we divide the experimented users into $1,254$ subgroups based on their pretreatment covariates, including gender, age, location, and degree of activeness. Each observation in Figure 1 represents one such subgroup. Figure 1(a) plots, for each subgroup, the true observed effect size in the experiment, as well as the calculated effect size, if we assume these two effects are linearly additive. It clearly reveals the *substantial gap* from the ground-truth to simply adopt the linear addition rule in policy evaluation with multiple experiments. Figure 1(b) illustrates that different user groups have drastically diverse responses toward the treatments, some with increasing marginal returns/losses and others with decreasing marginal returns/losses.

Given these observations, in this paper, our main research question is: *When conducting multiple A/B tests and observing only a small subset of treatment combinations, how does one estimate and infer the causal effect of all treatment combinations, and how does one identify the optimal treatment combination?* As discussed earlier, the most commonly adopted approach to solving this problem is to run individual experiments independently and infer the combined treatment effect by assuming linear additivity of treatment effects, an assumption not supported by our data and sensibly questionable in practice. An alternative solution is the factorial experiment design, which directly tests the causal effect of *each* treatment combination (Box et al. 1978, Wu and Hamada 2011, Dasgupta et al. 2015). However, for full factorial design, the user traffic required to obtain reliable estimation and inference results grows exponentially in the number of treatments, making this approach infeasible for a large-scale online platform that runs hundreds of experiments concurrently. Another tempting approach might be to predict the outcome of each user under each treatment combination via an end-to-end ML model, such as a deep neural network (DNN), in which one incorporates treatments as inputs. Such an approach is generally not amenable to the inference of causal effects. The inherent bias due to regulation or overfitting leads to an insufficient convergence to the true causal effects (i.e., average

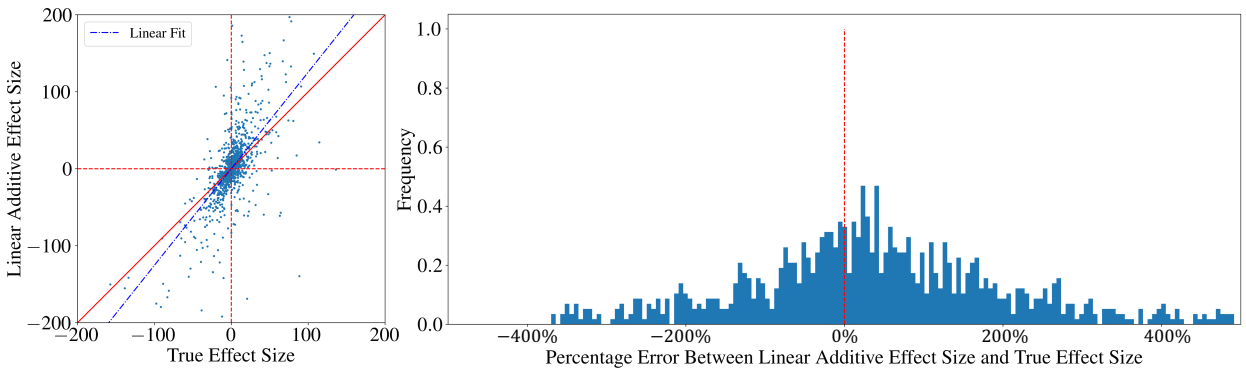


Figure 1 Heterogeneous Response to Two Experiments

Note: In the left subfigure, x-axis represents the true combined treatment effect of each subgroup, and y-axis represents the calculated treatment effect of each subgroup under the linear addition assumption. The 45-degree red solid line represents the equivalence of two effects, and the blue dashed line represents the linear fit of these two effects. The distinction between the blue dashed line and the red solid line elucidates that linear addition does not accurately recover the actual cumulative effect of the two treatments.

treatment effect, ATE) and, consequently, undesired statistical properties, which hinders effective inference (see, e.g., Chernozhukov et al. 2018).

The main goal of this paper is to develop a new theoretically sound and practically feasible method to estimate the causal effect of any treatment combination when observing the outcome of only a small subset of combinations. To this end, following the recent cutting-edge research of combining semiparametric statistics and deep learning (DL) for inference, namely the double/debiased machine learning (DML), we propose a statistical framework (called debiased deep learning, DeDL). We highlight the following contributions of our paper.

A DML-based modeling framework for multiple A/B tests with mild identification requirement and valid inference. We propose a DeDL framework with theoretical guarantees for researchers and practitioners to analyze treatment effects in concurrent experiments and identify the optimal treatment combination. Our framework, unlike factorial designs that require observing an exponential number of treatment combinations, requires observing only a linear number of treatment conditions. Our main theoretical contribution is modeling treatment effect interactions as a non-parametric function (i.e., DNN) of user characteristics and treatment vectors within the double machine learning framework (Chernozhukov et al. 2018, Farrell et al. 2020). We advocate for using a *generalized sigmoid link function* to map semi-parametrically the treatment vector and nuisance parameters, such as user characteristics, to the outcome. In Theorem 1, we establish a new approximation bound showing that the approximation error of our proposed link function, combined with

the non-parametric DNN for nuisance parameters, decreases at a square-root rate with the number of unique user characteristics, regardless of the data generation process for average treatment effects. Furthermore, we show that the abstract theoretical assumptions of the DML framework translate into easily verifiable and practically satisfied conditions in this context. With these conditions verified in practice, following Farrell et al. (2020), we demonstrate that our method produces asymptotically normal (and thus naturally $\sqrt{n}$-consistent) estimators, enabling valid inference.

Empirical validation of our framework. To demonstrate the practicality of our DeDL framework, we implement it in a large-scale multiple-experiment setting ($N >$ 2,000,000) on Platform O. We persuaded the company to conduct a costly full-factorial design with 3 treatment conditions (i.e., $2^3 = 8$ treatment combinations), allowing us to observe the ground-truth causal effects of any treatment combination for comparison. We compare our DeDL approach against a set of benchmarks, including linear regression and DL-based methods. Our results show that DeDL provides significantly more accurate estimates of the ATE and more precise identification of the optimal treatment combination. To the best of our knowledge, this is the first paper to validate the practical effectiveness of theoretically elegant DML methods through large-scale field experiments in the absence of unobserved endogenous variables. While recent literature (Gordon et al. 2023) highlights the limitations of the DML framework when omitted variables are present, we demonstrate its core strength: accurately approximating flexible treatment and outcome functions, by showcasing its superior performance in the absence of omitted variables. This evidence is increasingly relevant as more empirical researchers adopt DML methods, especially given the difficulty of verifying key technical assumptions, such as the first-stage convergence rate $o(n^{-1/4})$ of the ML algorithm. Furthermore, through comprehensive synthetic data in further simulation studies, we demonstrate the robustness of our DeDL framework in the presence of large biases in DNN training, a misspecified model layer, and highlight the covariate imbalance issue.

A Practitioner's Guide to DML-based Causal Inference with DNNs. Our work builds on the latest advances in the DML literature (Chernozhukov et al. 2018, Farrell et al. 2020). While this literature is theoretically elegant and powerful, it leaves many practical details and challenges for model implementation unaddressed. Our third contribution is to provide practical guidance and hands-on summaries for using DML in causal inference with DNNs. On the theoretical side, in Section 3, we outline the key steps for establishing a robust framework of using DML in various settings, including selecting an appropriate link function, training the DNN model, verifying the identifiability and convergence of nuisance parameters, constructing influence functions, and performing cross-fitting. On the practical implementation side, in Section 5, we provide critical checkpoints and guidance for successful empirical implementation of the DML algorithm. For example, we emphasize the importance of addressing covariate imbalance through stratified sampling and underscore the

value of DNN training error as a key indicator for evaluating goodness-of-fit when estimating nuisance parameters in practice. Finally, we also conduct comprehensive synthetic experiments to demonstrate the robustness of our proposed methods under various conditions (see Section 6). With these detailed demonstrations, we aim to bridge the gap between theory and practice, offering researchers and practitioners a clear roadmap for applying DML with DNNs to causal inference problems effectively and reliably.

The rest of this paper is organized as follows. In Section 2, we review the relevant literature. In Section 3, we present our DeDL framework to infer treatment effects and identify the optimal treatment combination. In Section 4, we apply the framework to analyze real-world experiments. In Section 5, drawing from our own learning, we provide a detailed discussion of some crucial issues in practical implementation, aiming to guide similar applications in other empirical contexts. In Section 6, we conduct comprehensive synthetic experiments to demonstrate the robust performance of our proposed framework. Section 7 concludes. All proofs are relegated to the Appendix. All codes regarding Section 6 can be accessed in the GitHub repository [3].

2. Literature Review

In this section, we review several streams of literature closely connected to our work.

The theory and applications of DML. The proposed DeDL framework stems from the recent advances in the semiparametric estimation and inference, the DML method in particular (e.g., Chernozhukov et al. 2018). Combining ML with Neyman orthogonality, the DML method performs remarkably well in estimating the parameters of interest in the presence of regularization and/or overfitting biases to estimate the nuisance parameter(s). In particular, Farrell et al. (2021) establish novel nonasymptotic high-probability bounds for nuisance parameter estimation with deep feedforward neural nets. Chiang et al. (2022) extend the DML framework by proposing a multiway cross-fitting algorithm suitable for multiway clustering sampled data such as panel data. Chernozhukov et al. (2022) develop an automatic DML framework using Lasso to learn the debiased term that often presents in the influence function directly from data. Combining DML with optimization further promotes its theoretical development in the context of an operation. For example, Qi et al. (2022) propose a personalized pricing algorithm by maximizing the expected revenue estimated using DML. We contribute to this literature by adapting the DML framework in the combinatorial experiment setting. Our DeDL framework is most related to Farrell et al. (2020), in which authors extend the partial linear model in Chernozhukov et al. (2018) into a more general semiparametric form (a pre-specified model layer on the top of neural networks) combining nuisance parameters and high dimension treatments.

[3] See https://github.com/zikunye2/deep_learning_based_causal_inference_for_combinatorial_experiments.git

While their approach is broadly applicable, the generality of their framework leaves the identification requirements under specific settings ambiguous. Their method presupposes the identification holds under the current observational data, enabling valid extrapolation on unobserved treatments, but this assumption is not directly applicable in our combinatorial experiment context. To address this gap, we tailor their method by explicitly identifying the easy-to-check technical conditions required for identification and the convergence of DNN for our application. The inference result of our ATE estimator constructed from the cross-fitting follows Chernozhukov et al. (2018) straightforwardly. Furthermore, we derive the estimator and inference result for the best treatment identification, the technique of which could be extended to broader decision-making problems based on DML estimators.

The DML is not solely the subject of extensive theoretical research; The DML has been extensively applied in many empirical settings for heterogeneous treatment effects. For example, Knaus (2020) employs DML to evaluate the effectiveness of four labor programs in Switzerland. Dube et al. (2020) utilize DML to obtain debiased estimators for the effects of rewards for MTurk workers on project duration to investigate monopsony in online labor markets. Farbmacher et al. (2022) apply DML combined with causal mediation analysis to study the effect of health insurance coverage on general health. Fan et al. (2022) explore the causal effect of maternal smoking on the birth weight of newborn babies via the DML estimator. Leveraging hundreds of experiments on Facebook, Gordon et al. (2023) find that DML implemented with observational data under the selection of ads for users may have substantial biases from the ground truth. While all applications beyond Gordon et al. (2023) use observational data, our research is the first to provide evidence on the performance of the DML method with the data from a set of large-scale field experiments such that the fundamental unconfoundedness assumption is guaranteed. Contrary to the conclusion form Gordon et al. (2023) that DML estimates are far from experimental results, we show that our debiased estimators are accurate and valid for inference, significantly outperforming other benchmarks. Our documented strong empirical validation and comprehensive discussions of the DML method can benefit both researchers and practitioners.

Estimation and inference with multiple experiments. Conventionally, researchers examine multiple-experiment settings through the lens of factorial design (i.e., full or fractional factorial designs). Interested readers are referred to Box et al. (1978) and Wu and Hamada (2011) for detailed discussions of these classical approaches. Recent works (e.g., Dasgupta et al. 2015, Pashley and Bind 2019) marry such design strategies with the potential outcome framework (Imbens and Rubin 2015) for the study of causal inference. However, factorial design is hardly applicable to large-scale A/B testing platforms, where the number of experiments m can potentially be hundreds or even thousands. It is next to impossible to obtain the 2^m treatment groups, as required by the full-factorial design. Even with the fractional factorial design, the sheer number of treatments implies that practically

one can only test $O(m)$ treatment combinations, suggesting that only $O(m)$ direct or interaction effects are identifiable. The vast majority of the effects are, however, aliased away. Therefore, factorial design methods are rarely employed on large-scale A/B testing platforms. Proposing a new strategy to deal with the inference problem in such settings, our work applies the DML framework to the empirical analysis in the multiple-experiment setting and requires looser identification conditions. With an appropriately specified form of the response function to the treatment for each individual, we only need to observe $m+2$ treatment combinations for the treatment effect inference of all 2^m combinations. We also apply this framework to a real multiple-experiment setting on Platform O and show the empirical success of the framework in this setting.

Causal inference and its applications to online platforms. Causal inference has long been a central topic in many fields, including economics, psychology, medical science, marketing, and operations (e.g., see Angrist and Pischke 2009, Wooldridge 2010). Recent advances in ML and high-dimensional statistics have enabled substantial development in this area. To name a few, Xie and Aurisset (2016) and Guo et al. (2021) propose variance-reduction techniques that use covariates to adjust estimators and obtain more precise ones with fewer data. Farias et al. (2021) compute debiased estimators of treatment effects under general intervention patterns, which subsume the synthetic control paradigm. Goli et al. (2022) propose a theoretical framework to overcome the bias due to the interference in a ranking experiment on travel websites. Kallus et al. (2018) use matrix factorization and bound the estimation errors for ATEs to reduce the noise and measurement error in covariates. Lee and Shen (2018) estimate the winner's-curse bias and use it to correct the final estimator for treatment effects. Athey et al. (2018) propose a two-stage approximate residual balancing algorithm to eliminate the bias in estimators obtained through sparse linear models. Arkhangelsky et al. (2021) propose a synthetic difference-in-differences estimator to deal with panel data, which possesses unbiasedness and consistency under regularity assumptions. Zhang and Politis (2022) improve the ridge regression estimator by adding a correction part to debias the original estimator. They use a wild-bootstrap algorithm to construct a confidence interval. An influential school of works combines ML methods with causal inference (Athey and Imbens 2016, Wager and Athey 2018). Among this literature, as discussed above, DML (Chernozhukov et al. 2018, Farrell et al. 2020) has received much attention. Furthermore, operations researchers also apply optimization techniques such as robust optimization (e.g., see Lim et al. 2006) to study the estimation in causal inference (e.g., Bertsimas et al. 2022).

In particular, our paper speaks to the applications of causal inference to online platforms. With a large amount of data available on online platforms, works in this area have proliferated in recent years. On the empirical side, field experiments on large-scale online platforms enable causal inference in a variety of business settings (e.g., Burtch et al. 2015, Cheung et al. 2017, Edelman et al. 2017,

Zhang et al. 2020, Zeng et al. 2022). On the theoretical side, researchers propose innovative methods to overcome challenges arising from online platforms, such as two-sided randomization (e.g., Nandy et al. 2021, Johari et al. 2022, Ye et al. 2022), sequential experiments (e.g., Song and Sun 2021, Bojinov et al. 2022, Xiong et al. 2022), and block randomization (Candogan et al. 2021). Whereas this literature typically focuses on the single-experiment setting, we study the inference problem with multiple experiments.

3. Debiased Deep Learning (DeDL) Framework

In this section, we introduce the DL-based framework following Chernozhukov et al. (2018) and Farrell et al. (2020) for estimating and inferring treatment effects in our multiple experiments setting.[4] As discussed in the introduction, one may view our work as a "playbook" for implementing DNN-backed DML for causal inference in broader settings, and in this section we outline the theoretical foundations using multiple treatment effect estimation as an example. Whereas the procedural details are specifically tailored to address our research questions in the multiple experiment setting, they underscore the critical elements that are broadly essential. For researchers and practitioners interested in other causal inference problems using DML, following a similar procedure is expected to yield the construction of desired estimators. We also emphasize that this section focuses on the theoretical side, and the guidelines for practical implementation are deferred to Section 5.

Our framework comprises 3 steps, starting with determining the setting and specifying the data generation process (DGP), as detailed in Section 3.1 and 3.2. Training the DL model is the second critical step, which requires careful design of a specialized model layer that enables accurate approximation of any ATE function with finite samples (Section 3.3). In the third step, we develop $\sqrt{n}$-consistent DML-type estimators for inference and optimal treatment combination identification (Section 3.4).

3.1. The Setup

Building on the recent advances ML-powered causal inference, we consider the DL-based inference framework for multiple experiments on a large platform. There are m concurrent field experiments on the platform, each with binary treatment levels, represented by $\boldsymbol{T} \in \{0,1\}^m$.[5] Without loss of generality, we focus on the binary treatment case, which is a common practice for A/B tests on large-scale online platforms, but our framework can be readily extended to continuous and discrete treatment

[4] We use "multiple experiments," "multiple treatments," and "combinatorial experiments" interchangeably.

[5] *On notations:* Throughout the paper, vectors and matrices are in boldface. Vectors are written as column vectors, and $\boldsymbol{v}'$ represents the transpose of vector $\boldsymbol{v}$. Random variables are represented by capital letters, and their realizations by lowercase letters. The L_2 norm of function $f(\cdot)$ is defined as $||f(\boldsymbol{x})||_{L_2(\boldsymbol{X})} := \mathbb{E}[f(\boldsymbol{X})^2]^{1/2}$. We use $\mathbb{E}_n$ to denote the sample average and $\boldsymbol{M} \succ 0$ to denote that matrix $\boldsymbol{M}$ is positive-definite.

levels. The platform can observe the individual-level response to the treatment $Y \in \mathbb{R}$,[6] along with the individual-level pretreatment covariates $\boldsymbol{X} \in \mathbb{R}^{d_{\boldsymbol{X}}}$ and treatment level $\boldsymbol{T}$. The treatment assignment mechanism is denoted by the conditional distribution $\nu(\cdot \,|\, \cdot)$, i.e., $\nu(\boldsymbol{t} \,|\, \boldsymbol{x}) = \mathbb{P}[\boldsymbol{T} = \boldsymbol{t} \,|\, \boldsymbol{X} = \boldsymbol{x}]$ for any $\boldsymbol{t} \in \{0,1\}^m$ given any $\boldsymbol{x}$.

In this setting, we address the following two essential questions of both academic and practical values: (a) What is the ATE for each treatment combination? (b) Which treatment combination is the most valuable for the platform (i.e., with the highest ATE)? We refer to the second question as the best treatment identification problem.

3.2. Structured Deep Learning: Specify the Data Generation Process (DGP)

In the first step, following Farrell et al. (2020), we postulate that the data-generating process (DGP) has the semiparametric form

$$\mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{T} = \boldsymbol{t}] = G(\boldsymbol{\theta}^*(\boldsymbol{x}), \boldsymbol{t}), \tag{1}$$

given any $\boldsymbol{x}$ and $\boldsymbol{t} \in \{0,1\}^m$. $G(\cdot,\cdot)$ is the known link function and $\boldsymbol{\theta}^*(\cdot) : \mathbb{R}^{d_{\boldsymbol{X}}} \mapsto \mathbb{R}^{d_{\boldsymbol{\theta}}}$ are the unknown nuisance parameters as functions of covariates $\boldsymbol{x}$.[7] In particular, $\boldsymbol{\theta}^*(\cdot)$ characterizes the heterogeneity in outcomes, and we shall predict them by ML models such as DNNs. The prespecified link function $G(\cdot,\cdot)$ allows for the flexibility and interpretability of the relationship between the outcome Y and the treatment combination $\boldsymbol{t}$. For example, if the link function is linear in $\boldsymbol{t}$ but with heterogeneous coefficients, i.e., $\mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{T} = \boldsymbol{t}] = \boldsymbol{\theta}^*(\boldsymbol{x})'\boldsymbol{t}$, the effect of any treatment combination equals the linear sum of each individual treatment effect therein. Combining the interpretability of the link function and the generalizability of ML, our framework not only provides practitioners with accurate inferences for the experimental outcome, but also delineates the interactions between treatments.

Before we explain the link function in detail, several remarks are in order. First, with slight adjustments, the DGP can be extended to general DML settings for causal inference. Specifically, treatments can still be represented by $\boldsymbol{t}$, accommodating multiple treatment levels or even continuous treatments depending on the application. The researcher selects the G function that fits the context, such as a logistic function for a binary outcome (Farrell et al. 2020) or a partially linear specification (Chernozhukov et al. 2018). In some cases, where the economic interpretation of the G function is well understood, this selection process is straightforward; in others, it requires careful consideration. Our subsequent discussion in this section offers an illustrative example. In particular, the choice of

[6] For expositional ease, we focus on the one-dimensional outcome setting throughout this paper. In practice, online platforms could be interested in multiple outcome metrics (e.g., the number of active users and revenue); and the extension of our framework to the case where $\boldsymbol{Y} \in \mathbb{R}^{d_{\boldsymbol{Y}}}$ $(d_{\boldsymbol{Y}} > 1)$ is straightforward.

[7] Throughout this paper, we make a regularity assumption commonly used in the DNN-estimation literature (e.g., Yarotsky 2017), i.e., Assumption 3 in Appendix A.2, which requires true parameter $\boldsymbol{\theta}^*(\cdot)$ is uniformly bounded and sufficiently smooth.

G function needs to balance the modeling power (see Sections 3.2.1 and 3.2.2) and nice statistical properties to reduce the difficulty of estimating nuisance parameters (see Proposition 1).

Second, it should be noted that in our setting, an alternative, more straightforward method would involve building a pure neural network without any parametric structure, i.e., $\mathsf{DNN}(\boldsymbol{x}, \boldsymbol{t})$, which, however, does not ensure valid inferences and is prone to overfitting the observed combinations. Consequently, a standard pure DNN may not work well because of its potentially poor out-of-sample performance, even with meticulous regularization. We demonstrate in Appendix D that standard regularizations such as dropout layer and Lasso regularization are ineffective for our. In essence, the G function we utilize is a variant of a regularized DNN that has been specifically tailored for our application, drawing upon prior knowledge of interaction patterns. Given that a suitably predefined G function is critical in our application, our study focuses primarily on selecting an appropriate G beyond Farrell et al. (2020), with theoretical underpinnings, strong approximation capabilities, and easy-to-validate and easy-to-satisfy identification conditions. Regarding estimation and inference, the procedures we adopt are straightforward and align with the standard DML (Chernozhukov et al. 2018).

Third, here we also give a brief overview of subsequent theoretical development: having specified the DGP, our framework involves the following two stages. In the first *training* stage, we adopt DL to obtain a consistent estimator of the unknown parameter $\boldsymbol{\theta}^*(\cdot)$, which is denoted by $\hat{\boldsymbol{\theta}}(\cdot)$. In the second *estimation and inference* stage, based on the trained parameter $\hat{\boldsymbol{\theta}}(\cdot)$, we construct asymptotically normal estimators for the quantities of managerial interest (e.g., ATE), thus yielding valid inferences.

3.2.1. Choose the Link Function. As discussed, we capture the richness of individual heterogeneity with the nonparametric function $\boldsymbol{\theta}^*(\cdot)$, while given $\boldsymbol{\theta}^*(\cdot)$, we essentially assume that the individual outcomes are fully structured and described by $G(\cdot,\cdot)$. In our setting, the link functions G may take many different forms. For example, the most straightforward choice of G is the linear function of $\boldsymbol{t}$ with heterogeneous coefficients, i.e., $G(\boldsymbol{\theta}(\boldsymbol{x}), \boldsymbol{t}) = \theta_0(\boldsymbol{x}) + \theta_1(\boldsymbol{x})t_1 + \cdots + \theta_m(\boldsymbol{x})t_m$. Although this simple linear form only requires $m+1$ linear independent $(1, \boldsymbol{t}')$ observed treatment vectors for identification, it clearly fails to capture the treatment interactions. Another extreme is $G(\boldsymbol{\theta}(\boldsymbol{x}), \boldsymbol{t}) = \theta_0(\boldsymbol{x}) + \sum_{i=1}^{m} \theta_i(\boldsymbol{x})t_i + \sum_{i\neq j} \theta_{ij}(\boldsymbol{x})t_i t_j + \ldots + \theta_{12\cdots m}(\boldsymbol{x})t_1 t_2 \cdots t_m$, which contains all heterogeneous high-order interaction terms. Although this is the most accurate link-function form for our application, it also requires the strongest identification condition, i.e., non-zero assignment probability for all treatment combinations, which is infeasible in practice. Thus, the choice of the link function should, on the one hand, reflect the economic nature of the multiple A/B tests business context and, on the other hand, be associated with the proper treatment assignment mechanisms to ensure the identifiability and convergence of the estimates $\hat{\boldsymbol{\theta}}(\cdot)$. To enhance the depth of discussion, we propose the following concrete link functions with clear economic interpretations.

ASSUMPTION 1 (LINK FUNCTIONS). *We consider the following link functions* $G(\boldsymbol{\theta}(\boldsymbol{x}), \boldsymbol{t})$ *where* $\boldsymbol{\theta}(\cdot) : \mathbb{R}^{d_{\boldsymbol{X}}} \mapsto \mathbb{R}^{d_{\boldsymbol{\theta}}}$.

(a) Multiplicative Form. $G(\boldsymbol{\theta}(\boldsymbol{x}), \boldsymbol{t}) = \theta_0(\boldsymbol{x})(1 + \theta_1(\boldsymbol{x})t_1)\ldots(1 + \theta_m(\boldsymbol{x})t_m)$, *and* $\mu \leq 1 + \theta_k(\boldsymbol{x}) \leq M$, $k = 1, \ldots, m$, *uniformly in* $\boldsymbol{x}$, *for some* $M > \mu > 0$.

(b) Standard Sigmoid Form. $G(\boldsymbol{\theta}(\boldsymbol{x}), \boldsymbol{t}) = a/(1+\exp(-(\theta_0(\boldsymbol{x})+\theta_1(\boldsymbol{x})t_1+\cdots+\theta_m(\boldsymbol{x})t_m))) + b$, *where* $a \neq 0$ *and* b *are known constants.*

(c) Generalized Sigmoid Form I. $G(\boldsymbol{\theta}(\boldsymbol{x}), \boldsymbol{t}) = \theta_{m+1}(\boldsymbol{x})/(1+\exp(-(\theta_1(\boldsymbol{x})t_1+\cdots+\theta_m(\boldsymbol{x})t_m)))$.

(d) Generalized Sigmoid Form II. $G(\boldsymbol{\theta}(\boldsymbol{x}), \boldsymbol{t}) = \theta_{m+1}(\boldsymbol{x})/(1 + \exp(-(\theta_0(\boldsymbol{x}) + \theta_1(\boldsymbol{x})t_1 + \cdots + \theta_m(\boldsymbol{x})t_m)))$.

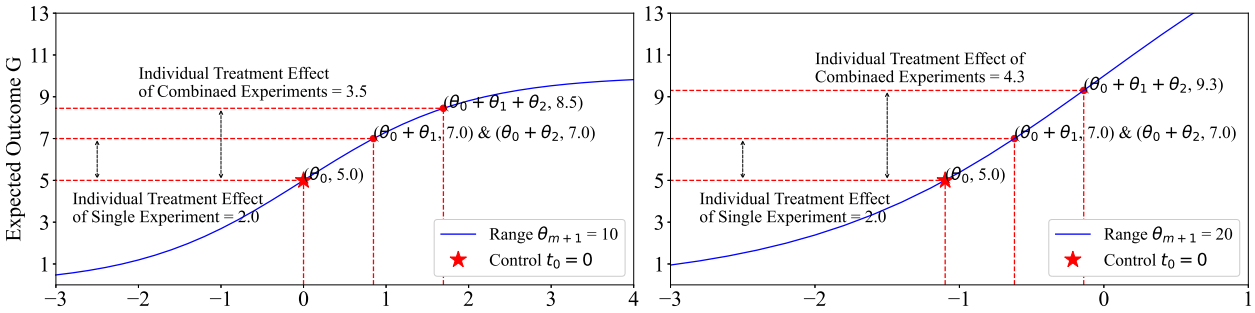


Figure 2 Illustration of Generalized Sigmoid Form II With Two Types Platform Users: Marginal Decreasing (Left) and Marginal Increasing (Right)

All four link functions in Assumption 1 capture the heterogeneity with respect to different covariates $\boldsymbol{x}$. The link function of the *Multiplicative Form* (Assumption 1(a)) assumes multiplicative *relative effect size* for different individual treatments, e.g., if each of two treatments increases the effect by 10%, then combined treatment increases by $(1 + 10\%)(1 + 10\%) - 1 = 21\%$. However, the *Multiplicative Form* can characterize only the increasing marginal effect, due to its global convexity.

The link function of the sigmoid forms (Assumption 1(b), (c), and (d)) leverages the convex-concave structure of the sigmoid function, thus capturing both increasing and decreasing marginal effects at the individual level. As a result, it is able to capture both marginal increasing and decreasing effects in ATE. We illustrate the coexistence of decreasing and increasing marginal effects with the *Generalized Sigmoid Form II* through an example. We consider two experiments ($\boldsymbol{t}_i \in \{0,1\}^2$) and two user types, as shown in Figure 2, where the y-axis represents an individual's average outcome $\mathbb{E}[Y_i|\boldsymbol{X}_i, \boldsymbol{T}_i]$ follows the *Generalized Sigmoid Form II* (Assumption 1(d)). Under the control condition (i.e., $\boldsymbol{t} = \boldsymbol{t}_0 = (0,0)'$), the expected outcome of both user types is 5 (i.e., $\mathbb{E}[Y_i|\boldsymbol{X}_i, \boldsymbol{T} = (0,0)'] = 5$ for all i), whereas the treatment effect of each experiment on each user type is 2 (i.e., $\mathbb{E}[Y_i|\boldsymbol{X}_i, \boldsymbol{T} = (1,0)'] = \mathbb{E}[Y_i|\boldsymbol{X}_i, \boldsymbol{T} = (0,1)'] = 7$ for all i). For the first (resp. second) user type, $\theta_{m+1} = 10$ (resp.

$\theta_{m+1} = 20$). Straightforward calculation implies that $\mathbb{E}[Y_i|\boldsymbol{X}_i, \boldsymbol{T} = (1,1)'] = 8.5$ (thus, the individual treatment effect is $8.5 - 5 = 3.5 < 2 + 2 = 4$, suggesting the decreasing marginal effects) for the first user type and $\mathbb{E}[Y_i|\boldsymbol{X}_i, \boldsymbol{T} = (1,1)'] = 9.3$ for the second user type (thus, the individual treatment effect is $9.3 - 5 = 4.3 > 4$, suggesting the increasing marginal effects). Therefore, if the first type of users takes more (resp. less) than 36% of the entire population, the platform will have a decreasing (resp. an increasing) marginal ATE.

The *Standard Sigmoid Form* with known constants a and b may be restrictive and cannot model the different outcome ranges across individuals. The *Generalized Sigmoid Forms I* and *II* resolve this issue by incorporating the parameter $\theta_{m+1}(\boldsymbol{x})$. Comparing three sigmoid link functions, one can notice that the *Standard Sigmoid Form* (Assumption 1(b)) and the *Generalized Sigmoid Form I* (Assumption 1(c)) are special cases of the *Generalized Sigmoid Form II* (Assumption 1(d)). Hence, we adopt the link function of *Generalized Sigmoid Form II* in our empirical study.

3.2.2. Justify the Link Function: Universal Approximation. To better illustrate the expressive power of our proposed *Generalized Sigmoid Form II* to capture ATEs, we present a counterintuitive example in Table 1, where the ATE of the individual treatments has a different sign from that of the combined treatment. There are two heterogeneous individuals and two individual experiments. We assume the individual response follows our *Generalized Sigmoid Form II* with the specific parameters detailed in the note of Table 1. It is apparent that, in this example, the ATEs for the first and second individual treatments are -0.04 and -0.06, respectively. However, the combined treatment effect has a different sign, i.e., 0.11. As a structured sigmoid function with only $O(m)$ parameters, our individual-level link function seems restrictive at the first glance, but the ATE averaging over the whole population can be quite flexible.

Furthermore, to highlight our proposed link function is indeed of practical interest, we develop the following bound on the approximation power of the *Generalized Sigmoid Form II* in Theorem 1, which shows that this link function can be used to approximate arbitrary ATEs with a provable finite-sample error bound. This result echoes the well-established expressive capabilities of the sigmoid function, as evidenced in both the theoretical and empirical studies of neural networks. This result confirms that the use of sigmoid function adds great versatility to our model. Interested readers may refer to the proof of the theorem for details in Appendix A.

THEOREM 1 (UNIVERSAL APPROXIMATION OF GENERALIZED SIGMOID LINK FUNCTION). *Given a population that contains $K \geq 1$ distinct feature vectors $\{\boldsymbol{x}_\ell \in \mathbb{R}^{d_{\boldsymbol{X}}} : \ell = 1, \ldots, K\}$ such that $\boldsymbol{X} = \boldsymbol{x}_\ell$ with probability $p(\ell) > 0$ for all $\ell = 1, \ldots, K$, where $\sum_{\ell=1}^{K} p(\ell) = 1$, and any outcome function $f(\boldsymbol{t}) : \{0,1\}^m \mapsto \mathbb{R}$, there exists a feature mapping $\{\boldsymbol{\theta}(\boldsymbol{x}_\ell) : \ell = 1, \ldots, K\}$ such that*

$$\left(\frac{1}{2^m} \sum_{t \in \{0,1\}^m} \left(f(\boldsymbol{t}) - \mathbb{E}_{\boldsymbol{X}} \left[G(\boldsymbol{\theta}(\boldsymbol{X}), \boldsymbol{t}) \right] \right)^2 \right)^{\frac{1}{2}} \lesssim \frac{1}{\sqrt{K}},$$

Table 1 Counterintuitive Example under Generalized Sigmoid Form II

Treatment Combination $\boldsymbol{t}'$	HTE of Individual 1	HTE of Individual 2	ATE
(1,0)	0.92	-1.00	-0.04
(0,1)	0.49	-0.60	-0.06
(1,1)	1.27	-1.05	0.11

Note: Parameter $\boldsymbol{\theta}' = (\theta_0, \theta_1, \theta_2, \theta_3)$ for the first individual is $(0, 1, 0.5, 4)$, $\boldsymbol{\theta}'$ for the second individual is $(-1, -3, -1, 4)$.

where $G(\cdot,\cdot)$ *is the Generalized Sigmoid Form II of individual response* $G(\boldsymbol{\theta}(\boldsymbol{x}), \boldsymbol{t}) = \theta_{m+1}(\boldsymbol{x})/(1 + \exp(-(\theta_0(\boldsymbol{x}) + \theta_1(\boldsymbol{x})t_1 + \cdots + \theta_m(\boldsymbol{x})t_m)))$.[8]

We comment on several observations. First, the omitted constant in Theorem 1 depends on the function $f(\cdot)$ and m and is independent of K. This result highlights how the approximation power of our proposed individual-level sigmoid model, i.e., *Generalized Sigmoid Form II*, improves as the number of distinct feature vectors K increases: it proves that one can use this link function to approximately construct any ATE function and captures arbitrary interactions among individual treatments, with an error of order $\mathcal{O}(K^{-1/2})$. We interpret K as the number of individuals with distinct features, suggesting that an increase in the number of distinct individual profiles in the population enhances the potential for a more precise approximation of the ATE function.

Second, this theoretical development is motivated by the classical idea of approximating continuous functions using sigmoid activation units in neural networks (e.g., see Hornik et al. 1989). Specifically, the result is deeply rooted in the classical mathematical theory of Fourier transformation. In our setting, where treatments are modeled as Boolean variables, the Fourier transformation of $f(\boldsymbol{t})$ can be expressed using indicator functions, which can be effectively approximated by sigmoid functions, given sufficient flexibility in the feature space. This underscores the modeling power of our link function—we are not aware of similar approximation results with other link functions, such as multiplicative or simple linear ones.

Third, we acknowledge that such an approximation result does not ensure the practical realizability of the nuisance parameters $\boldsymbol{\theta}(\cdot)$ proven to exist by Theorem 1, nor does it assure a precise estimate of the ATE. This limitation may be attributed to several factors. The first is that the link function $G(\cdot,\cdot)$ can indeed be misspecified. The model misspecification issue is further investigated and discussed in our synthetic experiments in Section 6. Second, as is well-documented in the computer science literature (e.g., Adcock and Dexter 2021), there exists a large gap between the theoretical approximation ability of DNNs and their actual performance, which may be resolved by designing better DNN architectures and training algorithms. The third factor is that the nuisance parameters may not be identifiable under the assignment mechanism of the A/B tests, especially when the focus is too narrow on specific treatment combinations without sufficient explorations. We discuss the

[8] The notation "$\lesssim$" means less than or equal to up to a constant independent of K.

detailed identification conditions of our proposed link functions, including the *Generalized Sigmoid Form II* following Proposition 1.

Finally, a closer examination on the proof of Theorem 1 reveals the deliberate and intricate design of our *Generalized Sigmoid Form II*, compared with the *Standard Sigmoid Form* and the *Standard Sigmoid Form I*. Incorporating a feature-dependent scale parameter, $\theta_{m+1}(\boldsymbol{x})$ and intercept, $\theta_0(\boldsymbol{x})$ appears crucial for realizing the expressive power emphasized in Theorem 1. Moreover, while we could generalize the link function to include more terms, such as $G(\boldsymbol{\theta}(\boldsymbol{x}),\boldsymbol{t}) = \theta_{m+1}(\boldsymbol{x})/(1+\exp(-(\theta_0(\boldsymbol{x})+\theta_1(\boldsymbol{x})t_1+\cdots+\theta_m(\boldsymbol{x})t_m)))+\theta_{m+2}(\boldsymbol{x})$, the proof also indicates that this additional complexity does not deliver significantly better approximations of the ATEs. From an empirical point of view, such additional complexity also makes identifying the nuisance parameters more challenging, which may not be justified by the increased challenge in identification.

3.3. Training Stage

In the second step – *training* stage, we use DNNs to approximate the unknown parameter $\boldsymbol{\theta}^*(\cdot)$, motivated by our DGP (1). We add a model layer to represent the link function on top of the DNNs to connect the outcome with the estimated parameters $\hat{\boldsymbol{\theta}}(\cdot)$ and treatment level $\boldsymbol{t}$. Figure 3 illustrates the whole model from $\boldsymbol{X}$ and $\boldsymbol{T}$ to Y (see also Figure 2 in Farrell et al. 2020). We use $\{(y_i,\boldsymbol{x}_i',\check{\boldsymbol{t}}_i')' : 1\le i\le n\}$ to denote the realization of random vector $(Y,\boldsymbol{X}',\boldsymbol{T}')'$, i.e., the observed data from experiments. We use the check symbol over the treatment $\check{\boldsymbol{t}}$ to represent the realized treatment level in experiments to avoid confusion by the notation $\boldsymbol{t}\in\{0,1\}^m$ representing any target treatment in ATE as defined in (3).

