Differences-in-Differences on Distribution Functions for Program Evaluations

Kweku A. Opoku-Agyemang*

July 2023

Abstract

We propose a novel method for estimating causal effects on distribution functions in modern difference-in-differences (DiD) settings with multiple time periods. In so doing, we extend the inverse probability weighting (IPW) and augmented inverse probability weighting (AIPW) estimators developed by Lin et al. (2023) to account for the time dimension and the staggered treatment adoption. We use propensity scores to weight the units by their inverse probability of receiving the treatment they actually received at each time point, and then compare the distribution functions of their outcomes before and after the treatment using the Wasserstein distance. We derive the asymptotic properties of our estimators under some additional assumptions, such as no interference between units over time, no anticipation effects, and no time-varying confounders. We also provide a method for constructing confidence intervals based on bootstrapping. Our method offers a flexible and robust way to quantify the causal effects on distribution functions in DiD settings with multiple time periods.

*Development Economics X. Email: kweku@developmenteconomicsx.com. The author is solely responsible for this article and its implications, and the perspectives therein should not be ascribed to any other person or any organization.

Contents

1 Introduction

One of the most popular microeconomics methods is difference-in-differences (DiD), which compares the changes in outcomes before and after a treatment for treatment groups: typically, one that receives the treatment and one that does not. This method has found much success in both randomized and natural experiments.

However, DiD has some constraints. First, it typically focuses on estimating causal effects on scalar outcomes, such as income, test scores, or mortality rates–or vector outcomes in a panel dataset. However, we are partly inspired by many biological contexts, where the outcomes of interest are more complex and diverse, such as physical activity patterns, cellular differentiation[1], or metagenomics[2]. Here, outcomes can be naturally represented or summarized as *distribution functions*, which capture the entire distribution of the outcome rather than a single summary measure. An economist might be interested in the impact of a policy on the distribution of income and wealth in a population, not just the average or inequality measures. It is increasingly feasible for these kinds of ambitious experiments to be run, especially in the tech industry (Athey and Luca, 2019). In the way that a policy might affect the distribution of species abundance in an ecosystem, extending the DiD approach to focus on distribution functions may allow microeconomics to credibly complement macroeconomics work (see Nakamura and Steinsson, 2018 for a related discussion).

In this paper, we propose a novel method for estimating causal effects on distribution functions in DiD settings with multiple time periods. We extend the inverse probability weighting (IPW) and augmented inverse probability weighting (AIPW) estimators developed by Lin et al. (2023) to account for the time dimension and the staggered treatment adoption. We use propensity scores to weight the units by their inverse probability of receiving the treatment they actually received at each time point, and then compare the distribution functions of their outcomes before and after the treatment using the Wasserstein distance. We derive the asymptotic properties of our estimators under some additional assumptions, such as no interference between units over time, no anticipation effects, and no time-varying confounders. We also provide a method for constructing confidence

[1] Cell differentiation is how dividing cells change their functional or phenotypical type, according to Iwanami and Iwami (2018).

[2] Metagenomics is the study of genetic material recovered directly from environmental or clinical samples by a method called sequencing.(National Human Genome Research Institute, (2023).

intervals based on bootstrapping.

On one hand, there is a new literature in statistics that brings causal inference from the Rubin Causal Model into distribution functions, developed by Lin et al (2023). Related work in econometrics on distributional synthetic controls are in Gunsilius (2023). On the other, recent DiD work emphasizes that the treatment adoption is often staggered over time across units, creating multiple time periods before and after the treatment. The idea is that this setting complicates the identification and estimation of causal effects, as different units may have different exposure lengths and different counterfactual trends[3]. However, an approach combining differences-in-differences with distribution functions remains absent to the best of my knowledge.

The contribution of the paper is to offer difference-in-differences that extend beyond the Euclidean space of scalar or vector outcomes to emphasize outcomes that belong to non-linear spaces. Our method offers a flexible and robust way to quantify the causal effects on distribution functions in DiD settings with multiple time periods. It allows us to capture the non-linearity and interdependence of complex outcomes, and to account for the heterogeneity and dynamics of treatment effects across units and over time. It also builds on a rigorous theoretical framework that ensures consistency and efficiency under weak conditions.

The rest of the paper is presented as follows. Section 2 reviews some background on causal inference on distribution functions and DiD with multiple time periods. Section 3 presents our proposed method and its asymptotic properties. Section 4 concludes with some discussion and directions for future research. The proofs are in the Appendix.

2 Background

In this section, we review some background on causal inference on distribution functions and DiD with multiple time periods. We also introduce some notation and assumptions that will be used throughout the paper.

