# Difference-in-Differences Meets Tree-based Methods: Heterogeneous Treatment Effects Estimation with Unmeasured Confounding

**Caizhi Tang** [*1] **Huiyuan Wang** [*2 3] **Xinyu Li** [2] **Cui Qing** [1] **Longfei Li** [1] **Jun Zhou** [1]

## Abstract

This study considers the estimation of conditional causal effects in the presence of unmeasured confounding for a balanced panel with treatment imposed at the last time point. To address this, we combine Difference-in-differences (DiD) and tree-based methods and propose a new identification assumption that allows for the violation of the (conditional) parallel trends assumption adopted by most existing DiD methods. Under this new assumption, we prove partial identifiability of the conditional average treatment effect on the treated group (CATT). Our proposed method estimates CATT through a tree-based causal approach, guided by a novel splitting rule that avoids model misspecification and unnecessary auxiliary parameter estimation. The splitting rule measures both the error of fitting observed data and the violation of conditional parallel trends simultaneously. We also develop an ensemble of multiple trees via gradient boosting to further enhance performance. Experimental results on both synthetic and real-world datasets validate the effectiveness of our proposed method.

## 1. Introduction

The identification and estimation of treatment effects is a fundamental and essential issue in the field of causal inference (Imbens & Rubin, 2016). In recent years, there has been a surge of application scenarios where treatment effects can be heterogeneous across different units based on their respective features or covariates. For instance, personalized medicine, socioeconomic policies evaluation, and large-scale recommender systems (Glass et al., 2013; Breslow & Johnson, 1993; Gilotte et al., 2018). The covariates that affect both the treatment and the outcome are referred to as confounders. In the practical applications described above, the collected data usually consist of repeated observations of the same units over a certain period, referred to as longitudinal or panel data. In such cases, both the confounders and the outcomes may vary with time. More importantly, some influential confounders may be unmeasured, or measured with covariate-dependent errors (Keller, 2014; Wacholder, 1995). The presence of unmeasured confounding leads to unidentifiability issues (Greenland & Robins, 1986), posing a significant threat to the estimation of heterogeneous causal effects in complex panel studies.

Several methods have been proposed to adjust for unmeasured confounders in longitudinal settings, such as instrumental-variable (IV) based methods and proximal causal inference, under additional assumptions on measured covariates (Tchetgen Tchetgen et al., 2018; Ying et al., 2021, among others). However, some issues need to be addressed in these methods. First, strong prior knowledge is required to determine valid IVs or treatment/outcome-inducing proxies for identification. Additionally, this line of work imposes restrictions on relationships between observable and unobservable covariates, and outcomes are only measured at the end of follow-up. For instance, Tchetgen Tchetgen et al. (2018) requires that time-varying IVs cannot interact with unmeasured confounders in an additive form, while assumptions in Ying et al. (2021) essentially entail a directed acyclic graph representation of time-varying variables. Moreover, a semiparametric marginal structural mean model (Robin, 1986) is specified for estimation, where the target causal parameter is essentially finite-dimensional. However, such a model specification is likely to be insufficient when the true relationship is complex (Ying et al., 2021).

Difference-in-Differences (DiD) methods provide an alternative to instrumental-variable (IV) and proxy-based methods for adjusting unmeasured confounders in longitudinal settings. DiD methods allow for time-varying outcomes and impose few restrictions on relationships between observable and unobservable covariates. The core identifi-

[*]Equal contribution [1]Ant Group, Zhejiang, China [2]School of Mathematical Sciences, Peking University, China [3]Department of Biostatistics and Epidemiology, University of Pennsylvania, 210 Blockley Hall, 423 Guardian Drive, Philadelphia, Pennsylvania 19104, U.S.A.. Correspondence to: Jun Zhou <jun.zhoujun@antgroup.com>.

*Proceedings of the $40^{th}$ International Conference on Machine Learning*, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

cation assumption in DiD methods is the parallel trends (PT) assumption, which assumes that the average outcome for the treatment and control groups would have followed parallel paths over time in the absence of the treatment. However, the PT assumption is fundamentally untestable and often fails in the presence of unmeasured time-varying confounders. One attempt to relax the PT assumption is to use the conditional parallel trends (CPT) assumption, which allows for adjustment of covariate-specific trends. Nonetheless, existing CPT-based methods, such as the outcome regression estimator (Heckman et al., 1997; 1998) and the inverse probability weighting (IPW) estimator (Abadie, 2005), suffer from inconsistency in the presence of model misspecification. Although some more recent work, such as the doubly robust estimator proposed by Zimmert (2018) and Sant'Anna & Zhao (2020) can avoid such a deficiency, they inevitably require estimation of a variety of auxiliary or nuisance parameters (e.g., propensity scores that characterize the treatment assignment mechanism). Moreover, this line of work smooths out individual information and mainly focuses on estimating the average treatment effect in the treated group (ATT) instead of the conditional version.

Even if ATT were our interest, the CPT assumption remains dubious. Roth & Sant'Anna (2020) showed that the parallel trends assumption is sensitive to specific functional forms unless additional structural conditions are imposed. Certain approaches for robust inference and sensitivity analysis are available (*e.g.*, Manski & Pepper, 2018; Keele et al., 2019). The core assumption behind this line of work is that unmeasured confounders that raise violations of parallel trends after the treatment are similar in magnitude to those before the treatment, under which ATT can be partially identified, and uniform confidence intervals can be further obtained (Rambachan & Roth, 2022). However, due to working model assumptions, this line of work cannot be used to estimate conditional ATT (CATT), either.

In this work, we develop a new DiD-based method to identify and estimate CATT. Our contribution is three-fold. First, we allow the violation of CPT such that the difference in trends is assumed to be in a pre-specified function set which can be viewed as an infinite-dimensional version of Rambachan & Roth (2022). Notably, our proposed assumption allows time-varying unmeasured confounders and outcomes, and historical outcomes can also be included as covariates. Second, under our proposed assumption, we further give the partial identification result of CATT in the sense that an identifiable function is close enough to the target CATT. Third, we propose the `DiDTree` to estimate CATT from observational data, whose splitting rule can measure not only the error of fitting observed data, but also the violation of CPT. By adopting such a splitting rule, we can avoid misspecification of outcome regression models and estimation of unnecessary auxiliary parameters (e.g., propensity scores). Moreover, to handle large-scale complex longitudinal datasets, we integrate multiple trees by gradient boosting, which we refer to as Gradient Boosting DiD-Tree.

## 2. Methodology

We first articulate our settings and related notations. Required assumptions will be stated below with discussion.

### 2.1. Problem Setup

Consider the case where we observe outcomes $Y_t \in \mathcal{Y} = \mathbb{R}$ for a unit among $T$ time periods, $t = 1, \ldots, T$. The unit is explained by $p$ covariates $\mathbf{X} \in \mathcal{X} = \mathbb{R}^p$, and will be exposed to a binary treatment, whose status is encoded by $D_t \in \{0, 1\}$. We consider a random design setting; that is, our samples are drawn from some unknown distribution.

**Assumption 1 (Random sampling)** *The observable* $\mathcal{O} = \{(Y_{i,1}, D_{i,1}, \ldots, Y_{i,T}, D_{i,T}, \mathbf{X}_i)\}_{i=1}^n$ *is independently and identically drawn from* $(Y_1, D_1, \ldots, Y_T, D_T, \mathbf{X})$.

Assumption 1 serves as an extension of Sant'Anna & Zhao (2020, Assumption 1(a)) for the case of multiple time periods, which specifies that the panel is balanced with time-invariant covariates [*]. To formalize the problem, additional assumptions are needed. We maintain the classic SUTVA assumption (Rubin, 1980) that no interference between units and no hidden variations of treatments occur. Moreover, we assume that the treatment is imposed at the last time point such that $D_t \equiv 0$ for all $t < T$. Thus, before the treatment, $Y_t = Y_t^{(0)}$ for $t = 1, \ldots, T-1$, and at the post-treatment period ($t = T$), $Y_T = D_T Y_T^{(1)} + (1 - D_T) Y_T^{(0)}$, where $Y_t^{(d)}$ denotes the potential outcome for the treatment $D_t = d \in \{0, 1\}$. The treatment assignment is allowed to depend on time-invariant observed covariates $\mathbf{X}$, unmeasured time-varying or time-invariant confounders $\mathbf{U}_t$, and the nearest historical outcome $Y_{T-1}$. To be rigorous, we impose the following assumption.