3.3.1. Specify the Loss Function. The loss function measures the quality of the estimated nuisance parameter, and plays an important role in the training stage. Depending on the application, different loss function can be used. For example, the maximum log-likelihood can used in case of binary outcomes. As a general guideline, certain smoothness assumption is needed (e.g., Assumption 1 in Farrell et al. 2020) to obtain a theoretical guarantee (e.g., Proposition 1 in our setting). Our DGP (1) suggests that the true parameter functions solve $\boldsymbol{\theta}^*(\cdot)\in\arg\min_{\boldsymbol{\theta}(\cdot)}\mathbb{E}[(Y-G(\boldsymbol{\theta}(\boldsymbol{X}),\boldsymbol{T}))^2]$. Consequently, we use the squared error, denoted by $\ell(y_i,\check{\boldsymbol{t}}_i',\boldsymbol{\theta}(\boldsymbol{x}_i)) = (y_i-G(\boldsymbol{\theta}(\boldsymbol{x}_i),\check{\boldsymbol{t}}_i'))^2$ as the per-observation loss function used to train the estimator $\hat{\boldsymbol{\theta}}(\cdot)$. The estimators of $\boldsymbol{\theta}^*(\cdot)$ can be obtained by minimizing the empirical loss on the training data set

$$\hat{\boldsymbol{\theta}}(\cdot) := \underset{\boldsymbol{\theta}(\cdot)\in\mathcal{F}_{\text{DNN}}}{\arg\min}\ \frac{1}{n}\sum_{i=1}^{n}\ell(y_i,\check{\boldsymbol{t}}_i,\boldsymbol{\theta}(\boldsymbol{x}_i)) := \frac{1}{n}\sum_{i=1}^{n}(y_i-G(\boldsymbol{\theta}(\boldsymbol{x}_i),\check{\boldsymbol{t}}_i))^2, \tag{2}$$

where $\mathcal{F}_{DNN}$ is the set of fully connected neural nets with bounded outputs. Note that we use DNNs to approximate $\boldsymbol{\theta}(\boldsymbol{x})$ for two reasons. (1) DNNs are easier to engineer with a general link function G, and one can leverage off-the-shelf packages (e.g., PyTorch and TensorFlow) to train models at scale.

(2) DNNs have better approximation and prediction power in general compared to other ML models. Our framework can be readily applied to other ML models to get $\hat{\boldsymbol{\theta}}(\boldsymbol{x})$ as long as the convergence rate is upperbounded by $o(n^{-1/4})$.

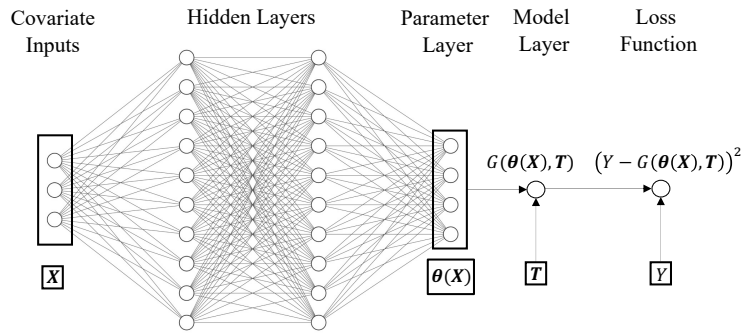


Figure 3 Illustration of Deep Neural Network with the Structured Model Layer

3.3.2. Identifiability and Convergence of Feature Mappings. With our proposed link functions, we are ready to show the identifiability and convergence rate of the DNN in our framework. The proof strategy largely builds on the approach in Farrell et al. (2020). Our primary theoretical contribution here is the introduction of practical, primitive conditions in out setting that ensure the identifiability of our model and facilitate the verification of the convergence conditions outlined in Farrell et al. (2020). Using multiple experiment treatment effect estimation as a test bed, we demonstrate the type of analysis required to provide theoretical guarantees for the DML framework. Specifically, for identifiability, we further assume the treatment assignment mechanism is sufficiently "regular" (see Assumption 4 in Appendix A.2). For convergence, the additional assumption on the data observations being *i.i.d.*, bounded, and $\boldsymbol{\theta}^*(\boldsymbol{x})$ being sufficiently smooth is also imposed (see Assumption 3 in Appendix A.2). Formally, we have the following proposition, the proof of which is relegated to Appendix A.4.

PROPOSITION 1 (IDENTIFIABILITY AND CONVERGENCE). *The following statements hold.*

(a) Under Assumptions 1 and 4 (in Appendix A.2), the parameter function $\boldsymbol{\theta}^*(\boldsymbol{x})$ can be nonparametrically identified in DGP (1).

(b) Under Assumptions 1, 3 (in Appendix A.2), and 4 (in Appendix A.2), if the structured DNN as illustrated in Figure 3 has width $H = O(n^{d_{\boldsymbol{X}}/2(p+d_{\boldsymbol{X}})} \log^2 n)$ and depth $L = O(\log n)$, there exists a positive constant C that depends on the fixed quantities in Assumption 3, such that with probability at least $1 - \exp(n^{-d_{\boldsymbol{X}}/(p+d_{\boldsymbol{X}})} \log^8 n)$, it holds that

$$\|\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*\|_{L_2(\boldsymbol{X})}^2 \lesssim n^{-\frac{p}{p+d_{\boldsymbol{X}}}} \log^8 n + \frac{\log\log n}{n},$$

for each $k \in [d_{\boldsymbol{\theta}}]$ when n is large enough.

The key step to prove Proposition 1 is translating the convergence of DNN estimation on outcomes Y into that of the parameter function estimates $\hat{\boldsymbol{\theta}}(\cdot)$ under the treatment assignment mechanism sufficient for identification. The convergence rate given by Proposition 1 may not be tight (see Farrell et al. 2021), but it is sufficiently fast for the subsequent inference in our setting if $p > d_{\boldsymbol{X}}$.

Another important implication of Proposition 1 is that, for our link functions in Assumption 1, it suffices to observe $m+2$ treatment combinations (see Assumption 4 in Appendix A.4 for details), which is orders of magnitude smaller than 2^m, to ensure the identifiability and sufficiently fast convergence. In other words, suppose there are 10 different treatments, and in turn $2^{10} = 1024$ possible treatment combinations. Our framework needs to observe only $10+2 = 12$ combinations to estimate the parameter function $\hat{\boldsymbol{\theta}}(\cdot)$ with sufficient convergence, which is only $12/1024 = 1.2\%$ of the total possible combinations. It should be noted that the requirement for $m+2$ combinations is not arbitrary. Take the *Generalized Sigmoid Form II* for example. We first stipulate that $\mathbb{E}[\tilde{\boldsymbol{T}}\tilde{\boldsymbol{T}}'|\boldsymbol{X}] \succ 0$ almost uniformly everywhere—that is, its smallest eigenvalue is uniformly lower bounded away from 0 across all $\boldsymbol{X}$, except on a set of zero probability. Specifically, this condition is readily met under a common and lenient treatment assignment rule: Each of the m individual treatment condition, as well as the full control condition, are assigned with a positive probability. In other words, we assign $\boldsymbol{t} = (0,0,\ldots,0)'$ and $(1,0,\ldots,0), (0,1,\ldots,0)', \ldots,\ (0,0,\ldots,1)$ each with a positive probability. This assumption is relatively mild. The $m+1$ linearly independent combinations $(1,\boldsymbol{t}')$ integrate seamlessly because the link function includes a linear component $\theta_0(\boldsymbol{x}) + \theta_1(\boldsymbol{x})t_1 + \cdots + \theta_m(\boldsymbol{x})t_{m+1}$. Beyond these $m+1$ combinations, we note that with a total of $m+2$ unknown parameters in the *Generalized Sigmoid Form II*, at least one additional treatment combination should be assigned with a positive probability for identification. For example, if two conditions $\boldsymbol{t} = (1,0,0,0)'$ and $(0,1,0,0)'$ are already assigned when $m = 4$, only one additional overlapping condition $(1,1,0,0)'$ is required to identify the nuisance parameter $\theta_{m+1}(\boldsymbol{x})$ in the numerator. Such modest requirement for identification stems from the well-structured nature of the proposed link function. More generally, proposing a more complex link function would entail a more demanding identification challenge.

3.4. Estimation and Inference Stage

In the *estimation and inference* stage, the key is to construct estimators for (a) the ATE of any treatment combination and (b) the improvement in ATE for the identified best treatment over any other treatment combination. We first define the *advantage function* of the treatment combination $\boldsymbol{t}^1 \in \{0,1\}^m$ over the treatment combination $\boldsymbol{t}^2 \in \{0,1\}^m$ as

$$H(\boldsymbol{x}, \boldsymbol{\theta}(\boldsymbol{x}); \boldsymbol{t}^1, \boldsymbol{t}^2) \ := \ G(\boldsymbol{\theta}(\boldsymbol{x}), \boldsymbol{t}^1) - G(\boldsymbol{\theta}(\boldsymbol{x}), \boldsymbol{t}^2).$$

Thus, the ground-truth ATE of any treatment combination $\boldsymbol{t} \in \{0,1\}^m$ can be written as

$$\mu(\boldsymbol{t}) \ = \ \mathbb{E}[G(\boldsymbol{\theta}^*(\boldsymbol{X}), \boldsymbol{t})] - \mathbb{E}[G(\boldsymbol{\theta}^*(\boldsymbol{X}), \boldsymbol{t}_0)] \ = \ \mathbb{E}[H(\boldsymbol{X}, \boldsymbol{\theta}^*(\boldsymbol{X}); \boldsymbol{t}, \boldsymbol{t}_0)]. \tag{3}$$

Denote $\hat{\mu}(\boldsymbol{t})$ as our proposed estimator for $\mu(\boldsymbol{t})$, we show that $\hat{\mu}(\cdot)$ is asymptotically normal and semiparametric efficient. Analogously, the ATE increment of the best treatment $\boldsymbol{t}^*$ over any $\boldsymbol{t} \in \{0,1\}^m$ is written as

$$\tau(\boldsymbol{t}) \;:=\; \mu(\boldsymbol{t}^*) - \mu(\boldsymbol{t}) \;=\; \mathbb{E}[G(\boldsymbol{\theta}^*(\boldsymbol{X}), \boldsymbol{t}^*)] - \mathbb{E}[G(\boldsymbol{\theta}^*(\boldsymbol{X}), \boldsymbol{t})] \;=\; \mathbb{E}[H(\boldsymbol{X}, \boldsymbol{\theta}^*(\boldsymbol{X}); \boldsymbol{t}^*, \boldsymbol{t})]. \tag{4}$$

Notice that identifying the best treatment boils down to identifying the ATE increment of each treatment combination. For the validity of inference of identified best treatment, we also test the one-sided hypothesis $\tau(\boldsymbol{t}) \geq 0$ for all $\boldsymbol{t}$. Letting $\hat{\tau}(\boldsymbol{t}) := \hat{\mu}(\hat{\boldsymbol{t}}^*) - \hat{\mu}(\boldsymbol{t})$ be an estimator of $\tau(\boldsymbol{t})$ for each $\boldsymbol{t} \in \{0,1\}^m$, we can also show that our proposed estimator $\hat{\tau}(\cdot)$ is also $\sqrt{n}$-consistent, and the empirical best treatment agrees with the true best treatment with probability approaching one. We mention that for brevity, this subsection focuses on ATE estimation and inference discussion, whereas the detailed discussion of best treatment identification is relegated to Appendix A.8.

3.4.1. Construct the Influence Function. One cannot directly use the plug-in estimator $\hat{\mu}(\boldsymbol{t}) = \frac{1}{n}\sum_{i=1}^{n} H(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}(\boldsymbol{x}_i); \boldsymbol{t}, \boldsymbol{t}_0)$ for ATE; this estimator is generally not $\sqrt{n}$-consistent due to the large bias of ML models. Intuitively, the issue with this naïve estimator is that the error of nuisance parameter is accounted for. To solve this, the inference with DML is mainly built upon the semiparametric technique—influence function (a.k.a. Neyman orthogonal score)—which implies the first order insensitive to perturbations in the nuisance parameters. We refer interested readers to Newey (1994) and Subsection 2.2.5 in Chernozhukov et al. (2018) for more discussion of influence functions and Neyman orthogonality. Similar to other works using influence function in semiparametric statistics, we make the following assumption.

Assumption 2. *For all $\boldsymbol{t} \in \{0,1\}^m$, the following conditions hold uniformly with respect to all $\boldsymbol{x}$: (i) The DGP (1) holds; (ii) $\boldsymbol{\Lambda}(\boldsymbol{x}) := 2\mathbb{E}[G_{\boldsymbol{\theta}}(\boldsymbol{\theta}(\boldsymbol{x}), \boldsymbol{T}) G_{\boldsymbol{\theta}}(\boldsymbol{\theta}(\boldsymbol{x}), \boldsymbol{T})' | \boldsymbol{X} = \boldsymbol{x}]$ is invertible with bounded inverse, where $G_{\boldsymbol{\theta}}(\boldsymbol{\theta}, \boldsymbol{t})$ is the gradient of $G(\boldsymbol{\theta}, \boldsymbol{t})$ with respect to $\boldsymbol{\theta}$; and (iii) The ATE $\mu(\boldsymbol{t})$ is identified and pathwise differentiable.*

We remark that Assumption 2 imposes several regularity conditions, which are standard and not restrictive in the literature on semiparametric statistics. The invertibility of $\boldsymbol{\Lambda}(\boldsymbol{x})$ is commonly assumed for deriving the influence function, and it can be easily satisfied in our setting. As shown in Appendix A.6, this invertibility condition can be translated into a lenient one under the *Generalized Sigmoid Form II*. Finally, the identification of $\mu(\boldsymbol{t})$ immediately follows Proposition 1(a), and the pathwise differentiability of $\mu(\boldsymbol{t})$ is a standard regularity condition. Such assumptions should be considered as standard in applications of DML. They facilitate the construction of the influence function, whose idea is attributed to Farrell et al. (2020), and we adapt it to our context.

PROPOSITION 2 (INFLUENCE FUNCTION). *Suppose Assumptions 1, 2, 3 (in Appendix A.2), and 4 (in Appendix A.2) hold, then, the influence function for* $\mu(\boldsymbol{t})$ *is* $\psi(\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\Lambda};\boldsymbol{t},\boldsymbol{t}_0)-\mu(\boldsymbol{t})$ *with,*

$$\psi(\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\Lambda};\boldsymbol{t},\boldsymbol{t}_0)=H(\boldsymbol{x},\boldsymbol{\theta}(\boldsymbol{x});\boldsymbol{t},\boldsymbol{t}_0)-H_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{\theta}(\boldsymbol{x});\boldsymbol{t},\boldsymbol{t}_0)'\boldsymbol{\Lambda}(\boldsymbol{x})^{-1}\ell_{\boldsymbol{\theta}}(y,\check{\boldsymbol{t}},\boldsymbol{\theta}(\boldsymbol{x})), \tag{5}$$

where $\boldsymbol{z}=(y,\boldsymbol{x}',\check{\boldsymbol{t}}')'$ *is observed data,* $\boldsymbol{\Lambda}(\boldsymbol{x}):=2\mathbb{E}[G_{\boldsymbol{\theta}}(\boldsymbol{\theta}(\boldsymbol{x}),\boldsymbol{T})G_{\boldsymbol{\theta}}(\boldsymbol{\theta}(\boldsymbol{x}),\boldsymbol{T})'|\boldsymbol{X}=\boldsymbol{x}]$*,* $G_{\boldsymbol{\theta}}$ *is gradient of* G *with respect to* $\boldsymbol{\theta}$*,* $H_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{\theta}(\boldsymbol{x});\boldsymbol{t},\boldsymbol{t}_0):=G_{\boldsymbol{\theta}}(\boldsymbol{\theta}(\boldsymbol{x}),\boldsymbol{t})-G_{\boldsymbol{\theta}}(\boldsymbol{\theta}(\boldsymbol{x}),\boldsymbol{t}_0)$ *is the gradient of* H *with respect to* $\boldsymbol{\theta}$*, and* $\ell_{\boldsymbol{\theta}}(y,\check{\boldsymbol{t}},\boldsymbol{\theta}(\boldsymbol{x})):=2G_{\boldsymbol{\theta}}(\boldsymbol{\theta}(\boldsymbol{x}),\check{\boldsymbol{t}})(G(\boldsymbol{\theta}(\boldsymbol{x}),\check{\boldsymbol{t}})-y)$ *is gradient of* ℓ *with respect to* $\boldsymbol{\theta}$*.*

The influence function defined by (5) contains a plug-in term $H(\boldsymbol{x},\boldsymbol{\theta}(\boldsymbol{x});\boldsymbol{t},\boldsymbol{t}_0)$ and a debiasing term $-H_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{\theta}(\boldsymbol{x});\boldsymbol{t},\boldsymbol{t}_0)'\boldsymbol{\Lambda}(\boldsymbol{x})^{-1}\ell_{\boldsymbol{\theta}}(y,\check{\boldsymbol{t}},\boldsymbol{\theta}(\boldsymbol{x}))$. Therefore, we call the framework *the debiased deep learning* (DeDL). The computation of $\boldsymbol{\Lambda}(\boldsymbol{x})$ is straighforward under the known treatment assignment mechanism $\nu(\cdot\,|\,\boldsymbol{x})$.

3.4.2. Complete the Framework: Asymptotic Normality. Based on this influence function and the cross-fitting technique (e.g., Chernozhukov et al. 2018, Farrell et al. 2020), we can construct estimators as illustrated in Algorithm 1. We refer interested readers to Appendix A.7 for the details of constructing the estimators $\hat{\mu}_{\mathsf{DeDL}}(\boldsymbol{t})$ by (33) and $\hat{\Psi}_{\mathsf{DeDL}}(\boldsymbol{t};\mu)$ by (34), and the confidence interval $\widehat{\mathcal{CI}}_{\mathsf{DeDL}}(\boldsymbol{t};\mu)$ by (35).

Algorithm 1 Implementing DeDL with Cross-fitting

1: (Cross-fitting) Split data samples into S nonoverlapping folds $\mathcal{S}_s$, $s=1,\ldots,S$.
2: (Training) For each fold s, use the complement of $\mathcal{S}_s$ to train DNN to get $\hat{\boldsymbol{\theta}}_s(\cdot)$ based on (2), and compute $\hat{\boldsymbol{\Lambda}}_s(\cdot)=2\mathbb{E}[G_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}_s(\boldsymbol{x}),\boldsymbol{T})G_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}_s(\boldsymbol{x}),\boldsymbol{T})'|\boldsymbol{X}=\boldsymbol{x}]$.
3: (ATE Estimation and Inference) For each $\boldsymbol{t}\in\{0,1\}^m$, leverage the influence function ψ and use data $\mathcal{S}$ to construct the ATE estimator $\hat{\mu}_{\mathsf{DeDL}}(\boldsymbol{t})$ and variance estimator $\hat{\Psi}_{\mathsf{DeDL}}(\boldsymbol{t};\mu)$. Conduct ATE inference based on $\hat{\mu}_{\mathsf{DeDL}}(\boldsymbol{t})$ and $\hat{\Psi}_{\mathsf{DeDL}}(\boldsymbol{t};\mu)$.
4: (Best Treatment Identification) Find empirical best treatment $\hat{\boldsymbol{t}}^*:=\arg\max_{\boldsymbol{t}}\hat{\mu}(\boldsymbol{t})$. Similarly, use influence function ψ and cross-fitting to construct estimators $\hat{\tau}_{\mathsf{DeDL}}(\boldsymbol{t})$ and $\hat{\Psi}_{\mathsf{DeDL}}(\boldsymbol{t};\tau)$ (see Appendix A.8) for the inference on best treatment identification.

We now present Theorem 2 on the asymptotic normality.

THEOREM 2 (ASYMPTOTIC NORMALITY). *Suppose Assumptions 1, 2, 3 (in Appendix A.2), and 4 (in Appendix A.2) hold and* $\hat{\boldsymbol{\Lambda}}_s(\boldsymbol{x}_i)$ *is uniformly invertible. Furthermore, we assume for all subsamples* $s=1,2,\ldots,S$*, the estimators obey* $\|\hat{\theta}_{sk}-\theta_k^*\|_{L_2(\boldsymbol{X})}=o(n^{-1/4})$*,* $k\in\{1,\ldots,d_{\boldsymbol{\theta}}\}$*, which holds under the assumptions and regularity conditions (for the structured DNN) of Proposition 1(b).*

(a) For any treatment level $\boldsymbol{t} \in \{0,1\}^m$,

$$\sqrt{n}\big(\hat{\Psi}_{\mathsf{DeDL}}(\boldsymbol{t};\mu)\big)^{-1/2}\big(\hat{\mu}_{\mathsf{DeDL}}(\boldsymbol{t}) - \mu(\boldsymbol{t})\big) \to_d \mathcal{N}(0,1).$$

(b) Furthermore suppose the best treatment $\boldsymbol{t}^* := \arg\max_{\boldsymbol{t}\in\{0,1\}^m} \mu(\boldsymbol{t})$ *is unique. We have* $\hat{\boldsymbol{t}}^* = \boldsymbol{t}^*$ *with probability approaching one as the sample size goes to infinity, and for any treatment level* $\boldsymbol{t} \in \{0,1\}^m$,

$$\sqrt{n}\big(\hat{\Psi}_{\mathsf{DeDL}}(\boldsymbol{t};\tau)\big)^{-1/2}\big(\hat{\tau}_{\mathsf{DeDL}}(\boldsymbol{t}) - \tau(\boldsymbol{t})\big) \to_d \mathcal{N}(0,1).$$

The formal proof of Theorem 2 can be found in Appendix A.8. Importantly, the probability of failing to identify the true best treatment vanishes as the sample size grows large. Therefore, Theorem 2 establishes the valid inference for ATE and best treatment identification under our DeDL framework. For the rest of this paper, we validate this framework with both experimental and synthetic data, and we demonstrate its superior performance over commonly used benchmarks.

As a concluding remark of this section, we summarize the key checkpoints for conducting DML-based causal inference using DNNs that are generically applicable to other settings. The first step involves specifying the DGP (Section 3.2). Particular attention should be given to designing the link function G (Section 3.2.1) to balance modeling power (Section 3.2.2) and the statistical properties that enhance the estimation of the nuisance parameters (see Proposition 1). In the second step, the DNN is trained using the available data and the link function as the top layer of the network (Section 3.3). Here, it is crucial to choose a loss function that possesses favorable theoretical properties for training (Section 3.3.1). If the link function G and the loss function are both appropriately specified, one can expect theoretical guarantees for the convergence of the nuisance parameters (Proposition 1). For robustness, cross-fitting is commonly applied (Algorithm 1). This technique splits the data into overlapping folds, with each fold's complement used to train a separate DNN estimator of the nuisance parameters. In the third step, where treatment effect estimation and inference are performed (Section 3.4), influence functions are constructed to mitigate the bias from machine learning errors (Proposition 2). Using these influence functions and the estimated nuisance parameters from cross-fitting, one construct the treatment effect estimator and the associated confidence interval (Algorithm 1). If each step is carefully carried out, one can show that the final estimator of the treatment effect is asymptotically normal, which enables statistical inference (Theorem 2).

4. Application to Field Experiment Data

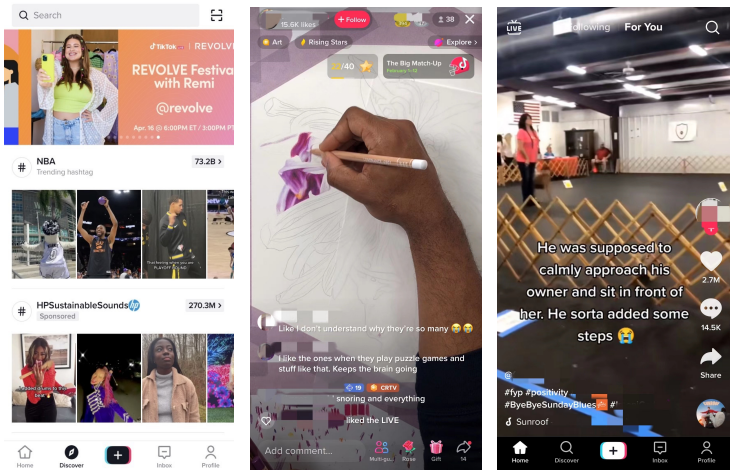
In this section, we conduct field experiments to test our theory. We apply our DeDL framework to the experimental data from Platform O. The empirical results highlight that, in the presence of unobserved treatment combinations, our approach more accurately estimates the ATE of any treatment combination, and identifies the optimal combination, than the commonly used benchmarks.

4.1. Field Setting, Experiments, and Data

In this section, we introduce the setup of our empirical setting. This empirical application highlights a unique experiment setting that allows us to verify the success of our DeDL framework with *observable* ground truth. To our knowledge, this is the first large-scale empirical validation based on large-scale field experiment.

4.1.1. Platform O and the Experimental Background. To empirically validate our proposed framework, we collaborate with Platform O, which features interactive short videos. Platform O, one of the largest short-video platforms, serves billions of users globally every day. Its users (referred to as "she" hereafter) may view the short videos on different product pages.

In our empirical analysis, we focus on three main pages of Platform O. To better illustrate, we refer to similar pages on TikTok: (i) the Discover Page (DP), (ii) the Live Page (LP), and (iii) the For You Page (FYP) (see Figure 4). On the DP, the platform generates trendy hashtags and videos based on users' preferences. On the LP, users are exposed to live streams. On the FYP, the platform recommends the best-performing videos (measured by, e.g., total click-throughs, total watch-time duration, like rate, and forward rate) that fit each user's idiosyncratic interests. Users of Platform O can easily switch to any of these pages at any time they are using the platform.



(a) Discover Page (DP) (b) Live Page (LP) (c) For You Page (FYP)

Figure 4 An Illustration of the Three Pages

Like all other large-scale UGC platforms such as Facebook and TikTok, Platform O runs hundreds of A/B tests daily to evaluate and optimize its product designs and recommendation algorithms. For most of these A/B tests, the platform's main objective is to improve user engagement, which can be well approximated by the amount of screen time a user spends on the platform per day. Each experiment is randomized based on a distinct hash function of user IDs, which ensures that the treatment assignment mechanisms from any two experiments are independent.

In this paper, we focus on a unique set of three A/B tests or treatments, each of which examines the effect of a major adjustment to the video recommendation algorithm on one of three main pages of Platform O (i.e., DP, LP, and FYP).[9] This set of experiments has one unique feature that the other sets of experiments on Platform O do not have: since these three experiments were run on the same population, the outcomes of the users under all $2^3 = 8$ possible treatment combinations are all *observable*. The main reason that the algorithm team in Platform O decided to test all three experiments on the same population is precisely that they want to understand how much money they have left on the table by running each experiment independently and not finding the best combination. Although we admit three is the smallest number of individual treatments that could be interesting, we highlight our orthogonal-experiment data set is of large scale and high quality, delivering trustworthy empirical evidence on the performance of double machine learning in the real setting. To validate the applicability of our framework to a larger number of experiments, we conduct simulations with synthetic data for $m = 10$ in Appendix D.2. If the number of experiments is even larger, e.g., $m = 100$ or 1000, we suggest first reducing m by filtering out insignificant treatments, which take a large proportion in practice. Another widely adopted practice is combining some similar experiments as one synthetic larger experiment.

With this unique setting, we can use this set of experiments to quantify the ground-truth ATE of any treatment combination, thus greatly facilitating our analysis to provide convincing validations of our DeDL framework against commonly adopted benchmarks. Throughout our empirical analysis, we use a three-dimensional binary vector $\boldsymbol{T} \in \mathcal{T} := \{0,1\}^3$ to represent the (random) treatment combination applied to a user, where the first component refers to whether the user is treated on DP, the second refers to whether she is treated in on LP, and the third refers to whether she is treated on FYP. Denote $\boldsymbol{t} \in \mathcal{T}$ as the realization of $\boldsymbol{T}$. For example, a treatment vector $\boldsymbol{t} = (1,0,0)'$ indicates that the user has received treatment on DP but is in the control group on LP and FYP.

This set of three experiments targeted 4,449,470 users in total between January 10, 2021, and February 1, 2021. Throughout our analysis, we use the total screen time of all three pages per day for each user as the outcome variable, consistent with the platform's primary objective to boost user engagement. To fully leverage the power of our DeDL framework, we have also collected the pretreatment covariate data for the users targeted by the experiments. The covariates adopted in our analysis include 16 discrete variables (such as gender, frequent residence area, age range, and the user's degree of activeness) and 10 continuous variables (such as the video-watching duration on each page per day in the 10 days right before the experiments). Table A1 (in Appendix B.1) describes all the covariates used in our analysis.

[9] For the sake of simplicity, we refer to the changes of several parameters in recommendation algorithm as major adjustment. In practice, the major adjustment could be adding weights for videos created by popular authors, increasing exposure of live streams viewed by nearby users, changing the degree of diversity of videos in users' feeds, and so on.

4.1.2. Treatment Assignment and Ground Truth Treatment Effects. For each experiment, any targeted user, regardless of her pretreatment covariates, was independently and randomly assigned to the treatment group (i.e., the new algorithm on the respective page was applied) with the probability of 0.6 and to the control group (i.e., the baseline algorithm was applied) with the probability 0.4 in each experiment. Therefore, because all treatment assignments are orthogonal, each targeted user with covariates $\boldsymbol{x}$ was assigned to the treatment combination $\boldsymbol{t} \in \{0,1\}^3$ with probability

$$\nu(\boldsymbol{t}|\boldsymbol{x}) = \mathbb{P}[\boldsymbol{T} = \boldsymbol{t}|\boldsymbol{X} = \boldsymbol{x}] = \prod_{k=1}^{3} \Big(0.4 \cdot \mathbb{I}[t_k = 0] + 0.6 \cdot \mathbb{I}[t_k = 1]\Big).$$

The ATE under treatment combination $\boldsymbol{t}$ is simply the expected outcome y under $\boldsymbol{t}$ compared to that under $\boldsymbol{t}_0 = (0,0,0)'$ over the same population $\boldsymbol{X}$. For a fair comparison of ATEs over different treatment combinations, we need to keep the user covariates similarly distributed under different treatment combinations in $\mathcal{T}$. Hence, we adopt stratified sampling to balance the covariates observed in different treatment groups, which is explained in detail in Section 5.1.

Table 2 documents the ground-truth ATE of all treatment combinations on the total screen time of all three pages per day benchmarked against the case where the baseline algorithm is applied in all three pages (i.e., $\mu(\boldsymbol{t})$ for all $\boldsymbol{t} \in \mathcal{T}$). To protect the sensitive data of Platform O, we report only the relative ATEs (see column (1) of Table 2). We emphasize that the orthogonal deployment of these three experiments enables us to observe the ground-truth ATE of *all* seven treatment combinations and, thus, to provide ground-truth ATEs for us to validate our DeDL framework.

To validate our DeDL framework, we assume that some treatment conditions are unobserved; we use our framework to recover these "unobserved" conditions, and we compare our results with the ground truth. In practice, different engineer and product teams launch the individual experiments independently (most likely in an asynchronous and uncoordinated fashion), and the centralized platform manager runs a back-test to check the treatment effect of the combined experiment (i.e., $\boldsymbol{t} = (1,1,1)'$) at the end. Following this business practice, we assume that the outcomes are observable for the baseline case, the three individual experiments, and the combined experiment, and unobservable otherwise (see column (2) of Table 2). We denote $\mathcal{T}_o := \{\boldsymbol{t} \in \mathcal{T} : t_1 + t_2 + t_3 \in \{0,1,3\}\}$ as the set of observable treatment combinations and $\mathcal{T}_u := \{\boldsymbol{t} \in \mathcal{T} : \boldsymbol{t} \notin \mathcal{T}_o\}$ as the set of unobservable treatment combinations. Table 2 shows that the ground-truth ATE of some unobserved treatment combination (e.g., $\boldsymbol{t} = (1,1,0)'$) is insignificant (at $\alpha = 0.05$).

4.2. DeDL Framework on Experimental Data

In this subsection, we present the key steps in applying our DeDL framework to estimate and infer the ATE of each treatment combination (i.e., $\mu(\boldsymbol{t})$ defined by (3) for all $\boldsymbol{t} \in \mathcal{T}$), whose ground-truth

Table 2 Ground-Truth ATE and Best Treatment Identification of Eight Treatment Combinations

Treatment Combination $\boldsymbol{t}'$	Relative Effect Size (1)	Observed or Not (2)	Number of Users (3)
(0, 0, 0)	0.000%	Observable	258,249
(0, 0, 1)	1.091%**	Observable	258,340
(0, 1, 0)	-0.267%	Observable	258,367
(1, 0, 0)	0.758%*	Observable	258,321
(1, 1, 1)	2.121%****	Observable	258,375
(1, 1, 0)	0.689%	Unobservable	258,480
(1, 0, 1)	2.299%****	Unobservable	258,305
(0, 1, 1)	1.387%***	Unobservable	258,172

Note: To protect sensitive data, ATE is proportionally rescaled to relative effect size. The optimal treatment combination (i.e., best treatment) is $\boldsymbol{t}^* = (1,0,1)'$. *p<0.05; **p<0.01; ***p<0.001; ****p<0.0001.

value is documented in Table 2 column (1). The implementation details are provided in Appendix 5.2. First, we consider the following model specification of DGP:

$$\mathbb{E}[Y|\boldsymbol{X}=\boldsymbol{x},\boldsymbol{T}=\boldsymbol{t}] = G(\boldsymbol{\theta}^*(\boldsymbol{x}),\boldsymbol{t}) = \frac{\theta_4^*}{1+\exp(-(\theta_0^*(\boldsymbol{x})+\theta_1^*(\boldsymbol{x})t_1+\theta_2^*(\boldsymbol{x})t_2+\theta_3^*(\boldsymbol{x})t_3))}. \tag{6}$$

The link function G, as a sigmoid function, can capture either "diminishing marginal return" or "increasing marginal return" of the experiments for different users, both of which we have observed in our data sample (see Figure 1). Here, we are using a simplified version of the *Generalized Sigmoid Form II* (Assumption 1(d)) to avoid overfitting. The parameter θ_4^* can be thought of as the maximum possible video-watching time of any user.

Figure 6 in Appendix 5.2 illustrates our DNN architecture. We use two DNNs with three hidden layers per network (20 nodes in each layer) to approximate the parameters $\theta_0^*(\cdot)$ and $\theta_k^*(\cdot)$, for $k \in \{1,2,3\}$, respectively. For each layer, the ReLU function (i.e., $\mathrm{ReLU}(x) = \max\{0,x\}$) is used as the activation function. We then concatenate the last layers of two DNNs, take the linear combination $(u = \theta_0(\boldsymbol{x}) + \theta_1(\boldsymbol{x})t_1 + \theta_2(\boldsymbol{x})t_2 + \theta_3(\boldsymbol{x})t_3)$ as the input of a sigmoid function layer, and add another linear layer (no intercept, the slope approximates θ_4^*) to output y. We implement our DNN in TensorFlow and train the DNN by Adam algorithm (Kingma and Ba 2014) under the mean squared loss. We emphasize that the details of cross-fitting and DNN implementation are critical for empirical success. Accordingly, we defer this discussion and dedicate Section 5.2 specifically to it.

After obtaining the fitted estimator $\hat{\boldsymbol{\theta}}(\cdot)$, we estimate and infer the ATE of each treatment combination $\boldsymbol{t}$ using the stratified sample with our influence function $\psi(\boldsymbol{z},\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\Lambda}};\boldsymbol{t},\boldsymbol{t}_0)$ defined by (5). Specifically, we apply the estimators $\hat{\mu}_{\mathsf{DeDL}}(\cdot)$ and $\widehat{\mathcal{CI}}_{\mathsf{DeDL}}(\cdot;\mu)$ defined by (33) and (35) to the stratified sample under each (observable or unobservable) treatment combination $\boldsymbol{t} \in \mathcal{T}$ to obtain the estimated value and confidence interval of $\mu(\boldsymbol{t})$. We remark that $\hat{\boldsymbol{\Lambda}}(\boldsymbol{x})$ as the estimator of $\boldsymbol{\Lambda}(\boldsymbol{x}) = \mathbb{E}[\ell_{\boldsymbol{\theta\theta}}(Y,\boldsymbol{T},\boldsymbol{\theta}(\boldsymbol{X}))|\boldsymbol{X}=\boldsymbol{x}]$ can be directly computed once $\hat{\boldsymbol{\theta}}(\cdot)$ is obtained, because the distribution of the treatment combination $\boldsymbol{T}$ is known. See Section 5.2 for details.

Finally, we apply our DeDL framework to identify the treatment combination with the highest ATE, $\boldsymbol{t}^* := \arg\max_{\boldsymbol{t}\in\mathcal{T}} \mu(\boldsymbol{t})$. Specifically, we identify the "best treatment" as $\hat{\boldsymbol{t}}^*_{\mathsf{DeDL}} := \arg\max_{\boldsymbol{t}\in\mathcal{T}} \hat{\mu}_{\mathsf{DeDL}}(\boldsymbol{t})$. Define $\hat{\mu}^*_{\mathsf{DeDL}} := \max_{\boldsymbol{t}\in\mathcal{T}} \hat{\mu}_{\mathsf{DeDL}}(\boldsymbol{t})$ as the ATE of the best treatment identified by our DeDL framework. We construct the estimators of ATE increment from treatment combination $\boldsymbol{t}$ to "best treatment" as $\hat{\tau}_{\mathsf{DeDL}}(\cdot) := \hat{\mu}^*_{\mathsf{DeDL}} - \hat{\mu}_{\mathsf{DeDL}}(\cdot)$ and $\widehat{\mathcal{CI}}_{\mathsf{DeDL}}(\cdot;\tau)$ defined by (36) and (40) to each treatment combination $\boldsymbol{t} \in \mathcal{T}$ and select the treatment combination(s) $\boldsymbol{t}$ with $\hat{\tau}_{\mathsf{DeDL}}(\boldsymbol{t})$ insignificant from 0 as best treatment(s).

4.3. Benchmarks

To evaluate the performance of our DeDL framework to (i) estimate and infer ATE and (ii) identify the optimal treatment combination, we consider five commonly used approaches as benchmarks: (a) the linear addition (LA) approach; (b) the linear regression (LR) approach; (c) the pure deep learning (PDL) approach; and (d) the structured deep learning (SDL) approach. For implementation details of all four benchmarks, see Appendix C.