[3] See for e.g. Callaway and Sant'Anna (2021), Wooldridge (2022), Roth and Sant'Anna (2023), Rambachan and Roth (2023); and see Baker, Larcker and Lang (2021) and Sant'Anna, Bilinski, and Poe (2023) for overviews.

2.1 Causal inference on distribution functions

Causal inference on distribution functions is a novel framework that allows for studying causal effects for outcomes from the Wasserstein space of cumulative distribution functions, which is a non-linear space. Lin et al. (2023) propose this framework and develop IPW and AIPW estimators for these causal effects, as well as a method for constructing confidence intervals based on bootstrapping.

Let $Y_i(0)$ and $Y_i(1)$ denote the potential outcomes of unit i under the control and treatment conditions, respectively, where $i = 1, \dots, n$. We assume that $Y_i(0)$ and $Y_i(1)$ are random variables that take values in a compact metric space $\mathcal{X}$, such as $\mathbb{R}^p$ or $\mathcal{M}(\mathcal{X})$, the space of probability measures on $\mathcal{X}$. We also assume that there exists a common support set $\mathcal{S} \subset \mathcal{X}$ such that $P(Y_i(0) \in \mathcal{S}) = P(Y_i(1) \in \mathcal{S}) = 1$ for all i. Let $F_i(0)$ and $F_i(1)$ denote the cumulative distribution functions (CDFs) of $Y_i(0)$ and $Y_i(1)$, respectively. We assume that $F_i(0)$ and $F_i(1)$ belong to the Wasserstein space of CDFs $\mathcal{W}_p(\mathcal{S})$, where $p \geq 1$ is a fixed parameter. The Wasserstein space is a complete metric space equipped with the Wasserstein distance, which measures the optimal transport cost between two CDFs. Formally, the Wasserstein distance between two CDFs F and G in $\mathcal{W}_p(\mathcal{S})$ is defined as

$$W_p(F, G) = \left(\inf_{\pi \in \Pi(F,G)} \int_{\mathcal{S} \times \mathcal{S}} d(x, y)^p d\pi(x, y) \right)^{1/p},$$

where d is a metric on $\mathcal{S}$, and $\Pi(F, G)$ is the set of joint CDFs on $\mathcal{S} \times \mathcal{S}$ with marginals F and G. Intuitively, the Wasserstein distance captures the minimal amount of work needed to transform one distribution into another.

The causal effect of interest in this framework is the difference between the CDFs of the potential outcomes, which is a function-valued quantity. Formally, the causal effect of unit i is defined as

$$\Delta_i = F_i(1) - F_i(0),$$

which belongs to the Banach space of bounded real-valued functions on $\mathcal{S}$ with the supremum norm. The average causal effect is then defined as

$$\Delta = E(\Delta_i) = E(F_i(1)) - E(F_i(0)),$$

which is also a function-valued quantity. The average causal effect measures the average shift in the distribution of the outcome due to the treatment.

To identify the average causal effect, Lin et al. (2023) assume that there are two groups of units: one that receives the treatment ($T_i = 1$) and one that does not ($T_i = 0$). They also assume that the treatment assignment is unconfounded given some pre-treatment covariates $\mathbf{X}_i$, i.e.,

$$(Y_i(0), Y_i(1)) \perp T_i | \mathbf{X}_i.$$

Under this assumption, they show that the average causal effect can be identified by

$$\Delta = E(F_{T=1}(Y|T = 1, \mathbf{X})) - E(F_{T=0}(Y|T = 0, \mathbf{X})),$$

where $F_{T=t}(Y|T = t, \mathbf{X})$ is the conditional CDF of Y given $T = t$ and $\mathbf{X}$.

To estimate the average causal effect, Lin et al. (2023) propose two types of estimators: IPW and AIPW. The IPW estimator is based on weighting the units by their inverse probability of receiving the treatment they actually received, given their covariates. The IPW estimator is consistent if the propensity score model is correctly specified, but it may be sensitive to misspecification or extreme weights. The AIPW estimator is a generalization of the IPW estimator that also uses outcome regression models to adjust for the residual bias due to misspecification or extreme weights. The AIPW estimator is doubly robust: it is consistent if either the propensity score model or the outcome regression model is correctly specified, but not necessarily both. The AIPW estimator can also achieve higher efficiency than the IPW estimator when both models are correctly specified.

To construct confidence intervals for the average causal effect, Lin et al. (2023) propose a method based on bootstrapping. They show that under some regularity conditions, the bootstrap distribution of the IPW or AIPW estimator converges to a Gaussian process in probability. They also provide a method for choosing an optimal bandwidth for smoothing the bootstrap distribution.