**Assumption 2 (Treatment assignment)** $\forall t = 1, \ldots, T-1$, $D_t \equiv 0$. *At the time point* $t = T$, $P(D_T = d \,|\, \mathbf{X}, Y_{T-1}) > 0$, *and* $Y_T^{(d)} \perp\!\!\!\perp D_T \,|\, \big(\mathbf{X}, Y_{T-1}, \{\mathbf{U}_t\}_{t=1}^T\big)$, $d \in \{0, 1\}$.

We include only $Y_{T-1}$ from all historical outcomes as covariates just for better demonstration. It can be readily extended to multiple historical outcomes. Note that it covers the case where historical outcomes have no treatment effects.

Our primary goal is to identify the conditional average treat-

[*]The term "balanced panel data" refers to a dataset in which all individuals are observed for the same length of time and no observations are missing for any subject.

ment effect on the treated group (CATT), defined as

$$\eta^*(\mathbf{x}, y) = E\big(Y_T^{(1)} - Y_T^{(0)} \mid D_T = 1, \mathbf{X} = \mathbf{x}, Y_{T-1} = y\big).$$

To ensure the mathematical well-posedness of $\eta^*(\cdot)$, we require that $E\big(Y_t^2 \mid D_t, \mathbf{X}, Y_{t-1}\big) < \infty$ almost surely for $t = 2, 3, \ldots, T$. For notational simplicity, we omit the subscript of $D_T$ and directly write it as $D$. Traditional difference-in-differences (DiD) estimators hinge on the so-called parallel trends assumption to adjust for unmeasured confounding. To identify CATT, a conditional version of parallel trends is required; that is, for each $\tau = 0, 1, 2, \ldots, T-2$ with $\mathbf{Z}_{T-\tau} = (\mathbf{X}, Y_{T-\tau-1})$,

$$\begin{aligned} &E\big(Y_T^{(0)} \mid D = 1, \mathbf{Z}_T\big) - E\big(Y_{T-\tau}^{(0)} \mid D = 1, \mathbf{Z}_{T-\tau}\big) \\ &= E\big(Y_T^{(0)} \mid D = 0, \mathbf{Z}_T\big) - E\big(Y_{T-\tau}^{(0)} \mid D = 0, \mathbf{Z}_{T-\tau}\big) \end{aligned} \tag{1}$$

holds almost surely. For simplicity, we write $\mathbf{z} = (\mathbf{x}, y) \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Let the conditional difference of control outcomes across treated/control groups be $\Delta_\tau^*(\mathbf{z}) = E\big(Y_{T-\tau}^{(0)} \mid D = 1, \mathbf{Z}_{T-\tau} = \mathbf{z}\big) - E\big(Y_{T-\tau}^{(0)} \mid D = 0, \mathbf{Z}_{T-\tau} = \mathbf{z}\big)$ and also let $m^{*(d)}(\mathbf{z}) = E\big(Y_T^{(d)} \mid D = d, \mathbf{Z}_T = \mathbf{z}\big), d \in \{0, 1\}$. Under (1), our target parameter can be identified by

$$\eta^*(\mathbf{z}) = m^{*(1)}(\mathbf{z}) - m^{*(0)}(\mathbf{z}) - \sum_{\tau=1}^{T-2} \omega_\tau \Delta_\tau^*(\mathbf{z}) \tag{2}$$

for any weight coefficients $\omega_\tau \geq 0, \sum_{\tau=1}^{T-2} \omega_\tau = 1$; see Lemma A.1 for details. From (2), it suffices to estimate functions at the right hand of (2) separately from $\mathcal{O}$, and aggregate them to produce an estimate for CATT. Under a proper choice of $\omega_\tau$s, this estimator can be viewed as a conditional version of the outcome regression estimator in Heckman et al. (1997; 1998).

However, several issues arise from this method. Most fundamentally, the additive form of (2) is sensitive to the conditional parallel trends assumption (1), especially the almost everywhere equality. Although individual-specific and time-specific variations in trends are adjusted for by covariates $\mathbf{X}$ and the adjacent outcome $Y_{T-\tau-1}$, such a parallel trends condition is more likely to hold approximately and in average, but not exactly or almost everywhere. Besides, it requires correct specifying all the models to consistently estimate these functions; any potential misspecification can introduce irreducible biases (Sant'Anna & Zhao, 2020). Moreover, collected samples are essentially divided into several sets for different goals, *i.e.*, $\{\mathbf{Z}_{i,T-\tau}, Y_{i,T-\tau}\}_{i=1}^n$ is only used for $\Delta_\tau^*(\mathbf{z})$ with $\tau = 1, \ldots, T-2$, leading to an inevitable power loss.

### 2.2. Partial Identification of CATT

We consider the case where the conditional parallel trends assumption is violated. We first deal with the issue of the almost everywhere equality (1). The high-level idea is that even though the conditional parallel trends assumption does not hold, we can always find the most common conditional difference before the treatment. That said, to obtain an identification result, we still require that pre-treatment conditional differences across treated/control groups $\Delta_\tau^*(\mathbf{z})$s are transferable to the post-treatment difference $\Delta_0^*(\mathbf{z})$.

**Assumption 3 (Approximate conditional parallel trends)** *For the function*

$$f^*(\cdot) = \underset{f \in \cap_{\tau=1}^{T-2} L_2(P_{T-\tau})}{\arg\min} (T-2)^{-1} \sum_{\tau=1}^{T-2} E_{P_{T-\tau}}\big\{\big|f(\mathbf{Z}) - \Delta_\tau^*(\mathbf{Z})\big|^2\big\}, \tag{3}$$

*where $P_{T-\tau}$ denotes the marginal distribution of $\mathbf{Z}_{T-\tau} = (\mathbf{X}, Y_{T-\tau-1})$, it holds that*

$$\begin{aligned} &\frac{1}{T-2} \sum_{\tau=1}^{T-2} E_{P_{T-\tau}}\big\{\big|f^*(\mathbf{Z}) - \Delta_\tau^*(\mathbf{Z})\big|^2\big\} \equiv \varepsilon_{\textit{history}} \geq 0, \\ &E_{P_T}\big\{\big|f^*(\mathbf{Z}) - \Delta_0^*(\mathbf{Z})\big|^2\big\} \leq \varepsilon_{\textit{history}}. \end{aligned} \tag{4}$$

Here, $\varepsilon_{\text{history}}$ measures the violation of parallel trends during the pre-treatment period, and $f^*(\cdot)$ can be interpreted as the *most common difference* of outcomes across treated/control groups before the treatment. Assumption 3 is essentially requiring that the most common difference during the pre-treatment period continues to capture the conditional difference after the treatment. For $\varepsilon_{\text{history}} = 0$, Assumption 3 is reduced to (1) with $f^* = \Delta_0^*$.

The exact almost-everywhere equality (1) is relaxed to an inequality in expectation, which allows a stronger disagreement with conditional parallel trends for points of smaller probability density. Thus, this relaxation can further adjust for the time-specific variations that introduce nonparallel trends. To understand this, when $P_{T-\tau}$ has the density function $p_{T-\tau}(\mathbf{z})$, under mild conditions, Lemma A.2 shows that $f^*(\mathbf{z})$ has an explicit form

$$f^*(\mathbf{z}) = \frac{1}{\sum_{\tau=1}^{T-2} p_{T-\tau}(\mathbf{z})} \sum_{\tau=1}^{T-2} p_{T-\tau}(\mathbf{z}) \Delta_\tau^*(\mathbf{z}),$$

which takes historical distribution of $\mathbf{Z}_{T-\tau} = (\mathbf{X}, Y_{T-\tau-1})$ into consideration. Recall that $\Delta_\tau^*(\cdot), \tau = 1, \ldots, T-2$ are identifiable from historical data; under Assumptions 1–3, we can also identify the target CATT, but only partially.