The LA approach assumes that ATE is linearly additive (i.e., $\mu(\boldsymbol{t}_1 + \boldsymbol{t}_2) = \mu(\boldsymbol{t}_1) + \mu(\boldsymbol{t}_2)$ for any $\boldsymbol{t}_1, \boldsymbol{t}_2 \in \mathcal{T}$) and, thus, predicts the ATE of an unobservable treatment combination using those of the observable individual experiments. This approach is intuitive, convenient, and scalable, thus widely adopted by most large-scale online platforms in practice, such as Platform O. The standard error of an LA estimator for any treatment combination is estimated by assuming that the estimators for individual experiments are independent. For best treatment identification, the LA approach is equivalent to selecting all treatments that have a positive significant ATE.

Under the LR approach, we first predict the unobservable outcomes using a linear regression model trained on the observed sample by regressing the outcome y on $\boldsymbol{t}$ and $\boldsymbol{x}$. The estimation and inference of ATE for each treatment combination $\boldsymbol{t}$ are based on the pair-wise t-test between the outcome *predictions* under $\boldsymbol{t}$ and those under the baseline combination $\boldsymbol{t}_0$. The LR approach identifies the best treatment by selecting the treatment combination with the highest ATE based on the *predicted* outcomes of all treatment combinations. We remark that the LA and LR approaches inherently assume that the treatment effects are linearly additive and homogeneous.

The PDL approach employs a DNN with a similar structure as DeDL that predicts the outcome variable y as a function of both $\boldsymbol{x}$ and $\boldsymbol{t}$. Unlike DeDL that has concrete link function to describe the relationship of $\boldsymbol{t}$ to y conditional on $\boldsymbol{x}$, PDL treats both $\boldsymbol{x}$ and $\boldsymbol{t}$ as network inputs and in turn, allows more flexible relationship from $\boldsymbol{t}$ and $\boldsymbol{x}$ to y. The PDL approach uses the same pair-wise t-test as the LR approach on the *predicted outcomes* for inference of ATE. The identification of the best treatment depends on the highest ATE among all treatment combinations. The PDL estimator fully leverages the predictive power of DNN (potentially more powerful than DeDL, which assumes a concrete link

function) but cannot use the influence function to debias the DL estimation as DeDL does. Therefore, the comparison between PDL and DeDL provides insights into the trade-off between the flexibility of DL models and the ability to construct influence functions with debiasing.

The SDL approach is exactly the same as the DeDL approach with only one distinction: Unlike DeDL, which uses the influence function to debias the estimation from DL, the SDL approach simply uses the prediction from the DL to construct plug-in estimator $\hat{\mu}(\boldsymbol{t}) = \frac{1}{n}\sum_{i=1}^{n} H(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}(\boldsymbol{x}_i); \boldsymbol{t}, \boldsymbol{t}_0)$. Similar to the LR and PDL estimators, the SDL approach estimates and infers the ATE for each treatment combination $\boldsymbol{t}$ by running the pair-wise t-test on the *predicted outcomes*. Likewise, the optimal treatment combination is identified as the one with the highest ATE based on the predicted outcomes of all treatment combinations. The comparison between SDL and PDL reveals the trade-off between economic interpretations and predictive power, whereas that between SDL and DeDL highlights the value of bias correction in our framework.

For any approach $\pi \in \{\mathsf{LA}, \mathsf{LR}, \mathsf{PDL}, \mathsf{SDL}, \mathsf{DeDL}\}$, we use $\hat{\mu}_\pi(\boldsymbol{t})$ (resp., $\widehat{\mathcal{CI}}_\pi(\boldsymbol{t}; \mu)$) to denote the ATE estimator (resp., the confidence interval of $\hat{\mu}_\pi(\boldsymbol{t})$) generated by π for the treatment combination $\boldsymbol{t} \neq \boldsymbol{t}_0$. Likewise, we denote $\hat{\tau}_\pi(\boldsymbol{t})$ (resp., $\widehat{\mathcal{CI}}_\pi(\boldsymbol{t}; \tau)$) to denote the estimator for the ATE difference between the optimal treatment $\boldsymbol{t}^*$ and the experiment combination $\boldsymbol{t} \neq \boldsymbol{t}_0$ (resp., the confidence interval for such ATE difference) generated by π.

4.4. Results on ATEs

We first compare the DeDL approach with the four benchmarks presented in Section 4.3 to estimate and infer the ATE of each "unobserved" treatment combination $\boldsymbol{t} \in \mathcal{T}^u$. As discussed above, the assignment mechanism of the three experiments on Platform O enables the observation of the ground-truth ATE of *each* treatment combination, based on which we assume that three treatment conditions are not observed by the algorithm and evaluate the performance of all approaches on these unobserved conditions. In particular, we document the following performance metrics:

- Correct Direction Ratio (CDRu). For any estimation and inference approach π, we denote $\mathcal{T}^{cd}(\pi)$ as the set of all treatment combinations with *correct direction identification*, i.e., the treatment combinations $\boldsymbol{t} \in \mathcal{T}$ whose ground-truth ATE $\mu(\boldsymbol{t})$ have been correctly identified by π in terms of both (i) sign and (ii) statistical significance. Define $\mathcal{T}_u^{cd}(\pi) := \mathcal{T}^{cd}(\pi) \cap \mathcal{T}_u$ as the set of unobservable treatment combinations with correct direction identification for π. Define the CDR for unobservable treatment combinations (CDRu) of π as $\mathcal{CDR}_u(\pi) = \frac{|\mathcal{T}_u^{cd}(\pi)|}{|\mathcal{T}_u|} \times 100\%$.
- Mean Absolute Percentage Error (MAPEu). The MAPE of any estimation and inference approach π is the average percentage error for unobserved treatment combinations with a significant ground-truth ATE. In other words, the MAPE of π for unobservable treatment combinations (MAPEu) is defined as $\mathcal{MAPE}_u(\pi) := \frac{1}{|\mathcal{T}_u^s|} \sum_{\boldsymbol{t} \in \mathcal{T}_u^s} \frac{|\mu(\boldsymbol{t}) - \hat{\mu}_\pi(\boldsymbol{t})|}{|\mu(\boldsymbol{t})|} \times 100\%$.

Table 3 Comparison of Different Estimators of ATE

Estimator	Unobserved Treatment Combinations				Estimator	All Treatment Combinations			
	CDRu (1)	MAPEu (2)	MSEu (3)	MAEu (4)		CDR (5)	MAPE (6)	MSE (7)	MAE (8)
LA	2/3	30.06%	18.597	4.032	LA	7/8	12.02%	7.966	1.728
LR	2/3	4.90%	5.303	1.855	LR	7/8	17.37%	6.551	2.348
PDL	2/3	6.86%	4.876	1.838	PDL	6/8	14.76%	4.962	2.031
SDL	2/3	14.71%	5.623	2.271	SDL	6/8	14.03%	3.840	1.804
DeDL	3/3	1.75%	4.095	1.343	DeDL	8/8	3.07%	1.845	0.737

Notes: The calculation of MSE, MSEu, MAE, and MAEu are based on the scaled outcome variable (see Table 2, column (1)). MSE and MSEu are scaled by multiplying a constant. MAE and MAEu are scaled by multiplying another constant.

- Mean Squared Error (MSEu). The MSE of π for unobservable treatment combinations (MSEu) is defined as $\frac{1}{|\mathcal{T}_u|}\sum_{\boldsymbol{t}\in\mathcal{T}_u}(\mu(\boldsymbol{t})-\hat{\mu}_\pi(\boldsymbol{t}))^2$.
- Mean Absolute Error (MAEu). The MAE of π for unobservable treatment combinations (MAEu) is defined analogously as $\frac{1}{|\mathcal{T}_u|}\sum_{\boldsymbol{t}\in\mathcal{T}_u}|\mu(\boldsymbol{t})-\hat{\mu}_\pi(\boldsymbol{t})|$.

To give a clear picture of the comparisons among *all* eight (observable and unobservable) treatment combinations and identify the best treatment of them (see Section 4.5), we present the estimated treatment effects of all treatment combinations in Figure 5. In Table 3, we also calculate the above four metrics evaluated on both *unobserved* treatment combinations and *all* treatment combinations.

Table 3 documents the comparison of our DeDL approach against the LA, LR, SL, PDL, and SDL benchmarks with respect to the four performance metrics described above. The punchline message of the empirical analysis is that our DeDL estimator substantially outperforms all five benchmarks with any of these performance metrics, providing evidence that the proposed framework accurately estimates and infers the ATE of any treatment combination for multiple A/B tests on a large-scale online platform. More notably, we highlight that, to our best knowledge, this is the first rigorous empirical validation of the DML methodology in a practical setting with data from large-scale field experiments.

As emphasized above, the unique concurrent orthogonal-design deployment of the three experiments on Platform O has made such validation possible by revealing the ground-truth ATE of each possible treatment combination. A key advantage of our DeDL approach to adopt DL in causal inference is its ability to capture the user-heterogeneity and nonlinearity for the combined treatment effect of multiple A/B tests on a large-scale online platform, the observation of which has motivated this study (see Figure 1). As expected, such an advantage stems from the strong predictive power of deep neural networks. As aforementioned, the comparison between our DeDL framework and the SDL estimator provides more insights into how bias correction (i.e., influence function) affects the final causal inference performance. To provide more insights, in Figure 5, we visualize these two approaches' ATE estimates and the 95% confidence interval for each treatment combination (i.e., $\hat{\mu}_\pi(\boldsymbol{t})$ and $\widehat{\mathcal{CI}}_\pi(\boldsymbol{t};\mu)$ for $\pi\in\{\mathsf{SDL},\mathsf{DeDL}\}$ and $\boldsymbol{t}\in\mathcal{T}$). Figure 5 shows that the bias correction can

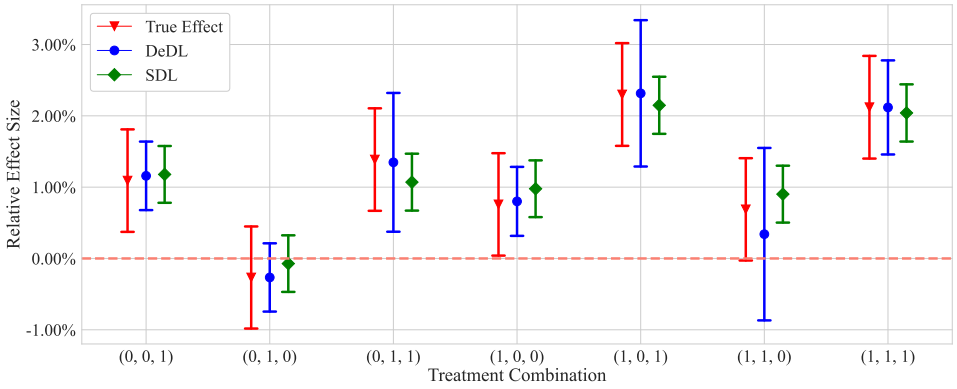


Figure 5 Detailed Comparisons Between the SDL and DeDL Estimators (bars represent confidence intervals)

help causal inference in two significant ways: First, by correcting the bias due to the variabilities of the training data from the plug-in estimator, the DeDL approach is able to accurately identify the confidence interval of the ATEs, while the SDL approach is always underestimating the standard error of ATE. This means that, without bias correction, the analysis leads to potentially more Type-I errors and higher false discovery rates for the platform. Second, DeDL provides a more accurate ATE estimate than SDL does, empirically confirming the advantage of the former for more accurate causal inference.

Moreover, the comparison between SDL and PDL in Table 3 shows that the pure DNN without assuming the link function can improve the performance of predicting the treatment effects (i.e., MAPEu reduced from 14.71% to 6.86%). This shows that, by specifying a concrete link function, we indeed sacrifice some prediction accuracy for the ability to derive bias correction terms and allow economic interpretation. However, since the performance of our DeDL estimator significantly outperform that of the PDL approach in all metrics, it shows that this sacrifice is justified in the context of causal inference. In other words, the benefit of assuming a concrete link function and applying the corresponding DML bias correction will outweigh the cost of a less flexible relationship between the treatment conditions and the outcome.

Last, we have also tested the standard logit model proposed by Farrell et al. (2020) such that $G(\boldsymbol{\theta}(\boldsymbol{x}), \boldsymbol{t}) = \frac{1}{1+\exp(-(\theta_0(\boldsymbol{x})+\theta_1(\boldsymbol{x})\sum_{i=1}^{m} t_i))}$. Despite the first-stage cross-validation losses being similar for both the standard logit model and our DeDL, with the scaled Mean Squared Error (MSE) values of 0.0087 and 0.0086, respectively, we observe that the latter significantly outperforms the former in ATE estimation. The limited flexibility of the standard logit function weakens its effectiveness in accurately estimating ATE.

Table 4 Comparison of Different Estimators of Best Treatment Identification

Estimator	CDR (1)	MAPE (2)	MSE (3)	MAE (4)
LA	7/8	21.92%	15.539	3.091
LR	7/8	11.86%	3.232	1.727
PDL	7/8	12.83%	4.053	1.839
SDL	8/8	17.45%	7.186	2.442
DeDL	8/8	5.97%	1.995	0.780

Notes: The calculations of MSE and MAE are based on the scaled outcome variable (see Table 2, column (4)). MSE is scaled by multiplying a constant. MAE is scaled by multiplying another constant.

4.5. Results on Best Treatment Identification

This subsection applies our DeDL framework to identify the optimal treatment combination with the highest ATE. By the ground-truth ATE estimates (Table 2 column (1)), the "true" optimal treatment combination is $\boldsymbol{t}^* = (1,0,1)'$. We compare different estimators for best treatment identification in Table 4, and we report the same set of performance metrics defined in Section 4.4. In particular, we focus on the comparison between $\tau(\boldsymbol{t})$ and $\hat{\tau}(\boldsymbol{t})$, the ground-truth treatment-effect increment and its estimator. Among eight treatment combinations, the treatment-effect increment of the "true" optimal treatment combination over both $(1,0,1)$ and $(1,1,1)$ insignificantly differ from 0. Column (1) in Table 4 indicates that only SDL and DeDL correctly identify both sign and statistical significance of all treatment-effect increments. Table 4 shows that DeDL estimators significantly outperform the LA, LR, PDL, and SDL benchmarks with respect to all performance metrics in identifying the best treatment combination.

4.6. Results on CATEs

To deepen our empirical analysis, we further validate the Conditional Average Treatment Effect (CATE) estimation across different demographic subgroups. Specifically, we classify users into four distinct groups based on two key demographic covariates: residential location (rural versus non-rural) and phone price tier (medium-priced versus non-medium priced[10]). This classification yields four distinct user groups: (1) Non-rural area with a medium-priced phone (NM), denoted as $\mathcal{X}_{NM}$; (2) Non-rural area with a non-medium-priced phone (NN), denoted as $\mathcal{X}_{NN}$; (3) Rural area with a medium-priced phone (RM), denoted as $\mathcal{X}_{RM}$; and (4) Rural area with a non-medium-priced phone (RN), denoted as $\mathcal{X}_{RN}$. This classification enables us to better understand the performance of DeDL across different demographic segments.

The implementation of CATE estimation follows the same four-fold cross-fitting described in Section 5.2. For each fold s, we construct the CATE estimator $\hat{\mu}_s(\boldsymbol{t}|\mathcal{X}_j)$ for treatment combination $\boldsymbol{t}$ conditional on demographic segements $\boldsymbol{X} \in \mathcal{X}_j$ $(j = NM, NN, RM, RN)$. This estimator is computed by averaging $\psi(\boldsymbol{z}, \hat{\boldsymbol{\theta}}_s, \hat{\boldsymbol{\Lambda}}_s; \boldsymbol{t}, \boldsymbol{t}_0)$ across all observations where $\boldsymbol{x} \in \mathcal{X}_j$. The final CATE estimator

[10] Medium-priced is defined as RMB 1000-2000.

Table 5 Comparison of Different Estimators of CATE

Estimator	Unobserved Treatment Combinations			Estimator	All Treatment Combinations		
	MAPEu (1)	MSEu (2)	MAEu (3)		MAPE (4)	MSE (5)	MAE (6)
LA	51.34%	11.197	29.959	LA	36.64%	8.895	20.807
LR	26.11%	5.898	18.129	LR	29.04%	4.500	17.732
PDL	17.03%	4.252	12.662	PDL	17.34%	2.593	11.142
SDL	26.47%	5.425	18.191	SDL	22.77%	3.292	14.177
DeDL	8.31%	1.528	6.614	DeDL	10.32%	1.062	6.486

Notes: The calculation of MSE, MSEu, MAE, and MAEu are group-size-weighted average. MSE and MSEu are scaled by multiplying a constant. MAE and MAEu are scaled by multiplying another constant.

Table 6 MAPE Comparison of Different Estimators of CATE for Four Groups

Estimator	Unobserved Treatment Combinations				Estimator	All Treatment Combinations			
	NM (774,326)	NN (460,330)	RM (451,494)	RN (380,465)		NM (774,326)	NN (460,330)	RM (451,494)	RN (380,465)
LA	48.39%	101.53%	15.46%	39.18%	LA	24.19%	101.53%	6.18%	19.58%
LR	18.46%	18.90%	42.69%	30.74%	LR	28.40%	18.90%	36.63%	33.46%
PDL	0.12%	12.52%	33.94%	36.89%	PDL	7.60%	12.51%	27.82%	30.57%
SDL	23.20%	5.66%	44.46%	36.96%	SDL	21.63%	5.66%	38.23%	27.43%
DeDL	0.04%	0.66%	15.23%	26.22%	DeDL	7.17%	0.66%	18.71%	18.46%

Notes: The sizes of the groups are indicated in parentheses.

$\mu(\boldsymbol{t}|\mathcal{X}_j) := \mathbb{E}[Y|\boldsymbol{T}=\boldsymbol{t}, X \in \mathcal{X}_j] - \mathbb{E}[Y|\boldsymbol{T}=\boldsymbol{t}_0, X \in \mathcal{X}_j]$ is obtained by taking the average of $\hat{\mu}_s(\boldsymbol{t}|\mathcal{X}_j)$ across four folds. The empirical results for CATE estimation are presented in Tables 5 and 6. Table 5 presents the group-size-weighted average metrics for CATE estimation on both unobserved treatment combinations and all treatment combinations across four distinct groups. Table 6 provides a detailed comparison of MAPE across different estimators for the CATE estimation with respect to four groups. Consistent with our main results on ATE estimation in Table 3, our method outperforms all other benchmarks in CATE estimation as well.

Our comprehensive empirical analysis, presented in Tables 3 through 6, demonstrates the superior performance of the DeDL estimator in three critical aspects: estimating ATEs of unobserved conditions, identifying optimal treatment combinations, and calculating CATEs across demographic segments. We highlight that this paper is among the first works to empirically investigate the accuracy of the DML framework to recover the ground-truth treatment effects through large-scale field experiments. Our empirical evidence also sheds light on how crucial bias correction is in estimating the causal effects, and it provides insights into the trade-off between more flexible models and the ability to derive bias correction terms. Last but not least, we demonstrate that specifying and training a good ML model can be essential for second-stage inference. Specifically, our carefully designed model layer within the DeDL framework, tailored for multiple-experiment settings, substantially outperforms standard unstructured deep learning (DL) approaches. We note that achieving these improvements requires extensive effort in model fine-tuning, consistent with well-documented challenges in deep learning implementation.

5. Key Checkpoints in the Empirical Implementation

Whereas Section 3 outlines the theoretical framework for applying DML-based causal inference with DNNs, our empirical success, as discussed in Section 4, hinges on several key steps of practical implementation. In this section, drawing from our own experiences to deploy the DeDL framework in the multiple experiment setting, we provide a detailed discussion of these crucial checkpoints, aiming to guide similar applications in other empirical contexts.

5.1. Covariate Balancing with Stratified Sampling

We observe from the DGP (1) that unconfoundedness is typically required when applying DML. However, in practice, the covariates may be often high-dimensional. Even if the treatment assignment is well-controlled and genuinely unconfounded, as in our setting (see the definition of $\nu(\cdot\,|\,\cdot)$ in Section 3.1), the empirical correlation between covariates and treatment assignment could deviate significantly from the intended design. This can obstruct effective empirical implementation.

Thus, for a fair comparison of ATEs over different treatment combinations and keeping the user covariates similarly distributed under different treatment combinations in $\mathcal{T}$, in our empirical implementation we adopt stratified sampling to randomly select $2,066,606$ users from those who are targeted by all three experiments.[11] We emphasize that although such a stratified sampling is not part of the requirement in DML, we find it still critical for the practical effectiveness.

Specifically, based on the moments and quantiles of the covariate distribution, we partition the covariate space into $69,111$ strata, and then randomly sample the same number of users whose covariates lie within the stratum for each treatment combination. After the stratified sampling, we construct a new dataset that has about $258,325$ users under each treatment combination (see column (3) of Table 2) and hence $2,066,606$ users in total. Hereafter, we call the data sample after stratified sampling the *stratified sample*, and all the empirical analysis from now on is performed on the stratified sample. For the stratified sample, it can be seen that $\mathbb{P}[T_k = 1] = \mathbb{P}[T_k = 0] = 0.5$ $(k = 1, 2, 3)$, independently distributed for different A/B tests. We detail the exact procedure of our stratified sampling in Appendix B.2. To confirm the success of randomization among our stratified sample of users, we compare users under different treatment combinations in their gender, activeness on the platform, frequent residence area, pre-experiment active days, pre-experiment screen time of DP, LP, and FYP, and pre-experiment app-usage duration. The complete randomization check results are

[11] Although in principle our randomization should guarantee balanced covariates, because there are many covariates and some covariates' distributions have long tails, the covariates are not in fact perfectly balanced across all 8 conditions. This leads to inconsistency to (3) and the theoretical discussion (Proposition 2) if one directly uses sample means as the ground truth to validate the treatment effects. One solution would be to redefine the treatment effect and rederive a rather complex new influence function that admits different covariate distributions, but we opt for a simpler approach, the stratified sampling.

listed in Table A2 in Appendix B.3. Table A2 shows that the seven treatment combinations have similar proportions of male users, high-active users, and users from the south as the baseline combination $\boldsymbol{t}_0 = (0,0,0)'$. Moreover, the summary statistics of the covariates during 10 days before the experiments further assure that there is no significant difference between the average active days, average page screen time, and average app-usage duration of the users under seven treatment combinations and those under baseline combination (all p-values > 0.05). Given the balanced user demographic and pre-experiment behavior covariates under different treatment combinations in our stratified sample, the difference between the outcome variables under different treatment combinations should be attributed to the experimental interventions, i.e., the implementation of new algorithms on different pages of Platform O. Furthermore, the randomization check reported in Table A2 also demonstrates that the covariate distributions are fairly similar with respect to different treatment combinations for the stratified sample, confirming that our DeDL framework can be applied with validity.

5.2. Building and Training the DNN with Care

In this section, we present the implementation details of using DeDL to estimate ATE. We emphasize that, from our experience, it takes substantial amount of effort to build and train the DNN in the DeDL framework. Failure to train a DNN with low cross-validation errors is likely to invalidate estimation and inference in the second stage, as detailed in Section 5.3.

We use the four-fold cross-fitting proposed in Chernozhukov et al. (2018) to obtain the estimators. We randomly partition the observed data of five observed treatment combinations with 1,291,652 data points into four folds $(I_s)_{s=1}^4$ such that each fold has approximately 322,913 observations. For each fold index s, we follow the steps described below to obtain an estimated ATE for each treatment combination. First, we use the other three folds $(I_i)_{i\in\{1,2,3,4\}\setminus s}$ as training data to fit a structured DNN (see Figure 6) and get an estimator of the unknown nuisance parameters $\hat{\boldsymbol{\theta}}_s(\cdot)$. To ascertain the architecture of DNN, we have explored one DNN for both $\theta_0^*(\cdot)$ and $\theta_k^*(\cdot)$ $(k \in \{1,2,3\})$ and two separate DNNs for $\theta_0^*(\cdot)$ and $\theta_k^*(\cdot)$ $(k \in \{1,2,3\})$. We also experimented with the number of layers equal to 3, 5, and 10 and the number of nodes equal to 10, 20, 50, and 100 per layer. We eventually adopted the DNN with a structure of a combined two 3-layer DNNs with 20 nodes per layer that achieves the minimal mean squared loss on the validation data. To fit the structured DNN, we tested a range of learning rates from 0.00001 to 0.1, applied batch size of 32, 100, 200, and 500. By conducting a grid search over the combinations of hyperparameters, we set $learning\ rate = 0.0001$ and $batch\ size = 100$ which achieves the lowest cross validation loss. For each treatment combination $\boldsymbol{t}$, we use fold I_s to construct the ATE estimator $\hat{\mu}_s(\boldsymbol{t})$ following (33).

We continue substituting $\boldsymbol{x}$ in fold I_s to the unknown parameters $\hat{\boldsymbol{\theta}}_s(\cdot)$ obtained by three other folds to calculate $\psi(\boldsymbol{z}, \hat{\boldsymbol{\theta}}_s, \hat{\boldsymbol{\Lambda}}_s; \boldsymbol{t}, \boldsymbol{t}_0)$ for each data point $\boldsymbol{z}$ in fold I_s. To be specific, after obtaining

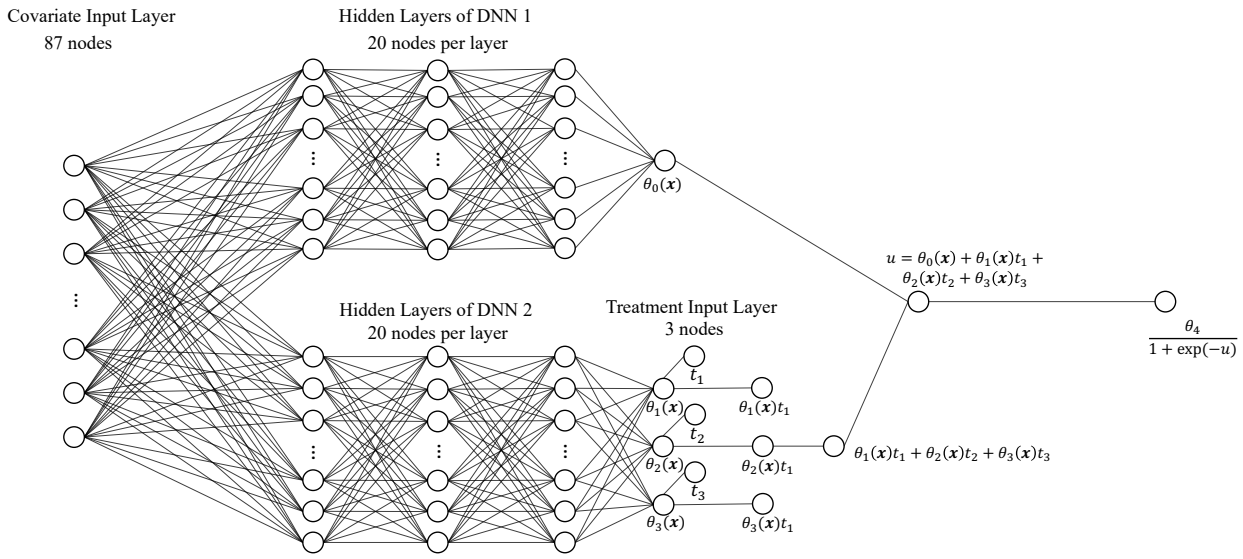


Figure 6 Structure of the Deep Neural Network Used in the Empirical Analysis

$\hat{\boldsymbol{\theta}}_s(\cdot)$, we calculate $\hat{\boldsymbol{\Lambda}}_s(\boldsymbol{x}) = \frac{2}{5}\sum_{\boldsymbol{t}\in\mathcal{T}_o} G_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}_s(\boldsymbol{x}),\boldsymbol{t})G_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}_s(\boldsymbol{x}),\boldsymbol{t})'$ because the distribution of observed treatment combination $\boldsymbol{t}\in\mathcal{T}_o$ is known in the stratified sample with $\mathbb{P}[\boldsymbol{T}=\boldsymbol{t}|\boldsymbol{X}=\boldsymbol{x}]=0.2$. We remark that $\lambda(\boldsymbol{x})$ may be non-invertible for certain values of $\boldsymbol{x}$ due to numerical precision issues in empirical implementation. We can add an identity matrix multiplied by a small regularization parameter (e.g., 0.001, 0.01, 1) to address this. The regularization parameter can be fine-tuned using grid search to improve performance and ensure invertibility in cases where direct inversion is affected by precision limitations. By averaging $\psi(\boldsymbol{z},\hat{\boldsymbol{\theta}}_s,\hat{\boldsymbol{\Lambda}}_s;\boldsymbol{t},\boldsymbol{t}_0)$, we obtain the estimated ATE of treatment combination $\boldsymbol{t}$ in fold index s as $\hat{\mu}_s(\boldsymbol{t})$.

The ATE estimate and the 95% confidence interval for each treatment combination in each fold are presented in the first four rows in each section in Table 7. After four-fold cross-fitting, we aggregate $\hat{\mu}_s(\hat{\boldsymbol{t}})$ by taking their average value as the final estimator $\hat{\mu}(\boldsymbol{t})$ for each treatment combination $\boldsymbol{t}$. We report the results of the final estimator in the last row of each section in Table 7.

The estimators for best treatment identification are obtained through similar implementation procedures. Each fold will generate its estimators for the advantage of the best treatment over any treatment combination $\boldsymbol{t}$, $\tau(\boldsymbol{t})$, for $\boldsymbol{t}\neq\boldsymbol{t}^*$. Aggregating the estimators from four folds generates the DeDL estimator for best treatment identification.

5.3. Using Training Error as the Compass Towards Success

Analyzing the theory behind our DeDL framework, it is clear that the convergence of nuisance parameter estimation, emphasized in Proposition 1(b), underpins the asymptotic normality established in Theorem 2 and subsequent valid inference. One may also refer to Appendix D.3 for further discussions regarding the impact of biased nuisance parameter estimation for DML. However, in practice,

Table 7 Detailed Results of four-fold DeDL Estimators

Treatment Combination	Ground-Truth ATE (1)	Fold (2)	Estimated ATE (3)	95% Confidence Interval for ATE Estimate (4)	APE (5)	SE (6)	AE (7)
(0, 0, 1)	1.091%**	1	0.974%	[0.484%, 1.464%]	10.76%	1.379	1.174
		2	1.277%	[0.780%, 1.774%]	17.01%	3.448	1.857
		3	1.337%	[0.895%, 1.780%]	22.55%	6.054	2.461
		4	1.045%	[0.552%, 1.538%]	4.22%	0.212	0.461
		Mean	1.158%	[0.678%, 1.639%]	6.14%	0.450	0.671
(0, 1, 0)	-0.267%	1	-0.117%	[−0.611%, 0.376%]	NA	2.242	1.174
		2	-0.358%	[−0.845%, 0.128%]	NA	0.835	0.914
		3	0.038%	[−0.399%, 0.475%]	NA	9.312	3.052
		4	-0.627%	[−1.125%, −0.129%]	NA	12.973	3.602
		Mean	-0.266%	[−0.745%, 0.212%]	NA	0.000	0.008
(1, 0, 0)	0.758%*	1	0.707%	[0.200%, 1.214%]	6.73%	0.260	0.510
		2	0.761%	[0.271%, 1.252%]	0.44%	0.000	0.033
		3	0.986%	[0.546%, 1.427%]	30.13%	5.216	2.284
		4	0.747%	[0.250%, 1.244%]	1.47%	0.012	0.111
		Mean	0.800%	[0.317%, 1.284%]	5.59%	0.180	0.424
(1, 1, 1)	2.121%****	1	2.457%	[1.785%, 3.128%]	15.84%	11.288	3.360
		2	2.304%	[1.630%, 2.978%]	8.63%	3.353	1.831
		3	1.254%	[0.630%, 1.878%]	40.85%	75.044	8.662
		4	2.457%	[1.788%, 3.127%]	15.88%	11.341	3.368
		Mean	2.118%	[1.458%, 2.778%]	0.12%	0.00	0.026
(1, 1, 0)	0.689%	1	-0.616%	[−2.801%, 1.568%]	NA	170.299	13.050
		2	0.900%	[0.259%, 1.541%]	NA	4.451	2.110
		3	0.640%	[0.136%, 1.271%]	NA	0.021	0.146
		4	0.376%	[−1.067%, 1.819%]	NA	9.766	3.125
		Mean	0.341%	[−0.868%, 1.550%]	NA	12.108	3.480
(1, 0, 1)	2.299%****	1	2.396%	[1.075%, 3.716%]	4.19%	0.926	0.962
		2	2.830%	[2.086%, 3.573%]	23.07%	28.129	5.304
		3	2.137%	[1.561%, 2.714%]	7.04%	2.619	1.619
		4	1.899%	[0.435%, 3.362%]	17.42%	16.048	4.006
		Mean	2.315%	[1.289%, 3.341%]	0.70%	0.026	0.160
(0, 1, 1)	1.387%***	1	0.702%	[−0.659%, 2.062%]	49.39%	46.919	6.850
		2	2.127%	[1.265%, 2.989%]	53.41%	54.849	7.406
		3	1.168%	[0.597%, 1.739%]	15.79%	4.792	2.189
		4	1.395%	[0.296%, 2.493%]	0.58%	0.00	0.080
		Mean	1.348%	[0.375%, 2.321%]	2.80%	0.151	0.388

Notes: The calculation of APE, SE, and AE is based on the scaled outcome variable (see column (1) of this table). SE is scaled by multiplying a constant. AE is scaled by multiplying another constant. The significance levels are encoded as *p<0.05; **p<0.01; ***p<0.001; ****p<0.0001.

the true values of the nuisance parameters $\theta^*(\boldsymbol{x})$ are unobservable, making it impossible to directly assess the quality of this estimation. This raises a key practical question: how can we be confident that we are on the right track? In this context, the readily accessible cross-validation error, which we refer to as the training error for simplicity, of the DNN can serve as a crucial proxy.

Specifically, we compare the estimation MAPE for all treatment combinations of the DeDL estimator with SDL and LR estimators under increasing DNN training epochs in Figure 7. As demonstrated in Figure 7, both DeDL and SDL estimators yield smaller MAPE, as DNN training mean squared loss is smaller. We highlight that when DNN is poorly trained (i.e., within 100 epochs), the DeDL and SDL estimators have similar estimation MAPE, which are dominated by the LR estimator. In this

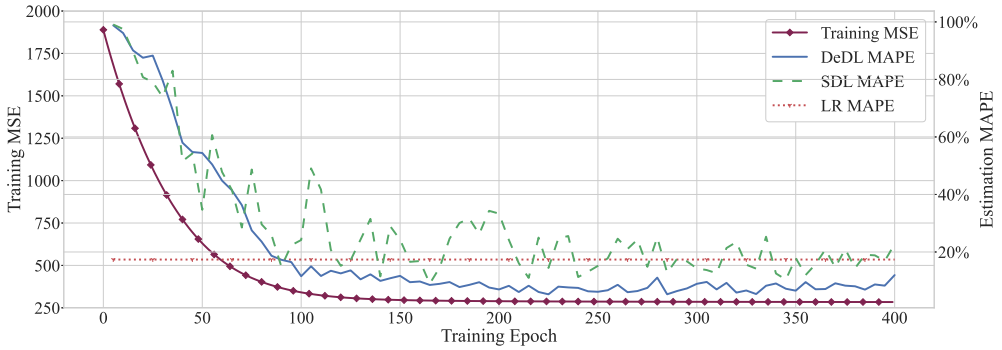


Figure 7 MAPE Comparison With DNN Training Epoch

case, debias does not benefit causal effect estimation. However, as the DNN converges, DeDL starts to show significant advantages in estimation accuracy compared to both SDL and LR estimators. This phenomenon is well connected to the classical literature on semiparametric statistics, which requires that the convergence rate of the estimated parameter functions $\hat{\theta}(\cdot)$ be sufficiently fast to obtain asymptotically unbiased estimators of the treatment effects (e.g., Chernozhukov et al. 2018). Finally, Figure 7 is also translated into an important actionable insight when adopting our DeDL framework that the DNN training error can be a useful indicator for the quality of the second-stage estimation and the effectiveness of debiasing via DML. We remark that there is no single recipe for judging whether a training error is sufficiently small for good estimation and inference, which requires context-specific judgmental calls based on domain knowledge. Additional discussions of the training error is provided in Appendix D.4.

6. Synthetic Experiments

Using synthetic experiments, we gauge the robustness of our approach under different scenarios. Due to the space limit, we defer the details of these discussions in the appendix and only provide a high-level summary here. We first validate our theory by varying the number of experiments with $m \in \{4, 6, 8, 10\}$ in Appendix D.2. This experiment shows that DeDL consistently outperforms across all m values. As m increases, simpler models like LA and LR experience rapid performance degradation due to their oversimplified linear extrapolation, which fails to capture complex treatment effects. In contrast, the performance of SDL and DeDL remains relatively stable. In Appendix D.3, we test the performance of our DeDL estimators with a potentially large bias to estimate $\hat{\boldsymbol{\theta}}(\cdot)$; we find that DeDL is fairly robust, with moderate biases. However, when this bias becomes excessively large, the effectiveness of DeDL diminishes. This highlights the importance of exercising careful DNN training during the first stage of the process. We also systematically assess the performance of DeDL under model misspecification, and shed light on how to test and select the link function in practice

(Appendix D.4). Resonating with our discussions in Section 5.3, this analysis highlights the crucial role of training or in-sample validation error in DML with DNNs. Furthermore, in Appendix D.5, we investigate a practical setting where the observed $\boldsymbol{X}$ distribution deviates from the population, and discuss how to use the rebalancing method to get trustworthy estimates.

7. Conclusion

In this paper, we leverage the DeDL framework to infer treatment effects for concurrent experiments and to identify the best treatment combination. We show the superior performance of our method using data from three A/B tests on a large-scale online platform. We also demonstrate the robustness of the DeDL method through synthetic experiments. Our framework can also be applied to analyze the individual heterogeneity of treatment effects with observational data under unconfoundedness.

We close by discussing two future directions. First, for multilevel discrete and continuous treatments, though our link functions can still be applied, researchers could design more-flexible link functions that better fit the richer treatment assignment mechanisms for the identification and convergence of parameter functions. Second, researchers could combine the current framework with classical causal inference methods such as instrumental variables and difference-in-differences for more general applications.

Acknowledgments

The authors thank Department Editor Prof. Vivek Farias, the anonymous associate editor, and three referees for their very helpful and constructive comments, which have led to significant improvements in both the content and exposition of this study. They also thank the industry partner for their support on sharing the data, implementing the algorithm, and conducting the experiment. Renyu Zhang is grateful for the financial support from the Hong Kong Research Grants Council General Research Fund [Grant 14502722, Grant 14504123, and Grant 14503224].

References

Adcock B, Dexter N (2021) The gap between theory and practice in function approximation with deep neural networks. *SIAM Journal on Mathematics of Data Science* 3(2):624–655.