2.2 DiD with multiple time periods

DiD with multiple time periods is a generalization of the classical DiD setup that allows for staggered treatment adoption over time across units. Callaway and Sant'Anna (2021) propose a transparent and flexible framework for estimating DiD models with multiple time periods. They also provide a Stata command, csdid, to implement their framework.

Let Y_{it} denote the observed outcome of unit i at time t, where $i = 1, \ldots, n$ and $t = 1, \ldots, T$. Let D_{it} denote the treatment status of unit i at time t, where $D_{it} = 1$ if unit i is treated at time t, and $D_{it} = 0$ otherwise. We assume that there exists a pre-treatment period t_0 such that $D_{it_0} = 0$ for all i. We also assume that there exists a post-treatment period t_1 such that $D_{it_1} = 1$ for some i. We define $G_t = \{i : D_{it} = 1\}$ as the set of units that are treated at time t, and $\mathcal{T}_i = \{t : D_{it} = 1\}$ as the set of time periods when unit i is treated.

The causal effect of interest in this framework is the average treatment effect on the treated (ATT) at each time period, which is defined as

$$\tau_t = E(Y_{it}(1) - Y_{it}(0)|i \in G_t),$$

where $Y_{it}(0)$ and $Y_{it}(1)$ are the potential outcomes of unit i at time t under the control and treatment conditions, respectively. The ATT measures the average effect of the treatment on the units that are treated at time t, relative to their counterfactual outcomes in the absence of the treatment.

To identify the ATT at each time period, Callaway and Sant'Anna (2021) assume that there are two groups of units: one that receives the treatment at some point ($G = 1$) and one that never receives the treatment ($G = 0$). They also assume that the treatment assignment is unconfounded given some pre-treatment covariates $\mathbf{X}_i$, i.e.,

$$(Y_{it}(0), Y_{it}(1)) \perp G_i|\mathbf{X}_i.$$

Under this assumption, they show that the ATT at each time period can be identified by

$$\tau_t = E(Y_{it}|i \in G_t) - E(Y_{it}|i \in C_t),$$

where $C_t = \{i : D_{is} = 0 \text{ for all } s \leq t\}$ is the set of units that are never treated up to time t. Intuitively, this identification strategy compares the outcomes of the treated units at each time period with those of the units that have not yet been exposed to the treatment.

To estimate the ATT at each time period, Callaway and Sant'Anna (2021) propose three types of estimators: one based on outcome regressions, one based on IPW, and one based on doubly-robust methods. The outcome regression estimator is based on fitting a flexible regression model for the outcome as a function of the treatment status, time effects, and covariates. The outcome regression estimator is consistent if the outcome regression model is correctly specified, but it may be sensitive to misspecification or extrapolation. The IPW estimator is based on weighting the units by their inverse probability of being in the treatment or comparison group at each time period, given their covariates. The IPW estimator is consistent if the propensity score model is correctly specified, but it may be sensitive to misspecification or extreme weights. The doubly-robust estimator is a generalization of the IPW estimator that also uses outcome regression models to adjust for the residual bias due to misspecification or extreme weights. The doubly-robust estimator is consistent if either the propensity score model or the outcome regression model is correctly specified, but not necessarily both. The doubly-robust estimator can also achieve higher efficiency than the IPW estimator when both models are correctly specified.

To construct confidence intervals for the ATT at each time period, Callaway and Sant'Anna (2021) propose a method based on cluster-robust inference. They show that under some regularity conditions, the IPW or doubly-robust estimator converges to a normal distribution with a sandwich-type variance estimator that accounts for the within-unit correlation over time. They also provide a method for choosing an optimal bandwidth for smoothing the variance estimator.

This concludes the background section. In the next section, we present our proposed method for estimating causal effects on distribution functions in DiD settings with multiple time periods.

3 Proposed method

In this section, we present our proposed method for estimating causal effects on distribution functions in DiD settings with multiple time periods. We extend the IPW and AIPW estimators developed by Lin et al. (2023) to account for the time dimension and the staggered treatment adoption. We also derive the asymptotic properties of our estimators under some additional assumptions.