**Proposition 1** *Under Assumptions 1–3, $\eta^*(\cdot)$ is partially identified in the sense that $E_{P_T}\{\eta^*(\mathbf{Z}) - \eta^{**}(\mathbf{Z})\}^2 \leq \varepsilon_{\textit{history}}$, where $\eta^{**}(\mathbf{z}) = m^{*(1)}(\mathbf{z}) - m^{*(0)}(\mathbf{z}) - f^*(\mathbf{z})$ is an identifiable function.*

The less different $\Delta_\tau^*$s are, the smaller $\varepsilon_{\text{history}}$ will be, and more accurately our target parameter can be recorved. Moreover, by (4), we can estimate $\varepsilon_{\text{history}}$ from historical data, which can be used to diagnose the validity of our partial identification. In subsequent sections, we consider how to estimate $\eta^{**}(\cdot)$ from finite samples.

### 2.3. Difference-in-Differences Trees

Tree-based methods are very popular in regression and classification tasks and demonstrate an additional superiority in heterogeneous treatment effects estimation, since the splitting procedure of trees can be understood as implicit matching/stratification. A regression tree $T(\mathbf{z}; \mathbf{Q}, \boldsymbol{\mu})$ consists of two components: a set of leafs $\mathbf{Q} = \{Q_1, \ldots, Q_q\}$ that partition the feature space, and the associated parameter $\boldsymbol{\mu} = \boldsymbol{\mu}(\mathbf{Q}) = (\mu(Q_1), \ldots, \mu(Q_q))^T$. Trees iteratively refine their partitions. The partition $\mathbf{Q}$ and its associated parameter $\boldsymbol{\mu}(\mathbf{Q})$ will be refined when such a refinement leads to significant performance improvement. We refer to the mechanism that determines whether a refinement is necessary as a *splitting rule*, *e.g.*, mean squared error for regression trees. The piece-wise constant nature of trees raises the universal approximation ability, and thus we can estimate the common trend $f^*$ and further the surrogate CATT $\eta^{**}$ using tree-based methods without the concern of misspecification. However, it still requires considering how to avoid explicitly estimating nuisance parameters.

We propose a new splitting rule for trees, which is carefully designed to directly estimate the surrogate CATT $\eta^{**}$. We state our procedure in detail with a given partition $\mathbf{Q} = \{Q_j\}_{j=1}^q$ of $\mathcal{Z}$ and $n$ tuples of i.i.d. samples and illustrate how to obtain a refinement of $\mathbf{Q}$. To begin with, we use $\widehat{\mu}_{\tau,j}^{(d)} = \left(1/n_{\tau,j}^{(d)}\right) \sum_{i=1}^n Y_{i,T-\tau} \mathbf{1}\{\mathbf{Z}_{i,T-\tau} \in Q_j, D_i = d\}$ to estimate $E(Y_{T-\tau} \,|\, \mathbf{Z}_{i,T-\tau} \in Q_j, D = d)$ for all $\tau = 0, 1, \ldots, T-2$, where $n_{\tau,j}^{(d)} = |\{i \colon \mathbf{Z}_{i,T-\tau} \in Q_j, D_i = d\}|$, $d = 0, 1$.

**Violation of conditional parallel trends.** We first measure the violation of conditional parallel trends under the given partition $\mathbf{Q}$. Replacing $P_{T-\tau}$ in (3) by the empirical version of the local distribution $P_{T-\tau,j} = P_{T-\tau}\mathbf{1}\{\mathbf{z} \in Q_j\}/P_{T-\tau}(Q_j)$, we then define the empirical risk for violating parallel trends at $Q_j$ as

$$R_{/\!/}(c_j, Q_j) = \frac{1}{T-2} \sum_{\tau=1}^{T-2} \sum_{i \colon \mathbf{Z}_{i,T-\tau} \in Q_j} \left| c_j - \widehat{\Delta}_\tau(\mathbf{Z}_{i,T-\tau}) \right|^2,$$

where $c_j$ is a constant to be determined and $\widehat{\Delta}_\tau(\mathbf{Z}_{i,T-\tau})$ denotes an imputed value of the pretreatment difference $\Delta_\tau(\mathbf{Z}_{i,T-\tau}) = E(Y_{T-\tau}|D = 1, \mathbf{Z}_{i,T-\tau}) - E(Y_{T-\tau}|D = 0, \mathbf{Z}_{i,T-\tau})$. To estimate $\Delta_\tau(\mathbf{Z}_{i,T-\tau})$, we adopt the effective idea of cross-imputation which performs well under imbalanced samples (Künzel et al., 2019), and impute the individualized value by $\widehat{\mu}_{\tau,j}^{(1)} - Y_{i,T-\tau}$ for the control group and $Y_{i,T-\tau} - \widehat{\mu}_{\tau,j}^{(0)}$ for the treated group, i.e., $(-1)^{D_i}\{\widehat{\mu}_{\tau,j}^{(1-D_i)} - Y_{i,T-\tau}\}$. Thus, we write $R_{/\!/}(c_j, Q_j)$ as

$$\begin{aligned} R_{/\!/}(c_j, Q_j) = & \frac{1}{T-2} \sum_{\tau=1}^{T-2} \sum_{i \colon \mathbf{Z}_{i,T-\tau} \in Q_j} \\ & \left| c_j - (-1)^{D_i}\{\widehat{\mu}_{\tau,j}^{(1-D_i)} - Y_{i,T-\tau}\} \right|^2, \end{aligned} \tag{5}$$

and we estimate $f^*$ by $\widehat{f}(\mathbf{z}) = \sum_{j=1}^q \mathbf{1}\{\mathbf{z} \in Q_j\}\widehat{c}_j$ with $\widehat{c}_j = \arg\min_{c_j \in \mathbb{R}} R_{/\!/}(c_j, Q_j)$. Moreover, the overall violation of parallel trends is measured by

$$R_{/\!/}(\mathbf{Q}) = \sum_{j=1}^q R_{/\!/}(\widehat{c}_j, Q_j). \tag{6}$$

A larger $R_{/\!/}(\mathbf{Q})$ suggests that the conditional trends are less parallel under the partition $\mathbf{Q}$.

**Error of fitting observational data.** We then measure the negative fidelity to the observed data at post-treatment periods under the given partition $\mathbf{Q}$. To give the final estimate for $\eta^{**}(\cdot)$, we also need to fit $Y_{i,T}^{(d)}$ using $\mathbf{Z}_{i,T}$ with $D_i = d$ for $d = 0, 1$. Choosing the loss function as $\ell(\cdot, \cdot)$, we then define the error of fitting data under the partition $\mathbf{Q}$ as

$$R_{\text{data}}(\mathbf{Q}) = \sum_{j=1}^q \sum_{i \colon X_i \in Q_j} \ell(\widehat{\mu}_{0,j}^{(1)} D_i + (1-D_i)\widehat{\mu}_{0,j}^{(0)}, Y_{i,T}) \tag{7}$$

A larger $R_{\text{data}}(\mathbf{Q})$ suggests under-fitting under the partition $\mathbf{Q}$.

The DiD Tree splitting rule is to find a partition $\widehat{\mathbf{Q}}$ iteratively, *i.e.*,

$$\widehat{\mathbf{Q}} \in \arg\min_{\mathbf{Q}} R_{/\!/}(\mathbf{Q}) + \lambda R_{\text{data}}(\mathbf{Q}), \tag{8}$$

where $\lambda$ is a positive hyper-parameter that balances the importance of data fidelity and violation of parallel trends. In summary, the splitting rule includes two terms, one measures the violation of conditional parallel trends, and the other measures the error of fitting observed data. The minimization of the former gives an accurate estimate of the common trend $f^*$, while the minimization of the latter leads to accurate estimates of $m^{*(d)}, d = 0, 1$. As a result, minimizing both terms can directly estimate $\eta^{**}$ and gives better performance.

### 2.4. Ensemble of Difference-in-Differences Trees

This section will introduce how to integrate multiple DiD-trees under the boosting framework. In general, the boosting

of $k$ trees are defined as additive form: $F(Z;\boldsymbol{\mathcal{Q}}_k,\boldsymbol{U}_k)=\sum_{i=1}^{k}T(Z;\boldsymbol{Q}_i,\boldsymbol{\mu}_i)$, where $\boldsymbol{\mathcal{Q}}_k=\{\boldsymbol{Q}_1,\ldots,\boldsymbol{Q}_k\}$ denotes the set of partitions of each tree, and $\boldsymbol{U}_k=\{\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_k\}$ denotes the set of corresponding parameters. At the $k$-th iteration, we have already obtained $k-1$ trees, $F(Z;\boldsymbol{Q}_{k-1};\boldsymbol{U}_{k-1}^{d})$, to fit $E[Y_T|Z=z,D=d]$ $d\in\{0,1\}$ and $k-1$ trees, $F(Z;\boldsymbol{Q}_{k-1};\boldsymbol{U}_{k-1,\tau}^{d})$, to fit $\{E[Y_{T-\tau}|Z=z,D=d]\}_{\tau=1}^{T-1}$. The former set of trees are used for learning the potential post-treatment outcomes and the latter set of trees focus on ensuring the concordance of parallel trends in pre-treatment periods. We remark that they share the same splitting regions.