Angrist JD, Pischke JS (2009) *Mostly harmless econometrics: An empiricist's companion* (Princeton university press).

Arkhangelsky D, Athey S, Hirshberg DA, Imbens GW, Wager S (2021) Synthetic difference-in-differences. *American Economic Review* 111(12):4088–4118.

Athey S, Imbens G (2016) Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences* 113(27):7353–7360.

Athey S, Imbens GW, Wager S (2018) Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 80(4):597–623.

Bertsimas D, Imai K, Li ML (2022) Distributionally robust causal inference with observational data. *arXiv preprint arXiv:2210.08326* .

Bojinov I, Simchi-Levi D, Zhao J (2022) Design and analysis of switchback experiments. *Management Science* forthcoming.

Box GE, Hunter WH, Hunter S, et al. (1978) *Statistics for experimenters*, volume 664 (John Wiley and sons New York).

Burtch G, Ghose A, Wattal S (2015) The hidden cost of accommodating crowdfunder privacy preferences: A randomized field experiment. *Management Science* 61(5):949–962.

Candogan O, Chen C, Niazadeh R (2021) Correlated cluster-based randomized experiments: Robust variance minimization. *Chicago Booth Research Paper* (21-17).

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

Chernozhukov V, Newey WK, Singh R (2022) Automatic debiased machine learning of causal and structural effects. *Econometrica* 90(3):967–1027.

Cheung WC, Simchi-Levi D, Wang H (2017) Dynamic pricing and demand learning with limited price experimentation. *Operations Research* 65(6):1722–1731.

Chiang HD, Kato K, Ma Y, Sasaki Y (2022) Multiway cluster robust double/debiased machine learning. *Journal of Business & Economic Statistics* 40(3):1046–1056.

Dasgupta T, Pillai NS, Rubin DB (2015) Causal inference from 2k factorial designs by using potential outcomes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 77(4):727–753.

Dube A, Jacobs J, Naidu S, Suri S (2020) Monopsony in online labor markets. *American Economic Review: Insights* 2(1):33–46.

Edelman B, Luca M, Svirsky D (2017) Racial discrimination in the sharing economy: Evidence from a field experiment. *American economic journal: applied economics* 9(2):1–22.

Fan Q, Hsu YC, Lieli RP, Zhang Y (2022) Estimation of conditional average treatment effects with high-dimensional data. *Journal of Business & Economic Statistics* 40(1):313–327.

Farbmacher H, Huber M, Lafférs L, Langen H, Spindler M (2022) Causal mediation analysis with double machine learning. *The Econometrics Journal* 25(2):277–300.

Farias V, Li A, Peng T (2021) Learning treatment effects in panels with general intervention patterns. *Advances in Neural Information Processing Systems* 34:14001–14013.

Farrell MH, Liang T, Misra S (2020) Deep learning for individual heterogeneity: an automatic inference framework. *arXiv preprint arXiv:2010.14694* .

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

Goli A, Lambrecht A, Yoganarasimhan H (2022) A bias correction approach for interference in ranking experiments. *Available at SSRN 4021266* .

Gordon BR, Moakler R, Zettelmeyer F (2023) Close enough? a large-scale exploration of non-experimental approaches to advertising measurement. *Marketing Science* 42(4):768–793.

Guo Y, Coey D, Konutgan M, Li W, Schoener C, Goldman M (2021) Machine learning for variance reduction in online experiments. *Advances in Neural Information Processing Systems* 34:8637–8648.

Hornik K, Stinchcombe M, White H (1989) Multilayer feedforward networks are universal approximators. *Neural networks* 2(5):359–366.

Imbens GW, Rubin DB (2015) *Causal inference in statistics, social, and biomedical sciences* (Cambridge University Press).

Johari R, Li H, Liskovich I, Weintraub GY (2022) Experimental design in two-sided platforms: An analysis of bias. *Management Science* 68(10):7069–7089.

Kallus N, Mao X, Udell M (2018) Causal inference with noisy and missing covariates via matrix factorization. *Advances in neural information processing systems* 31.

Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Knaus MC (2020) Double machine learning based program evaluation under unconfoundedness. *arXiv preprint arXiv:2003.03191* .

Kohavi R, Tang D, Xu Y (2020) *Trustworthy online controlled experiments: A practical guide to a/b testing* (Cambridge, UK: Cambridge University Press).

Kohavi R, Thomke S (2017) The surprising power of online experiments. *Harvard business review* 95(5):74–82.

Kushilevitz E, Mansour Y (1991) Learning decision trees using the fourier spectrum. *Proceedings of the twenty-third annual ACM symposium on Theory of computing*, 455–464.

Lee MR, Shen M (2018) Winner's curse: Bias estimation for total effects of features in online controlled experiments. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 491–499.

Li X, Ding P, Rubin DB (2020) Rerandomization in 2^k factorial experiments. *The Annals of Statistics* 48(1):43–63.

Lim AE, Shanthikumar JG, Shen ZM (2006) Model uncertainty, robust optimization, and learning. *Models, Methods, and Applications for Innovative Decision Making*, 66–94 (INFORMS).

Morgan KL, Rubin DB (2012) Rerandomization to improve covariate balance in experiments. *The Annals of Statistics* 40(2):1263–1282.

Nandy P, Venugopalan D, Lo C, Chatterjee S (2021) A/b testing for recommender systems in a two-sided marketplace. *Advances in Neural Information Processing Systems* 34:6466–6477.

Newey WK (1994) The asymptotic variance of semiparametric estimators. *Econometrica: Journal of the Econometric Society* 1349–1382.

Pashley NE, Bind MAC (2019) Causal inference for multiple treatments using fractional factorial designs. *arXiv preprint arXiv:1905.07596* .

Qi Z, Tang J, Fang E, Shi C (2022) Offline personalized pricing with censored demand. *Available at SSRN* .

Song Y, Sun T (2021) Ensembling experiments to optimize interventions along customer journey: A reinforcement learning approach. *Available at SSRN 3939073* .

Tang D, Agarwal A, O'Brien D, Meyer M (2010) Overlapping experiment infrastructure: More, better, faster experimentation. *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 17–26.

Wager S, Athey S (2018) Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association* 113(523):1228–1242.

Wooldridge JM (2010) *Econometric analysis of cross section and panel data* (MIT press).

Wu CJ, Hamada MS (2011) *Experiments: planning, analysis, and optimization* (John Wiley & Sons).

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

Xiong R, Chin A, Taylor S, Athey S (2022) Bias-variance tradeoffs for designing simultaneous temporal experiments .

Xiong T, Wang Y, Zheng S (2020) Orthogonal traffic assignment in online overlapping a/b tests. Technical report, EasyChair.

Yarotsky D (2017) Error bounds for approximations with deep relu networks. *Neural Networks* 94:103–114.

Ye Z, Zhang DJ, Zhang H, Zhang R, Chen X, Xu Z (2022) Cold start to improve market thickness on online advertising platforms: Data-driven algorithms and field experiments. *Management Science* 69(7):3838–3860.

Zeng Z, Dai H, Zhang DJ, Zhang H, Zhang RP, Xu Z, Shen ZJM (2022) The impact of social nudges on user-generated content for social network platforms. *Management Science* forthcoming.

Zhang DJ, Dai H, Dong L, Qi F, Zhang N, Liu X, Liu Z, Yang J (2020) The long-term and spillover effects of price promotions on retailing platforms: Evidence from a large randomized experiment on alibaba. *Management Science* 66(6):2589–2609.

Zhang Y, Politis DN (2022) Ridge regression revisited: Debiasing, thresholding and bootstrap. *The Annals of Statistics* 50(3):1401–1422.

Online Appendices

Deep-Learning-Based Causal Inference for Large-Scale Combinatorial Experiments

By Zikun Ye, Zhiqi Zhang, Dennis J. Zhang, Heng Zhang, Renyu Zhang

Appendix A: Technical Details

A.1. Proof of Theorem 1

Our proof is inspired by the analysis of universal approximation bounds of neural networks (e.g., Kushilevitz and Mansour 1991). At its core are the Fourier transformations of functions. Utilizing such transformations in our context, we develop a compact representation of the ATE as Boolean functions that allow us to connect to our *Generalized Sigmoid Form II*. The proof contains four steps. First, we give the Fourier representation of the function $f(\boldsymbol{t})$. Second, we rewrite the Fourier formula using indicator functions. Third, to remove the integral and obtain a finite population bound, we use a probabilistic method. Last, we approximate the indicator functions with our sigmoid link function.

Step 1 (Fourier representation). We first give the Fourier transformation of Boolean functions (e.g., Kushilevitz and Mansour 1991): For $f(\cdot):\{0,1\}^m \mapsto \mathbb{R}$, we always have

$$f(\boldsymbol{t}) \;=\; \sum_{\boldsymbol{w}\in\{0,1\}^m} \hat{f}(\boldsymbol{w})\chi_{\boldsymbol{w}}(\boldsymbol{t}), \quad \text{where} \;\; \hat{f}(\boldsymbol{w}) \;:=\; \frac{1}{2^m}\sum_{\boldsymbol{t}\in\{0,1\}^m} f(\boldsymbol{t})\chi_{\boldsymbol{w}}(\boldsymbol{t}),$$

where

$$\chi_{\boldsymbol{w}}(\boldsymbol{t}) \;=\; \begin{cases} 1 \;, & \text{if } \sum_{i=1}^m w_i t_i \mod 2 \;=\; 0 \\ -1 \;, & \text{if } \sum_{i=1}^m w_i t_i \mod 2 \;=\; 1. \end{cases}$$

Note that this implies

$$f(\boldsymbol{t}) \;=\; \sum_{\boldsymbol{w}\in\{0,1\}^m} \hat{f}(\boldsymbol{w})\chi_{\boldsymbol{w}}(\boldsymbol{t}) \;=\; \sum_{\boldsymbol{w}\in\{0,1\}^m} \left|\hat{f}(\boldsymbol{w})\right| \cos\left(\pi \boldsymbol{w}^\top \boldsymbol{t} + \delta(\boldsymbol{w})\right)$$

where $\delta(\boldsymbol{w}) = \pi \mathbb{1}\{\hat{f}(\boldsymbol{w}) < 0\}$.

Step 2 (Expansion with indicator functions). In the next, we expand the ATE, i.e., $f(\boldsymbol{t}) - f(\boldsymbol{0})$, using integrals and indicator functions.

$$\begin{aligned}
& f(\boldsymbol{t}) - f(\boldsymbol{0}) \\
=\; & \sum_{\boldsymbol{w}\in\{0,1\}^m} \left|\hat{f}(\boldsymbol{w})\right| \cos\left(\pi \boldsymbol{w}^\top \boldsymbol{t} + \delta(\boldsymbol{w})\right) - \cos\left(\delta(\boldsymbol{w})\right) \\
=\; & \sum_{\boldsymbol{w}\in\{0,1\}^m} \left|\hat{f}(\boldsymbol{w})\right| \cdot \left[-\pi \int_0^{\boldsymbol{w}^\top \boldsymbol{t}} \sin\left(\pi z + \delta(\boldsymbol{w})\right) \mathrm{d}z\right] \\
=\; & \sum_{\boldsymbol{w}\in\{0,1\}^m} \left|\hat{f}(\boldsymbol{w})\right| \cdot \left[-\pi \int_0^{\sqrt{m}\|\boldsymbol{w}\|} \mathbb{1}(\boldsymbol{w}^\top \boldsymbol{t} - z \geq 0) \sin\left(\pi z + \delta(\boldsymbol{w})\right) \mathrm{d}z + \pi \int_{-\sqrt{m}\|\boldsymbol{w}\|}^0 \mathbb{1}(\boldsymbol{w}^\top \boldsymbol{t} - z \leq 0) \sin\left(\pi z + \delta(\boldsymbol{w})\right) \mathrm{d}z\right],
\end{aligned} \tag{7}$$

where the last equality follows from the elementary inequality $|\boldsymbol{w}^\top \boldsymbol{t}| \leq \sqrt{m}\|\boldsymbol{w}\|$.

Step 3 (Sampling). Define $C_{f,m} := \sum_{\boldsymbol{w}\in\{0,1\}^m} \left|\hat{f}(\boldsymbol{w})\right| \sqrt{m}\|\boldsymbol{w}\|$. We use random sampling to remove the integrals in the ATE representation. Particularly, for each $\ell = 1,\dots,K$, we sample $(\boldsymbol{w}_\ell, z_\ell) \in \mathbb{R}^m \times \mathbb{R}$ in an *i.i.d* fashion as follows:

$$(\boldsymbol{w}_\ell, z_\ell) \;=\; \begin{cases} (0,0), & \text{with probability } \frac{1}{3}, \\ \{0,1\}^m \times (0, m\|\boldsymbol{w}\|], & \text{with probability density } \frac{|\hat{f}(\boldsymbol{w})|}{3C_{f,m}}, \\ \{0,1\}^m \times [-m\|\boldsymbol{w}\|, 0), & \text{with probability density } \frac{|\hat{f}(\boldsymbol{w})|}{3C_{f,m}}. \end{cases}$$

Also, define the function

$$g(\boldsymbol{w}, z; \boldsymbol{t}) \;:=\; \begin{cases} 3f(0)\mathbb{1}(\mathbf{0}^\top \boldsymbol{t} + 0 \geq 0), & \text{if } (\boldsymbol{w}, z) = \mathbf{0}, \\ -3\pi C_{f,m} \sin(\pi z + \delta(\boldsymbol{w}))\mathbb{1}(\boldsymbol{w}^\top \boldsymbol{t} - z \geq 0), & \text{if } (\boldsymbol{w}, z) \in \{0,1\}^m \times (\mathbf{0}, \sqrt{m}\|\boldsymbol{w}\|], \\ 3\pi C_{f,m} \sin(\pi z + \delta(\boldsymbol{w}))\mathbb{1}(-\boldsymbol{w}^\top \boldsymbol{t} + z \geq 0), & \text{if } (\boldsymbol{w}, z) \in \{0,1\}^m \times [-\sqrt{m}\|\boldsymbol{w}\|, \mathbf{0}). \end{cases}$$

Note that in the above we write $f(0) = f(0)\mathbb{1}(\mathbf{0}^\top \boldsymbol{t} + 0 \geq 0)$ for reasons that will be clear very soon. In view of (7), we notice that for any $\ell = 1,\dots,K$ and $\boldsymbol{t} \in \{0,1\}^m$, $\mathbb{E}_{(\boldsymbol{w}_\ell, z_\ell)}[g(\boldsymbol{w}_\ell, z_\ell; \boldsymbol{t})] = f(\boldsymbol{t})$. This implies that

$$\begin{aligned} \mathbb{E}_{\{(\boldsymbol{w}_\ell, z_\ell)\}_{\ell=1}^K} \left[\sum_{\boldsymbol{t}\in\{0,1\}^m} \left(f(\boldsymbol{t}) - \sum_{\ell=1}^K p(\ell) g(\boldsymbol{w}_\ell, z_\ell; \boldsymbol{t}) \right)^2 \right] \;&=\; \frac{1}{K} \sum_{\boldsymbol{t}\in\{0,1\}^m} \mathrm{Var}_{(\boldsymbol{w}_1, z_1)}[g(\boldsymbol{w}_1, z_1; \boldsymbol{t})] \\ &\leq\; \frac{1}{K} \sum_{\boldsymbol{t}\in\{0,1\}^m} \mathbb{E}_{(\boldsymbol{w}_1, z_1)}[g(\boldsymbol{w}_1, z_1; \boldsymbol{t})^2] \;\leq\; \frac{9 \cdot 2^m}{K} f^2(0) \vee C_{f,m}^2, \end{aligned}$$

where the first equality and the first inequality follow from straightforward computation, and the second inequality follows from the definition of $g(\cdot,\cdot;\cdot)$. Therefore, there exists $\{(\tilde{\boldsymbol{w}}_\ell, \tilde{z}_\ell) : \ell = 1,\dots K\}$ where each $(\tilde{\boldsymbol{w}}_\ell, \tilde{z}_\ell)$ is in the support of the random sampling distribution defined above such that

$$\frac{1}{2^m} \sum_{\boldsymbol{t}\in\{0,1\}^m} \left(f(\boldsymbol{t}) - \frac{1}{K} \sum_{\ell=1}^K g(\tilde{\boldsymbol{w}}_\ell, \tilde{z}_\ell; \boldsymbol{t}) \right)^2 \;\leq\; \frac{9}{K} f^2(0) \vee C_{f,m}^2.$$

Step 4 (Approximation with sigmoid functions). Note that regardless of the value of $(\boldsymbol{w}, z)$, $g(\boldsymbol{w}, z; t)$ can be written in the form $g_1(\boldsymbol{w}, z)\mathbb{1}(\boldsymbol{\alpha}(\boldsymbol{w})^\top t + \beta(z) \geq 0)$, where $g_1(\boldsymbol{w}, z)$, $\boldsymbol{\alpha}(\boldsymbol{w})$ and $\beta(z)$ are some functions of $\boldsymbol{w}$ and z. Further, such functions as $g_1(\boldsymbol{w}, z)\mathbb{1}(\boldsymbol{\alpha}(\boldsymbol{w})^\top t + \beta(z) \geq 0)$ can be well-approximated by scaled sigmoid functions. Indeed, $\frac{1}{1+e^{-\gamma z}} \to \mathbb{1}(z \geq 0)$ pointwise except for $z = 0$, when $\gamma \to \infty$. Thus, for any $\ell = 1,\dots,m$,

- if $(\tilde{\boldsymbol{w}}_\ell, \tilde{z}_\ell) = \mathbf{0}$, set $\theta_0(\boldsymbol{x}_\ell) = \gamma$, $\theta_1(\boldsymbol{x}_\ell) = \cdots = \theta_m(\boldsymbol{x}_\ell) = 0$, and $\theta_{m+1}(\boldsymbol{x}_\ell) = 3Kp(\ell)f(0)$;
- if $(\tilde{\boldsymbol{w}}_\ell, \tilde{z}_\ell) \in \{0,1\}^m \times (0, \sqrt{m}\|\tilde{\boldsymbol{w}}\|]$, set $\theta_0(\boldsymbol{x}_\ell) = -\gamma\tilde{z}_\ell$, $\theta_j(\boldsymbol{x}_\ell) = \gamma\tilde{w}_{\ell,j}$ for $j = 1,\dots,m$, and $\theta_{m+1}(\boldsymbol{x}_\ell) = -3\pi Kp(\ell)C_{f,m}\sin(\pi\tilde{z}_\ell + \delta(\tilde{\boldsymbol{w}}_\ell))$;
- and if $(\tilde{\boldsymbol{w}}_\ell, \tilde{z}_\ell) \in \{0,1\}^m \times [-\sqrt{m}\|\tilde{\boldsymbol{w}}\|, 0)$, set $\theta_0(\tilde{\boldsymbol{w}}_\ell) = \gamma\tilde{z}_\ell$, $\theta_j(\boldsymbol{x}_\ell) = -\gamma\tilde{w}_{\ell,j}$ for $j = 1,\dots,m$, and $\theta_{m+1}(\boldsymbol{x}_i) = 3\pi Kp(\ell)C_{f,m}\sin(\pi\tilde{z}_\ell + \delta(\tilde{\boldsymbol{w}}_\ell))$.

Here we set γ large enough so that

$$\left| p(\ell)G(\boldsymbol{\theta}(\boldsymbol{x}_\ell), \boldsymbol{t}) - \frac{1}{K} g(\tilde{\boldsymbol{w}}_\ell, \tilde{z}_\ell; \boldsymbol{t}) \right| \;\leq\; \frac{1}{K},$$

uniformly over all $\ell = 1,\dots,K$ and $\boldsymbol{t} \in \{0,1\}^m$. Thus,

$$\begin{aligned} &\frac{1}{2^m} \sum_{t\in\{0,1\}^m} \left(f(\boldsymbol{t}) - \sum_{\ell=1}^K p(\ell) G(\boldsymbol{\theta}(\boldsymbol{x}_i), \boldsymbol{t}) \right)^2 \\ &\leq \frac{2}{2^m} \sum_{t\in\{0,1\}^m} \left(f(\boldsymbol{t}) - \frac{1}{K} \sum_{\ell=1}^K g(\tilde{\boldsymbol{w}}_\ell, \tilde{z}_\ell; \boldsymbol{t}) \right)^2 + \frac{2K}{2^m} \sum_{t\in\{0,1\}^m} \sum_{\ell=1}^K \left(\frac{1}{K} g(\tilde{\boldsymbol{w}}_\ell, \tilde{z}_\ell; \boldsymbol{t}) - p(\ell) G(\boldsymbol{\theta}(\boldsymbol{x}_\ell), \boldsymbol{t}) \right)^2 \\ &\lesssim \frac{1}{K}(f^2(0) \vee C_{f,m}^2 + 1) \;\lesssim\; \frac{1}{K}, \end{aligned}$$

where the first inequality follows from repeated applications of the Young's inequality. We conclude the proof. □

A.2. Regularity Assumptions

We make the following regularity assumption throughout the theoretical analysis of this paper.

ASSUMPTION 3. *(a). $\boldsymbol{z}_i = (y_i, \boldsymbol{x}_i', \check{\boldsymbol{t}}_i')'$, $1 \le i \le n$, are i.i.d. copies from the population random variables $\boldsymbol{Z} = (Y, \boldsymbol{X}', \boldsymbol{T}')' \in \mathcal{Y} \times [-1,1]^{d_{\boldsymbol{X}}} \times \{0,1\}^m$, where $\mathcal{Y}$ is the bounded support of the outcome Y.*

(b). The parameter function $\boldsymbol{\theta}^*(\boldsymbol{x})$ is uniformly bounded. Furthermore, $\theta_k^*(\boldsymbol{x}) \in W^{p,\infty}([-1,1]^{d_{\boldsymbol{X}}})$, $k = 1, 2, \ldots, d_{\boldsymbol{\theta}}$, where for positive integers p, define the Sobolev ball $W^{p,\infty}([-1,1]^{d_{\boldsymbol{X}}})$ of functions $h : \mathbb{R}^{d_{\boldsymbol{X}}} \mapsto \mathbb{R}$ with smoothness $p \in \mathbb{N}_+$ as,

$$W^{p,\infty}([-1,1]^{d_{\boldsymbol{X}}}) := \Big\{ h : \max_{\boldsymbol{r}, |\boldsymbol{r}| < p} \operatorname*{ess\,sup}_{\boldsymbol{v} \in [-1,1]^{d_{\boldsymbol{X}}}} |D^{\boldsymbol{r}} h(\boldsymbol{v})| \le 1 \Big\},$$

where $\boldsymbol{r} = (r_1, \ldots, r_{d_{\boldsymbol{X}}})$, $|\boldsymbol{r}| = r_1 + \cdots + r_{d_{\boldsymbol{X}}}$ and $D^{\boldsymbol{r}} h$ is the weak derivative.

We remark that Assumption 3(a) implies that the DGP is bounded, whereas Assumption 3(b) ensures that the ground-truth parameter functions are uniformly bounded, and enjoy sufficient smoothness so they can be accurately approximated by DNNs. The smoothness assumptions (see, also, Assumption 2 in Farrell et al. 2020) are critical to deriving the sufficiently fast convergence rate of the estimator $\hat{\boldsymbol{\theta}}(\cdot)$.

We also make the following assumption throughout our analysis to ensure the identifiability and sufficient convergence rate of our model. Let $\boldsymbol{t}(S) = (t_1(S), \ldots, t_m(S))' \in \{0,1\}^m$ denote the treatment assignment such that $t_i(S) = \mathbf{1}\{i \in S\}$, where $S \subset \{1, 2, \ldots, m\}$, and define $\tilde{\boldsymbol{T}} := (1, \boldsymbol{T}')'$ and $\tilde{\boldsymbol{t}} := (1, \boldsymbol{t}')'$. Let $\lambda_{\min}(\cdot)$ denote the minimum eigenvalue of a symmetric matrix

ASSUMPTION 4. *Any of the following conditions hold:*

(a) $G(\cdot,\cdot)$ is of the Multiplicative Form, $\lambda_{\min}(\mathbb{E}[\tilde{\boldsymbol{T}}\tilde{\boldsymbol{T}}'|\boldsymbol{X}])$ and $|\theta_0^*(\boldsymbol{X})|$ are uniformly bounded away from zero;

(b) $G(\cdot,\cdot)$ is of the Standard Sigmoid Form, $\lambda_{\min}(\mathbb{E}[\tilde{\boldsymbol{T}}\tilde{\boldsymbol{T}}'|\boldsymbol{X}])$ is uniformly bounded away from zero;

(c) $G(\cdot,\cdot)$ is of the Generalized Sigmoid Form I, $\lambda_{\min}(\mathbb{E}[\boldsymbol{T}\boldsymbol{T}'|\boldsymbol{X}])$, $|\theta_{m+1}^*(\boldsymbol{X})|$, and $\nu(\boldsymbol{t}(\emptyset)|\boldsymbol{X})$ are uniformly bounded away from zero;

(d) $G(\cdot,\cdot)$ is of the Generalized Sigmoid Form II, $\lambda_{\min}(\mathbb{E}[\tilde{\boldsymbol{T}}\tilde{\boldsymbol{T}}'|\boldsymbol{X}])$ and $|\theta_{m+1}^*(\boldsymbol{x})|$ are uniformly bounded away from zero, and there exists a triplet (i, S_1, S_2) ($i \in \{1, \ldots, m\}$, $S_1, S_2 \subset \{1, 2, \ldots, m\}$) such that $i \notin S_1$, $i \notin S_2$, $S_1 \neq S_2$, and $\nu(\boldsymbol{t}(S_1 \cup \{i\})|\boldsymbol{X}) \cdot \nu(\boldsymbol{t}(S_1)|\boldsymbol{X}) \cdot \nu(\boldsymbol{t}(S_2 \cup \{i\})|\boldsymbol{X}) \cdot \nu(\boldsymbol{t}(S_2)|\boldsymbol{X})$ and $|G(\boldsymbol{\theta}^*(\boldsymbol{X}), \boldsymbol{t}(S_1 \cup \{i\}))G(\boldsymbol{\theta}^*(\boldsymbol{X}), \boldsymbol{t}(S_2)) - G(\boldsymbol{\theta}^*(\boldsymbol{X}), \boldsymbol{t}(S_2 \cup \{i\}))G(\boldsymbol{\theta}^*(\boldsymbol{X}), \boldsymbol{t}(S_1))|$ are both uniformly bounded away from zero.

Technically, we require that these conditions are satisfied uniformly on a set with probability one. For simplicity, we skip almost everywhere in these statements. Also, the assumptions $\lambda_{\min}(\mathbb{E}[\tilde{\boldsymbol{T}}\tilde{\boldsymbol{T}}'|\boldsymbol{X}])$ or $\lambda_{\min}(\mathbb{E}[\boldsymbol{T}\boldsymbol{T}'|\boldsymbol{X}])$ being uniformly bounded above zero is satisfied as long as all m individual treatment conditions and the full control condition are assigned with positive probability, i.e., $\nu((0,0,\ldots,0)'|\boldsymbol{x}) > c$ and $\nu((1,0,\ldots,0)'|\boldsymbol{x})$, $\nu((0,1,\ldots,0)'|\boldsymbol{x}), \ldots,$ $\nu((0,0,\ldots,1)'|\boldsymbol{x}) > c$, for some $c > 0$ almost everywhere, which is a fairly mild assumption. Therefore, these conditions state that $\Omega(m)$ treatment assignments are necessary.

At a high level, we indeed can extend our framework to cases with higher-order interactions between individual treatments, and Theorem 2 remains valid. However, the number of observable treatment combinations must increase. For instance, to ensure the valid identification and inference for the model with quadratic interactions mentioned above, one will need $\Omega(m^2)$ treatment combinations assigned with a positive probability. Following the same inference framework, we notice that the key is to guarantee the identifiability of the model and the sufficient curvature condition in the presence of higher-order interactions.

Let us first consider the case where all second-order treatment interactions are included in the link function. In this regard, upon inspection of the proof for the *Multiplicative Form*, *Standard Sigmoid Form*, and *Generalized Sigmoid Form I*, we note that only mild modifications of the assumptions on the data generation process ensure the validity of our results. More specifically, one only needs to redefine $\tilde{\boldsymbol{T}} = (1, \boldsymbol{T}', \bar{\boldsymbol{T}}')'$ where $\bar{\boldsymbol{T}} = (T_i T_j : \ \forall i < j, \ 1 \le i, j \le m)$. That is, we extend the definition of $\tilde{\boldsymbol{T}}$ to include all quadratic interactions. Then, all analyses of the *Multiplicative Form* and *Standard Sigmoid Form* go through unchanged. Particularly, we notice that in this case $\tilde{\boldsymbol{T}}$ is of dimension $\Theta(m^2)$, so to ensure that the minimum eigenvalue of $\mathbb{E}[\tilde{\boldsymbol{T}}\tilde{\boldsymbol{T}}'|\boldsymbol{X}]$ is uniformly bounded from below above zero, the support of treatment assignment distribution should have cardinality of order $\Omega(m^2)$. This is in sharp contrast with our results with linear terms only, which only requires $\Omega(m)$ treatment combinations with positive probability, highlighting the price for modeling the higher-order treatment interactions and estimating the associated nuisance parameters.

With the *Generalized Sigmoid Form II*, the analysis is more involved. Consider the following link function

$$G(\boldsymbol{\theta}(\boldsymbol{x}), \boldsymbol{t}) = \frac{\theta_{m+1}(\boldsymbol{x})}{1 + \exp(-(\theta_0(\boldsymbol{x}) + \sum_{i=1}^m \theta_i(\boldsymbol{x}) t_i + \sum_{i<j} \theta_{ij}(\boldsymbol{x}) t_i t_j)}.$$

As discussed above, we need the minimum eigenvalue of $\mathbb{E}[\tilde{\boldsymbol{T}}\tilde{\boldsymbol{T}}'|\boldsymbol{X}]$ is uniformly bounded from below above zero. Moreover, to estimate $\theta_{m+1}(\boldsymbol{x})$, fix any i, j with $1 \le i < j \le m$ and consider S and $\tilde{S}$ such that (1) $i, j \notin S, \tilde{S}$, (2) $S \ne \tilde{S}$, and (3) $\prod_{k=1}^8 \nu(\boldsymbol{t}(S_k)|\boldsymbol{X}) \ge \tilde{c}_1 > 0$ for some positive $\tilde{c}_1$, where $S_1 = S$, $S_2 = S \cup \{i\}$, $S_3 = S \cup \{j\}$, $S_4 = S \cup \{i, j\}$, $S_5 = \tilde{S}$, $S_6 = \tilde{S} \cup \{i\}$, $S_7 = \tilde{S} \cup \{j\}$, $S_8 = \tilde{S} \cup \{i, j\}$. Then it follows that

$$\frac{\left(\frac{\theta_{m+1}(\boldsymbol{X})}{G(\boldsymbol{\theta}(\boldsymbol{X}), \boldsymbol{t}(S_1))} - 1\right)\left(\frac{\theta_{m+1}(\boldsymbol{X})}{G(\boldsymbol{\theta}(\boldsymbol{X}), \boldsymbol{t}(S_4))} - 1\right)}{\left(\frac{\theta_{m+1}(\boldsymbol{X})}{G(\boldsymbol{\theta}(\boldsymbol{X}), \boldsymbol{t}(S_2))} - 1\right)\left(\frac{\theta_{m+1}(\boldsymbol{X})}{G(\boldsymbol{\theta}(\boldsymbol{X}), \boldsymbol{t}(S_3))} - 1\right)} = \exp(-(\theta_i(\boldsymbol{X}) + \theta_j(\boldsymbol{X}) + \theta_{ij}(\boldsymbol{X}))) \tag{8}$$

and

$$\frac{\left(\frac{\theta_{m+1}(\boldsymbol{X})}{G(\boldsymbol{\theta}(\boldsymbol{X}), \boldsymbol{t}(S_5))} - 1\right)\left(\frac{\theta_{m+1}(\boldsymbol{X})}{G(\boldsymbol{\theta}(\boldsymbol{X}), \boldsymbol{t}(S_8))} - 1\right)}{\left(\frac{\theta_{m+1}(\boldsymbol{X})}{G(\boldsymbol{\theta}(\boldsymbol{X}), \boldsymbol{t}(S_6))} - 1\right)\left(\frac{\theta_{m+1}(\boldsymbol{X})}{G(\boldsymbol{\theta}(\boldsymbol{X}), \boldsymbol{t}(S_7))} - 1\right)} = \exp(-(\theta_i(\boldsymbol{X}) + \theta_j(\boldsymbol{X}) + \theta_{ij}(\boldsymbol{X}))). \tag{9}$$

Under the assumption that $\theta_{m+1}(\boldsymbol{X}) \ne 0$, putting Eqns (8) and (9) together allows us to cancel out the term $\exp(-(\theta_i(\boldsymbol{X}) + \theta_j(\boldsymbol{X}) + \theta_{ij}(\boldsymbol{X})))$ and build a cubic equation of $\theta_{m+1}(\boldsymbol{X})$ with the products of $G(\boldsymbol{\theta}(\boldsymbol{X}), \boldsymbol{t}(S_i))$ $(i = 1, \ldots, 8)$ as the coefficients. This cubic equation admits at most 3 real solutions. As long as there are 3 such (i, j) pairs (and the correspondingly sets S and $\tilde{S}$), we can uniquely identify $\theta_{m+1}(\boldsymbol{X})$. Therefore, we need $\Omega(m^2)$ treatment combinations with positive probability in the treatment assignment mechanism to ensure identification.

The argument above aligns with our analysis of the identification condition for *Generalized Sigmoid Form II* detailed in Appendix A.3, yet it becomes significantly more complex in the presence of second-order treatment

interactions. More generally, with higher-order interactions, we can establish equations similar to (8) and (9) to formulate primitive conditions for identifying $\theta_{m+1}(\boldsymbol{X})$, which makes the identification condition more restrictive and practically infeasible. Consequently, carefully trading off the generality, tractability, and practicality of our framework, we have decided to focus on the *Generalized Sigmoid Form II* link function.

A.3. Influence Function

The following lemma from Farrell et al. (2020) formally states a generic result regarding the influence function, which proves useful to derive the influence function in our setting (Proposition 2). The proof follows the pathwise derivative method originated in Newey (1994).

LEMMA 1 **(Theorem 2 in Farrell et al. (2020))**. *For all $\boldsymbol{t} \in \{0,1\}^m$, suppose the following conditions hold uniformly in the given conditioning elements. (i) (1) holds and identifies $\boldsymbol{\theta}^*(\cdot)$. (ii) $\mathbb{E}[\ell_{\boldsymbol{\theta}}(Y,\check{\boldsymbol{t}},\boldsymbol{\theta}^*(\boldsymbol{x}))|\boldsymbol{X}=\boldsymbol{x},\boldsymbol{T}=\check{\boldsymbol{t}}]=0$. (iii) $\boldsymbol{\Lambda}(x) := \mathbb{E}[\ell_{\boldsymbol{\theta}\boldsymbol{\theta}}(Y,\boldsymbol{T},\boldsymbol{\theta}(\boldsymbol{x}))|\boldsymbol{X}=\boldsymbol{x}]$ is invertible with uniformly bounded inverse. (iv) Parameter $\mu(\boldsymbol{t})$ is identified, pathwise differentiable, and H and ℓ are thrice continuously differentiable in $\boldsymbol{\theta}$. (v) $H(\boldsymbol{X},\boldsymbol{\theta}^*(\boldsymbol{X}),\boldsymbol{t},\boldsymbol{t}_0)$ and $\ell_{\boldsymbol{\theta}}(Y,\boldsymbol{T},\boldsymbol{\theta}^*(\boldsymbol{X}))$ possess $q>4$ finite absolute moments and positive variance. Then for the treatment effect $\mu(\boldsymbol{t})$, the Neyman orthogonal score is $\psi(\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\Lambda};\boldsymbol{t},\boldsymbol{t}_0)-\mu(\boldsymbol{t})$, where*

$$\psi(\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\Lambda};\boldsymbol{t},\boldsymbol{t}_0) = H(\boldsymbol{x},\boldsymbol{\theta}(\boldsymbol{x});\boldsymbol{t},\boldsymbol{t}_0) - H_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{\theta}(\boldsymbol{x});\boldsymbol{t},\boldsymbol{t}_0)'\boldsymbol{\Lambda}(\boldsymbol{x})^{-1}\ell_{\boldsymbol{\theta}}(y,\check{\boldsymbol{t}},\boldsymbol{\theta}(\boldsymbol{x})), \tag{10}$$

where $\ell_{\boldsymbol{\theta}}, H_{\boldsymbol{\theta}}$ are $d_{\boldsymbol{\theta}}$-dimensional vectors of first order derivatives, and $\ell_{\boldsymbol{\theta}\boldsymbol{\theta}}$ is the $d_{\boldsymbol{\theta}}\times d_{\boldsymbol{\theta}}$ Hessian matrix of ℓ, with $\{k_1,k_2\}$ element defined by $\partial^2\ell/\partial\theta_{k_1}\partial\theta_{k_2}$.

A.4. Proof of Proposition 1

We first present without proof a key convergence result inherited from Farrell et al. (2020).

LEMMA 2 **(Theorem 1 in Farrell et al. (2020))**. *Suppose Assumption 3, and the following regularity assumptions hold,*

(a). (NONPARAMETRIC IDENTIFIABILITY) *The arameter function $\boldsymbol{\theta}^*(\boldsymbol{x})$ can be nonparametrically identified in DGP (1).*

(b). (LIPSCHITZ CONTINUITY) *There exists a positive constant C_ℓ such that, for any $\boldsymbol{\theta}(\cdot)$, $\tilde{\boldsymbol{\theta}}(\cdot)$ and $\boldsymbol{x}$,*

$$|\,\ell(y,\boldsymbol{t},\boldsymbol{\theta}(\boldsymbol{x})) - \ell(y,\boldsymbol{t},\tilde{\boldsymbol{\theta}}(\boldsymbol{x}))| \le C_\ell\|\boldsymbol{\theta}(\boldsymbol{x})-\tilde{\boldsymbol{\theta}}(\boldsymbol{x})\|_2, \tag{11}$$

(c). (SUFFICIENT CURVATURE) *There exist positive constants c_1 and c_2 such that, for any $\boldsymbol{\theta}(\cdot)\in\mathcal{F}_{DNN}$,*

$$c_1\mathbb{E}[\|\boldsymbol{\theta}(\boldsymbol{X})-\boldsymbol{\theta}^*(\boldsymbol{X})\|_2^2] \le \mathbb{E}[\ell(Y,\boldsymbol{T},\boldsymbol{\theta}(\boldsymbol{X}))] - \mathbb{E}[\ell(Y,\boldsymbol{T},\boldsymbol{\theta}^*(\boldsymbol{X}))] \le c_2\mathbb{E}[\|\boldsymbol{\theta}(\boldsymbol{X})-\boldsymbol{\theta}^*(\boldsymbol{X})\|_2^2]. \tag{12}$$

With the structured DNN of width $H = O(n^{\frac{d_{\boldsymbol{X}}}{2(p+d_{\boldsymbol{X}})}}\log^2 n)$ and depth $L = O(\log n)$ as illustrated in Figure 3, there exists a constant C such that

$$\|\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*\|^2_{L_2(\boldsymbol{X})} \lesssim n^{-\frac{p}{p+d_{\boldsymbol{X}}}}\log^8 n + \frac{\log\log n}{n}$$

and

$$\mathbb{E}_n\big[(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*)^2\big] \lesssim n^{-\frac{p}{p+d_{\boldsymbol{X}}}}\log^8 n + \frac{\log\log n}{n}$$

for n large enough with probability at least $1-\exp(n^{-\frac{d_{\boldsymbol{X}}}{p+d_{\boldsymbol{X}}}}\log^8 n)$, for $k=1,\ldots,d_{\boldsymbol{\theta}}$.