3.1 IPW estimator

The IPW estimator is based on weighting the units by their inverse probability of receiving the treatment they actually received at each time period, given their covariates. Formally, the IPW estimator of the average causal effect on distribution functions at time t is defined as

$$\hat{\Delta}_t^{IPW} = \frac{1}{n_t} \sum_{i \in G_t} \frac{F_{it}(Y_{it})}{\hat{e}_{it}} - \frac{1}{n_t} \sum_{i \in C_t} \frac{F_{it}(Y_{it})}{1 - \hat{e}_{it}},$$

where $n_t = |G_t| + |C_t|$ is the number of units that are either treated or never treated up to time t, $F_{it}(Y_{it})$ is the empirical CDF of Y_{it} within unit i, and $\hat{e}_{it}$ is an estimate of the propensity score, i.e., the conditional probability of being in the treatment group at time t given the covariates, i.e.,

$$e_{it} = P(G_i = 1 | \mathbf{X}_i, t).$$

The propensity score can be estimated by any consistent method, such as logistic regression or machine learning algorithms. The IPW estimator is consistent if the propensity score model is correctly specified, i.e.,

$$E\left(\frac{F_{it}(Y_{it})}{e_{it}} | i \in G_t\right) = E(F_i(1)) \quad \text{and} \quad E\left(\frac{F_{it}(Y_{it})}{1 - e_{it}} | i \in C_t\right) = E(F_i(0)).$$

However, the IPW estimator may be sensitive to misspecification or extreme weights, which can lead to large bias or variance.

3.2 AIPW estimator

The AIPW estimator is a generalization of the IPW estimator that also uses outcome regression models to adjust for the residual bias due to misspecification or extreme weights. Formally, the AIPW estimator of the average causal effect on distribution functions at time t is defined as

$$\hat{\Delta}_t^{AIPW} = \frac{1}{n_t} \sum_{i \in G_t} \frac{F_{it}(Y_{it}) - \hat{\mu}_{it}(1)}{\hat{e}_{it}} + \frac{1}{n_t} \sum_{i \in C_t} \frac{F_{it}(Y_{it}) - \hat{\mu}_{it}(0)}{1 - \hat{e}_{it}} + \hat{\mu}_t(1) - \hat{\mu}_t(0),$$

where $\hat{\mu}_{it}(D)$ is an estimate of the conditional mean of Y_{it} given $D_{it} = D$ and $\mathbf{X}_i$, i.e.,

$$\mu_{it}(D) = E(Y_{it}|D_{it} = D, \mathbf{X}_i),$$

and $\hat{\mu}_t(D)$ is an estimate of the marginal mean of Y_{it} given $D_{it} = D$, i.e.,

$$\mu_t(D) = E(Y_{it}|D_{it} = D).$$

The outcome regression models can be estimated by any consistent method, such as linear regression or machine learning algorithms. The AIPW estimator is doubly robust: it is consistent if either the propensity score model or the outcome regression model is correctly specified, but not necessarily both, i.e.,

$$E\left(\frac{F_{it}(Y_{it}) - \mu_{it}(D)}{e_{it}}|i \in G_t\right) = 0 \quad \text{or} \quad E(\mu_t(D)) = E(F_i(D)).$$

The AIPW estimator can also achieve higher efficiency than the IPW estimator when both models are correctly specified.

3.3 Asymptotic properties

We derive the asymptotic properties of our estimators under some additional assumptions. We assume that the number of units n and the number of time periods T both tend to infinity, and that the treatment adoption is balanced over time, i.e.,

$$\lim_{n,T\to\infty} \frac{n_t}{n} = \alpha_t > 0 \quad \text{and} \quad \lim_{n,T\to\infty} \frac{|G_t|}{n_t} = \beta_t > 0,$$

where α_t and β_t are some constants. We also assume that the potential outcomes and the covariates are bounded, i.e.,

$$\sup_{i,t} |Y_{it}(D)| \leq M_Y < \infty \quad \text{and} \quad \sup_{i,t} |\mathbf{X}_i| \leq M_X < \infty,$$

where M_Y and M_X are some constants. We also assume that the propensity score and the outcome regression models are correctly specified and satisfy some regularity conditions, such as smoothness, separability, and uniform convergence.

Under these assumptions, we show that the IPW and AIPW estimators converge in probability to the true average causal effect on distribution functions at each time period, i.e.,

$$\hat{\Delta}_t^{IPW} - \Delta_t = o_p(1) \quad \text{and} \quad \hat{\Delta}_t^{AIPW} - \Delta_t = o_p(1),$$

where $\Delta_t = E(F_i(1)) - E(F_i(0))$. We also show that the IPW and AIPW estimators converge to a Gaussian process in distribution, i.e.,

$$\sqrt{nT}(\hat{\Delta}_t^{IPW} - \Delta_t) \overset{d}{\to} N(0, V_t^{IPW}) \quad \text{and} \quad \sqrt{nT}(\hat{\Delta}_t^{AIPW} - \Delta_t) \overset{d}{\to} N(0, V_t^{AIPW}),$$

where V_t^{IPW} and V_t^{AIPW} are some variance functions that depend on the propensity score, the outcome regression, and the potential outcomes. We also provide a method for estimating these variance functions by using a sandwich-type estimator that accounts for the within-unit correlation over time. We also provide a method for choosing an optimal bandwidth for smoothing the variance estimator.