**Violation of conditional parallel trends for multiple trees.** Similar to a single tree, we also develop a metric to measure the violation of conditional parallel trends in the multiple trees case. At $k$-th iteration, if we have obtained $k-1$ trees $F(Z_i;\boldsymbol{\mathcal{Q}}_{k-1},\boldsymbol{U}_{k-1,\tau}^{(1-D_i)})$, we then measure the violation of conditional parallel trends with respect to the residual, e.g., $\widetilde{Y}_{i,T-\tau}=Y_{i,T-\tau}-F(Z_i;\boldsymbol{\mathcal{Q}}_{k-1},\boldsymbol{U}_{k-1,\tau}^{(1-D_i)})+(-1)^{D_i}\widehat{C}_{k-1,i}$, where $\widehat{C}_{k-1,i}=\sum_{r=1}^{k-1}\widehat{c}_{r,i}$, is the individual gap learned from the previous $k-1$ trees. We then by replacing $\widehat{\mu}_{\tau,j}^{(1-d)}$ with $\widehat{\mu}_{\tau,k,j}^{(1-d)}$ in (5), define the empirical risk for violating parallel trends at $k$-th iteration of the $j-$th partition $Q_{k,j}$ as

$$\mathcal{R}_{/\!/}(c_{k,j},Q_{k,j})=\sum_{i:Z_i\in Q_{k,j}}\sum_{\tau=1}^{T-1}\Big\{c_{k,j}-(-1)^{D_i}\big(\widehat{\mu}_{\tau,k,j}^{(1-D_i)}-\widetilde{Y}_{i,T-\tau}\big)\Big\}^2, \tag{9}$$

and we estimate the $\widehat{\mu}_{\tau,k,j}^{(1-d)}$ by $E(\widetilde{Y}_{T-\tau}|D=d,\mathbf{Z}_{i,T-\tau})$ through the cross-imputation idea as well. Given a partition of $k$-th tree $\boldsymbol{Q}_k$, the overall violation of parallel trends is measured by

$$\mathcal{R}_{/\!/}(\boldsymbol{Q}_k)=\sum_{j=1}^{|\boldsymbol{Q}_k|}\mathcal{R}_{/\!/}(\widehat{c}_{k,j},Q_{k,j}) \tag{10}$$

where $\widehat{c}_{k,j}=\arg\min_c\mathcal{R}_{/\!/}(c,Q_{k,j})$.

**The objective function of multiple trees.** In addition to the violation of parallel trends loss ($\mathcal{R}_{/\!/}(Q_k)$), under the given partition $\mathbf{Q}_k$, we also should consider the data fidelity in the post-treatment period. We have used $k-1$ trees to estimate the $E(Y_{i,T}|Z_i,D_i)$, e.g., $\widehat{Y}_{i,T}=\sum_{d=0}^{1}\mathbf{1}(d=D_i)F(Z;\boldsymbol{\mathcal{Q}}_{k-1};\boldsymbol{U}_{k-1,T}^{(d)})$. Then, at the next iteration, the loss of the fidelity of observational data is

$$\mathcal{R}_{\text{data}}(\boldsymbol{Q}_k)=\sum_{j=1}^{|\boldsymbol{Q}_k|}\sum_{i:Z_i\in Q_j}\ell(\widehat{Y}_{i,T}+\widehat{y}_{i,T},Y_{i,T}),$$

where $\widehat{y}_{i,T}$ is the prediction of current tree, like $\widehat{\mu}_{0,j}^{(1)}D_i+(1-D_i)\widehat{\mu}_{0,j}^{(0)}$. Finally, the optimal partition of $k$-th tree, $\mathbf{Q}_k^*$, is obtained by minimizing the trade-off between the violation of parallel trends loss and the data fidelity loss by as follows,

$$\mathbf{Q}_k^*=\arg\min_{\mathbf{Q}}\mathcal{R}_{/\!/}(\mathbf{Q})+\lambda\mathcal{R}_{\text{data}}(\mathbf{Q}). \tag{11}$$

## 3. Experiments

In this section, we conduct experiments on both simulated and real-world datasets to verify the effectiveness of our method. We evaluate DiD Causal Tree against state-of-the-art causal inference algorithms (All the baselines' code are implemented by third-party libraries [†]: (i) meta-learners (Künzel et al., 2019) including `TLearner` and `SLearner`; (ii) causal forests, which are forest-based methodologies to model the treatment effect, including generalized causal forest (abbr. `GRF`, Athey et al., 2019) and bayesian causal forest (abbr. `BCF`, Hahn et al., 2020); (iii) naive-DiD, which uses $2\times T$ regression models to fit $f(X;\theta_t^{(d)})=E(Y_t^{(d)}|X),1\le t\le T,d\in\{0,1\}$, and then estimates CATE by the naive difference-in-differences estimator, $f(X;\theta_T^{(1)})-f(X;\theta_T^{(0)})-\frac{1}{T-1}\sum_{t=1}^{T-1}f(X;\theta_T^{(1)})-f(Z;\theta_T^{(0)})$; (iv) balanced representation methods including `CFR-MMD` and `CFR-WASS` (Johansson et al., 2016; Shalit et al., 2017), which learn a latent representation that balances the distributions of the treated and control groups.

For baselines, both covariates $X$ and pre-treatment outcomes are used as input features. In contrast, our proposed `DiDTree` employs an alternative way of exploiting historical controls, *i.e.*, fitting models on $X$ while treating pre-treatment outcomes as labels to measure the violation of conditional parallel trends. All the hype parameters are the same. For example, the max number of trees in ensemble models (including boosting and bagging) is 500, the subsample ratios of instance and feature are 0.8 and 0.8, and the learning rate is 0.05. The max depth of each tree in forest-based (`GRF` and `BCF`) and boosting-based (meta-learners and `DiDTree`) methods are 10 and 3 respectively, where it is worth noting that the trees in random forests are generally deeper due to the bagging and boosting frameworks' respective characteristics.

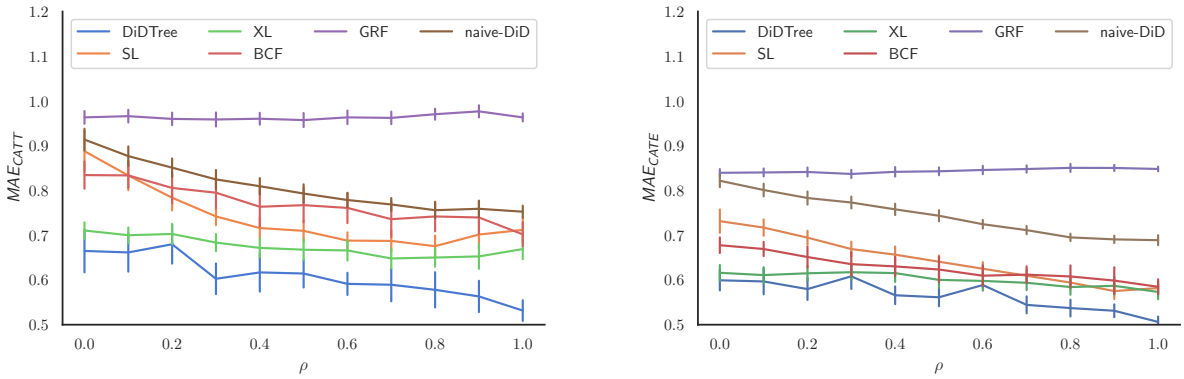
### 3.1. Simulation Data

**Data generation process** We introduce the generation process of the simulation data. The covariates consist of two parts: (i) time-invariant covariates represented by

[†]The BCF is from https://github.com/socket778/XBCF, GRF is from https://github.com/grf-labs/grf and the others are from https://github.com/microsoft/EconML.