Assumptions in Lemma 2 are natural and common in the nonparametric M-estimation literature, which consists of three parts: (a) the nonparametric identifiability of $\boldsymbol{\theta}^*(\cdot)$, (b) the Lipschitz continuity of loss (i.e., (11)), and (c) the sufficient curvature of loss (i.e., (12)). Whereas the Lipschitz continuity condition is mild and easy to check in our setting, the identifiability of $\boldsymbol{\theta}^*(\cdot)$ and the sufficient curvature condition are non-trivial and should be verified carefully. In particular, the sufficient curvature condition (12) is usually implied by a proper choice of the link function $G(\cdot,\cdot)$ and the treatment assignment mechanism $\nu(\cdot\,|\,\cdot)$. This condition helps translate the convergence of outcomes Y into that of parameter functions $\boldsymbol{\theta}(\cdot)$.

To obtain the convergence results in Proposition 1, it suffices to verify the assumptions in Lemma 2 are satisfied. The Lipschitz condition in Assumption (b) can be easily satisfied by for all our proposed link functions in Section 3.3 because all G functions are sufficiently smooth with bounded $\boldsymbol{X}$, $\boldsymbol{T}$, and $\boldsymbol{\theta}$. Since the square loss function and our link functions are differentiable, the second inequality in the curvature condition is satisfied for all of our link functions. In particular, we note the identity

$$\mathbb{E}[\ell(Y,\boldsymbol{T},\boldsymbol{\theta}(\boldsymbol{X}))]-\mathbb{E}[\ell(Y,\boldsymbol{T},\boldsymbol{\theta}^*(\boldsymbol{X}))]=\mathbb{E}\left[\left(G(\boldsymbol{\theta}(\boldsymbol{X}),\boldsymbol{T})-G(\boldsymbol{\theta}^*(\boldsymbol{X}),\boldsymbol{T})\right)^2\right] \tag{13}$$

by DGP (1), which then implies by the mean value theorem

$$\mathbb{E}[\ell(Y,\boldsymbol{T},\boldsymbol{\theta}(\boldsymbol{X}))]-\mathbb{E}[\ell(Y,\boldsymbol{T},\boldsymbol{\theta}^*(\boldsymbol{X}))]=\mathbb{E}[(\boldsymbol{\theta}(\boldsymbol{X})-\boldsymbol{\theta}^*(\boldsymbol{X}))'G_{\boldsymbol{\theta}}(\tilde{\boldsymbol{\theta}}(\boldsymbol{X}),\boldsymbol{T})G_{\boldsymbol{\theta}}(\tilde{\boldsymbol{\theta}}(\boldsymbol{X}),\boldsymbol{T})'(\boldsymbol{\theta}(\boldsymbol{X})-\boldsymbol{\theta}^*(\boldsymbol{X}))], \tag{14}$$

where $\tilde{\boldsymbol{\theta}}(\boldsymbol{X})$ is such that $\tilde{\theta}_i(\boldsymbol{X})\in[\theta_i^*(\boldsymbol{X}),\theta_i(\boldsymbol{X})]$ for all component i. Since all variables are bounded, $G_{\boldsymbol{\theta}}(\tilde{\boldsymbol{\theta}}(\boldsymbol{X}),\boldsymbol{T})G_{\boldsymbol{\theta}}(\tilde{\boldsymbol{\theta}}(\boldsymbol{X}),\boldsymbol{T})'$ is also uniformly bounded and the claim follows.

Consequently, it suffices to verify the nonparametric identifiability and the first inequality in the curvature condition in Lemma 2 for different forms of G functions to guarantee the convergence of structured DNNs. To simplify the notation, we define the one-dimension sigmoid function as $S(x):=1/(1+\exp(-x))$. In the following, we prove that, for each proposed link functions in Assumption 1, the conditions of Lemma 2 hold. We also remark that the constants may be different for different forms of the link function.

Standard Sigmoid Form. In this part, we start with the standard sigmoid form,

$$G(\boldsymbol{\theta}(\boldsymbol{x}),\boldsymbol{t}):=\frac{a}{1+\exp\Big(-(\theta_0(\boldsymbol{x})+\theta_1(\boldsymbol{x})t_1+\cdots+\theta_m(\boldsymbol{x})t_m)\Big)}+b, \tag{15}$$

where the constants $a\neq 0,\ b$ are known.

Proof. Because the sigmoid function is invertible, it suffices to verify the identifiability and sufficient curvature conditions for the linear link function $\boldsymbol{\theta}(\boldsymbol{x})'\tilde{\boldsymbol{t}}$. Suppose $\mathbb{E}[Y|\boldsymbol{X}=\boldsymbol{x},\boldsymbol{T}=\boldsymbol{t}]=G(\hat{\boldsymbol{\theta}}(\boldsymbol{x})'\tilde{\boldsymbol{t}})=G(\boldsymbol{\theta}^*(\boldsymbol{x})'\tilde{\boldsymbol{t}})$. Next, we show $\boldsymbol{\theta}^*(\cdot)$ can be nonparametrically identified, i.e., $\hat{\boldsymbol{\theta}}(\boldsymbol{X})=\boldsymbol{\theta}^*(\boldsymbol{X})$. For all $\tilde{\boldsymbol{t}}$, we have,

$$0=\hat{\boldsymbol{\theta}}(\boldsymbol{X})'\tilde{\boldsymbol{t}}-\boldsymbol{\theta}^*(\boldsymbol{X})'\tilde{\boldsymbol{t}}=(\hat{\boldsymbol{\theta}}(\boldsymbol{X})-\boldsymbol{\theta}^*(\boldsymbol{X}))'\tilde{\boldsymbol{t}},$$

which implies,

$$0=\mathbb{E}[((\hat{\boldsymbol{\theta}}(\boldsymbol{X})-\boldsymbol{\theta}^*(\boldsymbol{X}))'\tilde{\boldsymbol{T}})^2|\boldsymbol{X}]=(\hat{\boldsymbol{\theta}}(\boldsymbol{X})-\boldsymbol{\theta}^*(\boldsymbol{X}))'\mathbb{E}[\tilde{\boldsymbol{T}}\tilde{\boldsymbol{T}}'|\boldsymbol{X}](\hat{\boldsymbol{\theta}}(\boldsymbol{X})-\boldsymbol{\theta}^*(\boldsymbol{X})).$$

Because $\mathbb{E}[\tilde{\boldsymbol{T}}\tilde{\boldsymbol{T}}'|\boldsymbol{X}]$ is positive definite uniformly with respect to $\boldsymbol{X}$, it must hold that $\hat{\boldsymbol{\theta}}(\boldsymbol{X})-\tilde{\boldsymbol{\theta}}(\boldsymbol{X})=0$ almost everywhere, which concludes the proof of identifiability.

We show that the first inequality in the sufficient curvature condition is satisfied. Similar to (14), it holds that

$$\begin{aligned}\mathbb{E}[\ell(Y,\boldsymbol{T},\boldsymbol{\theta}(\boldsymbol{X}))]-\mathbb{E}[\ell(Y,\boldsymbol{T},\boldsymbol{\theta}^*(\boldsymbol{X}))] &= \mathbb{E}\left[\left((\boldsymbol{\theta}(\boldsymbol{X})-\boldsymbol{\theta}^*(\boldsymbol{X}))'\tilde{\boldsymbol{T}}\right)^2\cdot S^2(w(\boldsymbol{X},\tilde{\boldsymbol{T}}))\left(1-S(w(\boldsymbol{X},\tilde{\boldsymbol{T}}))\right)^2\right] \quad (16)\\ &\geq \tilde{c}_1\mathbb{E}\left[\left((\boldsymbol{\theta}(\boldsymbol{X})-\boldsymbol{\theta}^*(\boldsymbol{X}))'\tilde{\boldsymbol{T}}\right)^2\right] \quad (17)\end{aligned}$$

where $w(\boldsymbol{X},\tilde{\boldsymbol{T}})$ is some function such that $w(\boldsymbol{X},\tilde{\boldsymbol{T}})\in[\boldsymbol{\theta}^*(\boldsymbol{X})'\tilde{\boldsymbol{T}},\ \boldsymbol{\theta}(\boldsymbol{X})'\tilde{\boldsymbol{T}}]$ and the inequality follows since all of the variables are bounded. Then, under the assumption $\mathbb{E}[\tilde{\boldsymbol{T}}\tilde{\boldsymbol{T}}'|\boldsymbol{X}]$ is uniformly bounded away from zero and following the law of total expectation, we derive that,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{X},\boldsymbol{T}}\left[((\boldsymbol{\theta}(\boldsymbol{X})'\tilde{\boldsymbol{T}}-\boldsymbol{\theta}^*(\boldsymbol{X})'\tilde{\boldsymbol{T}})^2\right] &= a^2\mathbb{E}_{\boldsymbol{X}}[((\boldsymbol{\theta}(\boldsymbol{X})-\boldsymbol{\theta}^*(\boldsymbol{X}))'\mathbb{E}[\tilde{\boldsymbol{T}}\tilde{\boldsymbol{T}}'|\boldsymbol{X}](\boldsymbol{\theta}(\boldsymbol{X})-\boldsymbol{\theta}^*(\boldsymbol{X}))]\\ &\geq \tilde{c}_2\mathbb{E}_{\boldsymbol{X}}[(\boldsymbol{\theta}(\boldsymbol{X})-\boldsymbol{\theta}^*(\boldsymbol{X}))'(\boldsymbol{\theta}(\boldsymbol{X})-\boldsymbol{\theta}^*(\boldsymbol{X}))],\end{aligned}$$

for some $\tilde{c}_2>0$, which concludes the proof of sufficient curvature, i.e.,

$$\mathbb{E}[\ell(Y,\boldsymbol{T},\boldsymbol{\theta}(\boldsymbol{X}))]-\mathbb{E}[\ell(Y,\boldsymbol{T},\boldsymbol{\theta}^*(\boldsymbol{X}))]\ \geq\ \tilde{c}_1\tilde{c}_2\mathbb{E}_{\boldsymbol{X}}[(\boldsymbol{\theta}(\boldsymbol{X})-\boldsymbol{\theta}^*(\boldsymbol{X}))'(\boldsymbol{\theta}(\boldsymbol{X})-\boldsymbol{\theta}^*(\boldsymbol{X}))] \quad (18)$$

This concludes the proof. □

Multiplicative Form. Consider the link function

$$G(\boldsymbol{\theta}(\boldsymbol{x}),\boldsymbol{t})=\theta_0(\boldsymbol{x})\big(1+\theta_1(\boldsymbol{x})t_1\big)\ldots\big(1+\theta_m(\boldsymbol{x})t_m\big), \quad (19)$$

where $\mu\leq 1+\theta_k(\boldsymbol{x})\leq M,\ k=1,\ldots,m$, uniformly in $\boldsymbol{x}$ by assumption.

Proof. For simplicity of the proof, let us assume throughout $\theta_0(\boldsymbol{x})>0$ for $\boldsymbol{x}$ almost everywhere. Other cases can be proved similarly. These conditions guarantee that $\log\big(G(\boldsymbol{\theta}(\boldsymbol{x}),\boldsymbol{t})\big)$ is well-defined. The proof of nonparametric identifiability and the curvature conditions can be verified as shown in the following:

$$\begin{aligned}\log G &= \log\theta_0(\boldsymbol{x})(1+\theta_1(\boldsymbol{x})t_1)\ldots(1+\theta_m(\boldsymbol{x})t_m)\\ &= \log\theta_0(\boldsymbol{x})+\log(1+\theta_1(\boldsymbol{x})t_1)+\cdots+\log(1+\theta_m(\boldsymbol{x})t_m)\\ &= \log\theta_0(\boldsymbol{x})+\log(1+\theta_1(\boldsymbol{x}))t_1+\cdots+\log(1+\theta_m(\boldsymbol{x}))t_m),\end{aligned}$$

where the last equality follows from that $t_i=0,1$ for all $i=1,2,...,m$. Hence, the *Multiplicative Form* is equivalent to $G(\boldsymbol{\theta}(\boldsymbol{x}),\boldsymbol{t})=\exp(a(\boldsymbol{x})+b_1(\boldsymbol{x})t_1+\cdots+b_m(\boldsymbol{x})t_m)$, with $a(\boldsymbol{x})=\log\theta_0(\boldsymbol{x})$, and $b_k(\boldsymbol{x})=\log(1+\theta_k(\boldsymbol{x}))$. Because the exponential function is monotone and smooth, to satisfy the identification and sufficient curvature condition, we only need that $\mathbb{E}[\tilde{\boldsymbol{T}}\tilde{\boldsymbol{T}}'|\boldsymbol{X}=\boldsymbol{x}]$ are invertible uniformly in $\boldsymbol{x}$, where $\tilde{\boldsymbol{T}}'=(1,\boldsymbol{T}')$, i.e., Assumption 4(a). The proof follows from the same argument as the *Standard Sigmoid Form*. □

Generalized Sigmoid Form I. Next, we consider

$$G(\boldsymbol{\theta}(\boldsymbol{x}),\boldsymbol{t}):=\frac{\theta_{m+1}(\boldsymbol{x})}{1+\exp\Big(-(\theta_1(\boldsymbol{x})t_1+\cdots+\theta_m(\boldsymbol{x})t_m)\Big)}, \quad (20)$$

where $\theta_{m+1}(\boldsymbol{x})\in\mathbb{R}$ can capture the range of the expected outcome, and the sign of $\theta_i(\boldsymbol{x})$ $(i=1,2,...,m)$ represents whether the experiment i has a positive or negative effect.

Proof. Let $\boldsymbol{\theta} = (\boldsymbol{\theta}_-(\boldsymbol{x})', \theta_{m+1}(\boldsymbol{x}))'$, where $\boldsymbol{\theta}_-(\boldsymbol{x}) = (\theta_1(\boldsymbol{x}), \ldots, \theta_m(\boldsymbol{x}))'$. Hence, we rewrite $G(\boldsymbol{\theta}(\boldsymbol{x}), \boldsymbol{t}) = \theta_{m+1}(\boldsymbol{x}) S(\boldsymbol{\theta}_-(\boldsymbol{x})'\boldsymbol{t})$.

First, we show that $\boldsymbol{\theta}^*(\cdot)$ can be nonparametrically identified. Assume that $\mathbb{E}[Y|\boldsymbol{X}, \boldsymbol{T}] = \hat{\theta}_{m+1}(\boldsymbol{x}) S(\hat{\boldsymbol{\theta}}_-(\boldsymbol{X})'\boldsymbol{T}) = \theta^*_{m+1}(\boldsymbol{X}) S(\boldsymbol{\theta}^*_-(\boldsymbol{X})'\boldsymbol{T})$. Because $\nu(\mathbf{0} \,|\, \boldsymbol{X}) = \mathbb{P}[\boldsymbol{T} = \mathbf{0}|\boldsymbol{X}] > 0$ uniformly and $S(0) = 0.5 > 0$, we have $\hat{\theta}_{m+1}(\boldsymbol{X}) = \theta^*_{m+1}(\boldsymbol{X})$. Because $\theta^*_{m+1}(\boldsymbol{X}) S(\boldsymbol{\theta}^*_-(\boldsymbol{X})'\boldsymbol{T}) \neq 0$ by assumption and $\hat{\theta}_{m+1}(\boldsymbol{X}) = \theta^*_{m+1}(\boldsymbol{X})$, we have $S(\hat{\boldsymbol{\theta}}_-(\boldsymbol{X})'\boldsymbol{T}) = S(\boldsymbol{\theta}^*_-(\boldsymbol{X})'\boldsymbol{T})$. Also, because $S(\cdot)$ is continuously invertible, following the similar argument for the *Standard Sigmoid Form*, we derive that

$$0 = \mathbb{E}[(\hat{\boldsymbol{\theta}}_-(\boldsymbol{X})'\boldsymbol{T} - \boldsymbol{\theta}^*_-(\boldsymbol{X})'\boldsymbol{T})^2 | \boldsymbol{X}] = (\hat{\boldsymbol{\theta}}_-(\boldsymbol{X}) - \boldsymbol{\theta}^*_-(\boldsymbol{X}))' \mathbb{E}[\boldsymbol{T}\boldsymbol{T}'|\boldsymbol{X}] (\hat{\boldsymbol{\theta}}_-(\boldsymbol{X}) - \boldsymbol{\theta}^*_-(\boldsymbol{X})).$$

Since $\mathbb{E}[\boldsymbol{T}\boldsymbol{T}'| \ \boldsymbol{X}] \succ 0$, it follows that $\hat{\boldsymbol{\theta}}_-(\boldsymbol{X}) = \boldsymbol{\theta}^*_-(\boldsymbol{X})$. Together with $\hat{\theta}_{m+1}(\boldsymbol{x}) = \theta^*_{m+1}(\boldsymbol{x})$, we conclude the proof of identifiability.

Next, we show the sufficient curvature condition. Using (13), we obtain $\mathbb{E}[\ell(Y, \boldsymbol{T}, \boldsymbol{\theta}(\boldsymbol{X}))] - \mathbb{E}[\ell(Y, \boldsymbol{T}, \boldsymbol{\theta}^*(\boldsymbol{X}))] = \mathbb{E}[(\theta_{m+1}(\boldsymbol{X}) S(\boldsymbol{\theta}_-(\boldsymbol{X})'\boldsymbol{T}) - \theta^*_{m+1}(\boldsymbol{X}) S(\boldsymbol{\theta}^*_-(\boldsymbol{X})'\boldsymbol{T}))^2]$. Hence, the sufficient curvature condition is equivalent to that there exists a constant $c_1 > 0$ such that

$$c_1 \mathbb{E}[\sum_{i=1}^{m+1} (\theta_i(\boldsymbol{X}) - \theta^*_i(\boldsymbol{X}))^2] \leq \mathbb{E}[(\theta_{m+1}(\boldsymbol{X}) S(\boldsymbol{\theta}_-(\boldsymbol{X})'\boldsymbol{T}) - \theta^*_{m+1}(\boldsymbol{X}) S(\boldsymbol{\theta}^*_-(\boldsymbol{X})'\boldsymbol{T}))^2].$$

Since $\mathbb{P}[\boldsymbol{T} = \mathbf{0}|\boldsymbol{X}]$ is uniformly bounded away from zero, which implies that $|G(\boldsymbol{\theta}^*(\boldsymbol{X})'\boldsymbol{t}(\emptyset))| = |\theta^*_{m+1}(\boldsymbol{X}) S(0)| = |\theta_{m+1}(\boldsymbol{X})|/2$ with probability uniformly bounded away from zero for all $\boldsymbol{X}$, there must exist a constant $\tilde{c}_1 > 0$ such that,

$$\mathbb{E}_{\boldsymbol{X},\boldsymbol{T}}\left[\left(\theta_{m+1}(\boldsymbol{X}) S(\boldsymbol{\theta}_-(\boldsymbol{X})'\boldsymbol{T}) - \theta^*_{m+1}(\boldsymbol{X}) S(\boldsymbol{\theta}^*_-(\boldsymbol{X})'\boldsymbol{T})\right)^2\right] \geq \tilde{c}_1 \mathbb{E}_{\boldsymbol{X}}\left[\left(\theta_{m+1}(\boldsymbol{X}) - \theta^*_{m+1}(\boldsymbol{X})\right)^2\right]. \tag{21}$$

With the condition $\mathbb{E}[\boldsymbol{T}\boldsymbol{T}'|\boldsymbol{X}]$ uniformly bounded away from zero, using the argument similar to (18), we can show that there exists a constant $\tilde{c}_2 > 0$, the following inequality holds

$$\mathbb{E}_{\boldsymbol{X},\boldsymbol{T}}[(S(\boldsymbol{\theta}_-(\boldsymbol{X})'\boldsymbol{T}) - S(\boldsymbol{\theta}^*_-(\boldsymbol{X})'\boldsymbol{T}))^2] \geq \tilde{c}_2 \mathbb{E}_{\boldsymbol{X}}\left[\sum_{i=1}^{m} (\theta_i(\boldsymbol{X}) - \theta^*_i(\boldsymbol{X}))^2\right]. \tag{22}$$

Then, for any realized $\boldsymbol{x}$ and $\boldsymbol{t}$, we have the following decomposition

$$\begin{aligned}
&\left|\theta^*_{m+1}(\boldsymbol{x}) S(\boldsymbol{\theta}^*_-(\boldsymbol{x})'\boldsymbol{t}) - \theta_{m+1}(\boldsymbol{x}) S(\boldsymbol{\theta}_-(\boldsymbol{x})'\boldsymbol{t})\right| \\
\geq &\left|\theta^*_{m+1}(\boldsymbol{x}) S(\boldsymbol{\theta}^*_-(\boldsymbol{x})'\boldsymbol{t}) - \theta^*_{m+1}(\boldsymbol{x}) S(\boldsymbol{\theta}_-(\boldsymbol{x})'\boldsymbol{t})\right| - \left|\theta^*_{m+1}(\boldsymbol{x}) S(\boldsymbol{\theta}_-(\boldsymbol{x})'\boldsymbol{t}) - \theta_{m+1}(\boldsymbol{x}) S(\boldsymbol{\theta}_-(\boldsymbol{x})'\boldsymbol{t})\right| \\
= &|\theta^*_{m+1}(\boldsymbol{x})| \cdot \left|S(\boldsymbol{\theta}^*_-(\boldsymbol{x})'\boldsymbol{t}) - S(\boldsymbol{\theta}_-(\boldsymbol{x})'\boldsymbol{t})\right| - |\theta^*_{m+1}(\boldsymbol{x}) - \theta_{m+1}(\boldsymbol{x})| \cdot S(\boldsymbol{\theta}_-(\boldsymbol{x})'\boldsymbol{t}).
\end{aligned}$$

Since $|\theta^*_{m+1}(\boldsymbol{x})| > 0$, rearranging the terms implies that

$$|S(\boldsymbol{\theta}^*_-(\boldsymbol{x})'\boldsymbol{t}) - S(\boldsymbol{\theta}_-(\boldsymbol{x})'\boldsymbol{t})| \leq \frac{|\theta^*_{m+1}(\boldsymbol{x}) S(\boldsymbol{\theta}^*_-(\boldsymbol{x})'\boldsymbol{t}) - \theta_{m+1}(\boldsymbol{x}) S(\boldsymbol{\theta}_-(\boldsymbol{x})'\boldsymbol{t})| + |\theta^*_{m+1}(\boldsymbol{x}) - \theta_{m+1}(\boldsymbol{x})| S(\boldsymbol{\theta}_-(\boldsymbol{x})'\boldsymbol{t})}{|\theta^*_{m+1}(\boldsymbol{x})|}.$$

If $|A| \leq |B| + |C|$, one can get $|A|^2 \leq |B|^2 + |C|^2 + 2|BC| \leq 2|B|^2 + 2|C|^2$. Using this elementary identity, together with the uniform boundedness of $\theta^*_{m+1}(\boldsymbol{x})$, we arrive at that there exists some constant $\tilde{c}_3 > 0$ such that

$$\begin{aligned}
\tilde{c}_3 \mathbb{E}\left[\left(S(\boldsymbol{\theta}_-(\boldsymbol{X})'\boldsymbol{T}) - S(\boldsymbol{\theta}^*_-(\boldsymbol{X})'\boldsymbol{T})\right)^2\right] \ &\leq \ \mathbb{E}\left[\left(\theta_{m+1}(\boldsymbol{X}) S(\boldsymbol{\theta}_-(\boldsymbol{X})'\boldsymbol{T}) - \theta^*_{m+1}(\boldsymbol{X}) S(\boldsymbol{\theta}^*_-(\boldsymbol{X})'\boldsymbol{T})\right)^2\right] \\
&\quad + \mathbb{E}\left[\left(\theta^*_{m+1}(\boldsymbol{X}) - \theta_{m+1}(\boldsymbol{X})\right)^2\right] \\
&\leq \ \left(1 + \frac{1}{\tilde{c}_1}\right) \cdot \mathbb{E}[(\theta_{m+1}(\boldsymbol{X}) S(\boldsymbol{\theta}_-(\boldsymbol{X})'\boldsymbol{T}) - \theta^*_{m+1}(\boldsymbol{X}) S(\boldsymbol{\theta}^*_-(\boldsymbol{X})'\boldsymbol{T}))^2] \\
&= \ \left(1 + \frac{1}{\tilde{c}_1}\right) \cdot \mathbb{E}\left[\left(G(\boldsymbol{\theta}(\boldsymbol{X}), \boldsymbol{T}) - G(\boldsymbol{\theta}^*(\boldsymbol{X}), \boldsymbol{T})\right)^2\right],
\end{aligned} \tag{23}$$

where the second inequality follows from (21). Putting the inequalities (21), (22), and (23) together, we conclude that

$$c_1\mathbb{E}\Big[\|\boldsymbol{\theta}(\boldsymbol{X})-\boldsymbol{\theta}^*(\boldsymbol{X})\|_2^2\Big]\le\mathbb{E}\Big[\Big(G(\boldsymbol{\theta}(\boldsymbol{X}),\boldsymbol{T}))-G(\boldsymbol{\theta}^*(\boldsymbol{X}),\boldsymbol{T}))\Big)^2\Big],$$

where $c_1=\tilde{c}_1\tilde{c}_2\tilde{c}_3/(1+\tilde{c}_1+\tilde{c}_2\tilde{c}_3)$, which concludes the proof.

□

Generalized Sigmoid Form II. We consider

$$G(\boldsymbol{\theta}(\boldsymbol{x}),\boldsymbol{t}):=\frac{\theta_{m+1}(\boldsymbol{x})}{1+\exp\Big(-(\theta_0(\boldsymbol{x})+\theta_1(\boldsymbol{x})t_1+\cdots+\theta_m(\boldsymbol{x})t_m)\Big)}. \tag{24}$$

Recall that $\boldsymbol{t}(S)$ represents the treatment assignment vector such that $t_i(S)=1$ if and only if $i\in S$, where $S\subset\{1,2,\ldots,m\}$. We restate the following sufficient conditions of Assumption 4(d): Uniformly in $\boldsymbol{X}$,

- the matrix $\mathbb{E}[\tilde{\boldsymbol{T}}\tilde{\boldsymbol{T}}'|\boldsymbol{X}]\succ 0$, where $\tilde{\boldsymbol{T}}'=(1,\boldsymbol{T}')$,
- $|\theta^*_{m+1}(\boldsymbol{X})|>0$,
- and there exists a triplet $i\in\{1,\ldots,m\}$, S_1, $S_2\subset\{1,2,\ldots,m\}$ such that $i\notin S_1$, $i\notin S_2$, $S_1\neq S_2$, and there exists a constant $\tilde{c}_1$,$\tilde{c}_2$ such that

$$\nu(\boldsymbol{t}(S_1\cup\{i\})|\boldsymbol{X})\cdot\nu(\boldsymbol{t}(S_1)|\boldsymbol{X})\cdot\nu(\boldsymbol{t}(S_2\cup\{i\})|\boldsymbol{X})\cdot\nu(\boldsymbol{t}(S_2)|\boldsymbol{X})\ge\tilde{c}_1>0,$$

and

$$|G(\boldsymbol{\theta}^*(\boldsymbol{X}),\boldsymbol{t}(S_1\cup\{i\}))G(\boldsymbol{\theta}^*(\boldsymbol{X}),\boldsymbol{t}(S_2))-G(\boldsymbol{\theta}^*(\boldsymbol{X}),\boldsymbol{t}(S_2\cup\{i\}))G(\boldsymbol{\theta}^*(\boldsymbol{X}),\boldsymbol{t}(S_1))|\ge\tilde{c}_2>0. \tag{25}$$

We show that these conditions are sufficient to guarantee the identification and sufficient curvature near the ground truth, i.e., Proposition 1 holds for the *Generalized Sigmoid Form II*.

Proof. To flesh out the analysis, given each $\boldsymbol{\theta}(\cdot)$ and $\boldsymbol{X}$, let us define

$$g_1(\boldsymbol{\theta},\boldsymbol{X})=G(\boldsymbol{\theta}(\boldsymbol{X}),\boldsymbol{t}(S_1)),\ g_2(\boldsymbol{\theta},\boldsymbol{X})=G(\boldsymbol{\theta}(\boldsymbol{X}),\boldsymbol{t}(S_2)),\ g_3(\boldsymbol{\theta},\boldsymbol{X})=G(\boldsymbol{\theta}(\boldsymbol{X}),\boldsymbol{t}(S_1\cup\{i\})),$$

$$g_4(\boldsymbol{\theta},\boldsymbol{X})=G(\boldsymbol{\theta}(\boldsymbol{X}),\boldsymbol{t}(S_2\cup\{i\})),\ \text{and}\ \ \boldsymbol{g}(\boldsymbol{\theta},\boldsymbol{X})=(g_1(\boldsymbol{\theta},\boldsymbol{X}),g_2(\boldsymbol{\theta},\boldsymbol{X}),g_3(\boldsymbol{\theta},\boldsymbol{X}),g_4(\boldsymbol{\theta},\boldsymbol{X}))'.$$

Then, note that for any fixed $\boldsymbol{t}$

$$\frac{\theta_{m+1}(\boldsymbol{X})}{G(\boldsymbol{\theta}(\boldsymbol{X}),\boldsymbol{t})}-1=\exp\Big(-(\theta_0(\boldsymbol{X})+\theta_1(\boldsymbol{X})t_1+\cdots+\theta_m(\boldsymbol{X})t_m)\Big).$$

As a result, we have

$$\frac{\dfrac{\theta_{m+1}(\boldsymbol{X})}{G(\boldsymbol{\theta}(\boldsymbol{X}),\boldsymbol{t}(S_1\cup\{i\}))}-1}{\dfrac{\theta_{m+1}(\boldsymbol{X})}{G(\boldsymbol{\theta}(\boldsymbol{X}),\boldsymbol{t}(S_1))}-1}=\frac{\dfrac{\theta_{m+1}(\boldsymbol{X})}{G(\boldsymbol{\theta}(\boldsymbol{X}),\boldsymbol{t}(S_2\cup\{i\}))}-1}{\dfrac{\theta_{m+1}(\boldsymbol{X})}{G(\boldsymbol{\theta}(\boldsymbol{X}),\boldsymbol{t}(S_2))}-1}=\exp(-\theta_i(\boldsymbol{X})),$$

which implies that

$$\theta_{m+1}(\boldsymbol{X})\cdot\Bigg[\left(\frac{1}{g_3(\boldsymbol{\theta},\boldsymbol{X})g_2(\boldsymbol{\theta},\boldsymbol{X})}-\frac{1}{g_4(\boldsymbol{\theta},\boldsymbol{X})g_1(\boldsymbol{\theta},\boldsymbol{X})}\right)\cdot\theta_{m+1}(\boldsymbol{X})$$
$$-\frac{1}{g_3(\boldsymbol{\theta},\boldsymbol{X})}-\frac{1}{g_2(\boldsymbol{\theta},\boldsymbol{X})}+\frac{1}{g_4(\boldsymbol{\theta},\boldsymbol{X})}+\frac{1}{g_1(\boldsymbol{\theta},\boldsymbol{X})}\Bigg]=0.$$

By assumption, $\theta_{m+1}(\boldsymbol{X}) \neq 0$, it then holds that

$$\begin{aligned}&\left(g_4(\boldsymbol{\theta},\boldsymbol{X})g_1(\boldsymbol{\theta},\boldsymbol{X}) - g_3(\boldsymbol{\theta},\boldsymbol{X})g_2(\boldsymbol{\theta},\boldsymbol{X})\right)\theta_{m+1}(\boldsymbol{X})\\ =\ & g_1(\boldsymbol{\theta},\boldsymbol{X})g_2(\boldsymbol{\theta},\boldsymbol{X})g_4(\boldsymbol{\theta},\boldsymbol{X}) + g_1(\boldsymbol{\theta},\boldsymbol{X})g_3(\boldsymbol{\theta},\boldsymbol{X})g_4(\boldsymbol{\theta},\boldsymbol{X}) - g_1(\boldsymbol{\theta},\boldsymbol{X})g_2(\boldsymbol{\theta},\boldsymbol{X})g_3(\boldsymbol{\theta},\boldsymbol{X}) - g_2(\boldsymbol{\theta},\boldsymbol{X})g_3(\boldsymbol{\theta},\boldsymbol{X})g_4(\boldsymbol{\theta},\boldsymbol{X}).\end{aligned} \tag{26}$$

Assume that there exist $\tilde{\boldsymbol{\theta}}(\boldsymbol{X})$ such that

$$G(\tilde{\boldsymbol{\theta}}(\boldsymbol{X}),\boldsymbol{T}) = G(\boldsymbol{\theta}^*(\boldsymbol{X}),\boldsymbol{T}),$$

with the assignment mechanism $\nu(\cdot\,|\,\cdot)$. Under our assignment mechanism assumption that

$$\mathbb{P}\Big[\boldsymbol{T}=\boldsymbol{t}(S_1\cup\{i\})\Big|\boldsymbol{X}\Big]\cdot\mathbb{P}\Big[\boldsymbol{T}=\boldsymbol{t}(S_1)\Big|\boldsymbol{X}\Big]\cdot\mathbb{P}\Big[\boldsymbol{T}=\boldsymbol{t}(S_2\cup\{i\})\Big|\boldsymbol{X}\Big]\cdot\mathbb{P}\Big[\boldsymbol{T}=\boldsymbol{t}(S_2)\Big|\boldsymbol{X}\Big] > \tilde{c}_1 > 0, \tag{27}$$

for some $\tilde{c}_1 > 0$, we have that $G(\tilde{\boldsymbol{\theta}}(\boldsymbol{X}),\boldsymbol{t}) = G(\boldsymbol{\theta}^*(\boldsymbol{X}),\boldsymbol{t})$ for $\boldsymbol{t} = \boldsymbol{t}(S_1\cup\{i\})$, $\boldsymbol{t}(S_1)$, $\boldsymbol{t}(S_2\cup\{i\})$, $\boldsymbol{t}(S_2)$, i.e., $\boldsymbol{g}(\boldsymbol{\theta},\boldsymbol{X}) = \boldsymbol{g}(\boldsymbol{\theta}^*,\boldsymbol{X})$. And clearly, $\tilde{\theta}_{m+1}(\boldsymbol{X}) \neq 0$, because $\tilde{\theta}^*_{m+1}(\boldsymbol{X}) \neq 0$. Then we have $\tilde{\theta}_{m+1}(\boldsymbol{X}) = \theta_{m+1}(\boldsymbol{X})$, which is implied by (26). The rest of the proof of identifiability is the same as the *Generalized Sigmoid Form I*.