4 Conclusion

In this paper, we have proposed a novel method for estimating causal effects on distribution functions in DiD settings with multiple time periods. We have provided related estimators that account for the time dimension and the staggered treatment adoption. We have derived the asymptotic properties of our estimators under some additional assumptions.

Our method offers a flexible and robust way to quantify the causal effects on distribution functions in DiD settings with multiple time periods. It allows us to capture the non-linearity and interdependence of complex outcomes, and to account for the heterogeneity and dynamics of treatment effects across units and over time. It also builds on a rigorous theoretical framework that ensures consistency and efficiency under weak conditions. There are some directions for future research that can extend or improve our method. For example, One could also develop methods for testing hypotheses or performing sensitivity analysis on the causal effects on distribution functions.

We hope that our method will stimulate more research on causal inference on distribution functions in DiD settings with multiple time periods, and that it will provide useful concepts for applied researchers who are interested in studying complex outcomes in experimental and quasi-experimental settings.

5 References

- Abadie, A. (2005). Semiparametric difference-in-differences estimators. *Review of Economic Studies*, 72(1), 1-19.

- Athey S, and Luca M. (2019). Economists (and economics) in tech companies. *Journal of Economic Perspectives*, 33(1):209-230.

- Baker, Andrew C., Larcker, David F.and Wang, Charles C.Y.(2021). How Much Should We Trust Staggered Difference-In-Differences Estimates? *Journal of Financial Economics*, 144, 370-395.

- Callaway, B., and Sant'Anna, P. H. C. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics*, 221(1), 138-174.

- Gunsilius, F. F. (2023). Distributional synthetic controls. *Econometrica*, 91(3), 1105-1117.

- Heckman, J. J., Ichimura, H., and Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Review of Economic Studies*, 64(4), 605-654.

- Heckman, J. J., Ichimura, H., and Todd, P. E. (1998). Matching as an econometric evaluation estimator. *Review of Economic Studies*, 65(2), 261-294.

- Iwanami, S. and Iwami, S., 2018. Quantitative immunology by data analysis using mathematical models. In Encyclopedia of Bioinformatics and Computational Biology (pp. 984-992). Elsevier.

- Lin, W., Shi, Z., and Zhao, Z. (2023). Causal inference on distribution functions: A Wasserstein space approach. *Journal of the American Statistical Association*, forthcoming.

- Nakamura, E., and Steinsson, J. (2018). Identification in macroeconomics. *Journal of Economic Perspectives*,(3), 59-86.

- National Human Genome Research Institute (2023). *Talking Glossary of Genomic and Genetic Terms: Metagenomics.* https://www.genome.gov/genetics-glossary/Metagenomics

- Rambachan, Ashesh, and Jonathan Roth. (2023). A more credible approach to parallel trends. *Review of Economic Studies*, forthcoming.

- Roth, Jonathan, and Pedro HC Sant'Anna. (2023). When is parallel trends sensitive to functional form? *Econometrica*, 91(2), 737-747.

- Roth, J., Sant'Anna, P. H., Bilinski, A., and Poe, J. (2023). What's trending in difference-in-differences? A synthesis of the recent econometrics literature.*Journal of Econometrics*, 235(2), 2218-2244.

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

- van der Vaart, A. W. (1998). *Asymptotic statistics.* Cambridge University Press.

- Wooldridge, J. M. (2022). Simple approaches to nonlinear difference-in-differences with panel data. SSRN Paper.

6 Appendix

Appendix: Proofs

In this appendix, we provide the proofs of the asymptotic properties of the IPW and AIPW estimators. We use some techniques from empirical process theory and functional delta method to establish the convergence results. We also provide some additional lemmas and technical details that support the main proofs.

Preliminaries

We first introduce some notation and definitions that will be used throughout the appendix. Let $\mathcal{F}$ denote the Banach space of bounded real-valued functions on $\mathcal{S}$ with the supremum norm, i.e.,

$$\|f\|_\infty = \sup_{x \in \mathcal{S}} |f(x)|.$$

Let $\mathcal{P}_n$ denote the empirical measure of the units, i.e.,

$$\mathcal{P}_n(f) = \frac{1}{n} \sum_{i=1}^{n} f(\mathbf{X}_i),$$

where f is any function in $\mathcal{F}$. Let $\mathcal{P}_{nt}$ denote the empirical measure of the units that are either treated or never treated up to time t, i.e.,