*Table 1.* The mean absolute error MAE (mean± s.d.) of each algorithm on simulation data. Scenario I represents the presence of unmeasured confounding, whereas II represents the absence of unmeasured confounding.

| | $\varphi$ | DiDs | | Meta-Learners | | Causal Forests | | Balanced Representation | |
|---|---|---|---|---|---|---|---|---|---|
| | | DiDTree | Naive-DiD | S | X | BCF | GRF | CFR-MMD | CFR-WASS |
| I | 0.05 | $0.95 \pm 0.06$• | $1.30 \pm 0.06$ | $1.27 \pm 0.05$ | $1.32 \pm 0.09$ | $1.21 \pm 0.09$ | $1.84 \pm 0.06$ | $2.92 \pm 0.11$ | $3.04 \pm 0.07$ |
| | 0.1 | $0.93 \pm 0.04$• | $1.14 \pm 0.06$ | $1.22 \pm 0.06$ | $1.07 \pm 0.06$ | $1.10 \pm 0.08$ | $1.59 \pm 0.05$ | $1.93 \pm 0.10$ | $2.66 \pm 0.09$ |
| | 0.5 | $0.71 \pm 0.04$• | $0.90 \pm 0.03$ | $1.05 \pm 0.02$ | $0.73 \pm 0.01$ | $0.77 \pm 0.02$ | $1.22 \pm 0.03$ | $1.05 \pm 0.10$ | $1.30 \pm 0.07$ |
| | 0.9 | $0.72 \pm 0.03$• | $1.09 \pm 0.03$ | $1.02 \pm 0.04$ | $0.78 \pm 0.02$ | $0.80 \pm 0.04$ | $1.26 \pm 0.06$ | $0.86 \pm 0.07$ | $1.67 \pm 0.11$ |
| | 0.95 | $0.76 \pm 0.04$• | $1.20 \pm 0.06$ | $1.09 \pm 0.05$ | $0.86 \pm 0.07$ | $0.84 \pm 0.06$ | $1.45 \pm 0.07$ | $1.01 \pm 0.08$ | $1.84 \pm 0.12$ |
| II | 0.05 | $0.91 \pm 0.04$• | $1.28 \pm 0.05$ | $1.18 \pm 0.05$ | $1.21 \pm 0.09$ | $1.19 \pm 0.13$ | $1.17 \pm 0.02$ | $2.93 \pm 0.08$ | $3.07 \pm 0.07$ |
| | 0.1 | $0.90 \pm 0.04$• | $1.13 \pm 0.06$ | $1.10 \pm 0.04$ | $0.98 \pm 0.06$ | $1.00 \pm 0.07$ | $1.12 \pm 0.02$ | $1.81 \pm 0.12$ | $2.56 \pm 0.19$ |
| | 0.5 | $0.65 \pm 0.02$• | $0.88 \pm 0.01$ | $0.87 \pm 0.02$ | $0.66 \pm 0.01$ | $0.69 \pm 0.02$ | $1.01 \pm 0.01$ | $0.83 \pm 0.03$ | $1.30 \pm 0.07$ |
| | 0.9 | $0.69 \pm 0.02$• | $1.07 \pm 0.03$ | $0.89 \pm 0.03$ | $0.83 \pm 0.03$ | $0.77 \pm 0.03$ | $0.99 \pm 0.02$ | $0.88 \pm 0.08$ | $1.61 \pm 0.21$ |
| | 0.95 | $0.76 \pm 0.04$• | $1.24 \pm 0.07$ | $0.94 \pm 0.05$ | $0.98 \pm 0.08$ | $0.84 \pm 0.07$ | $1.01 \pm 0.02$ | $1.09 \pm 0.14$ | $1.80 \pm 0.10$ |




(a) presence of unmeasured confounders (b) absence of unmeasured confounders

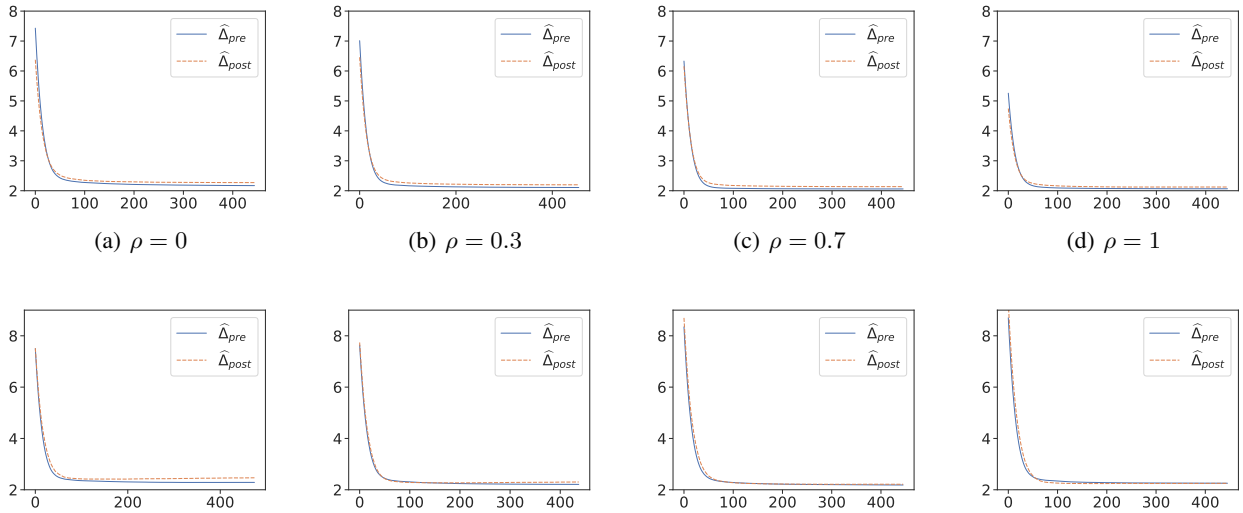
*Figure 1.* The mean absolute error MAE (mean with 2 × s.d. error bars) of each algorithm on multiple simulation datasets. The left sub-figure represents the presence of unmeasured confounding, whereas the right represents the absence of unmeasured confounding.

$X = (\widetilde{X}, U)$, and (ii) unmeasured time factors represented by $\Lambda_t$. The $X$ is a $p$-dimensional vector (the dimensions of $X, U$ are $p_{\tilde{x}}, p_u$ respectively) generated by one of the $S$ Gaussian distributions and the hidden variable $g$ indicates which group they belong to, that is $X \sim \sum_{s=1}^{S} \mathbb{I}(g = s)\mathcal{N}(\mu_s, \Sigma_s)$. To add to that, $\widetilde{X}$ represents covariates that can definitely be observed, while $U$ represents covariates that may not be observable. The time-varying factors $\Lambda_t$, such as time and holidays, etc., are independent of individuals. We then generate the potential outcomes by the following processes, for each time step $1 \leq t \leq T$:

$$\begin{aligned} Y_t^0 &= \alpha(X) + \lambda(\Lambda_t) + \upsilon_t + \epsilon, \\ \upsilon_t &= \rho\upsilon_{t-1} + (1-\rho)\upsilon(X, \Lambda_t), \\ Y_T^1 &= Y_T^0 + \tau(X) + \epsilon_1, \end{aligned} \tag{12}$$

where $\alpha, \lambda, \upsilon, \tau$ are individual-specific effects function, time-specific effects function, individual-transitory effects function, and treatment effects function, respectively, and $\epsilon \sim \mathcal{N}(0, 2), \epsilon_1 \sim \mathcal{N}(0, 0.3)$ are zero-mean Gaussian noise. To ensure the samples are drawn from the stationary distribution of the data-generating process, we generate 100 time steps and choose the latest $T$ time steps as observations. Besides, to control the parallelism between the treated and control groups concerning time, we leverage an exponential moving average form, by a coefficient parameter $\rho \in [0, 1]$, to smooth the individual time-varying term. In particular, $\rho = 1$ means the trend of all individuals over time is homogeneous, that is, strictly meeting the conditional parallel trends assumption. We set $S = 2, k = 4, p_{\tilde{x}} = 15, p_u = 5$ and generate a total of 20000 instances which are randomly split into training and validation sets by 10 times. The assignment of treatment is by a function $ps(X, \varphi)$, where $\varphi$ controls the ratio of the treated instances, *i.e.*, $\varphi = \frac{\#control}{\#treated + \#control}$.