Next, we move on to the first inequality of the sufficient curvature condition. Due to Assumption 3 and the bounded output of the neural network, there must exists a constant $\tilde{c}_3$ such that $G(\boldsymbol{\theta}(\boldsymbol{x}),\boldsymbol{t}) \leq \tilde{c}_3$ for some $\tilde{c}_3 > 0$ for all $\boldsymbol{\theta}(\cdot) \in \mathcal{F}_{\text{DNN}}$, $\boldsymbol{X}$ and $\boldsymbol{t}$. Let

$$\mathcal{E} = \Big\{\boldsymbol{X}:\ |g_\ell(\boldsymbol{\theta},\boldsymbol{X}) - g_\ell(\boldsymbol{\theta}^*,\boldsymbol{X})| \leq \epsilon\ \forall \ell = 1,\ldots,4,\ \text{ and } \theta_{m+1}(\boldsymbol{X}) \neq 0\Big\}$$

where $\epsilon = \tilde{c}_2/(8\tilde{c}_3)$. Let us assume from now on that $\boldsymbol{X} \in \mathcal{E}$. Define function $h(\boldsymbol{g}) = g_4g_1 - g_3g_2$. By assumption, we have $|h(\boldsymbol{g}(\boldsymbol{\theta}^*,\boldsymbol{X}))| \geq \tilde{c}_2$. Then, by mean value theorem, it must be that

$$h\left(\boldsymbol{g}(\boldsymbol{\theta},\boldsymbol{X})\right) = h\left(\boldsymbol{g}(\boldsymbol{\theta}^*,\boldsymbol{X})\right) + \nabla h(\hat{\boldsymbol{g}}(\boldsymbol{X}))'\Big(\boldsymbol{g}(\boldsymbol{\theta},\boldsymbol{X}) - \boldsymbol{g}(\boldsymbol{\theta}^*,\boldsymbol{X})\Big) \geq h\left(\boldsymbol{g}(\boldsymbol{\theta}^*,\boldsymbol{X})\right) - 4\tilde{c}_3\epsilon \geq h\left(\boldsymbol{g}(\boldsymbol{\theta}^*,\boldsymbol{X})\right)/2 \geq \tilde{c}_2/2 > 0, \tag{28}$$

where $\hat{\boldsymbol{g}}(\boldsymbol{X})$ is such that $\hat{g}_\ell(\boldsymbol{X}) \in [g_\ell(\boldsymbol{\theta}^*,\boldsymbol{X}), g_\ell(\boldsymbol{\theta},\boldsymbol{X})]$ for $\ell = 1,\ldots,4$ and the first inequality holds because $|\hat{g}_\ell(\boldsymbol{X}) - g_\ell(\boldsymbol{\theta}^*,\boldsymbol{X})| \leq \epsilon$ for all $\ell = 1,\ldots,4$ and the second inequality holds due to the definition of ϵ. Therefore, if we define

$$f(\boldsymbol{g}) = (g_1g_2g_4 + g_1g_3g_4 - g_1g_2g_3 - g_1g_3g_4)/h(\boldsymbol{g}),$$

then given (26) it holds that $\theta_{m+1}(\boldsymbol{X}) = f(\boldsymbol{g}(\boldsymbol{\theta},\boldsymbol{X}))$, which is well-defined since we have already shown that $h(\boldsymbol{g}(\boldsymbol{\theta},\boldsymbol{X})) > 0$. Similarly, we have $\theta^*_{m+1}(\boldsymbol{X}) = f(\boldsymbol{g}(\boldsymbol{\theta}^*,\boldsymbol{X}))$. Then, we have

$$\left(\boldsymbol{\theta}_{m+1}(\boldsymbol{X}) - \boldsymbol{\theta}^*_{m+1}(\boldsymbol{X})\right)^2 = \left[\nabla f(\tilde{\boldsymbol{g}}(\boldsymbol{X}))'(\boldsymbol{g}(\boldsymbol{\theta},\boldsymbol{X}) - \boldsymbol{g}(\boldsymbol{\theta}^*,\boldsymbol{X}))\right]^2, \tag{29}$$

by the mean value theorem, where $\tilde{g}_\ell(\boldsymbol{X}) \in [g_\ell(\boldsymbol{\theta}^*,\boldsymbol{X}), g_\ell(\boldsymbol{\theta},\boldsymbol{X})]$ for all $\ell = 1,\ldots,4$. Note that every term in $\nabla f(\tilde{\boldsymbol{g}}(\boldsymbol{X}))$ is of the form $\mathsf{poly}(\tilde{\boldsymbol{g}}(\boldsymbol{X}))/h^2(\tilde{\boldsymbol{g}}(\boldsymbol{X}))$, where $\mathsf{poly}(\tilde{\boldsymbol{g}}(\boldsymbol{X}))$ is a polynomial of $\tilde{\boldsymbol{g}}(\boldsymbol{X})$ and uniformly bounded by assumption. By exactly the same argument to (28), we can still show that $h(\tilde{\boldsymbol{g}}(\boldsymbol{X})) > \tilde{c}_2/2$. Therefore, expanding the quadratic term on the right-hand side of (29), we obtain that for some constant $\tilde{c}_3$,

$$\left(\boldsymbol{\theta}_{m+1}(\boldsymbol{X}) - \boldsymbol{\theta}^*_{m+1}(\boldsymbol{X})\right)^2 \leq \tilde{c}_4 \cdot \left\|\boldsymbol{g}(\boldsymbol{\theta},\boldsymbol{X}) - \boldsymbol{g}(\boldsymbol{\theta}^*,\boldsymbol{X}))\right\|^2. \tag{30}$$

If $\boldsymbol{X} \in \mathcal{E}_1 = \left\{\boldsymbol{X} : \ \theta_{m+1}(\boldsymbol{X}) = 0\right\}$, it must be that $g(\boldsymbol{\theta}, \boldsymbol{X}) = 0$ by definition. Also, by the uniform boundedness assumption, there exists constant $\tilde{c}_5$ such that $g_\ell(\boldsymbol{\theta}^*, \boldsymbol{X}) \geq \tilde{c}_5$. Therefore, it must be that there is constant $\tilde{c}_6$ such that

$$\left(\boldsymbol{\theta}_{m+1}(\boldsymbol{X}) - \boldsymbol{\theta}^*_{m+1}(\boldsymbol{X})\right)^2 = \boldsymbol{\theta}^*_{m+1}(\boldsymbol{X})^2 \leq \tilde{c}_6 \cdot \|\boldsymbol{g}(\boldsymbol{\theta}^*, \boldsymbol{X}))\|^2 = \tilde{c}_6 \cdot ||\boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{X}) - \boldsymbol{g}(\boldsymbol{\theta}^*, \boldsymbol{X})\|^2 .$$

If $\boldsymbol{X} \notin \mathcal{E} \cup \mathcal{E}_1$, it must be that there is $\ell \in \{1,2,3,4\}$, such that $|g_\ell(\boldsymbol{\theta}, X) - g_\ell(\boldsymbol{\theta}^*, X)| > \epsilon$. This clearly implies that for some $\tilde{c}_7 > 0$,

$$\left(\boldsymbol{\theta}_{m+1}(\boldsymbol{X}) - \boldsymbol{\theta}^*_{m+1}(\boldsymbol{X})\right)^2 \leq \tilde{c}_7 \cdot \|\boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{X}) - \boldsymbol{g}(\boldsymbol{\theta}^*, \boldsymbol{X})\|^2 .$$

by the boundedness of $\boldsymbol{\theta}_{m+1}(\boldsymbol{X})$ and $\boldsymbol{\theta}^*_{m+1}(\boldsymbol{X})$. Combined, we have that

$$\mathbb{E}_{\boldsymbol{X}}\left[\left(\boldsymbol{\theta}_{m+1}(\boldsymbol{X}) - \boldsymbol{\theta}^*_{m+1}(\boldsymbol{X})\right)^2\right] \leq \max\{\tilde{c}_4, \tilde{c}_6, \tilde{c}_7\} \cdot \mathbb{E}_{\boldsymbol{X}}\left[\|\boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{X}) - \boldsymbol{g}(\boldsymbol{\theta}^*, \boldsymbol{X})\|^2\right].$$

Note that (27) implies there exists a constant $\tilde{c}_7 > 0$ such that,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{X},\boldsymbol{T}}\left[(G(\boldsymbol{\theta}(\boldsymbol{X}), \boldsymbol{T}) - G(\boldsymbol{\theta}^*(\boldsymbol{X}), \boldsymbol{T}))^2\right] &\geq \tilde{c}_8 \mathbb{E}_{\boldsymbol{X}}\left[\left(G(\boldsymbol{\theta}(\boldsymbol{X}), \boldsymbol{t}(S_1 \cup \{i\})) - G(\boldsymbol{\theta}^*(\boldsymbol{X}), \boldsymbol{t}(S_1 \cup \{i\}))\right)^2\right] \\
\mathbb{E}_{\boldsymbol{X},\boldsymbol{T}}\left[(G(\boldsymbol{\theta}(\boldsymbol{X}), \boldsymbol{T}) - G(\boldsymbol{\theta}^*(\boldsymbol{X}), \boldsymbol{T}))^2\right] &\geq \tilde{c}_8 \mathbb{E}_{\boldsymbol{X}}\left[\left(G(\boldsymbol{\theta}(\boldsymbol{X}), \boldsymbol{t}(S_1)) - G(\boldsymbol{\theta}^*(\boldsymbol{X}), \boldsymbol{t}(S_1))\right)^2\right] \\
\mathbb{E}_{\boldsymbol{X},\boldsymbol{T}}\left[(G(\boldsymbol{\theta}(\boldsymbol{X}), \boldsymbol{T}) - G(\boldsymbol{\theta}^*(\boldsymbol{X}), \boldsymbol{T}))^2\right] &\geq \tilde{c}_8 \mathbb{E}_{\boldsymbol{X}}\left[\left(G(\boldsymbol{\theta}(\boldsymbol{X}), \boldsymbol{t}(S_2 \cup \{i\})) - G(\boldsymbol{\theta}^*(\boldsymbol{X}), \boldsymbol{t}(S_2 \cup \{i\}))\right)^2\right] \\
\mathbb{E}_{\boldsymbol{X},\boldsymbol{T}}\left[(G(\boldsymbol{\theta}(\boldsymbol{X}), \boldsymbol{T}) - G(\boldsymbol{\theta}^*(\boldsymbol{X}), \boldsymbol{T}))^2\right] &\geq \tilde{c}_8 \mathbb{E}_{\boldsymbol{X}}\left[\left(G(\boldsymbol{\theta}(\boldsymbol{X}), \boldsymbol{t}(S_2)) - G(\boldsymbol{\theta}^*(\boldsymbol{X}), \boldsymbol{t}(S_2))\right)^2\right].
\end{aligned} \tag{31}$$

Then,

$$\mathbb{E}_{\boldsymbol{X},\boldsymbol{T}}\left[(G(\boldsymbol{\theta}(\boldsymbol{X}), \boldsymbol{T}) - G(\boldsymbol{\theta}^*(\boldsymbol{X}), \boldsymbol{T}))^2\right] \geq \frac{\tilde{c}_8}{4} \cdot \frac{1}{\max\{\tilde{c}_4, \tilde{c}_6, \tilde{c}_7\}} \cdot \mathbb{E}_{\boldsymbol{X}}\left[\left(\boldsymbol{\theta}_{m+1}(\boldsymbol{X}) - \boldsymbol{\theta}^*_{m+1}(\boldsymbol{X})\right)^2\right]. \tag{32}$$

The rest of the proof is the same as that for the *Generalized Sigmoid Form I.* □

A.5. Proof of Proposition 2

Proof. The results directly follow from Theorem 2 of Farrell et al. (2020), which we also restate as Lemma 1 in Appendix A.3.

First, we derive the first-order derivative $\ell_{\boldsymbol{\theta}}$ and the Hessian matrix $\ell_{\boldsymbol{\theta\theta}}$. With the loss function $\ell(y, \boldsymbol{t}, \boldsymbol{\theta}(\boldsymbol{x})) = (y - G(\boldsymbol{\theta}(\boldsymbol{x}), \boldsymbol{t}))^2$, we can compute to get

$\ell_{\boldsymbol{\theta}} = (\partial\ell/\partial\theta_1, \ldots, \partial\ell/\partial\theta_{d_{\boldsymbol{\theta}}})' = 2(G - y)G_{\boldsymbol{\theta}}$,

$\ell_{\boldsymbol{\theta\theta}} = 2G_{\boldsymbol{\theta}}G'_{\boldsymbol{\theta}} + 2G_{\boldsymbol{\theta\theta}}(G - y)$,

$\boldsymbol{\Lambda}(\boldsymbol{x}) = \mathbb{E}[\ell_{\boldsymbol{\theta\theta}}|\boldsymbol{X} = \boldsymbol{x}] = 2\mathbb{E}_{\boldsymbol{T}|\boldsymbol{x}}[G_{\boldsymbol{\theta}}G'_{\boldsymbol{\theta}}|\boldsymbol{x}] + 2\mathbb{E}_{\boldsymbol{T}|\boldsymbol{x}}\left[G_{\boldsymbol{\theta\theta}}\mathbb{E}_{y|\boldsymbol{t},\boldsymbol{x}}[(G - y)|\boldsymbol{T} = \boldsymbol{t}, \boldsymbol{X} = \boldsymbol{x}]|\boldsymbol{X} = \boldsymbol{x}\right] = 2\mathbb{E}[G_{\boldsymbol{\theta}}G'_{\boldsymbol{\theta}}|\boldsymbol{X} = \boldsymbol{x}]$.

We now verify the assumptions of Lemma 1. The first condition, identification of $\boldsymbol{\theta}^*$, is satisfied under Assumptions 1 and 4, which is shown in Proposition 1(a). The second condition holds by Assumption 2(i) (i.e., DGP (1) holds) and the mean squared loss function. The third condition follows from Assumption 2(ii). The fourth condition regarding the pathwise differentiable follows from Assumption 2(iii), and the thrice continuously differentiability holds by the form of G defined in Assumption 1 and the mean squared loss with sufficient smoothness in $\boldsymbol{\theta}$. The fourth condition holds because the link functions G in Assumption 1 are sufficiently smooth together with the boundedness of $\boldsymbol{\theta}^*(\boldsymbol{X})$ given by Assumption 3. Thus, Proposition 2 follows immediately from Lemma 1. □

A.6. Verification of Assumption 2(ii)

In the following, we illustrate that, if the treatment assignment mechanism $\nu(\cdot\,|\,\cdot)$ satisfies Assumption 4, Assumption 2(ii) (i.e., $\mathbf{\Lambda}(\boldsymbol{x}) \succ 0$) is easy to satisfy and can be translated into very lenient conditions. $\mathbf{\Lambda}(\boldsymbol{x}) \succ 0$ is equivalent to that the vectors $\{G_{\boldsymbol{\theta}}(\boldsymbol{\theta}(\boldsymbol{x}), \boldsymbol{t})\}_{\boldsymbol{t}}$ are nondegenerate. Without loss of generality, we take the *Generalized Sigmoid Form II* for illustration. The other forms of link function can be verified similarly.

By definition, we have $\mathbf{\Lambda}(\boldsymbol{x}) = 2\mathbb{E}[G_{\boldsymbol{\theta}}(\boldsymbol{\theta}(\boldsymbol{x}), \boldsymbol{T}) G_{\boldsymbol{\theta}}(\boldsymbol{\theta}(\boldsymbol{x}), \boldsymbol{T})' | \boldsymbol{X} = \boldsymbol{x}]$, where

$$G_{\boldsymbol{\theta}}(\boldsymbol{\theta}(\boldsymbol{x}), \boldsymbol{t})' = \Big(\frac{\theta_{m+1}(\boldsymbol{x}) \exp(-(\theta_0(\boldsymbol{x}) + \cdots + \theta_m(\boldsymbol{x}) t_m))}{(1 + \exp(-(\theta_0(\boldsymbol{x}) + \cdots + \theta_m(\boldsymbol{x}) t_m)))^2} (1, t_1, \ldots, t_m),\ \frac{1}{1 + \exp(-(\theta_0(\boldsymbol{x}) + \cdots + \theta_m(\boldsymbol{x}) t_m))} \Big).$$

Thus, to verify that $\mathbf{\Lambda}(\boldsymbol{x})$ is invertible, it suffices to show that the matrix constructed by vectors $\{G_{\boldsymbol{\theta}}(\boldsymbol{\theta}(\boldsymbol{x}), \boldsymbol{t})\}_{\boldsymbol{t}}$ has full rank $m+2$. For ease of exposition, we drop the dependence on $\boldsymbol{x}$, which will not cause any confusion.

Next, we verify that $\mathbf{\Lambda}(\boldsymbol{x}) \succ 0$ for the treatment assignment mechanism in Assumption 4(b-iv). All other assignment rules in Assumption 4(b) can be translated into similar lenient conditions, so we omit their verification for brevity. Specifically, to ensure that $\mathbb{E}[\tilde{\boldsymbol{T}}\tilde{\boldsymbol{T}}' | \boldsymbol{X} = \boldsymbol{x}] \succ 0$, we consider the assignment mechanism $\nu(\boldsymbol{t}(\emptyset)|\boldsymbol{x}) > 0,\ \nu(\boldsymbol{t}(\{1\})|\boldsymbol{x}) > 0,\ \nu(\boldsymbol{t}(\{2\})|\boldsymbol{x}) > 0, \ldots,\ \nu(\boldsymbol{t}(\{m\})|\boldsymbol{x}) > 0$, together with the overlapping assignment $\nu(\boldsymbol{t}(S_1)|\boldsymbol{x}) > 0,\ nu(\boldsymbol{t}(S_1 \cup \{i\})|\boldsymbol{x}) > 0$, $\nu(\boldsymbol{t}(S_2)|\boldsymbol{x}) > 0$, and $\nu(\boldsymbol{t}(S_2 \cup \{i\})|\boldsymbol{x}) > 0$. Define $e_0 := \exp(-\theta_0)$, $e_i := \exp(-\theta_0 - \theta_i)$, for $i = 1, 2, \ldots, m$, and $e_{m+1} := \exp(-\boldsymbol{\theta}' \boldsymbol{t}(S_1 \cup \{i\}))$. We have

$$rank(\mathbf{\Lambda}) \overset{(a)}{\geq} rank \begin{pmatrix} \frac{\theta_{m+1} e_0}{1+e_0}(1,0,0,0,\ldots,0) & \frac{1}{1+e_0} \\ \frac{\theta_{m+1} e_1}{1+e_1}(1,1,0,0,\ldots,0) & \frac{1}{1+e_1} \\ \frac{\theta_{m+1} e_2}{1+e_2}(1,0,1,0,\ldots,0) & \frac{1}{1+e_2} \\ \frac{\theta_{m+1} e_3}{1+e_3}(1,0,0,1,\ldots,0) & \frac{1}{1+e_3} \\ \vdots & \vdots \\ \frac{\theta_{m+1} e_m}{1+e_m}(1,0,0,0,\ldots,1) & \frac{1}{1+e_m} \\ \frac{\theta_{m+1} e_{m+1}}{1+e_{m+1}}(1, \boldsymbol{t}(S_1 \cup \{i\})') & \frac{1}{1+e_{m+1}} \end{pmatrix} \overset{(b)}{=} rank \begin{pmatrix} \frac{e_0}{1+e_0}(1,0,0,0,\ldots,0) & \frac{1}{1+e_0} \\ \frac{e_1}{1+e_1}(1,1,0,0,\ldots,0) & \frac{1}{1+e_1} \\ \frac{e_2}{1+e_2}(1,0,1,0,\ldots,0) & \frac{1}{1+e_2} \\ \frac{e_3}{1+e_3}(1,0,0,1,\ldots,0) & \frac{1}{1+e_3} \\ \vdots & \vdots \\ \frac{e_m}{1+e_m}(1,0,0,0,\ldots,1) & \frac{1}{1+e_m} \\ \frac{e_{m+1}}{1+e_{m+1}}(1, \boldsymbol{t}(S_1 \cup \{i\})') & \frac{1}{1+e_{m+1}} \end{pmatrix}$$

$$\overset{(c)}{=} rank \begin{pmatrix} 1\ 0\ 0 & 0 & \ldots\ 0 & \frac{1}{e_0} \\ 1\ 1\ 0 & 0 & \ldots\ 0 & \frac{1}{e_1} \\ 1\ 0\ 1 & 0 & \ldots\ 0 & \frac{1}{e_2} \\ 1\ 0\ 0 & 1 & \ldots\ 0 & \frac{1}{e_3} \\ \vdots & & & \vdots\ \ \vdots \\ 1\ 0\ 0 & 0 & \ldots\ 1 & \frac{1}{e_m} \\ 1 & \boldsymbol{t}(S_1 \cup \{i\})' & & \frac{1}{e_{m+1}} \end{pmatrix} \overset{(d)}{=} rank \begin{pmatrix} 1\ 0\ 0\ 0\ \ldots\ 0 & \frac{1}{e_0} \\ 0\ 1\ 0\ 0\ \ldots\ 0 & \frac{1}{e_1} - \frac{1}{e_0} \\ 0\ 0\ 1\ 0\ \ldots\ 0 & \frac{1}{e_2} - \frac{1}{e_0} \\ 0\ 0\ 0\ 1\ \ldots\ 0 & \frac{1}{e_3} - \frac{1}{e_0} \\ \vdots & \vdots \\ 0\ 0\ 0\ 0\ \ldots\ 1 & \frac{1}{e_m} - \frac{1}{e_0} \\ 0\ 0\ 0\ 0\ \ldots\ 0 & \frac{1}{e_{m+1}} - \frac{1}{e_0} - (\frac{1}{e_1} - \frac{1}{e_0}) - (\frac{1}{e_i} - \frac{1}{e_0}) \end{pmatrix}.$$

The inequality (a) follows from the fact the right-hand-side matrix is expanded only by a partial of vectors $\{G_{\boldsymbol{\theta}}(\boldsymbol{\theta}(\boldsymbol{x}), \boldsymbol{t})\}_{\boldsymbol{t}}$. The equality (b) follows from $\theta_{m+1} \neq 0$. The equality (c) follows from $1 + e_i \neq 0$ and $e_i \neq 0$ for all $i = 1, 2, \ldots, m$. The equality (d) follows from some subtraction operations by rows. To guarantee the full rank condition that $rank(\mathbf{\Lambda}) = m + 2$, a sufficient condition is $\frac{1}{e_{m+1}} - \frac{1}{e_1} - \frac{1}{e_i} + \frac{1}{e_0} \neq 0$, i.e., the bottom-right entry of the last matrix is nonzero. This is indeed a very weak condition.

For other assignment mechanisms in Assumption 4(b), one can also translate the invertibility of $\mathbf{\Lambda}$ into very lenient conditions. We omit the details for brevity.

A.7. ATE Estimator Construction via Cross-Fitting

In the following, we introduce the sample-splitting/cross-fitting and estimation procedure from Chernozhukov et al. (2018). The entire DeDL framework is given by Algorithm 1. In our setting with the known distribution of $\boldsymbol{T}$, we require only the two-way splitting, but for unknown $\boldsymbol{T}$ distribution, three-way splitting is required, with an additional portion of samples used for obtaining $\hat{\boldsymbol{\Lambda}}(\boldsymbol{x})$.

One can split the data samples $\{1,2,\ldots,n\}$ into S nonoverlapping copies $\mathcal{S}_s \subset \{1,2,\ldots,n\},\ s=1,2\ldots,S$ with the cardinality $|\mathcal{S}_s|$ being proportionally to the sample size n, and let $\mathcal{S}_s^c$ be the complement of $\mathcal{S}_s$. First, we use $\mathcal{S}^c$ to get estimators $\hat{\boldsymbol{\theta}}_s(\cdot)$ of parameters $\boldsymbol{\theta}^*(\cdot)$, and compute $\hat{\boldsymbol{\Lambda}}_s(\cdot)$ given the estimators $\hat{\boldsymbol{\theta}}_s(\cdot)$ and distribution of $\boldsymbol{T}$. Then, one can use the other samples to construct an estimator of $\mu(\boldsymbol{t})$, for any $\boldsymbol{t}\in\{0,1\}^m$ as

$$\hat{\mu}_{\mathsf{DeDL}}(\boldsymbol{t}) = \frac{1}{S}\hat{\mu}_s(\boldsymbol{t}), \qquad \hat{\mu}_s(\boldsymbol{t}) = \frac{1}{|\mathcal{S}_s|}\sum_{i\in\mathcal{S}_s}\psi(\boldsymbol{z}_i,\hat{\boldsymbol{\theta}}_s(\boldsymbol{x}_i),\hat{\boldsymbol{\Lambda}}_s(\boldsymbol{x}_i);\boldsymbol{t},\boldsymbol{t}_0). \tag{33}$$

Similarly, the variance estimator can be constructed as

$$\hat{\Psi}_{\mathsf{DeDL}}(\boldsymbol{t};\mu) = \frac{1}{S}\hat{\Psi}_s(\boldsymbol{t}), \qquad \hat{\Psi}_s(\boldsymbol{t}) = \frac{1}{|\mathcal{S}_s|}\sum_{i\in\mathcal{S}_s}\Big(\psi(\boldsymbol{z}_i,\hat{\boldsymbol{\theta}}_s(\boldsymbol{x}_i),\hat{\boldsymbol{\Lambda}}_s(\boldsymbol{x}_i);\boldsymbol{t},\boldsymbol{t}_0)-\hat{\mu}_{\mathsf{DeDL}}(\boldsymbol{t})\Big)^2. \tag{34}$$

The asymptotic normality of $\hat{\mu}_{\mathsf{DeDL}}(\boldsymbol{t})$ in Theorem 2(a) directly follows from Chernozhukov et al. (2018), and a detailed proof can also be found in Farrell et al. (2020).

Therefore, the $(1-\alpha)$-confidence interval of $\hat{\mu}_{\mathsf{DeDL}}(\boldsymbol{t})$ is given by

$$\widehat{\mathcal{CI}}_{\mathsf{DeDL}}(\boldsymbol{t};\mu) = \left[\hat{\mu}_{\mathsf{DeDL}}(\boldsymbol{t}) - \frac{1}{\sqrt{n}}\cdot\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\cdot\hat{\Psi}_{\mathsf{DeDL}}(\boldsymbol{t};\mu), \hat{\mu}_{\mathsf{DeDL}}(\boldsymbol{t}) + \frac{1}{\sqrt{n}}\cdot\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\cdot\hat{\Psi}_{\mathsf{DeDL}}(\boldsymbol{t};\mu)\right], \tag{35}$$

where $\Phi^{-1}(\cdot)$ is the inverse cumulative distribution function of a standard normal random variable.

A.8. Estimation and Inference for Best Treatment Identification

Details for the best treatment identification. After obtaining the asymptotically normal estimators of ATE $\hat{\mu}(\boldsymbol{t})$ for all experiment combinations $\boldsymbol{t}\in\{0,1\}^m$, the next step is identifying the best experiment combination, which is defined as $\boldsymbol{t}^* := \arg\max_{\boldsymbol{t}\in\{0,1\}^m}\mu(\boldsymbol{t})$.

Following the common practice, one can search for the best treatment combination by searching for the highest ATE estimation. Formally, we define the empirical best treatment level as $\hat{\boldsymbol{t}}^* := \arg\max_{\boldsymbol{t}\in\{0,1\}^m}\hat{\mu}_{\mathsf{DeDL}}(\boldsymbol{t})$. The remaining job is verifying whether $\hat{\boldsymbol{t}}^*$ is the best treatment level with significant improvements over all other treatment levels, i.e., we test the one-sided hypothesis on whether the ATE of $\hat{\boldsymbol{t}}^*$ is significantly better than other treatments:

$$H_0 : \tau(\boldsymbol{t}) > 0, \text{for all } \boldsymbol{t}\in\{0,1\}^m\backslash\{\hat{\boldsymbol{t}}^*\},$$

where

$$\tau(\boldsymbol{t}) := \mu(\hat{\boldsymbol{t}}^*) - \mu(\boldsymbol{t}) = \mathbb{E}[G(\boldsymbol{\theta}^*(\boldsymbol{X}),\hat{\boldsymbol{t}}^*)] - \mathbb{E}[G(\boldsymbol{\theta}^*(\boldsymbol{X}),\boldsymbol{t})],$$

is the improvement in ATE of the empirically optimal $\hat{\boldsymbol{t}}^*$ over treatment level $\boldsymbol{t}\in\{0,1\}^m$. Notice that the $\tau(\boldsymbol{t})$ can be rewritten as,

$$\tau(\boldsymbol{t}) = \mathbb{E}[H(\boldsymbol{x},\boldsymbol{\theta}^*(\boldsymbol{x});\hat{\boldsymbol{t}}^*,\boldsymbol{t})],$$

which is similar in structure to $\mu(\boldsymbol{t}) = \mathbb{E}[H(\boldsymbol{X}, \boldsymbol{\theta}^*(\boldsymbol{X}); \boldsymbol{t}, \boldsymbol{t}_0)]$ with a change of the inputs in the advantage function H.

Applying an influence function similar to (10) developed in Section 3.4, i.e., $\psi(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\Lambda}; \hat{\boldsymbol{t}}^*, \boldsymbol{t}) = H(\boldsymbol{x}, \boldsymbol{\theta}(\boldsymbol{x}); \hat{\boldsymbol{t}}^*, \boldsymbol{t}) - H_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{\theta}(\boldsymbol{x}); \hat{\boldsymbol{t}}^*, \boldsymbol{t})' \boldsymbol{\Lambda}(\boldsymbol{x})^{-1} \ell_{\boldsymbol{\theta}}(y, \check{\boldsymbol{t}}, \boldsymbol{\theta}(\boldsymbol{x}))$ and the two-way splitting procedure described above, we construct the estimators

$$\hat{\tau}_{\mathsf{DeDL}}(\boldsymbol{t}) := \hat{\mu}_{\mathsf{DeDL}}(\hat{\boldsymbol{t}}^*_{\mathsf{DeDL}}) - \hat{\mu}_{\mathsf{DeDL}}(\boldsymbol{t}), \tag{36}$$

and the variance estimate $\hat{\Psi}_{\mathsf{DeDL}}(\boldsymbol{t}; \tau)$ for $\tau_{\mathsf{DeDL}}(\boldsymbol{t})$.

Proof of Theorem 2

By Proposition 1(b), it is straightforward to verify that $\|\hat{\theta}_{sk} - \theta^*_k\|_{L_2(\boldsymbol{X})} = o(n^{-1/4})$ holds if $p > d_{\boldsymbol{X}}$. By Theorem 3 of Farrell et al. (2020), we have the desired asymptotic normality of the proposed estimators $\hat{\mu}_{\mathsf{DeDL}}(\boldsymbol{t})$ and $\hat{\tau}_{\mathsf{DeDL}}(\boldsymbol{t})$. This proves part (a). On the other hand, we also highlight that this DNN convergence rate is inherited from Lemma 1 (Theorem 2 in Farrell et al. 2020), which might not be tight as well. As establishing a tighter rate for DNN convergence is beyond the scope of our paper, we instead follow the standard $o(n^{-1/4})$ decay rate (as claimed in Proposition 1).

For Part (b), the empirically optimal treatment combination $\hat{\boldsymbol{t}}^* := \arg\max_{\boldsymbol{t} \in \{0,1\}^m} \hat{\mu}(\boldsymbol{t})$ depends on the samples used for the training and inference of ATE, which is a critical challenge in our proof below. For the rest of our proof, we drop the subscript DeDL to simplify the notation. First, we construct the following estimator $\hat{\zeta}(\boldsymbol{t}) := \hat{\mu}(\boldsymbol{t}^*) - \hat{\mu}(\boldsymbol{t})$, where $\hat{\mu}(\cdot)$ is given by (33), for the true advantage of the optimal treatment combination $\zeta(\boldsymbol{t}) := \mu(\boldsymbol{t}^*) - \mu(\boldsymbol{t})$. Similar to (34), one can construct the variance estimator $\hat{\Psi}(\boldsymbol{t}; \zeta)$ for $\hat{\zeta}(\boldsymbol{t})$. Importantly, $\hat{\zeta}(\boldsymbol{t})$ is a virtual estimator not available in practice, because the ground-truth optimal treatment combination $\boldsymbol{t}^*$ is unknown.

First, Theorem 3 in Farrell et al. (2020) establishes the asymptotic normality of $\hat{\zeta}(\boldsymbol{t})$:

$$\sqrt{n}\big(\hat{\Psi}(\boldsymbol{t}; \zeta)\big)^{-1/2}(\hat{\zeta}(\boldsymbol{t}) - \zeta(\boldsymbol{t})) = \sum_{i=1}^{n} \big(\hat{\Psi}(\boldsymbol{t}; \zeta)\big)^{-1/2}\big(\psi(\boldsymbol{z}_i, \boldsymbol{\theta}^*(\boldsymbol{x}_i), \boldsymbol{\Lambda}(\boldsymbol{x}_i), \boldsymbol{t}^*, \boldsymbol{t}) - \zeta(\boldsymbol{t})\big)/\sqrt{n} + o_p(1) \to_d \mathcal{N}(0, 1). \tag{37}$$

Next, we show that, with probability going to 1, the optimal treatment combination is correctly identified, i.e., $\lim_{n \to \infty} \mathbb{P}[\hat{\boldsymbol{t}}^* = \boldsymbol{t}^*] = 1$, when the dependence on sample size n is dropped for brevity. By Part (a), we have $\hat{\mu}(\boldsymbol{t}^*) \to_p \mu(\boldsymbol{t}^*)$ and $\hat{\mu}(\boldsymbol{t}) \to_p \mu(\boldsymbol{t})$ for any $\boldsymbol{t} \neq \boldsymbol{t}^*$. Because the max operator is continuous with respect to the L_1 norm, we have

$$|\hat{\mu}(\boldsymbol{t}^*) - \mu(\boldsymbol{t}^*)| \vee |\hat{\mu}(\boldsymbol{t}) - \mu(\boldsymbol{t})| \to_p 0,$$

where "$\vee$" denote the maximum of two real numbers. Since the optimal treatment combination $\boldsymbol{t}^*$ is unique, as the sample size n goes to infinity,

$$|\hat{\mu}(\boldsymbol{t}^*) - \mu(\boldsymbol{t}^*)| \vee |\hat{\mu}(\boldsymbol{t}) - \mu(\boldsymbol{t})| \leq \frac{\mu(\boldsymbol{t}^*) - \mu(\boldsymbol{t})}{4}, \tag{38}$$

with a probability going to 1.

Furthermore, (38) implies,

$$\hat{\mu}(\boldsymbol{t}^*) - \hat{\mu}(\boldsymbol{t}) = [\mu(\boldsymbol{t}^*) - (\mu(\boldsymbol{t}^*) - \hat{\mu}(\boldsymbol{t}^*))] - [\mu(\boldsymbol{t}) - (\mu(\boldsymbol{t}) - \hat{\mu}(\boldsymbol{t}))] \geq \frac{\mu(\boldsymbol{t}^*) - \mu(\boldsymbol{t})}{2} > 0.$$

Therefore, with the probability going to 1, we have $\hat{\mu}(\boldsymbol{t}^*) > \hat{\mu}(\boldsymbol{t})$ as the sample size n goes to infinity. Taking the union bound over $(2^m - 1)$ treatment combinations, we have $\hat{\mu}(\boldsymbol{t}^*) > \max_{\boldsymbol{t} \neq \boldsymbol{t}^*} \hat{\mu}(\boldsymbol{t})$, i.e., $\hat{\boldsymbol{t}}^* = \arg\max_{\boldsymbol{t}} \hat{\mu}(\boldsymbol{t}) = \boldsymbol{t}^*$, with a probability going to 1 as the sample size n goes to infinity.

If $\hat{\boldsymbol{t}}^* = \boldsymbol{t}^*$, by definition,

$$\sqrt{n}\big(\hat{\Psi}(\boldsymbol{t};\zeta)\big)^{-1/2}(\hat{\zeta}(\boldsymbol{t}) - \zeta(\boldsymbol{t})) - \sqrt{n}\big(\hat{\Psi}(\boldsymbol{t};\tau)\big)^{-1/2}(\hat{\tau}(\boldsymbol{t}) - \tau(\boldsymbol{t})) = 0$$

Hence, we have that for any $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}[|\sqrt{n}\big(\hat{\Psi}(\boldsymbol{t};\zeta)\big)^{-1/2}(\hat{\zeta}(\boldsymbol{t}) - \zeta(\boldsymbol{t})) - \sqrt{n}\big(\hat{\Psi}(\boldsymbol{t};\tau)\big)^{-1/2}(\hat{\tau}(\boldsymbol{t}) - \tau(\boldsymbol{t}))| < \varepsilon] = 1. \tag{39}$$

Combining (39) with the asymptotic normality (37), by Slutsky's theorem, we have

$$\sqrt{n}\big(\hat{\Psi}(\boldsymbol{t};\tau)\big)^{-1/2}(\hat{\tau}(\boldsymbol{t}) - \tau(\boldsymbol{t})) = \sum_{i=1}^{n} \big(\hat{\Psi}(\boldsymbol{t};\tau)\big)^{-1/2}\big(\psi(\boldsymbol{z}_i, \boldsymbol{\theta}^*(\boldsymbol{x}_i), \boldsymbol{\Lambda}(\boldsymbol{x}_i), \boldsymbol{t}^*, \boldsymbol{t}) - \tau(\boldsymbol{t})\big)/\sqrt{n} + o_p(1) \to_d \mathcal{N}(0,1),$$

which concludes the proof. □

Given the asymptotic normality, the $(1-\alpha)$-confidence interval for $\hat{\tau}_{\mathsf{DeDL}}(\boldsymbol{t})$ is given by

$$\widehat{\mathcal{CI}}_{\mathsf{DeDL}}(\boldsymbol{t};\tau) = \left[\hat{\tau}_{\mathsf{DeDL}}(\boldsymbol{t}) - \frac{1}{\sqrt{n}} \cdot \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \cdot \hat{\Psi}_{\mathsf{DeDL}}(\boldsymbol{t};\tau), \hat{\tau}_{\mathsf{DeDL}}(\boldsymbol{t}) + \frac{1}{\sqrt{n}} \cdot \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \cdot \hat{\Psi}_{\mathsf{DeDL}}(\boldsymbol{t};\tau)\right]. \tag{40}$$

Appendix B: Empirical Analysis With Field Experiment Data

B.1. User Covariates Used in Section 4

Table A1 presents all the user covariates used in our empirical analysis in Section 4.

B.2. Stratified Sampling Details

We employ a three-step stratified sampling to keep the user covariates balanced with respect to eight different treatment combinations. First, we categorize 10 continuous variables in Table A1 by their quantile ranges $[0\%, 25\%)$, $[25\%, 50\%)$, $[50\%, 75\%)$, and $[75\%, 100\%]$. Specifically, we assign 1, 2, 3, and 4 as new values for values in each quantile interval respectively for each continuous variable. After all the variables are discretized, we proceed to divide the population into subpopulations according to imbalanced covariates. To check for diversely distributed covariates, we utilize a pairwise t-test between the baseline combination $(0,0,0)$ and seven treatment combinations for covariates in Table A1. Among the 26 covariates, 16 discrete covariates show no significant difference between the baseline combination $(0,0,0)$ and the other seven treatment combinations, while 10 continuous variables are imbalanced-distributed. Therefore, we divide users into 69,111 strata by the value of imbalanced covariates. Namely, users with the same value in all imbalanced covariates are grouped together. For each stratum, we set the minimum number of users among all treatment combinations as its target size and then randomly sample this many individuals for each combination as the stratified sample. Therefore, the stratified sample has a similar number of users in each treatment combination, and the treatment assignment mechanism satisfies $\mathbb{P}[T_i = 1|\boldsymbol{X} = \boldsymbol{x}] = \mathbb{P}[T_i = 0|\boldsymbol{X} = \boldsymbol{x}] = 0.5$ $(i = 1,2,3)$ for any $\boldsymbol{x}$.

B.3. Complete Randomization Check Results

Table A2 presents complete randomization check results of the stratified samples used in our empirical analysis in Section 4.