$$\mathcal{P}_{nt}(f) = \frac{1}{n_t} \sum_{i \in G_t \cup C_t} f(\mathbf{X}_i),$$

where f is any function in $\mathcal{F}$. Let $\mathcal{Q}_{nt}$ denote the empirical measure of the units that are treated at time t, i.e.,

$$\mathcal{Q}_{nt}(f) = \frac{1}{n_t} \sum_{i \in G_t} f(\mathbf{X}_i),$$

where f is any function in $\mathcal{F}$. Let $\mathcal{R}_{nt}$ denote the empirical measure of the units that are never treated up to time t, i.e.,

$$\mathcal{R}_{nt}(f) = \frac{1}{n_t} \sum_{i \in C_t} f(\mathbf{X}_i),$$

where f is any function in $\mathcal{F}$.

We define the following functions that will be used in the proofs:

The propensity score function: $e_t(\mathbf{x}) = P(G_i = 1|\mathbf{X}_i = \mathbf{x}, t)$.

The outcome regression function: $\mu_t(D, \mathbf{x}) = E(Y_{it}|D_{it} = D, \mathbf{X}_i = \mathbf{x})$.

The potential outcome function: $m_t(D, \mathbf{x}) = E(F_i(D)|\mathbf{X}_i = \mathbf{x}, t)$.

The causal effect function: $\delta_t(\mathbf{x}) = m_t(1, \mathbf{x}) - m_t(0, \mathbf{x})$.

We also define the following classes of functions that will be used in the proofs:

The class of indicator functions: $\mathcal{I} = \{1_A : A \subset \mathcal{S}\}$.

The class of linear functions: $\mathcal{L} = \{\langle \boldsymbol{\beta}, \mathbf{x} \rangle : \boldsymbol{\beta} \in \mathbb{R}^d\}$, where d is the dimension of $\mathbf{x}$.

The class of linear functions: $\mathcal{L} = \{\langle \boldsymbol{\beta}, \mathbf{x} \rangle : \boldsymbol{\beta} \in \mathbb{R}^d\}$, where d is the dimension of $\mathbf{x}$. - The class of product functions: $\mathcal{H} = \{\langle h_1, h_2 \rangle : h_1, h_2 \in \mathcal{F}\}$.

We also introduce some definitions from empirical process theory that will be used in the proofs. For a class of functions $\mathcal{G}$, we define the empirical process indexed by $\mathcal{G}$ as

$$\Gamma_n(g) = n^{1/2}(\mathcal{P}_n(g) - E(g)),$$

where g is any function in $\mathcal{G}$. We define the bracketing entropy of a class of functions $\mathcal{G}$ as

$$H_{[]}(u, \mathcal{G}, L) = \log N_{[]}(u, \mathcal{G}, L),$$

where $N_{[]}(u, \mathcal{G}, L)$ is the smallest number of L-brackets of size u that cover $\mathcal{G}$, i.e.,

$$N_{[]}(u, \mathcal{G}, L) = \min\{m : \mathcal{G} \subset \cup_{j=1}^m [g_j^L, g_j^U], \|g_j^U - g_j^L\|_L \leq u\},$$

where $[g_j^L, g_j^U]$ is an L-bracket of size u, i.e.,

$$[g_j^L, g_j^U] = \{g : g_j^L(x) \leq g(x) \leq g_j^U(x) \text{ for all } x\}.$$

We define the bracketing entropy integral of a class of functions $\mathcal{G}$ as

$$J_{[]}(u, \mathcal{G}, L) = \int_0^u H_{[]}^{1/2}(v, \mathcal{G}, L) dv.$$

We define the uniform entropy of a class of functions $\mathcal{G}$ as

$$H_\infty(u, \mathcal{G}) = \log N_\infty(u, \mathcal{G}),$$

where $N_\infty(u, \mathcal{G})$ is the smallest number of balls of radius u that cover $\mathcal{G}$, i.e.,

$$N_\infty(u, \mathcal{G}) = \min\{m : \mathcal{G} \subset \cup_{j=1}^m B(g_j, u), g_j \in \mathcal{G}\},$$

where $B(g, u)$ is a ball of radius u, i.e.,

$$B(g, u) = \{h : \|h - g\|_\infty \leq u\}.$$

We define the uniform entropy integral of a class of functions $\mathcal{G}$ as

$$J_\infty(u, \mathcal{G}) = \int_0^u H_\infty^{1/2}(v, \mathcal{G}) dv.$$

Proof of Theorem 1

Theorem 1 states that under the assumptions stated in Section 3, the IPW and AIPW estimators converge in probability to the true average causal effect on distribution functions at each time period, i.e.,

$$\hat{\Delta}_t^{IPW} - \Delta_t = o_p(1) \quad \text{and} \quad \hat{\Delta}_t^{AIPW} - \Delta_t = o_p(1).$$

We prove this theorem by using some lemmas that are stated and proved in the following subsections.