**Evaluation on heterogeneous treatment effect** For heterogeneous treatment effect evaluation, we report the mean




(a) $\rho = 0$ (b) $\rho = 0.3$ (c) $\rho = 0.7$ (d) $\rho = 1$

(e) $\varphi = 0.05$ (f) $\varphi = 0.2$ (g) $\varphi = 0.8$ (h) $\varphi = 0.95$

*Figure 2.* The parallelism error on the validation data under different numbers of trees in the unmeasured confounding scenario.

absolute error (MAE) of the conditional average treatment effects (CATE) when all confounders are observed, *i.e.*,

$$\mathrm{MAE}_{\mathrm{CATE}} = \sum_{i=1}^{n} |\hat{\eta}(Z_i) - \eta_i|/n,$$

where $\eta_i$ is the true treatment effect of $i$-th unit. In addition to the absence of unmeasured confounders scenario, the MAE of conditional average treatment effects on the treated group (CATT) is also reported when not all confounders are fully observed, *i.e.*,

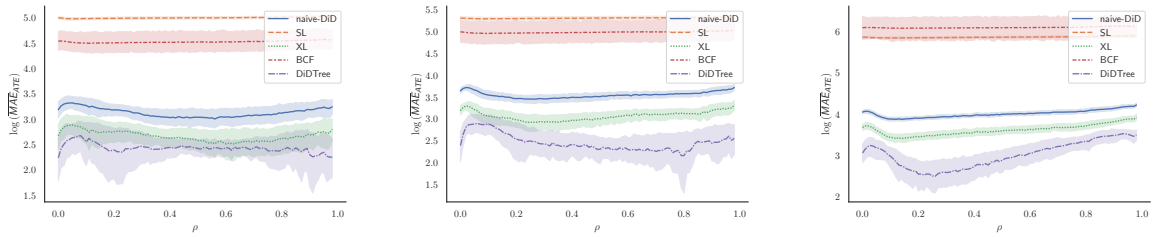
$$\mathrm{MAE}_{\mathrm{CATT}} = (\sum_{i=1}^{n} D_i)^{-1} \sum_{i=1}^{n} D_i \,|\, \hat{\eta}(Z_i) - \eta_i \,|\,.$$

In practice, especially when the number of observed instances is limited, the imbalance between the treated and control populations usually affects the performance of estimation methods. We vary the parameter $\varphi \in \{0.05, 0.1, 0.5, 0.9, 0.95\}$ to generate datasets from extreme imbalance ($\varphi = 0.05$ or $0.95$) to complete balance ($\varphi = 0.5$), and we set the parameter $\rho = 0$ (without exponential moving average). First, we compare our method with baselines on the series of datasets when not all confounders are observed; we report the results in scenario I of Table 1. For estimation and evaluation, we discard the 'unmeasured' confounding vector (e.g., $U$) and retain the rest (e.g., $\widetilde{X}$). Second, we also consider the scenario where all confounders are observed (including $\widetilde{X}$ and $U$), which is shown in scenario II of Table 1. As the degree of imbalance in treatment allocation gradually increases, except for `DiDTree`, other methods generally lead to severe bias and significant performance degradation. We reach the same conclusion in both scenarios I and II. The results indicate that `DiDTree` is less sensitive to the imbalance between the treated and control populations.

**Evaluation on the violation of conditional parallel trends** The core idea of this paper depends on the approximate conditional parallel trends that the parallelism in pre-treatment periods can be transferred to the post-treatment period. To assess the robustness of our method, we conduct the following two experiments.

In the first experiment, we gradually increase the parameter $\rho$ from 0 to 1. Especially when $\rho = 1$, the generated datasets strictly satisfy the conditional parallel trends assumption; when $\rho = 0$, it degenerates into an ordinary time series relationship. On a series of data sets generated in this way, we compare each method in both the absence of unmeasured confounding and the presence of unmeasured confounding. The result of all methods is shown in Figure 1. We can see that: (i) `DiDTree` outperforms others baselines in all scenarios; (ii) as $\rho$ gradually approaches 1, the error of all methods except `GRF` is generally decreasing and in particular, the performance of `DiDTree` has improved most significantly with the increase of $\rho$; (iii) the results demonstrate the important role of the parallelism loss ($R_{//}$) in making the estimates less sensitive to the violation of conditional parallel trends.

In the second experiment, we demonstrate whether the conditional parallel trends obtained by minimizing the parallel loss ($R_{//}$) in pre-treatment periods, can be extended to the post-treatment period. Specifically, if we have trained




(a) treatment $= 2k$ (b) treatment $= 3k$ (c) treatment $= 6k$

*Figure 3.* The result of MAE on the real-world dataset. The X-axis is the threshold of propensity score varying from 0 to 1 and Y-axis is the logarithmic $\text{MAE}_{\text{ATE}}$.

$k$ DiD causal trees, $F(X; \boldsymbol{\mathcal{Q}}_k, \boldsymbol{U}_k^{(d)})$, we consider the $\widehat{Y}_{i,t} = F(Z_i; \mathcal{Q}_k, \mathbf{U}_t^{1-D_i})$ as its parallel prediction and $\widehat{C}_{k,i}$ as the individual gap. We use the parallel loss in (9) to measure the degree of violation of conditional parallel trends in pre-treatment periods, *i.e.*, $\widehat{\Delta}_{pre} = n^{-1}\mathcal{R}_{//}(\boldsymbol{Q}_k)$ for the $k$-th tree. Similarly, we use $\widehat{\Delta}_{post}$ to measure the degree of violation of conditional parallel trends in post-treatment periods. It should be noted that, in the post-treatment periods, the gap between the treated and control groups also includes treatment effects. The $\widehat{\Delta}_{post}$ can be calculated by $\widehat{\Delta}_{post} = \left\{n^{-1}\sum_{i=1}^{n}\left[(-1)^{D_i}(\widehat{Y}_{i,T} - Y_{i,T}) - \widehat{C}_{k,i} - \eta_i\right]^2\right\}^{\frac{1}{2}}$. Specifically, we conduct the experiments, recording the $\widehat{\Delta}_{pre}$ and $\widehat{\Delta}_{post}$ at each iteration, to validate it on the simulation datasets, since the parallelism in post-treatment periods is unavailable for real-world data. The result shown in Figure 2 demonstrates that as the iteration increases, the parallelism before and after treatment would decrease simultaneously. This fully demonstrates that Assumption 3, *i.e.*, the nearly common trends during the pre-treatment period continues to exist after the treatment, is reasonable and the proposed loss $\mathcal{R}_{//}$ is effective in partitioning sub-spaces in which the approximate conditional parallel trend holds.

### 3.2. Real-world Data

The dataset comes from a randomized controlled trial (RCT) by a commercial finance company aimed at assessing users' heterogeneous responses to increasing credit lines of credit card [‡](Tang et al., 2022). The trial employs a stratified random assignment design with strata based on risk, dividing users into low-risk and medium-risk. Within each stratum, users are randomly assigned to one of four treatment groups: increasing credit line by 0, 2000, 3000, or 6000 converted to some currency, where 0 corresponds to the control group. In order to assess the performance of different methods in the real-world scenario, we artificially construct biased observational data by retaining only medium-risk users with no credit line increase and low-risk users with credit line increase. To assess variability, we randomly split the entire samples into two folds and repeat this step ten times, each time using one fold as unbiased test data and the others as the training data. See Table 2 for details of the dataset.