Table A1 User Covariates Used in the DeDL Framework

	Variable	Description
Discrete Var.	Age Range	Age range of user: young, mid-age, senior, old, or unknown
	Gender	Gender of user: male, female, or unknown
	Operating System	OS of user's device: Android, IOS, or other
	Product Version	Version used: lite, express, or regular
	Phone Price Range	Price range of device: luxury, expensive, affordable, or unknown
	User Activeness Degree	Activeness of user: high-, mid-, low-active, or new user
	User Active-Deepness Degree	Active-deepness of user: deep-, shallow-active, or new user
	Feed Mode	Preferred mode of app interface: video stream, video cover stream, or unknown
	Number of Followers	Interval of the user's number of followers: <10, 10 - 10k, 10k - 100k, >100k
	Influencer Level	Mainly determined by the number of followers: micro, midtier, or macro influencer
	Number of Mutual Followers	Interval of the user's number of mutual followers (friends): <10, 10 - 10k, 10k - 100k, >100k
	Sociableness Level	Mainly determined by the number of mutual followers (friends): low-, mid-, or high-sociable
	Frequent Residence Area	Region in which the user is frequently on the platform: South, North, or unknown
	Frequent Residence City Level	Level of city in which the user is frequently on the platform: large city, big city, medium city, small city, or unknown
	Frequent Residence City Type	Type of city in which the user is frequently on the platform: city, town, rural, or unknown
	Active Engagement City Level	Level of city in which the user is always active: large city, big city, medium city, small city, or unknown
Continuous Var.	Average App Usage Duration	User's average usage duration on platform per day
	Average Video Watching Time	User's average time on watching videos on platform per day
	Average Live Watching Time	User's average time on watching live on platform per day
	Average DP Video Watching Time	User's average time on watching videos on Discover Page per day
	Average LP Video Watching Time	User's average time on watching videos on Live Page per day
	Average FYP Video Watching Time	User's average time on watching videos on For You Page per day
	Average FP Video Watching Time	User's average time on watching videos on Following Page per day
	Average DP Screen Time	User's average time on Discover Page per day
	Average LP Screen Time	User's average time on Live Page per day
	Average FP Screen Time	User's average time on Following Page per day

Table A2 Randomization Check

	Treatment Combination $\boldsymbol{t}'$	$(0,0,0)$	$(0,0,1)$	$(0,1,0)$	$(1,0,0)$	$(1,1,1)$	$(1,1,0)$	$(1,0,1)$	$(0,1,1)$
User Demographics	Proportion of Male Users	60.51%	60.63% (0.41)	60.31% (0.13)	60.49% (0.85)	60.40% (0.40)	60.36% (0.26)	60.30% (0.11)	60.31% (0.15)
	Proportion of High-Active Users	29.67%	29.65% (0.88)	29.60% (0.55)	29.73% (0.62)	29.77% (0.44)	29.65% (0.85)	29.80% (0.34)	29.66% (0.91)
	Proportion of Users from the South	39.83%	39.64% (0.16)	39.90% (0.63)	39.80% (0.88)	39.85% (0.90)	39.68% (0.26)	39.69% (0.29)	39.74% (0.48)
User Behaviors 10 Days Prior to the Experiments	Average Active Days per User	-0.0003	-0.0010 (0.78)	0.0012 (0.59)	-0.0006 (0.91)	-0.0011 (0.75)	0.0001 (0.88)	0.0011 (0.61)	0.0008 (0.69)
	Average DP Screen Time per User	0.0173	0.0144 (0.31)	0.0151 (0.45)	0.0152 (0.46)	0.0140 (0.25)	0.0145 (0.34)	0.0152 (0.46)	0.0157 (0.58)
	Average LP Screen Time per User	0.0126	0.0091 (0.21)	0.0146 (0.49)	0.0104 (0.42)	0.0136 (0.75)	0.0145 (0.52)	0.0103 (0.42)	0.0159 (0.25)
	Average FYP Screen Time per User	0.0048	0.0022 (0.36)	0.0033 (0.62)	0.0044 (0.91)	0.0010 (0.18)	0.0033 (0.61)	0.0015 (0.25)	0.0024 (0.41)
	Average App Usage Duration per User	0.0129	0.0111 (0.55)	0.0114 (0.62)	0.0114 (0.61)	0.0104 (0.39)	0.0103 (0.37)	0.0120 (0.76)	0.0117 (0.69)

Note: p-values of t-tests between the baseline combination $\boldsymbol{t}_0 = (0,0,0)'$ and other treatment combinations are reported in parentheses. To protect sensitive data, the reported metrics of active days, screen time, and app-usage duration are rescaled.

Appendix C: Implementation Details of Benchmarks

In this section, we present the implementation details of using linear addition (LA), linear regression (LR), pure deep learning (PDL), and structured deep learning (SDL) as our benchmark to estimate ATE. For LR, PDL, and SDL estimation, we follow the same four-fold cross-fitting described in Appendix 5.2.

C.1. Linear Addition (LA) Estimator

To obtain the LA estimator for each treatment level, we use the ATE of five observed treatment combinations to calculate the ATE of unobserved treatment combinations. For observed treatment combinations, their estimated ATE and significance level are the same as the ground truth. For unobserved treatment combinations, we use the linear addition of ATE of individual observed treatment combinations as the estimated ATE. Furthermore, we estimate the standard error of the estimated ATE for unobserved treatment combinations by assuming that the estimators for individual experiments are independent.

We report the LA estimators in the first row in each section of treatment combination in Table A3. The top four treatment combinations are observable, and therefore the LA estimator yields zero error (see columns 5–7 in Table A3). The estimators for the ATE of the bottom-three treatment combinations use the ATE of individual treatment combinations, i.e., $(0,0,1)$, $(0,1,0)$, and $(1,0,0)$, to calculate the final results $\hat{\mu}(\boldsymbol{t}) = \hat{\mu}(\boldsymbol{t}_1) + \hat{\mu}(\boldsymbol{t}_2)$, where $\boldsymbol{t} = \boldsymbol{t}_1 + \boldsymbol{t}_2$. The standard deviation of the estimator follows $\hat{\sigma}(\hat{\mu}(\boldsymbol{t})) = \sqrt{\hat{\sigma}(\hat{\mu}(\boldsymbol{t}_1))^2 + \hat{\sigma}(\hat{\mu}(\boldsymbol{t}_2))^2}$.

C.2. Linear Regression (LR) Estimator

The LR estimator uses the regression coefficients as the estimated ATE. The regression is defined as

$$y = \beta_1 t_1 + \beta_2 t_2 + \beta_3 t_3 + \boldsymbol{\alpha}\boldsymbol{x}, \tag{41}$$

where t_1, t_2, and t_3 denote three experiments for pages DP, LP, and FYP, respectively, and $\boldsymbol{x}$ denotes the user covariates in Table A1. We fit the linear regression model (41) with 1,291,652 data points with the five observed treatment combinations. The estimator for β_i, $\hat{\beta}_i$, captures the ATE of treatment t_i. Similar to LA estimators, LR estimators assume the linear additivity of the ATEs. To maintain a fair comparison between the estimators, we adopt a four-fold cross-fitting described in Appendix 5.2 to obtain the LR estimator as well. Specifically, we fit the linear regression model with 75% of the training data. For the second-stage inference, we predict the potential outcome of each user given the covariates $\boldsymbol{x}$ under treatment combination $\boldsymbol{t} \in \mathcal{T}$ using the trained linear regression. The estimation and inference of each ATE is obtained through the standard pairwise t-test for each treatment combination with the predicted outcomes on the last data fold. The average values of estimated ATEs for each treatment combination are presented in the second row in Table A3.

C.3. Pure Deep Learning (PDL) Estimator

Similar to the implementation of DeDL approach in Appendix 5.2, we apply the four-fold cross-fitting procedure to reduce overfitting when implementing the PDL estimators. Specifically, we evenly partition the observed data of five treatment combinations into four random folds. For each data fold, we use the data from three other folds as the training data to approximate $y = f(\boldsymbol{x}, \boldsymbol{t})$, where $f(\cdot,\cdot)$ is a fully-connected

three-layer DNN with 20 nodes in each layer. We call this approach the "pure" deep learning estimator because no constraints are imposed on the DNN $f(\cdot,\cdot)$. The DNN is trained with the Adam optimizer and the mean squared error as the loss function. We then use the trained DNN $\hat{f}(\cdot,\cdot)$ to predict the potential outcomes under eight treatment combinations for each user the covariates $\boldsymbol{x}$ on the last fold. Similar to the LR estimator, pairwise t-tests with the predicted outcomes between any treatment combination $\boldsymbol{t}$ and the control $\boldsymbol{t}_0$ provide the ATE estimate and its corresponding standard error, whereas the final estimates are the average value of all four data folds. The fourth row for each treatment combination in Table A3 shows the estimated ATE of the PDL approach.

C.4. Structured Deep Learning (SDL) Estimator

The SDL approach adopts the same structured DNN and four-fold cross-fitting as DeDL as documented in Appendix 5.2, but without the debiasing term. For each treatment combination $\boldsymbol{t} \in \mathcal{T}$, we use the trained structured DNN to predict the potential outcome given the covariates $\boldsymbol{x}$: $H(\boldsymbol{x}, \hat{\theta}_s(\boldsymbol{x}); \boldsymbol{t}, \boldsymbol{t}_0) := G(\hat{\boldsymbol{\theta}}_s(\boldsymbol{x}), \boldsymbol{t}) - G(\hat{\boldsymbol{\theta}}_{\boldsymbol{s}}(\boldsymbol{x}), \boldsymbol{t}_0)$. Similar to the LA and PDL approaches, t-tests with the predicted outcomes provide the ATE estimates under SDL. We remark that the SDL approach underestimates the standard errors because it ignores the noises in fitting $\hat{\boldsymbol{\theta}}(\cdot)$ caused by the variations in the training data. The estimation results of SDL for each treatment combination are reported in the last row in Table A3.

Appendix D: Details of Synthetic Experiments

In this section, we provide details of the synthetic experiments designed to assess several factors we have identified that could potentially influence the performance of the DeDL estimators in practice. We first validate our theory by varying the number of experiments $m \in \{4, 6, 8, 10\}$ in Appendix D.2. In Appendix D.3, we test the performance of our DeDL estimators with a potentially large bias of estimating $\hat{\boldsymbol{\theta}}(\cdot)$; we find that DeDL is fairly robust, with moderate biases. We also systematically assess the performance of DeDL under model misspecification, and we shed light on how to test and select the link function in practice in Appendix D.4. Furthermore, in Appendix D.5, we investigate a practical setting where the observed $\boldsymbol{X}$ distribution deviates from the population, and discuss how to use the rebalancing method to get trustworthy estimates.

D.1. Simulation Setup

Throughout Sections D.2 and D.3, we assume that the link function G is correctly specified. Consistent with our empirical study in Section 4, we use the *Generalized Sigmoid Form II*, i.e., for each data point i, we have

$$y_i = \frac{\theta^*_{m+1}}{1 + \exp(\theta^*_0(\boldsymbol{x}_i) + \theta^*_1(\boldsymbol{x}_i)t_{i1} + \theta^*_2(\boldsymbol{x}_i)t_{i2} + \cdots + \theta^*_m(\boldsymbol{x}_i)t_{im})} + \epsilon_i, \tag{42}$$

where ϵ_i is the *i.i.d.* random noise with zero mean.

We assume there are m concurrent field experiments, each with a binary treatment. We sample θ^*_{m+1} from the uniform distribution $\mathcal{U}(10, 20)$ throughout, while generating $\theta_j(\boldsymbol{x})$, $j = 0, 1, \ldots, m$, differently in different subsequent sections. We generate data points $\boldsymbol{z}_i = (y_i, \boldsymbol{x}_i', \boldsymbol{t}_i')'$ as *i.i.d.* copies of the random vector $\boldsymbol{Z} = (Y, \boldsymbol{X}', \boldsymbol{T}')'$. The random perturbation ϵ follows a uniform distribution $\mathcal{U}(-0.05, 0.05)$. Without loss of generality, we generate covariates $\boldsymbol{x}$ as follows: (1) the dimension of covariates $\boldsymbol{X}$ satisfies $d_{\boldsymbol{X}} = 10$; (2) the

Table A3 Detailed Results of Benchmark Estimators

Treatment Combination	Ground-Truth ATE (1)	Estimator (2)	Estimated ATE (3)	CD (4)	APE (5)	SE (6)	AE (7)
(0, 0, 1)	1.091%**	LA	1.091%	1	0.00%	0.000	0.000
		LR	1.329%	1	21.79%	5.657	2.379
		PDL	1.247%	1	14.24%	2.417	1.555
		SDL	1.179%	1	8.00%	0.763	0.873
(0, 1, 0)	-0.267%	LA	-0.267%	1	NA	0.000	0.000
		LR	-0.013%	1	NA	6.460	2.542
		PDL	-0.036%	0	NA	5.353	2.314
		SDL	-0.072%	0	NA	3.792	1.947
(1, 0, 0)	0.758%*	LA	0.758%	1	0.00%	0.000	0.000
		LR	1.079%	1	42.29%	10.028	3.206
		PDL	1.043%	1	37.60%	8.126	2.851
		SDL	0.978%	1	28.95%	4.816	2.195
(1, 1, 1)	2.121%****	LA	2.121%	1	0.00%	0.000	0.000
		LR	2.395%	1	12.95%	7.539	2.746
		PDL	2.326%	1	9.67%	4.209	2.052
		SDL	2.040%	1	3.78%	0.642	0.801
(1, 1, 0)	0.689%	LA	0.491%	1	NA	3.902	1.975
		LR	1.066%	0	NA	14.233	3.773
		PDL	1.030%	0	NA	11.625	3.410
		SDL	0.902%	0	NA	4.543	2.132
(1, 0, 1)	2.299%****	LA	1.850%	1	19.56%	20.229	4.498
		LR	2.408%	1	4.72%	1.178	1.085
		PDL	2.333%	1	1.46%	0.112	0.336
		SDL	2.148%	1	6.59%	2.297	1.516
(0, 1, 1)	1.387%***	LA	0.824%	0	40.58%	31.670	5.628
		LR	1.316%	1	5.08%	0.495	0.704
		PDL	1.217%	1	12.26%	2.890	1.700
		SDL	1.070%	1	22.84%	10.030	3.167

Notes: The calculation of APE, SE, and AE is based on the scaled outcome variable (see column (1) of this table). SE is scaled by multiplying a constant. AE is scaled by multiplying another constant. The significance levels are encoded as *p<0.05; **p<0.01; ***p<0.001; ****p<0.0001.

different components of $\boldsymbol{X}$ are *i.i.d.* following the uniform distribution $\mathcal{U}(0,1)$. We remark that larger random perturbations, higher dimensional covariates, and/or more complicated joint distributions of $\boldsymbol{X}$ can be easily incorporated into the simulation. We adopt the current setting for ease of model training and efficiency of experiments.

To ensure that the identifiability and sufficient curvature conditions (i.e., Assumption 4) are met, we adopt the following treatment assignment mechanism $\nu(\cdot\,|\,\cdot)$. We assume the independence between $\boldsymbol{X}$ and $\boldsymbol{T}$, whose distribution we denote as $\nu(\boldsymbol{t}) = \mathbb{P}[\boldsymbol{T} = \boldsymbol{t}]$. Furthermore, in the training stage, with equal probability, we randomly assign each experimental unit to one of $m+2$ different treatment combinations with equal probability, i.e., $\nu(\boldsymbol{t}) = 1/(m+2)$ if $\boldsymbol{t} \in \{\boldsymbol{t} \in \{0,1\}^m : \sum_{i=1}^m t_i = 0 \text{ or } 1\} \cup \{(1,1,0,\ldots,0)'\}$ and $\nu(\boldsymbol{t}) = 0$ for other treatment combinations. In other words, we assume the *partially observed outcome* setting with only

$m+2$ observable treatment combinations, while other treatment outcomes are masked for gauging the performance of estimators.

Our neural network structure to estimate $\theta_j^*(\boldsymbol{x}),\ j=0,1,\ldots,m$, is prespecified as two-layer perceptrons with ReLU activation functions and 10 nodes in hidden layers in all experiments throughout this section. Because there are $m+2$ unknown parameters, which linearly scale with the number of experiments m, we generate $500m$ *i.i.d.* experimental data points $\boldsymbol{z}_i$ for the DNN training, and another independent $500m$ experimental data points for the inference stage. During the DNN training stage, we randomly split this data set with 70% for training and the rest 30% for cross-validation. We adopt the MSE loss function and Adam algorithm (Kingma and Ba 2014). In most experiments, we stop training when the loss on cross-validation data is less than a fixed threshold of 0.3. We empirically tested various thresholds and found that the gain from a smaller stopping threshold is marginal. Thus, we picked this threshold based on the computational efficiency consideration. In the experiments with misspecified link functions and imbalanced data, which we will discuss later, the cross-validation loss tends to increase. Hence, we adjust the threshold accordingly in such cases. We also experiment with popular training strategies such as dropout or regularizing the weights, but the gain is marginal, so we do not include them in this discussion.

At the inference stage, we independently generate a data sample with the same size as the training data. To avoid the rare case where the empirical estimate $\hat{\boldsymbol{\Lambda}}(\boldsymbol{x})$ is not invertible (e.g., $\boldsymbol{\theta}_i(\boldsymbol{x})=\boldsymbol{\theta}_j(\boldsymbol{x})$ for all $i,j\in\{1,2,\ldots,m\}$), we add a small regularization to $\hat{\boldsymbol{\Lambda}}(\boldsymbol{x})$ so that $(\hat{\boldsymbol{\Lambda}}(\boldsymbol{x})+0.0005\boldsymbol{I}_{d_{m+2}})^{-1}$ is well-defined. Similar regularization or trimming techniques based on the propensity score are quite common in practice for numerical stability.

To calculate the true ATE over the population, we use the sample average of 2,000 independent samples for each treatment combination, and use the standard t-test with a significance level of 0.05 to determine whether the ATE of an experiment combination is statistically significant. To derive statistical metrics such as confidence intervals, we replicate all experiments 200 times.

D.2. Validation of the DeDL Estimator under Large m

In this section, we aim at empirically validating the theoretical results in Section 3 and further demonstrating the superior performance of our DeDL estimator in practice. Such experiments are necessary due to the gaps between theory and practice. In particular, there are two potential inconsistencies between the underlying theory (e.g., see Section 3) and practical settings. First, the key theoretical result Proposition 1 is proved under the assumption that one obtains the estimator $\hat{\boldsymbol{\theta}}(\boldsymbol{x})$ by (almost) minimizing the empirical loss. However, the loss of DNN is difficult to optimize globally, especially given our novel structured architecture with a model layer. Second, the theoretical DNN width $O(n^{d_{\boldsymbol{X}}/2(p+d_{\boldsymbol{X}})}\log^2 n)$ and depth $O(\log n)$ required in Proposition 1 are clearly too large for practical applications. Practitioners often fine-tune these hyperparameters at a much smaller scale. Given these practical issues, we find it necessary to conduct synthetic experiments to demonstrate the performance of DeDL in general settings.

To generate the functions $\boldsymbol{\theta}^*(\boldsymbol{x})$ in this subsection, we first define the coefficient matrix $\boldsymbol{A}\in\mathbb{R}^{(m+1)\times d_{\boldsymbol{X}}}$ with each component independently drawn from the uniform distribution $\mathcal{U}(-0.5,0.5)$ and write the j the row of $\boldsymbol{A}$ as row vector $\boldsymbol{A}_{[j]}$. Then, we let $\theta_j^*(\boldsymbol{x})=(\boldsymbol{A}_{[j+1]}\boldsymbol{x})^3,\ j=0,2,\ldots,m$. As mentioned, the parameter θ_{m+1}^* is

randomly generated from the uniform distribution $\mathcal{U}(10, 20)$. To facilitate numerical experiments, we choose a relatively simple structure for the neural network, as discussed previously. However, the more complex function $\boldsymbol{\theta}^*(\boldsymbol{x})$ can be readily incorporated and tested in our synthetic experiments, and DNN training-related hyperparameters (e.g., width, depth, and training algorithm) should be fine-tuned to accommodate such cases.

We summarize the main result here while deferring more details into Appendices D.2.1 and D.2.2. There, we report complete simulation results in Table A4 and Table A5, along with more observations made. In Figure A1(a), we evaluate the performance of different estimators including LA, LR, PDL, SDL, and DeDL with varying numbers of A/B tests, i.e., $m \in \{4, 6, 8, 10\}$. We train these estimators with partially observed outcomes, as discussed above. We note that the performance of MAE is highly correlated with other performance metrics, so to keep the discussion simple, we mainly report and visualize the performance comparison under MAE along with the 95% confidence interval, shown in Figure A1. We compute the confidence interval using the 200 instances for each parameter combination. One can observe from panel (a) that DeDL has the best performance under all $m \in \{4, 6, 8, 10\}$. Increased values of m lead to quick degradation of the performance of LA and LR. Such simple models relying on linear extrapolation are unable to capture the rich treatment effects, but the performance of SDL and DeDL are relatively stable.

To highlight the issue that PDL can easily overfit the observed data and result in large biases in unobserved treatments, we conduct a more detailed synthetic experiment focusing on PDLs with different network size specifications and partially/fully observed data. Figure A1(b) displays the performance of different variants of PDL estimators under different m values. For better visualization, we report the performance of PDL with a different scale on the y-axis than that in Figure A1(a). Among these different PDL estimators, we use subscripts s (small) and l (large) to represent different widths of neural nets, with 10 (small) and 40 (large) hidden nodes for all three hidden layers, respectively. All DNNs have three linear layers followed by ReLU activation layers. The subscripts p and a represent the training samples generated from *partially observed treatments* and *all* treatments, respectively. For a fair comparison to SDL and DeDL, we focus on PDL_s_p with partially observed treatments and a similar DNN size to SDL. We observe that, with partially observed combinations, both PDL_s_p and PDL_l_p perform poorly, mainly driven by the bad performance under the unobserved treatments. Increased network size does not help much. However, when we incorporate data from all treatment combinations for training, PDL_s_a and PDL_l_a obtain comparable performance with the best estimator DeDL. Therefore, a structured DNN model allows us to capture the interaction of experiments in practical settings, while the model flexibility in PDL helps only in unrealistic scenarios where a large proportion of experimental combinations are observed. We refer interested readers to Appendix **??** for more detailed comparisons, discussions, and additional simulations of incorporating regularizers into PDL.

Because all performance metrics show similar patterns under different m values, we maintain $m = 4$ for computation efficiency in subsequent sections. Also, due to the inferior performance of PDL with partially observed treatments, we do not report its performance in the following sections.

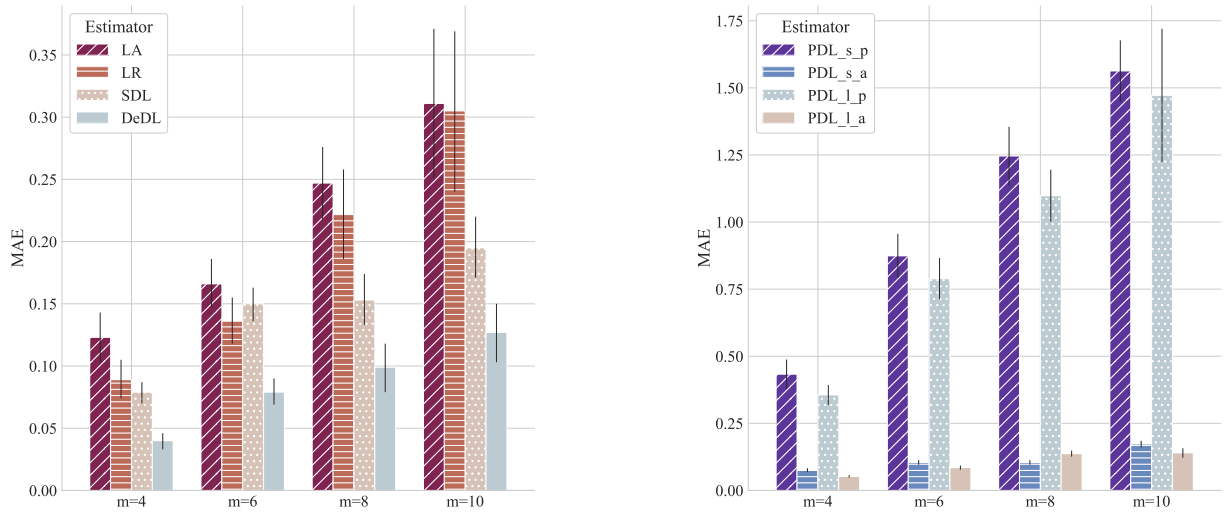


Figure A1 **MAE comparison among estimators under the increasing number of experiments m values. Panel (a) shows the performance of LA,LR,SDL, and DeDL. Panel (b) presents the performance of PDLs with different network size specifications and partially/fully observed data. Subscript s represents a small 10-width DNN while l represents a large 40-width DNN. Subscript p represents that the training set contains partially observed treatment while a represents that the training set contains all 2^m treatments.**

D.2.1. Detailed Results of the Comparison Among Estimators. We document the complete simulation results of Appendix D.2 in Table A4 and Table A5. We evaluate the performance of the estimators under $m \in \{4, 6, 8, 10\}$ A/B tests and report the estimation results in Panels A to D. The first column "Estimator" describes which estimator is tested. The second column "CDR" shows the proportion of treatment combinations whose estimated ATE significance levels and signs are consistent with the ground truth. The third column "MAPE" gives the mean absolute percentage error of ATE estimates over all treatment combinations whose real average treatment effects are significant. In other words, we rule out insignificant treatment combinations when calculating MAPE. Otherwise, those insignificant treatment effects would result in a close-to-zero value in the denominators to calculate APE, resulting in an undesired metric. Similarly, column "MSE" and column "MAE" represent squared error and absolute error, respectively. In these two columns, we do not exclude those insignificant combinations. Indeed, unlike MAPE, close-to-zero treatment effects do not cause a problem for MSE and MAE. As a result, MAE and MSE over all combinations can supplement MAPE. However, to better understand the scales of MSE and MAE errors, we also report in the table notes 95% confidence intervals of average absolute treatment effects over all combinations. Finally, the column "BTI" presents the results on best treatment identification. For each replication of the experiment, we verify whether different methods can successfully identify the best treatment combination. The value in this column shows the proportion of replications in which the optimal combination is correctly identified.

We observe the following consistent patterns across all panels in Table A4. First, DeDL outperforms LA, LR, PDL_s_p, and SDL under all metrics, which validates our theory and provides strong evidence for the advantage of our method. In particular, DeDL increases the success rate of both CDR and BTI, and decreases MAPE, MSE, and MAE compared to SDL by a significant margin. This demonstrates the value of debiasing

Table A4 Learning Estimator Validation Result

Panel A: Comparison of Different Estimators Under $m = 4$					
Estimator	CDR	MAPE	MSE	MAE	BTI
LA	92.40% (91.02%, 93.78%)	22.64% (20.10%, 25.17%)	0.070 (0.023, 0.118)	0.123 (0.103, 0.143)	79.5% (73.9%, 85.1%)
LR	95.34% (94.46%, 96.23%)	15.30% (13.56%, 17.04%)	0.038 (0.0179, 0.0582)	0.089 (0.074, 0.105)	82.0% (76.6%, 87.3%)
PDL_s_p	88.59% (87.02%, 90.17%)	68.83% (60.76%, 76.89%)	0.867 (0.572, 1.163)	0.433 (0.378, 0.488)	56.0% (49.1%, 62.9%)
SDL	95.12% (94.17%, 96.08%)	16.12% (14.57%, 17.66%)	0.018 (0.011, 0.024)	0.079 (0.070, 0.087)	90.5% (86.4%, 94.6%)
DeDL	97.53% (96.90%, 98.15%)	7.20% (6.45%, 7.95%)	0.008 (0.004, 0.012)	0.040 (0.033, 0.046)	93.5% (90.1%, 96.9%)
Panel B: Comparison of Different Estimators Under $m = 6$					
Estimator	CDR	MAPE	MSE	MAE	BTI
LA	92.05% (90.70%, 93.40%)	27.35% (23.81%, 30.89%)	0.080 (0.056, 0.105)	0.166 (0.147, 0.186)	73.0% (66.8%, 79.2%)
LR	94.88% (94.10%, 95.67%)	18.42% (16.61%, 20.24%)	0.073 (0.038, 0.108)	0.136 (0.118, 0.155)	75.5% (69.4%, 81.6%)
PDL_s_p	84.23% (82.54%, 85.93%)	111.70% (98.07%, 125.33%)	2.172 (1.736, 2.607)	0.874 (0.793, 0.956)	28.0% (21.7%, 34.3%)
SDL	93.07% (92.17%, 93.96%)	27.22% (25.16%, 29.29%)	0.051 (0.037, 0.064)	0.150 (0.136, 0.163)	68.0% (61.5%, 74.5%)
DeDL	95.28% (94.61%, 95.96%)	13.33% (11.99%, 14.68%)	0.023 (0.012, 0.033)	0.079 (0.069, 0.090)	87.0% (82.3%, 91.7%)
Panel C: Comparison of Different Estimators Under $m = 8$					
Estimator	CDR	MAPE	MSE	MAE	BTI
LA	93.83% (92.74%, 94.92%)	28.37% (24.98%, 31.75%)	0.196 (0.135, 0.256)	0.247 (0.218, 0.276)	65.0% (58.3%, 71.7%)
LR	95.51% (94.87%, 96.14%)	22.49% (20.03%, 24.96%)	0.209 (0.128, 0.289)	0.222 (0.186, 0.258)	75.6% (69.2%, 82.0%)
PDL_s_p	78.89% (77.01%, 80.77%)	140.60% (128.28%, 152.91%)	3.825 (3.059, 4.591)	1.246 (1.136, 1.355)	11.5% (7.0%, 16.0%)
SDL	94.89% (94.25%, 95.53%)	22.18% (20.20%, 24.15%)	0.072 (0.025, 0.120)	0.153 (0.133, 0.174)	67.5% (61.0%, 74.0%)
DeDL	95.53% (94.92%, 96.15%)	13.49% (11.82%, 15.15%)	0.049 (0.007, 0.092)	0.099 (0.079, 0.118)	82.0% (76.6%, 87.4%)
Panel D: Comparison of Different Estimators Under $m = 10$					
Estimator	CDR	MAPE	MSE	MAE	BTI
LA	95.01% (93.96%, 96.05%)	31.54% (24.23%, 38.84%)	0.326 (0.152, 0.500)	0.311 (0.251, 0.371)	58.8% (47.7%, 69.8%)
LR	95.51% (94.87%, 96.14%)	25.28% (21.73%, 28.83%)	0.359 (0.146, 0.572)	0.305 (0.240, 0.369)	62.0% (52.3%, 71.7%)
PDL_s_p	77.27% (75.50%, 79.03%)	163.85% (151.47%, 176.22%)	5.360 (4.527, 6.193)	1.563 (1.448, 1.677)	5.0% (2.0%, 8.0%)
SDL	94.27% (93.21%, 95.33%)	26.37% (22.97%, 29.78%)	0.077 (0.054, 0.100)	0.195 (0.171, 0.220)	56.2% (45.1%, 67.4%)
DeDL	94.54% (93.48%, 95.60%)	16.59% (13.39%, 19.78%)	0.046 (0.026, 0.067)	0.127 (0.103, 0.150)	78.8% (69.6%, 87.9%)

Notes: All experiments are replicated 200 times, with 95% $\mathcal{CI}$s reported in parentheses. 95% $\mathcal{CI}$s of average absolute treatment effects are $(0.68, 1.06)$, $(1.09, 1.30)$, $(1.45, 1.71)$, and $(1.54, 2.03)$, respectively, in Panel A, Panel B, Panel C, and Panel D.

Table A5 Performance of PDL Estimators

Panel A: Comparison of Different Estimators Under $m = 4$					
Estimator	CDR	MAPE	MSE	MAE	BTI
PDL_s_p	88.59% (87.02%, 90.17%)	68.83% (60.76%, 76.89%)	0.867 (0.572, 1.163)	0.433 (0.378, 0.488)	56.0% (49.1%, 62.9%)
PDL_s_a	95.16% (94.27%, 96.04%)	16.01% (14.49%, 17.53%)	0.012 (0.009, 0.016)	0.075 (0.068, 0.083)	78.0% (72.2%, 83.8%)
PDL_l_p	89.12% (87.61%, 90.64%)	60.15% (53.09%, 67.20%)	0.493 (0.375, 0.610)	0.356 (0.319, 0.393)	56.0% (49.1%, 62.9%)
PDL_l_a	96.34% (95.57%, 97.12%)	12.55% (11.20%, 13.89%)	0.006 (0.004, 0.007)	0.052 (0.048, 0.057)	86.5% (81.7%, 91.3%)
Panel B: Comparison of Different Estimators Under $m = 6$					
Estimator	CDR	MAPE	MSE	MAE	BTI
PDL_s_p	84.23% (82.54%, 85.93%)	111.70% (98.07%, 125.33%)	2.172 (1.736, 2.607)	0.874 (0.793, 0.956)	28.0% (21.7%, 34.3%)
PDL_s_a	95.04% (94.43%, 95.65%)	20.03% (18.39%, 21.67%)	0.021 (0.017, 0.024)	0.104 (0.096, 0.113)	70.5% (64.1%, 76.9%)
PDL_l_p	83.87% (82.33%, 85.40%)	111.37% (99.17%, 123.57%)	1.748 (1.357, 2.138)	0.789 (0.712, 0.866)	25.0% (18.9%, 31.1%)
PDL_l_a	95.73% (95.05%, 96.42%)	16.91% (14.93%, 18.89%)	0.015 (0.012, 0.019)	0.085 (0.077, 0.093)	68.0% (61.5%, 74.5%)
Panel C: Comparison of Different Estimators Under $m = 8$					
Estimator	CDR	MAPE	MSE	MAE	BTI
PDL_s_p	78.89% (77.01%, 80.77%)	140.60% (128.28%, 152.91%)	3.825 (3.059, 4.591)	1.246 (1.136, 1.355)	11.5% (7.0%, 16.0%)
PDL_s_a	95.04% (94.43%, 95.65%)	20.03% (18.39%, 21.67%)	0.021 (0.017, 0.024)	0.104 (0.096, 0.113)	70.5% (64.1%, 76.9%)
PDL_l_p	80.80% (79.15%, 82.44%)	125.38% (114.46%, 136.29%)	3.059 (2.347, 3.770)	1.098 (1.001, 1.195)	10.0% (5.8%, 14.2%)
PDL_l_a	95.43% (94.90%, 95.95%)	20.24% (18.36%, 22.13%)	0.035 (0.028, 0.041)	0.137 (0.126, 0.148)	57.5% (50.6%, 64.4%)
Panel D: Comparison of Different Estimators Under $m = 10$					
Estimator	CDR	MAPE	MSE	MAE	BTI
PDL_s_p	77.27% (75.50%, 79.03%)	163.85% (151.47%, 176.22%)	5.360 (4.527, 6.193)	1.563 (1.448, 1.677)	5.0% (2.0%, 8.0%)
PDL_s_a	95.14% (94.69%, 95.59%)	24.87% (22.64%, 27.11%)	0.052 (0.044, 0.061)	0.173 (0.160, 0.185)	43.0% (36.1%, 49.9%)
PDL_l_p	76.03% (73.04%, 79.01%)	159.58% (124.10%, 195.06%)	5.436 (3.533, 7.339)	1.472 (1.224, 1.720)	6.0% (1.2%, 10.8%)
PDL_l_a	95.67% (94.82%, 96.53%)	19.66% (17.00%, 22.31%)	0.033 (0.025, 0.041)	0.140 (0.122, 0.157)	56.0% (41.7%, 70.3%)

Note: All experiments are replicated 200 times, with 95% $\mathcal{CI}$s reported in parentheses.

in the influence function (5). Second, LA and LR perform worse than SDL in general, demonstrating the advantage of neural networks over linear methods, although SDL may still be asymptotically biased.

Comparing across different panels, we observe that the performance of all estimators becomes worse when m grows larger. In particular, the performance of BTI worsens quickly due to the exponentially increased number of combinations. Even so, DeDL can still successfully identify the best treatment among the $1,024$ combinations with a relatively high probability of 78.8% when $m = 10$.

We also remark on the underlying mechanisms of these degenerating performances. On one hand, intuitively, due to the sigmoid link function setup, when the number of field experiments gets larger, the treatment effect becomes more nonlinear. It creates difficulty for the LA estimator. Therefore, the performance of LA worsens due to its inherent lack of model richness. On the other hand, with the correct link function specification, one may expect that the performance of SDL and DeDL should be relatively stable because of their strong modeling power. However, since we fix the same DNN structure with a constant number of 10 hidden nodes across all experiments for a fair comparison with LA, the model complexity of neural networks is limited by design. We have verified that increasing the number of hidden nodes for $m = 6, 8, 10$ helps achieve similar performance to that of $m = 4$. Indeed, the challenge of estimation and inference from an increased m can be mitigated by increasing the size and complexity of the DNN. Therefore, without loss of generality, for all experiments in this section, we keep $m = 4$ for computational efficiency.

D.2.2. A Closer Look at PDL. We report the performance of the PDL estimators in a separate table, i.e., Table A5, to provide an anatomy of their bad performance. The subscripts s and l represent different widths of neural nets, with 10 and 40 hidden nodes for hidden layers, resulting in the DNN structure $(d_{\boldsymbol{x}} + d_{\boldsymbol{t}} + 1)$-10-ReLU-10-ReLU-10-ReLU-1 and $(d_{\boldsymbol{x}} + d_{\boldsymbol{t}} + 1)$-40-ReLU-40-ReLU-40-ReLU-1, respectively The subscripts p and a represent the training samples generated from partially observed treatments or all treatments. For a fair comparison to estimators in Table A4, we should focus on PDL_s_p with partially observed treatments and similar DNN size as SDL. The training stopping criterion is set the same as SDL with 0.3 validation loss threshold.

With partially observed combinations, both PDL_s_p and PDL_l_p perform worse than all other estimators under all metrics. This is due to the bad performance of the out-of-sample test under the unobserved treatments, rather than the bad approximation ability of DNN. When we increase the DNN width from 10 to 40, the performance only slightly increases. Also, PDL indeed has much better in-sample tests. Because when we incorporate data from all treatment combinations in the training, the resulting estimators PDL_s_a and PDL_l_a have comparable performance with the best estimator DeDL.

It is within our expectation that PDL with partially observed treatments has such bad performance because it can only access the data generated by base treatment level, m single experiment data, and only one treatment level with interaction $(1, 1, 0, \ldots, 0)$. It means that it is almost impossible for PDL to learn interaction between different experiments. Unlike LR and SDL, which impose parametric structures of experimental interaction, PDL aims only at increasing the performance of in-sample outcome prediction, totally ignoring the out-of-sample performance. One may argue that this is due to the over-fitting of PDL.

To explore in more detail why PDL performs badly, we investigate further. To simplify the discussion, we conduct an extra synthetic experiment with $m = 3$; we report the result in Figure A2. Each point in the scatter plots represents the values of real ATE (x-axis) and predicted ATE (y-axis). From left to right, we visualize the performance of LR, PDL-base (i.e., PDL_s_p), PDL-dropout (i.e., PDL_s_p with $p = 0.1$ dropout regularizer after each activation layer), and PDL-L1 (i.e., PDL_s_p with L_1 regularization loss over DNN parameters with fine-tuned weight 0.05).

Table A6 Performance of PDL With Regularizers

Estimator	CDR	MAPE	MSE	MAE	BTI
LA	93.56% (92.06%, 95.07%)	27.44% (22.60%, 32.29%)	0.705 (0.504, 0.905)	0.408 (0.359, 0.457)	89.0% (84.6%, 93.4%)
LR	95.50% (94.37%, 96.63%)	26.01% (20.56%, 31.45%)	0.654 (0.449, 0.859)	0.389 (0.335, 0.443)	91.5% (87.6%, 95.4%)
PDL-base	91.62% (90.22%, 93.03%)	52.68% (44.27%, 61.09%)	2.930 (2.104, 3.756)	0.762 (0.675, 0.849)	73.0% (66.8%, 79.2%)
PDL-dropout	90.81% (89.17%, 92.46%)	35.74% (31.43%, 40.04%)	0.967 (0.739, 1.195)	0.504 (0.453, 0.555)	63.5% (56.8%, 70.2%)
PDL-L1	92.00% (90.25%, 93.75%)	33.64% (27.96%, 39.32%)	1.274 (0.513, 2.035)	0.426 (0.354, 0.498)	84.5% (79.4%, 89.6%)

Note: All experiments are replicated 200 times under $m = 3$, with 95% $\mathcal{CI}$s reported in parentheses.