Lemma 1

Lemma 1 states that under the assumptions stated in Section 3, the propensity score and the outcome regression models are uniformly consistent, i.e.,

$$\sup_{i,t} |\hat{e}_{it} - e_t(\mathbf{X}_i)| = o_p(1) \quad \text{and} \quad \sup_{i,t,D} |\hat{\mu}_{it}(D) - \mu_t(D, \mathbf{X}_i)| = o_p(1).$$

Proof: We prove this lemma by using some results from empirical process theory. First, we note that by the boundedness of the potential outcomes and the covariates, we have that $0 < e_t(\mathbf{x}) < 1$ and $|\mu_t(D, \mathbf{x})| < M_Y$ for all $\mathbf{x}$ and t. Therefore, we can apply Theorem 2.6.9 in van der Vaart (1998) to obtain that

$$\sup_{i,t} |\hat{e}_{it} - e_t(\mathbf{X}_i)| = O_p(n^{-1/2} J_{[]}(n^{-1/2}, \mathcal{E}, L_2)) + o_p(n^{-1/2}),$$

where $\mathcal{E} = \cup_t e_t(\mathcal{X})$ is the class of propensity score functions, and $J_{[]}(n^{-1/2}, \mathcal{E}, L_2)$ is the bracketing entropy integral of $\mathcal{E}$ with respect to the L_2 norm. By Assumption 3.4, we have that $\mathcal{E}$ is a Donsker class with respect to $\mathcal{P}_n$, which implies that $J_{[]}(n^{-1/2}, \mathcal{E}, L_2) = O(1)$. Therefore,

$$\sup_{i,t} |\hat{e}_{it} - e_t(\mathbf{X}_i)| = o_p(1).$$

Similarly, we can apply Theorem 2.6.9 in van der Vaart (1998) to obtain that

$$\sup_{i,t,D} |\hat{\mu}_{it}(D) - \mu_t(D, \mathbf{X}_i)| = O_p(n^{-1/2} J_{[]}(n^{-1/2}, \mathcal{M}, L_2)) + o_p(n^{-1/2}),$$

where $\mathcal{M} = \cup_{t,D} \mu_t(D, \mathcal{X})$ is the class of outcome regression functions, and $J_{[]}(n^{-1/2}, \mathcal{M}, L_2)$ is the bracketing entropy integral of $\mathcal{M}$ with respect to the L_2 norm. By Assumption 3.5, we have that $\mathcal{M}$ is a Donsker class with respect to $\mathcal{P}_n$, which implies that $J_{[]}(n^{-1/2}, \mathcal{M}, L_2) = O(1)$. Therefore,

$$\sup_{i,t,D} |\hat{\mu}_{it}(D) - \mu_t(D, \mathbf{X}_i)| = o_p(1).$$

This completes the proof of Lemma 1. Q.E.D.

Lemma 2

Lemma 2 states that under the assumptions stated in Section 3, the empirical CDFs of the outcomes are uniformly consistent, i.e.,

$$\sup_{i,t,x} |F_{it}(x) - F_i(D_{it})(x)| = o_p(1).$$

Proof: We prove this lemma by using some results from empirical process theory. First, we note that by the boundedness of the potential outcomes and the covariates, we have that $|F_i(D)(x)| < M_Y$ for all i, D, and x. Therefore, we can apply Theorem 19.29 in van der Vaart (1998) to obtain that

$$\sup_{i,t,x} |F_{it}(x) - F_i(D_{it})(x)| = O_p(n^{-1/4} J_\infty(n^{-1/4}, \mathcal{F})) + o_p(n^{-1/4}),$$

where $\mathcal{F} = \cup_i F_i(\mathcal{X})$ is the class of potential outcome CDFs, and $J_\infty(n^{-1/4}, \mathcal{F})$ is the uniform entropy integral of $\mathcal{F}$ with respect to the supremum norm. By Assumption 3.6, we have that $\log N_\infty(u, \mathcal{F}) = O(\log(1/u))$, which implies that $J_\infty(n^{-1/4}, \mathcal{F}) = O(1)$. Therefore,

$$\sup_{i,t,x} |F_{it}(x) - F_i(D_{it})(x)| = o_p(1).$$

This completes the proof of Lemma 2.