One can not evaluate the MAE of CATE directly, since there is no ground truth of the individual level in real-world applications. Considering the treated instances are from the RCT, we can approximately compute the true ATE, denoted as $\eta_{\text{ATE}}$. In light of this, to evaluate the quality of CATE estimation, we use propensity score to stratify individuals and then evaluate the performance of methods on MAE of ATE in each stratum, *i.e.*, $\text{MAE}_{\text{ATE}}(H) = \left| |H|^{-1}\sum_{i\in H}\hat{\eta}(Z_i) - \eta_{\text{ATE}}\right|$, where $H$ denotes the set of units. Specifically, given a threshold $\Psi \in [0,1]$ and propensity score $\psi_i$, we split the test dataset into two subgroups: $H_{>\Psi} = \{i \,|\, \psi_i > \Psi\}$ and $H_{\leq\Psi} = \{i \,|\, \psi_i \leq \Psi\}$, and then report the weighted mean of $\text{MAE}_{\text{ATE}}$ on two subgroups, *i.e.*, $\overline{\text{MAE}}_{\text{ATE}} = \frac{|H_{>\Psi}|}{|H_{>\Psi}|+|H_{\leq\Psi}|}\text{MAE}_{\text{ATE}}(H_{>\Psi}) + \frac{|H_{\leq\Psi}|}{|H_{\leq\Psi}|+|H_{>\Psi}|}\text{MAE}_{\text{ATE}}(H_{\leq\Psi})$; the results are summarized in Figure 3. It is impressive to observe that `DiDTree` significantly outperforms the benchmarks in terms of mean error and standard deviation. These findings are in line with those of the prior simulation studies.

## 4. Conclusion

In this work, we develop a novel DiD-motivated framework to identify and estimate CATT with observational panel data: (i) we propose a new splitting rule to partition feature space into multiple sub-spaces guaranteeing each of them satisfy the approximate conditional parallel trends; (ii) we then estimate the conditional treatment effect by a DiD-based estimator in each sub-space. Our method allows the

[‡]1. The data set does not contain any Personal Identifiable Information (PII); 2. The data set is desensitized and encrypted;3. Adequate data protection was carried out during the experiment to prevent the risk of data copy leakage, and the data set was destroyed after the experiment;4. The data set is only used for academic research, it does not represent any real business situation.

violation of conditional parallel trends and the presence of time-varying unmeasured confounders. These advantages enable our methods to be widely used in many complex real-world scenarios. Nonetheless, our method may not fully exploit the time structures of time-varying covariates, which we plan to address in future work.

## References


Abadie, A. Semiparametric difference-in-difference estimators. *Review of Economic Studies*, 72:1–19, 2005.

Athey, S., Tibshirani, J., and Wager, S. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019.

Breslow, L. and Johnson, M. California's proposition 99 on tobacco, and its impact. *Annual Review of Public Health*, 14(1):585–604, 1993.

Gilotte, A., Calauzènes, C., Nedelec, T., Abraham, A., and Dollé, S. Offline A/B testing for recommender systems. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pp. 198–206, 2018.

Glass, T. A., Goodman, S. N., Hernán, M. A., and Samet, J. M. Causal inference in public health. *Annual review of public health*, 34:61–75, 2013.

Greenland, S. and Robins, J. M. Identifiability, exchangeability, and epidemiological confounding. *International Journal of Epidemiology*, 15(3):413–419, 1986.

Hahn, P. R., Murray, J. S., and Carvalho, C. M. Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, 15(3):965–1056, 2020.

Heckman, J. J., Ichimura, H., and Todd, P. E. Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Review of Economic Studies*, 64:605–654, 1997.

Heckman, J. J., Ichimura, H., Smith, J. A., and Todd, P. E. Characterizing selection bias using experimental data. *Econometrica*, 66:1017–1098, 1998.

Imbens, G. W. and Rubin, D. B. *Causal inference for statistics, social, and biomedical sciences: An introduction*. Taylor & Francis, 2016.

Johansson, F., Shalit, U., and Sontag, D. Learning representations for counterfactual inference. In *International Conference on Machine Learning*, pp. 3020–3029, 2016.

Keele, L. J., Small, D. S., Hsu, J. Y., and Fogarty, C. B. Patterns of ef- fects and sensitivity analysis for differences-in-differences. *arXiv:1901.01869*, 2019.

Keller, M. C. Gene × environment interaction studies have not properly controlled for potential confounders: The problem and the (simple) solution. *Biological Psychiatry*, 75(1):18–24, 2014. Temperament: Genetic and Environmental Factors.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019.

Manski, C. F. and Pepper, J. V. How do right-to-carry laws affect crime rates? coping with ambiguity using bounded-variation assumptions. *Review of Economics and Statistics*, 100(2):232–244, 2018.

Rambachan, A. and Roth, J. A more credible approach to parallel trends. *Review of Economic Studies*, Forthcoming, 2022.

Robin, J. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9-12):1393–1512, 1986.

Roth, J. and Sant'Anna, P. H. When is parallel trends sensitive to functional form? *arXiv preprint arXiv:2010.04814*, 2020.

Rubin, D. B. Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593, 1980.

Sant'Anna, P. H. C. and Zhao, J. Doubly robust difference-in-differences estimators. *Journal of Econometrics*, 219 (1):101–122, 2020.

Shalit, U., Johansson, F. D., and Sontag, D. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pp. 3076–3085, 2017.

Tang, C., Wang, H., Li, X., Cui, Q., Zhang, Y.-L., Zhu, F., Li, L., Zhou, J., and Jiang, L. Debiased causal tree: Heterogeneous treatment effects estimation with unmeasured confounding. In *Advances in Neural Information Processing Systems*, 2022.

Tchetgen Tchetgen, E. J., Michael, H., and Cui, Y. Marginal structural models for time-varying endogenous treatments: A time-varying instrumental variable approach. *arXiv e-prints*, pp. arXiv–1809, 2018.

Wacholder, S. When measurement errors correlate with truth: Surprising effects of nondifferential misclassification. *Epidemiology*, 6(2):157–161, 1995. ISSN 10443983.

Ying, A., Miao, W., Shi, X., and Tchetgen, E. J. T. Proximal causal inference for complex longitudinal studies. *arXiv preprint arXiv:2109.07030*, 2021.

Zimmert, M. Efficient difference-in-differences estimation with high-dimensional common trend confounding. *arXiv preprint arXiv:1809.01643*, 2018.

# A. Proofs of theoretical results

**Lemma A.1** *Under Assumptions 1 and 2, if further (1) holds, then for any weights coefficients $(\omega_1, \ldots, \omega_{T-2})^T$ such that $\omega_\tau \geq 0, \sum_{\tau=1}^{T-2} \omega_\tau = 1$, we obtain (2).*

**Proof 1 (Proof of Lemma A.1)** *By the definition of $m^{*(d)}, d = 0, 1$, the right hand of Equation 2 is*

$$\begin{aligned}
&m^{*(1)}(\mathbf{z}) - m^{*(0)}(\mathbf{z}) - \sum_{\tau=1}^{T-2} \omega_\tau \Delta_\tau^*(\mathbf{z}) \\
&= E[Y_T^{(1)}|D = 1, \mathbf{Z}_T = \mathbf{z}] - E[Y_T^{(0)}|D = 0, \mathbf{Z}_T = \mathbf{z}] - \sum_{\tau=1}^{T-2} \omega_\tau \Delta_\tau^*(\mathbf{z}) \\
&= E[Y_T^{(1)}|D = 1, \mathbf{Z}_T = \mathbf{z}] - E[Y_T^{(0)}|D = 1, \mathbf{Z}_T = \mathbf{z}] \\
&+ E[Y_T^{(0)}|D = 1, \mathbf{Z}_T = \mathbf{z}] - E[Y_T^{(0)}|D = 0, \mathbf{Z}_T = \mathbf{z}] - \sum_{\tau=1}^{T-2} \omega_\tau \Delta_\tau^*(\mathbf{z}) \\
&\equiv \eta^*(\mathbf{z}) + \Delta_0^*(\mathbf{z}) - \sum_{\tau=1}^{T-2} \omega_\tau \Delta_\tau^*(\mathbf{z}).
\end{aligned}$$

*Under the conditional parallel trends assumption, Equation 1 holds and thus $\Delta_0^*(\mathbf{z}) = \Delta_1^*(\mathbf{z}) = \cdots = \Delta_{T-2}^*(\mathbf{z})$. Choosing $\omega_1, \ldots, \omega_{T-2}$ to be any non-negative constants with $\sum_{\tau=1}^{T-2} \omega_\tau = 1$, we conclude that*

$$m^{*(1)}(\mathbf{z}) - m^{*(0)}(\mathbf{z}) - \sum_{\tau=1}^{T-2} \omega_\tau \Delta_\tau^*(\mathbf{z}) = \eta^*(\mathbf{z}) + \Delta_0^*(\mathbf{z}) - \sum_{\tau=1}^{T-2} \omega_\tau \Delta_0^*(\mathbf{z}) = \eta^*(\mathbf{z}),$$