Four subfigures in the upper panel show the in-sample performance, while lower panel subfigures show the performance under unobserved treatment combinations. Notice that PDL-base has better in-sample performance than LR, as the data points are more concentrated around true line $y = x$, but PDL-base has the worst out-of-sample performance. Although LR assumes linear extrapolation of treatment effects, in our example, most scatter points in the out-of-sample test are still well concentrated around $y = x$ except clear patterns of overestimate when absolute ATE increases. When applying regularizations, PDL-dropout and PDL-L1 have deteriorated in-sample performance with points less concentrated around the true line while improving out-of-sample performance.

Generally, we do not know the out-of-sample ground-truth ATE in practice, which makes it difficult to guide the selection of regularizers. We also try other regularizers, such as early stopping, L_2 parameter weights, and smaller network sizes, which, however, still perform badly comparable to LR. Due to the bad performance of PDL with partially observed treatments, we do not report it in the following simulations.

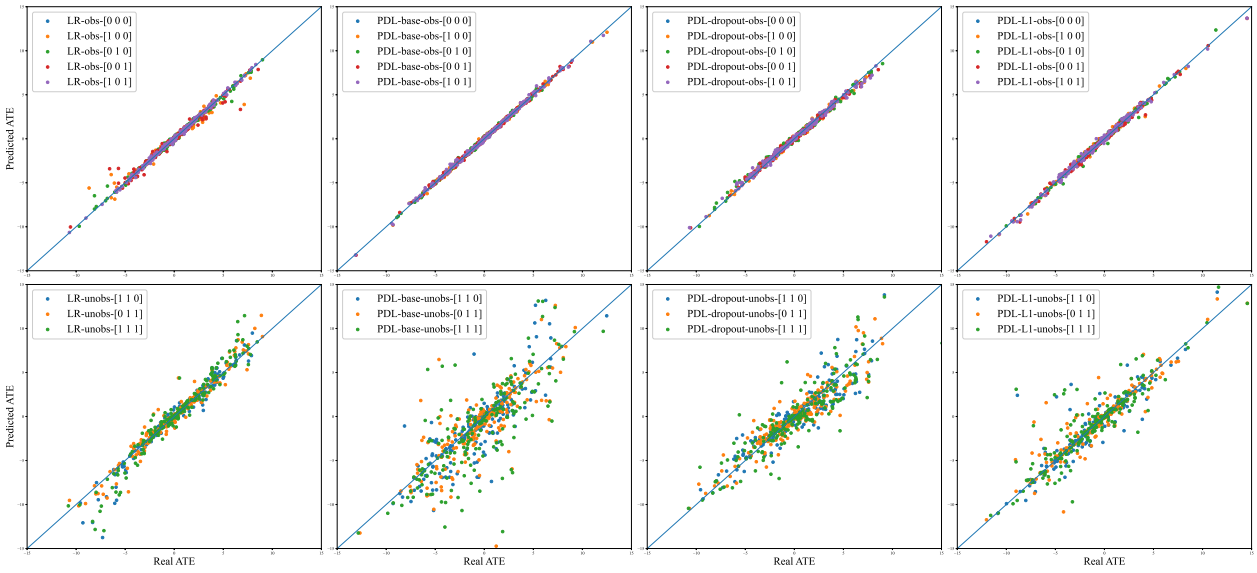


Figure A2 LR and PDL Estimator Comparison

D.3. Robustness to the DNN Convergence Rate

As pointed out by the semiparametric estimation literature (e.g., Chernozhukov et al. 2018), in practice, it is sometimes questionable whether DNN estimators $\hat{\boldsymbol{\theta}}(\cdot)$ can achieve the $o(n^{-1/4})$ convergence rate required for the inference stage. To systematically illustrate the performance of the debiasing technique in learning the treatment effects, we artificially control the biases in DNN estimators to evaluate the impact of such biases in the second-stage inference.

Specifically, we use an approach with a similar spirit to Chernozhukov et al. (2018) by constructing $\hat{\boldsymbol{\theta}}(\boldsymbol{x})$ with manually controlled biases. First, parameter functions are defined as $\theta_j^*(\boldsymbol{x}) = \boldsymbol{A}_{[j]}\boldsymbol{x}$, for $j = 0, 1, \ldots, m$, where each component in the coefficient matrix $\boldsymbol{A} \in \mathbb{R}^{(m+1)\times d_{\boldsymbol{X}}}$ is generated under independent and uniform distribution $\mathcal{U}(-0.5, 0.5)$. The parameter θ_{m+1}^* is randomly generated from the uniform distribution $\mathcal{U}(10, 20)$. Next, instead of training a DNN for estimation, we manually set the biased estimator $\hat{\theta}_j(\boldsymbol{x}) = (1 + err_j)\theta_j^*(\boldsymbol{x}),\ j = 0, 1, \ldots, m+1$, where all $err_j,\ j = 0, 1, \ldots, m+1$ terms independently follow the uniform distribution $\mathcal{U}(-\delta, \delta)$. We set different levels of the bias range coefficient $\delta \in \{0.1, 0.2, 0.3\}$ to investigate the effectiveness of DeDL with different levels of biases of the estimators $\hat{\boldsymbol{\theta}}$. In the following discussion, we focus on the MAE performance metric. We refer interested readers to Appendix **??** for a more detailed comparison and discussion.

As documented in Figure A3, DeDL has much smaller MAEs than SDL, implying a significant performance improvement with adding debias term. In all settings, DeDL performs better than LA and LR despite that SDL may generate MAEs much larger than LA and LR, in particular, when $\delta = 0.3$. This shows that the DeDL estimator is fairly robust, with moderate biases in $\hat{\boldsymbol{\theta}}$. However, we also point out that when the bias from training is large, debiasing may not improve the performance, as we illustrated in Figure 7 in Section 4. In our simulation, we have a similar observation that, when δ increases to 1.0, DeDL may perform worse than SDL under the MAPE metric. In such cases, it is critical to improve DeDL by training a better DNN model, for example, through using a larger network. For completeness, we also document the detaild results in Table A7, the columns of which are the same as Table A4. Insights similar to above can be infered from the table.

D.4. Link Function Misspecification

In this subsection, we first investigate how a misspecified link function impairs the effectiveness of our debiased estimator. Indeed, a key assumption of our framework is that the link function G is correctly specified. In practice, however, it can be challenging to select the best link function. On the positive side, as we will discuss, one may test the efficacy of the link function by checking the cross-validation errors in the DNN training stage. For this subsection, we adjust our true DGP as,

$$y_i = \frac{\theta_{m+1}^*}{1 + \exp(\theta_0^*(\boldsymbol{x}_i) + \theta_1^*(\boldsymbol{x}_i)t_{i1} + \cdots + \theta_m^*(\boldsymbol{x}_i)t_{im})} + \gamma\Big(\beta_0^*(\boldsymbol{x}_i) + \beta_1^*(\boldsymbol{x}_i)t_{i1} + \cdots + \beta_m^*(\boldsymbol{x}_i)t_{im}\Big) + \epsilon_i, \tag{43}$$

where parameter functions are defined as $\theta_j^*(\boldsymbol{x}) = \boldsymbol{A}_{[j]}\boldsymbol{x}$, and $\beta_j^*(\boldsymbol{x}) = \boldsymbol{B}_{[j]}\boldsymbol{x}$, for $j = 0, 1, \ldots, m$. The parameter $\gamma \geq 0$ captures the extent to which the true link function deviates from the *Generalized Sigmoid Form II*, which we still adopt in our DeDL framework for estimating the treatment effect. The larger the γ, the more misspecified our model is. We generate both the coefficient matrices $\boldsymbol{A} \in \mathbb{R}^{(m+1)\times d_{\boldsymbol{X}}}$ and $\boldsymbol{B} \in \mathbb{R}^{(m+1)\times d_{\boldsymbol{X}}}$ with

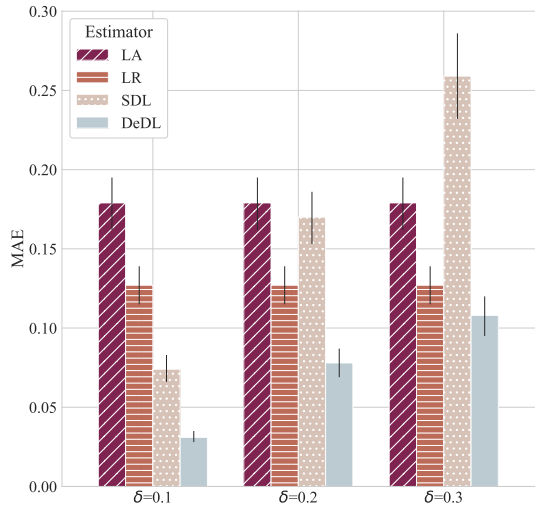


Figure A3 MAE Comparison Under Decreasing DNN Convergence Rate

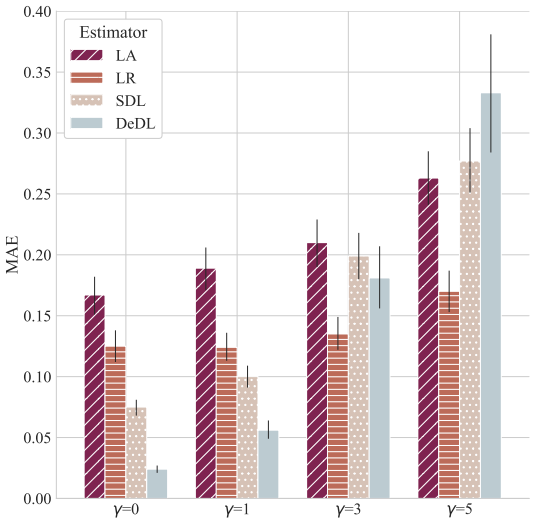


Figure A4 MAE Comparison Under Link Function Misspecification

Table A7 DNN Convergence Violation Result

Panel A: Performance of Benchmarks					
Estimator	CDR	MAPE	MSE	MAE	BTI
LA	94.56% (93.69%, 95.44%)	19.93% (17.94%, 21.93%)	0.085 (0.066, 0.104)	0.179 (0.162, 0.195)	92.0% (88.2%, 95.8%)
LR	96.88% (96.21%, 97.54%)	13.59% (12.11%, 15.08%)	0.046 (0.036, 0.056)	0.127 (0.115, 0.139)	94.0% (90.7%, 97.3%)
Panel B: Comparison of Different Estimators Under $\delta = 0.1$					
Estimator	CDR	MAPE	MSE	MAE	BTI
SDL	96.97% (96.32%, 97.62%)	4.98% (4.58%, 5.38%)	0.013 (0.010, 0.015)	0.074 (0.066, 0.083)	92.5% (88.8%, 96.2%)
DeDL	98.56% (98.12%, 99.01%)	5.04% (4.32%, 5.76%)	0.003 (0.002, 0.003)	0.031 (0.028, 0.035)	97.0% (94.6%, 99.4%)
Panel C: Comparison of Different Estimators Under $\delta = 0.2$					
Estimator	CDR	MAPE	MSE	MAE	BTI
SDL	94.19% (93.19%, 95.19%)	11.06% (10.28%, 11.83%)	0.061 (0.049, 0.072)	0.170 (0.153, 0.186)	85.5% (80.6%, 90.4%)
DeDL	96.31% (95.50%, 97.12%)	11.10% (9.46%, 12.74%)	0.016 (0.012, 0.020)	0.078 (0.069, 0.087)	95.5% (92.6%, 98.4%)
Panel D: Comparison of Different Estimators Under $\delta = 0.3$					
Estimator	CDR	MAPE	MSE	MAE	BTI
SDL	93.44% (92.48%, 94.39%)	15.53% (14.33%, 16.73%)	0.145 (0.118, 0.173)	0.259 (0.232, 0.286)	80.0% (74.4%, 85.6%)
DeDL	94.63% (93.74%, 95.51%)	12.78% (11.05%, 14.50%)	0.032 (0.024, 0.039)	0.108 (0.095, 0.120)	92.5% (88.8%, 96.2%)

Notes: All simulations are replicated 200 times, with 95% $\mathcal{CI}$s reported in parentheses. 95% $\mathcal{CI}$s of average absolute treatment effect are $(1.45, 1.65)$ in all panels.

independent entries, and each component follows the uniform distribution $\mathcal{U}(-0.5, 0.5)$. The parameter θ^*_{m+1} is again randomly generated from the uniform distribution $\mathcal{U}(10, 20)$.

We run synthetic experiments under different levels of model misspecification $\gamma \in \{0, 1, 3, 5\}$. To illustrate how the link function is skewed, we plot in Figure A5 the histograms of outcome y in (43). The x-axis represents the experimental outcome y among the population in all 2^m treatment combinations with equal probability. Observe that when γ is large, e.g., $\gamma = 5$, the bias from using the sigmoid link function to approximate the experimental outcome is likely to be large. Because the DGP defined by (43) may be too restrictive in the sense of ATE representation, further in Appendix D.4.2, we conduct another synthetic experiment where the true DGP has explicit higher-order interaction terms. The results show that debiasing through Neyman orthogonality is fairly robust under such model misspecification.

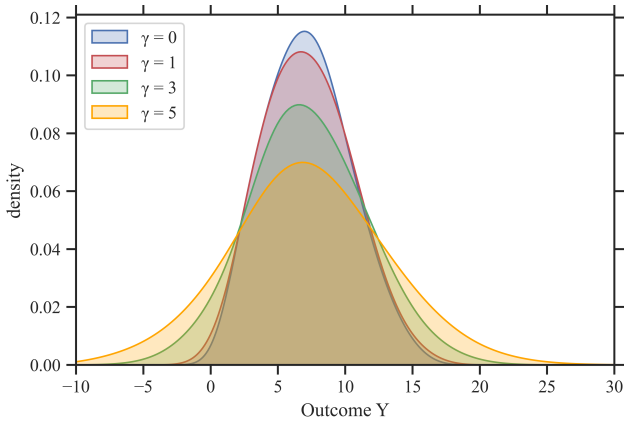


Figure A5 Experimental Outcome of Misspecified Model

Defferring reporting the complete results in Appendix D.4.1, here we emphasize the main insights. We first plot the comparison between different methods under model misspecification in Figure A4. This comparison reveals that when the link function is more specified (i.e., larger γ), SDL and DeDL estimators perform substantially worse, whereas the performances of the LA and LR estimators are relatively stable. Indeed, the LA and LR estimators face no significant increases in MAPE as γ gets larger. Also, when γ is increased to 5, DeDL performs worse than SDL, implying that debiasing via Neyman orthogonality hurts the performance when the link function is not correctly specified.

In practice, however, it is difficult, if not impossible, to verify the true link function. Fortunately, one may detect link function misspecification through large training errors. In other words, we can resort to checking the cross-validation errors in the DNN training stage. To shed light on this point, we report the training errors from our experiments. Specifically, we compare the errors induced by pure DNN without a sigmoid link function (a three-layer perceptron in our case) and structured DNN (a two-layer perceptron followed by a link function layer), respectively. Pure DNN takes treatment level $\boldsymbol{t}$ together with covariates $\boldsymbol{x}$ as inputs to the first linear layer. In contrast, the structured DNN takes only covariates $\boldsymbol{x}$ to the first linear layer and uses treatment level $\boldsymbol{t}$ in the final link function layer.

For a fair comparison, we set the pure DNN structure with a similar width and depth, i.e., $(d_{\boldsymbol{x}} + d_{\boldsymbol{t}} + 1)$–10–ReLU–10–ReLU–10–ReLU–1. Equipped with higher flexibility, pure DNN is generally good at fitting individual responses in the partially observed treatment combinations. Hence, we can use the pure DNN as a

benchmark for the in-sample comparison to check whether the link function is reasonable. Adopting the same Adam algorithm and training samples, we obtain the following 95% $\mathcal{CI}$s of cross-validation mean squared errors under different misspecification levels: (a) $[0.044, 0.053]$, $[0.086, 0.112]$, $[0.116, 0.159]$, $[0.142, 0.197]$ for pure DNN under $\gamma = 0, 1, 3, 5$ respectively; and (b) $[0.012, 0.014]$, $[0.019, 0.026]$, $[0.076, 0.106]$, $[0.215, 0.279]$ for structured DNN under $\gamma = 0, 1, 3, 5$ respectively. One can observe that when $\gamma \in \{0, 1, 3\}$, the structured DNN has a smaller or comparable in-sample loss than the pure DNN, which indicates the reasonable performance of the generalized sigmoid link function to approximate the outcome variable. However, when the misspecification level $\gamma = 5$, the in-sample error from the generalized sigmoid form grows larger than that of the pure DNN. Correspondingly, in Figure A4, we observe severe performance degradation for both SDL and DeDL and a significant negative impact from the debiasing term. In this case, we suggest experimenting with a different link function in the first DNN training stage and/or not using the debiased estimator. More generally, as long as the in-sample loss of the structured DNN is on par with that of a pure DNN with similar depth and width, we recommend adopting this structured DNN with debiasing.

D.4.1. Complete Results of the Experiments in Appendix D.4. Table A8 shows the complete simulation results. Comparing across different panels representing different values of γ, one can observe that when the link function gets more misspecified, SDL and DeDL estimators get worse under all metrics. While the MAPE, MSE, and MAE performances of LA and LR are not significantly deteriorated, considering the increased absolute treatment effects listed in table notes. When the model is not misspecified or marginally misspecified, i.e., $\gamma \in \{0, 1, 3\}$, SDL and DeDL work no worse than LA in all performance metrics. When $\gamma = 1$, DeDL can only marginally improve the performance of SDL for all metrics. However, when γ is increased to 3, we can observe that DeDL has worse MSE performance than SDL. Even worse, when γ is increased to 5, DeDL is not better than SDL under all performance metrics. In summary, we still recommend using DeDL when $\gamma \in \{0, 1, 3\}$.

D.4.2. Experiment on Model Misspecification under Higher-Order Treatment Interactions. We conduct a new set of synthetic experiments on the model misspecification with explicit higher-order interaction terms. Specifically, the true data-generating process with all higher-order terms is defined as

$$\begin{aligned} y_i =& \theta_0^*(\boldsymbol{x}_i) + \theta_1^*(\boldsymbol{x}_i)t_1 + \theta_2^*(\boldsymbol{x}_i)t_2 + \theta_3^*(\boldsymbol{x}_i)t_3 + \theta_4^*(\boldsymbol{x}_i)t_4 \\ &+ \theta_5^*(\boldsymbol{x}_i)t_1t_2 + \theta_6^*(\boldsymbol{x}_i)t_1t_3 + \theta_7^*(\boldsymbol{x}_i)t_1t_4 + \theta_8^*(\boldsymbol{x}_i)t_2t_3 + \theta_9^*(\boldsymbol{x}_i)t_2t_4 + \theta_{10}^*(\boldsymbol{x}_i)t_3t_4 \\ &+ \theta_{11}^*(\boldsymbol{x}_i)t_1t_2t_3 + \theta_{12}^*(\boldsymbol{x}_i)t_2t_3t_4 + \theta_{13}^*(\boldsymbol{x}_i)t_3t_4t_1 + \theta_{14}^*(\boldsymbol{x}_i)t_4t_1t_2 \\ &+ \theta_{15}^*(\boldsymbol{x}_i)t_1t_2t_3t_4 + \epsilon_i, \end{aligned} \tag{44}$$

where all parameters $\theta_i^*(\boldsymbol{x}) = \boldsymbol{c}_i'\boldsymbol{x}$ $(i = 0, 1, \ldots, 15)$ are linear in the covariate vector. We generate the coefficient $\boldsymbol{c} \in \mathbb{R}^{d_{\boldsymbol{X}}}$ with independent entries, and each component follows the uniform distribution $\mathcal{U}(0, 0.5)$. Note that our proposed *Generalized Sigmoid Form II* is largely misspecified compared to the above true DGP. We run synthetic experiments under different levels of model misspecification by incorporating different terms into the DGP. Specifically, we conduct four different tests. In the first test, we add all terms, i.e., 1st, 2nd, 3rd, and 4th-order interaction terms. In the second test, we only put 1st, 2nd, and 3rd-order terms in the true DGP. In the third test, we have 1st and 2nd-order terms, while the last test only has linear terms.

Table A8 Model Misspecification Result

Panel A: Comparison of Different Estimators Under $\gamma = 0$					
Estimator	CDR	MAPE	MSE	MAE	BTI
LA	94.78% (93.80%, 95.76%)	20.69% (18.30%, 23.08%)	0.074 (0.056, 0.091)	0.167 (0.151, 0.182)	94.5% (91.3%, 97.7%)
LR	97.38% (96.81%, 97.94%)	12.81% (11.53%, 14.08%)	0.043 (0.033, 0.053)	0.125 (0.112, 0.138)	94.5% (91.3%, 97.7%)
SDL	97.71% (97.18%, 98.26%)	10.07% (8.92%, 11.21%)	0.014 (0.010, 0.018)	0.075 (0.068, 0.081)	96.5% (93.9%, 99.1%)
DeDL	99.28% (98.99%, 99.57%)	3.23% (2.68%, 3.78%)	0.002 (0.001, 0.003)	0.024 (0.021, 0.027)	99.0% (97.6%, 100.4%)
Panel B: Comparison of Different Estimators Under $\gamma = 1$					
Estimator	CDR	MAPE	MSE	MAE	BTI
LA	95.56% (94.77%, 96.35%)	21.76% (19.29%, 24.23%)	0.091 (0.069, 0.112)	0.189 (0.171, 0.206)	87.5% (82.9%, 92.1%)
LR	97.38% (96.79%, 97.96%)	12.85% (11.45%, 14.24%)	0.042 (0.033, 0.052)	0.124 (0.113, 0.136)	88.0% (83.5%, 92.5%)
SDL	97.22% (96.57%, 97.86%)	11.55% (10.40%, 12.70%)	0.027 (0.019, 0.035)	0.100 (0.091, 0.109)	90.5% (86.4%, 94.6%)
DeDL	98.81% (98.33%, 99.29%)	5.67% (4.87%, 6.47%)	0.014 (0.008, 0.020)	0.056 (0.049, 0.064)	92.5% (88.8%, 96.2%)
Panel C: Comparison of Different Estimators Under $\gamma = 3$					
Estimator	CDR	MAPE	MSE	MAE	BTI
LA	94.62% (93.71%, 95.54%)	18.07% (16.18%, 19.96%)	0.112 (0.075, 0.148)	0.210 (0.190, 0.229)	89.0% (84.6%, 93.4%)
LR	97.09% (96.46%, 97.73%)	12.10% (10.77%, 13.43%)	0.050 (0.036, 0.063)	0.135 (0.122, 0.149)	88.0% (83.5%, 92.5%)
SDL	97.00% (96.33%, 97.67%)	15.10% (13.54%, 16.66%)	0.123 (0.088, 0.159)	0.199 (0.180, 0.218)	90.5% (86.4%, 94.6%)
DeDL	97.50% (96.77%, 98.22%)	13.44% (10.87%, 16.00%)	0.151 (0.096, 0.207)	0.181 (0.156, 0.207)	92.5% (88.8%, 96.2%)
Panel D: Comparison of Different Estimators Under $\gamma = 5$					
Estimator	CDR	MAPE	MSE	MAE	BTI
LA	95.34% (94.47%, 96.22%)	19.93% (17.12%, 22.74%)	0.155 (0.125, 0.185)	0.263 (0.241, 0.285)	88.0% (83.5%, 92.5%)
LR	96.97% (96.21%, 97.73%)	12.05% (10.28%, 13.81%)	0.077 (0.053, 0.102)	0.170 (0.153, 0.187)	92.5% (88.8%, 96.2%)
SDL	95.90% (95.09%, 96.72%)	17.24% (15.21%, 19.26%)	0.241 (0.183, 0.299)	0.277 (0.251, 0.304)	88.5% (84.0%, 93.0%)
DeDL	95.46% (94.43%, 96.50%)	23.20% (18.31%, 28.09%)	0.559 (0.375, 0.743)	0.333 (0.284, 0.381)	84.5% (79.4%, 89.6%)

Notes: All simulations are replicated 200 times, with 95% $\mathcal{CI}$s reported in parentheses. 95% $\mathcal{CI}$s of average absolute treatment effects are $(1.45, 1.65)$, $(1.52, 1.72)$, $(2.06, 2.36)$, and $(2.71, 3.07)$ respectively in Panel A, Panel B, Panel C, and Panel D.

The complete result are reported in Table A9, where each panel shows a test result with different terms in the DGP as defined above. Because there are no significant differences in the performance metrics BTI and CDR, we only document the MAPE, MSE, and MAE measures. One can observe that DeDL performs best when there exist interactions among the treatment effects of different experiments; see Panel A, B, and C in Table A9. It means that even the link function is misspecified, DeDL can still capture some level of interactions and make a relatively accurate estimation. Comparing DeDL and SDL, Table A9 shows that

DeDL improves the performance over SDL as long as higher-order interaction terms are incorporated. Even if the DGP is linear (i.e., Panel D), the performance of DeDL is on par with that of SDL, suggesting that debiasing with Neyman orthogonality is fairly robust under model misspecification.

Table A9 Model Misspecification Result with Explicit Individual-Level Higher Order Terms

Panel A: Comparison of Different Estimators Under 1st, 2nd, 3rd, 4th - Order Terms			
Estimator	MAPE	MSE	MAE
LA	34.06% (33.77%, 34.35 %)	18.732 (18.426, 19.038)	2.582 (2.559, 2.604)
LR	29.30% (28.98%, 29.63 %)	12.225 (11.967, 12.482)	1.883 (1.860, 1.907)
SDL	21.78% (21.41%, 22.14 %)	11.453 (11.168, 11.739)	1.671 (1.646, 1.696)
DeDL	15.15% (14.80%, 15.50%)	8.013 (7.706, 8.321)	1.310 (1.277, 1.343)
Panel B: Comparison of Different Estimators Under 1st, 2nd, 3rd - Order Terms			
Estimator	MAPE	MSE	MAE
LA	33.65% (33.35%, 33.96 %)	16.560 (16.297, 16.824)	2.493 (2.472, 2.514)
LR	28.81% (28.49%, 29.14 %)	10.437 (10.246, 10.629)	1.802 (1.784, 1.821)
SDL	21.51% (21.17%, 21.85 %)	9.653 (9.437, 9.870)	1.591 (1.571, 1.611)
DeDL	14.88% (14.56%, 15.19%)	6.476 (6.216, 6.736)	1.230 (1.200, 1.260)
Panel C: Comparison of Different Estimators Under Under 1st, 2nd - Order Terms			
Estimator	MAPE	MSE	MAE
LA	31.04% (30.74%, 31.33 %)	7.659 (7.485, 7.833)	1.874 (1.853, 1.896)
LR	25.09% (24.80%, 25.39 %)	3.466 (3.362, 3.570)	1.166 (1.146, 1.185)
SDL	17.32% (16.99%, 17.64 %)	2.903 (2.786, 3.019)	0.948 (0.927, 0.968)
DeDL	10.07% (9.74%, 10.40%)	1.328 (1.223, 1.433)	0.591 (0.565, 0.617)
Panel D: Comparison of Different Estimators Under Under 1st - Order Terms			
Estimator	MAPE	MSE	MAE
LA	1.56% (1.45%, 1.67 %)	0.002 (0.002, 0.003)	0.036 (0.034, 0.039)
LR	0.75% (0.68%, 0.82 %)	0.001 (0.001, 0.001)	0.018 (0.017, 0.020)
SDL	6.32% (6.08%, 6.55 %)	0.157 (0.145, 0.169)	0.215 (0.206, 0.224)
DeDL	6.67% (6.41%, 6.92%)	0.220 (0.210, 0.231)	0.212 (0.195, 0.230)

Notes: All simulations are replicated 200 times, with 95% $\mathcal{CI}$s reported in parentheses. 95% $\mathcal{CI}$s of average absolute treatment effects are $(5.03, 5.10)$, $(4.97, 5.04)$, $(4.34, 4.41)$, and $(2.48, 2.55)$ respectively in Panel A, Panel B, Panel C, and Panel D.

D.5. Imbalanced Covariates

This set of simulations investigates a practical setting where the observed $\boldsymbol{X}$ distribution deviates from the population. We call this setting *imbalanced covariates*, i.e., $\boldsymbol{X}$ covariates are imbalanced between any treatment level $\boldsymbol{t} \in \{0,1\}^m$ and population $\boldsymbol{X}$ distribution. The imbalanced covariates affect both training and inference stages, invalidating the DeDL estimators. Below, we discuss how to rebalance the covariates to get precise and trustworthy ATE estimates.

We adopt exactly the same simulation setup to Appendix D.3 except that the observed covariates follow a different distribution. Specifically, the last dimension of $\boldsymbol{x}$ follows the exponential distribution with rate $\lambda \in \{2.0, 1.0, 0.5\}$ instead of the uniform distribution $\mathcal{U}(0,1)$. The ground truth ATEs are still calculated using the uniform distributed $\boldsymbol{x}$. Although this setup is relatively simple because the observed covariates uniformly deviate from the true distribution across all treatment combinations while in practice the observed covariates $\boldsymbol{x}$ may even follow different distributions under different $\boldsymbol{t}$, this simulation result still demonstrates the importance of rebalancing covariates.

To reconcile the imbalanced covariates, one can do stratified sampling on sampled units to match the covariate distribution over the population, as we implement in our empirical study. There are also many re-randomization techniques to improve covariate balance in experiments. We refer interested readers to Morgan and Rubin (2012) and Li et al. (2020).

In this simulation study, to keep the discussion simple, we use the same stratified sampling procedure in our empirical study. Specifically, we focus on the imbalanced covariate dimension with exponential distribution. First, we do stratified sampling to keep only the data with $\boldsymbol{x}$ in the support $[0,1]^{d_{\boldsymbol{X}}}$. Then we do stratified sampling to make sure the imbalanced dimension of $\boldsymbol{x}$ is rebalanced in the sense that the sample sizes in $[0,1]^{d_{\boldsymbol{X}}-1} \times [0, 0.5)$ and $[0,1]^{d_{\boldsymbol{X}}-1} \times [0.5, 1]$ are the same. Through stratified sampling, we may discard some samples but sacrifice efficiency. We also conducted the stratified sampling with higher accuracy, e.g., the numbers of samples in five buckets with 0.2 bandwidth are the same.

We report the MAEs for different estimators with balanced and imbalanced covariates in Figure A6(a) and (b), respectively. Using the imbalanced covariates for both training and inference, Figure A6(a) reports the comparison results. (We present the complete results in Table A10.) After the stratified sampling, we use the rebalanced covariates for both training and inference; we show the result in Figure A6(b). $\lambda \in \{2.0, 1.0, 0.5\}$ indicates the growing imbalance level. The key observation is that when the data is too imbalanced with $\lambda \in \{1.0, 0.5\}$, DeDL is not precise. After the stratified sampling, debiasing is still trustworthy, reducing MAE compared to SDL.

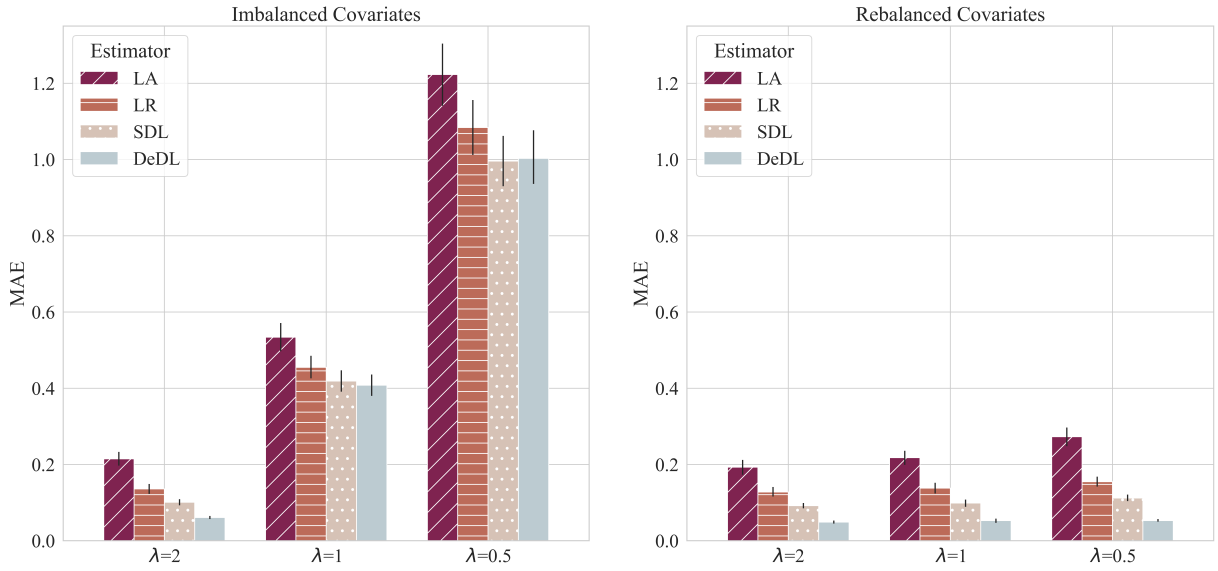


Figure A6 MAE Comparison Among Estimators Under the Imbalanced Covariates Setting. Panel (a) shows the performance of LA,LR,SDL, **and** DeDL **before rebalancing. Panel (b) presents the performance after rebalancing.**

Table A10 Imbalanced Covariates Result

Estimator	CRD	MAPE	MSE	MAE	BTI
Panel A: Comparison of Different Estimators Under $\lambda = 2.0$					
	Before Covariates Rebalancing				
LA	91.22% (89.99%, 92.45%)	32.03% (28.21%, 35.85%)	0.104 (0.081, 0.127)	0.215 (0.197, 0.233)	87.5% (82.9%, 92.1%)
LR	96.00% (95.32%, 96.68%)	18.92% (16.87%, 20.97%)	0.047 (0.036, 0.059)	0.136 (0.123, 0.149)	91.0% (87.0%, 95.0%)
SDL	96.12% (95.42%, 96.83%)	15.84% (14.09%, 17.60%)	0.024 (0.018, 0.031)	0.101 (0.093, 0.109)	95.0% (92.0%, 98.0%)
DeDL	96.44% (95.74%, 97.13%)	11.10% (9.71%, 12.48%)	0.007 (0.006, 0.008)	0.061 (0.057, 0.065)	95.0% (92.0%, 98.0%)
	After Covariates Rebalancing				
LA	93.94% (92.85%, 95.02%)	27.24% (23.10%, 31.38%)	0.102 (0.074, 0.129)	0.193 (0.173, 0.212)	92.0% (88.2%, 95.8%)
LR	96.44% (95.75%, 97.13%)	18.36% (15.69%, 21.03%)	0.044 (0.033, 0.056)	0.128 (0.116, 0.141)	93.5% (90.1%, 96.9%)
SDL	96.62% (95.95%, 97.30%)	14.47% (12.66%, 16.28%)	0.019 (0.015, 0.023)	0.092 (0.085, 0.099)	96.0% (93.3%, 98.7%)
DeDL	97.66% (97.11%, 98.20%)	7.86% (6.66%, 9.05%)	0.006 (0.004, 0.007)	0.049 (0.045, 0.053)	98.0% (96.0%, 100.0%)
Panel B: Comparison of Different Estimators Under $\lambda = 1.0$					
	Before Covariates Rebalancing				
LA	87.84% (86.47%, 89.21%)	82.23% (73.27%, 91.18%)	0.519 (0.445, 0.593)	0.534 (0.497, 0.571)	71.5% (65.2%, 77.8%)
LR	90.44% (89.36%, 91.51%)	70.13% (62.66%, 77.60%)	0.371 (0.321, 0.420)	0.455 (0.426, 0.485)	72.0% (65.7%, 78.3%)
SDL	90.53% (89.41%, 91.66%)	71.28% (63.36%, 79.20%)	0.304 (0.260, 0.348)	0.419 (0.391, 0.447)	76.0% (70.0%, 82.0%)
DeDL	90.59% (89.51%, 91.68%)	68.73% (61.16%, 76.30%)	0.291 (0.248, 0.334)	0.408 (0.380, 0.436)	75.5% (69.5%, 81.5%)
	After Covariates Rebalancing				
LA	91.44% (90.29%, 92.58%)	32.66% (28.67%, 36.66%)	0.109 (0.086, 0.133)	0.218 (0.199, 0.236)	88.0% (83.5%, 92.5%)
LR	95.84% (95.17%, 96.52%)	20.03% (16.01%, 24.04%)	0.056 (0.036, 0.076)	0.138 (0.124, 0.152)	94.5% (91.3%, 97.7%)
SDL	96.00% (95.23%, 96.77%)	15.04% (12.57%, 17.51%)	0.027 (0.018, 0.036)	0.099 (0.089, 0.108)	96.0% (93.3%, 98.7%)
DeDL	96.47% (95.84%, 97.10%)	9.16% (7.23%, 11.08%)	0.008 (0.004, 0.012)	0.053 (0.047, 0.058)	96.0% (93.3%, 98.7%)
Panel C: Comparison of Different Estimators Under $\lambda = 0.5$					
	Before Covariates Rebalancing				
LA	78.41% (76.31%, 80.51%)	186.87% (167.63%, 206.10%)	2.644 (2.289, 2.999)	1.223 (1.141, 1.304)	43.5% (36.6%, 50.4%)
LR	80.84% (78.92%, 82.77%)	167.21% (150.48%, 183.94%)	2.097 (1.811, 2.383)	1.084 (1.012, 1.156)	47.0% (40.0%, 54.0%)
SDL	81.50% (79.51%, 83.49%)	162.02% (144.89%, 179.15%)	1.737 (1.503, 1.972)	0.996 (0.930, 1.062)	51.0% (44.0%, 58.0%)
DeDL	81.34% (79.36%, 83.33%)	163.63% (146.21%, 181.06%)	1.767 (1.528, 2.005)	1.003 (0.936, 1.070)	51.5% (44.5%, 58.5%)
	After Covariates Rebalancing				
LA	88.78% (87.40%, 90.17%)	42.56% (37.80%, 47.32%)	0.158 (0.126, 0.191)	0.273 (0.248, 0.297)	84.0% (78.9%, 89.1%)
LR	94.78% (93.85%, 95.72%)	22.60% (20.11%, 25.09%)	0.055 (0.045, 0.065)	0.155 (0.142, 0.168)	88.5% (84.0%, 93.0%)
SDL	94.41% (93.49%, 95.32%)	19.69% (17.41%, 21.97%)	0.026 (0.022, 0.030)	0.112 (0.104, 0.121)	92.0% (88.2%, 95.8%)
DeDL	94.75% (93.87%, 95.63%)	9.27% (8.32%, 10.21%)	0.005 (0.005, 0.006)	0.053 (0.050, 0.057)	93.5% (90.1%, 96.9%)

Notes: All simulations are replicated 200 times, with 95% $\mathcal{CI}$s reported in parentheses. 95% $\mathcal{CI}$s of average absolute treatment effects are $(1.45, 1.65)$.