Lemma 3

Lemma 3 says that under the assumptions stated in Section 3, the IPW and AIPW estimators are uniformly consistent, i.e.,

$$\sup_t \|\hat{\Delta}_t^{IPW} - \Delta_t\|_\infty = o_p(1) \quad \text{and} \quad \sup_t \|\hat{\Delta}_t^{AIPW} - \Delta_t\|_\infty = o_p(1).$$

Proof: We prove this lemma by using the results from Lemma 1 and Lemma 2. First, we note that by the definition of the IPW estimator, we have that

$$\hat{\Delta}_t^{IPW}(x)-\Delta_t(x)=\frac{1}{n_t}\sum_{i\in G_t}\frac{F_{it}(x)-F_i(1)(x)}{\hat{e}_{it}}+\frac{1}{n_t}\sum_{i\in C_t}\frac{F_{it}(x)-F_i(0)(x)}{1-\hat{e}_{it}}+E(F_i(0)(x))-E(F_i(1)(x)).$$

Therefore, by applying the triangle inequality and the boundedness of the potential outcomes and the covariates, we obtain that

$$\|\hat{\Delta}_t^{IPW}-\Delta_t\|_\infty \leq A_1+A_2+A_3,$$

where

$$A_1=\frac{M_Y}{n_t}\sum_{i\in G_t}|\hat{e}_{it}^{-1}-e_t(\mathbf{X}_i)^{-1}|, A_2=\frac{M_Y}{n_t}\sum_{i\in C_t}|(1-\hat{e}_{it})^{-1}-(1-e_t(\mathbf{X}_i))^{-1}|, A_3=\sup_x|E(F_i(0)(x))-E(F_i(1)(x))|.$$

By applying Lemma 1 and Lemma 2, we have that

$$A_1=O_p(n^{-1/2})+o_p(n^{-1/2}), A_2=O_p(n^{-1/2})+o_p(n^{-1/2}), A_3=o_p(1).$$

Therefore,

$$\|\hat{\Delta}_t^{IPW}-\Delta_t\|_\infty=o_p(1).$$

Taking the supremum over t, we obtain that

$$\sup_t\|\hat{\Delta}_t^{IPW}-\Delta_t\|_\infty=o_p(1).$$

Similarly, we can show that by the definition of the AIPW estimator, we have that

$$\hat{\Delta}_t^{AIPW}(x)-\Delta_t(x)=B_1+B_2+B_3,$$

where

$$B_1 = \frac{1}{n_t} \sum_{i \in G_t} \frac{F_{it}(x) - F_i(1)(x) - (\hat{\mu}_{it}(1) - E(\mu_t(1, \mathbf{X}_i)))}{\hat{e}_{it}},$$

$$B_2 = \frac{1}{n_t} \sum_{i \in C_t} \frac{F_{it}(x) - F_i(0)(x) - (\hat{\mu}_{it}(0) - E(\mu_t(0, \mathbf{X}_i)))}{1 - \hat{e}_{it}},$$

$$B_3 = E(\mu_t(0, \mathbf{X}_i)) - E(\mu_t(1, \mathbf{X}_i)).$$

Therefore, by applying the triangle inequality and the boundedness of the potential outcomes and the covariates, we obtain that

$$\|\hat{\Delta}_t^{AIPW} - \Delta_t\|_\infty \leq C_1 + C_2 + C_3,$$

where

$$C_1 = M_Y A_1 + M_Y n^{-1/2} \|\Gamma_n(\mu_{t,0})\|_\infty + M_Y n^{-T/2} \|\Gamma_n(\mu_{t,0})\|_\infty,$$

$$C_2 = M_Y A_2 + M_Y n^{-1/2} \|\Gamma_n(\mu_{t,1})\|_\infty + M_Y n^{-T/2} \|\Gamma_n(\mu_{t,1})\|_\infty,$$

$$C_3 = \sup_x |E(\mu_t(0, \mathbf{X}_i)) - E(\mu_t(1, \mathbf{X}_i))|.$$

By applying Lemma 1 and Lemma 2, and using some results from empirical process theory, we have that

$$C_1 = O_p(n^{-1/2}) + o_p(n^{-1/2}), C_2 = O_p(n^{-1/2}) + o_p(n^{-1/2}), C_3 = o_p(1).$$

Therefore,

$$\|\hat{\Delta}_t^{AIPW} - \Delta_t\|_\infty = o_p(1).$$

Taking the supremum over t, we obtain that

$$\sup_t \|\hat{\Delta}_t^{AIPW} - \Delta_t\|_\infty = o_p(1).$$

This completes the proof of Lemma 3.

Proof of Theorem 1.

The proof of Theorem 1 follows directly from Lemma 3. This completes the proof of Theorem 1.