*completing the proof.*

**Lemma A.2** *Suppose that each $P_{T-\tau}$, the marginal distribution of $\mathbf{Z}_{T-\tau} = (\mathbf{X}, Y_{T-\tau-1})$, has the density function $p_{T-\tau}(\mathbf{z})$ with respect to the Lebesgue measure on $\mathcal{X} \times \mathcal{Y}$ for $\tau = 0, 1, \ldots, T-1$. Then, the most common difference of control outcomes across treated/control groups before the treatment (i.e., $f^*$ defined in (3)) has an explicit form*

$$f^*(\mathbf{z}) = \frac{1}{\sum_{\tau=1}^{T-2} p_{T-\tau}(\mathbf{z})} \sum_{\tau=1}^{T-2} p_{T-\tau}(\mathbf{z}) \Delta_\tau^*(\mathbf{z}),$$

*almost surely with respect to the distribution $(T-2)^{-1} \sum_{\tau=1}^{T-2} P_{T-\tau}$.*

**Proof 2 (Proof of Lemma A.2)** *Let*

$$\mathcal{L}(f) = \frac{1}{T-2} \sum_{\tau=1}^{T-2} E_{P_{T-\tau}} \left\{ \left| f(\mathbf{Z}) - \Delta_\tau^*(\mathbf{Z}) \right|^2 \right\}.$$

*For notational simplicity, we write $\langle f, g \rangle_{P_{T-\tau}} = E_{P_{T-\tau}}\{f(\mathbf{Z})g(\mathbf{Z})\} = \int_{\mathcal{Z}} f(\mathbf{z})g(\mathbf{z})p_{T-\tau}(\mathbf{z})d\mathbf{z}$. Then, $\mathcal{L}(f) = (T-2)^{-1} \sum_{\tau=1}^{T-2} 2\langle f - \Delta_\tau^*, f - \Delta_\tau^* \rangle_{P_{T-\tau}}$. Then, fix $f \in \cap_{\tau=1}^{T-2} L_2(P_{T-\tau})$, and choose any $\delta \in \cap_{\tau=1}^{T-2} L_2(P_{T-\tau})$, we obtain*

$$(T-2)\big\{\mathcal{L}(f+\delta) - \mathcal{L}(f)\big\} = \sum_{\tau=1}^{T-2} \Big\{ \langle f - \Delta_\tau^*, \delta \rangle_{P_{T-\tau}} + \langle \delta, \delta \rangle_{P_{T-\tau}} \Big\}.$$

*If we let $P_{history} = (T-2)^{-1} \sum_{\tau=1}^{T-2} P_{T-\tau}$, we can further write*

$$\mathcal{L}(f+\delta) - \mathcal{L}(f) = \int_{\mathcal{Z}} \delta(\mathbf{z}) \frac{2}{T-2} \sum_{\tau=1}^{T-2} \big(f(\mathbf{z}) - \Delta_\tau^*(\mathbf{z})\big) p_{T-\tau}(\mathbf{z}) d\mathbf{z} + \int_{\mathcal{Z}} \delta^2(\mathbf{z}) \frac{1}{T-2} \sum_{\tau=1}^{T-2} p_{T-\tau}(\mathbf{Z}) d\mathbf{z}.$$

*If* $f(\mathbf{z}) \neq \{\sum_{\tau=1}^{T-2} p_{T-\tau}(\mathbf{z})\}^{-1} \sum_{\tau=1}^{T-2} p_{T-\tau}(\mathbf{z})\Delta_\tau^*(\mathbf{z}) \equiv f_0(\mathbf{z})$ *almost surely with respect to* $P_{history}$*, we can always choose* $\delta = -2(f - f_0)$ *so that*

$$\mathcal{L}(f+\delta) - \mathcal{L}(f) = -\int_{\mathcal{Z}} \{f_0(\mathbf{z}) - f(\mathbf{z})\}^2 \frac{1}{T-2} \sum_{\tau=1}^{T-2} p_{T-\tau}(\mathbf{Z}) d\mathbf{z} < 0.$$

*This suggest that* $f_0$ *is the global minimizer of* $\mathcal{L}(f)$ *over* $f \in \cap_{\tau=1}^{T-2} L_2(P_{T-\tau})$.

**Proof 3 (Proof of Proposition 1)** *Let* $\mathbf{Z}_{T-\tau} = (\mathbf{X}, Y_{T-\tau-1}), \tau = 1, \ldots, T-2$. *First note that* $m^{*(d)}(\mathbf{z}) = E\big(Y_T^{(d)} \mid D = d, \mathbf{Z}_T = \mathbf{z}\big), d = 0, 1$ *and* $\Delta_\tau^*(\mathbf{z}) = E\big(Y_{T-\tau}^{(0)} \mid D = 1, \mathbf{Z}_{T-\tau} = \mathbf{z}\big) - E\big(Y_{T-\tau}^{(0)} \mid D = 0, \mathbf{Z}_{T-\tau} = \mathbf{z}\big), \tau = 1, \ldots, T-2$ *are all identifiable; that is, they can be uniquely determined from the joint distribution of observable random variables* $(Y_1, D_1, \ldots, Y_T, D_T, \mathbf{X})$.

*Next, we show that* $f^*$ *exists and can be uniquely determined by* $\Delta_1^*, \ldots, \Delta_{T-1}^*$ *and* $P_{T-1}, \ldots, P_2$. *Let*

$$\mathcal{L}(f) = \frac{1}{T-2} \sum_{\tau=1}^{T-2} E_{P_{T-\tau}} \{\big|f(\mathbf{Z}) - \Delta_\tau^*(\mathbf{Z})\big|^2\}.$$

*We remark that* $f^*$ *exists because* $\mathcal{L}(f)$ *is lower bounded. Assume that there exists two* $f_1$ *and* $f_2$*, both of which can minimize* $\mathcal{L}(f)$. *Then, consider* $\widetilde{f} = (f_1 + f_2)/2$. *By the Cauchy–Schwarz inequality and the assumption* $f_1 \neq f_2$*, it holds that*

$$\mathcal{L}(\widetilde{f}) = \frac{1}{4(T-2)} \sum_{\tau=1}^{T-2} E_{P_{T-\tau}} \{\big|f_1(\mathbf{Z}) - \Delta_\tau^*(\mathbf{Z}) + f_2(\mathbf{Z}) - \Delta_\tau^*(\mathbf{Z})\big|^2\} < \frac{1}{2}\{\mathcal{L}(f_1) + \mathcal{L}(f_2)\},$$

*which contradicts with the optimality of* $f_1$ *and* $f_2$. *We then conclude that* $f^*$ *is unique.*

*Thus, the function* $\eta^{**}(\mathbf{z}) = m^{*(1)}(\mathbf{z}) - m^{*(0)}(\mathbf{z}) - f^*(\mathbf{z})$ *is identifiable.*

*It suffices to bound the difference* $\eta^*(\mathbf{z}) - \eta^{**}(\mathbf{z})$*, which directly follows Assumption 3; that is,* $\eta^*(\mathbf{z}) = m^{*(1)}(\mathbf{z}) - m^{*(0)}(\mathbf{z}) - \Delta_0^*(\mathbf{z}) = \eta^{**}(\mathbf{z}) + f^*(\mathbf{z}) - \Delta_0^*(\mathbf{z})$*, and*

$$E_{P_T}\{\eta^*(\mathbf{z}) - \eta^{**}(\mathbf{z})\}^2 = E_{P_T}\{\Delta_0^*(\mathbf{z}) - f^*(\mathbf{z})\}^2 \leq \varepsilon_{history}.$$

## B. Description of The Credit Card Balance Dataset

*Table 2.* Description of the credit card balance dataset. The biased observational dataset only consists of samples among the underlined instances. The features consist of the time-invariant covariates and the balance of credit card account over the last eight months.

| Risk | Number of Instances | | | | Number of Features |
|---|---|---|---|---|---|
| | 0 | 2k | 3k | 6k | |
| medium-risk | <u>389477</u> | 390928 | 391211 | 391215 | 87+8 |
| low-risk | 456773 | <u>459762</u> | <u>459443</u> | <u>460352</u> | |