Causal Inference under Dynamic Selection and Latent Heterogeneity *

Weisheng Zhang†

November 5, 2026

Click here for the latest version

Abstract

I study dynamic treatment effects in panel data under staggered adoption when treatment timing depends jointly on unobserved time-invariant heterogeneity and time-varying pretreatment covariates, including lagged outcomes. Untreated potential outcomes follow a nonparametric dynamic panel model that allows flexible interactions between time-varying covariates and latent heterogeneity. I use pretreatment outcome histories to find individuals with similar time-invariant latent factors, and the key requirement is that these histories are sufficiently informative about those latent factors. I develop an identification strategy for the dynamic average treatment effect on the treated (ATT) and propose kernel-based doubly robust estimators for the dynamic ATT. I further combine double cross-fitting with undersmoothing and show that, under suitable regularity conditions, the proposed estimators are $\sqrt{n}$-consistent, asymptotically normal, and asymptotically unbiased. The simulation study demonstrates that the proposed method provides accurate inference across a wide range of data-generating processes. I illustrate the method with an application to the U.S. family planning program studied by Bailey (2012) and reestimate its effect on fertility rates.

Keywords: Dynamic selection, nonparametric dynamic panel, double cross-fitting

*I thank Richard Blundell, Pedro Carneiro, Andrew Chesher, Ben Deaner, Aureo de Paula, Dennis Kristensen, Ge Sun, Martin Weidner, Andrei Zeleneev, and the participants in the UCL Econometrics Brownbag Seminar, the Cemmap Seminar, and the 2026 PhD Research Day for their valuable comments.

†University College London: weisheng.zhang.21@ucl.ac.uk.

1 Introduction

Self-selection is a fundamental challenge in causal inference because treated and untreated units may be systematically different. These differences include time-invariant characteristics, such as individual ability, as well as *time-varying* characteristics, such as lagged outcomes. Causal inference methods using panel data can help address self-selection by accounting for selection on unobserved time-invariant heterogeneity (fixed effects). These causal panel methods, including difference-in-differences (DiD), synthetic control, and matrix completion, have become increasingly popular in empirical research.

In many empirical applications, treated and untreated units may also differ in time-varying covariates, such as lagged outcomes. A well-known empirical example is that workers often experience a decline in earnings before participating in a job-training program (Ashenfelter and Card, 1985). In these applications, controlling for pretreatment time-varying covariates is necessary to disentangle the treatment effect from the dynamic effects arising from lagged outcomes.

However, the validity of most existing causal panel methods builds on the assumption that treatment selection depends *only* on time-invariant characteristics. For example, Ghanem, Sant'Anna and Wüthrich (2022) and Marx, Tamer and Tang (2024*b*) show that, except in a few special cases, the parallel trends assumption in DiD fails when treatment depends on both time-invariant characteristics and lagged outcomes. This concern is not limited to DiD but arises in other causal panel methods when treatment assignment responds to time-varying pretreatment covariates (Arkhangelsky and Imbens, 2024).

In this paper, I propose a new causal inference method using panel data when treatment selection depends jointly on time-invariant latent factors and time-varying pretreatment covariates. I establish nonparametric identification of the dynamic average treatment effect on the treated (ATT), develop a doubly robust estimator, and provide a corresponding inference procedure. The method is designed to accommodate staggered adoption, which is common in empirical applications.

The proposed method allows the outcome process to depend on time-varying covariates, including lagged outcomes. Such dynamic dependence is ubiquitous in economics (e.g., due to intertemporal decision-making, adjustment costs, and persistent shocks) but is often abstracted away in causal panel methods (Arkhangelsky and Imbens, 2024). Instead, I allow past outcomes to affect current outcomes, capturing dynamic decision-making in staggered adoption settings and connecting the proposed framework to the broader panel econometrics literature (e.g., Arellano and Bond, 1991) that emphasizes the importance of dynamic dependence.

In addition, I do not impose functional-form restrictions on untreated potential outcomes. Many popular causal panel methods rely on linearity assumptions that may be too restrictive in applications. For example, DiD relies on the parallel trends assumption, and most synthetic control and matrix completion methods require that untreated potential outcomes follow a linear factor model. In contrast, my method builds on a nonparametric model in which outcomes depend flexibly on their past outcomes and time-invariant latent factors, eliminating concerns about the functional-form specification of untreated potential outcomes in empirical applications.

The key technical challenge is that time-invariant latent factors are unobserved. Since untreated outcomes follow a possibly nonlinear model, these factors can neither be differenced out as in DiD nor estimated using standard linear factor methods. Feng (2023), Deaner, Hsiang and Zeleneev (2025), and Athey and Imbens (2025) address this challenge in a pure nonlinear factor model without dynamic dependence. They construct pairwise pseudo-distances from long pretreatment outcome histories. Each pseudo-distance measures the similarity between two individuals' pretreatment histories and serves as a proxy for the distance between latent factors. However, their analysis does not directly extend to dynamic models, because in such settings, outcome histories reflect not only the effects of time-invariant heterogeneity but also the dynamic effects of time-varying covariates.

To accommodate dynamic dependence, I extend the pseudo-distance approach to dynamic panel models and show that the pairwise pseudo-distance remains informative about latent heterogeneity under weak dependence. The intuition is that the dynamic effects decay over time, whereas individual latent factors continue to affect the entire outcome path. As a result, individuals with similar latent factors exhibit similar long-run pretreatment histories, and the pseudo-distance remains informative about latent similarity even in the presence of dynamic dependence.

The identification of the dynamic ATT uses matching based on time-varying covariates and the pseudo-distance. The key assumption is the informativeness condition, which requires that if the pseudo-distance between two individuals is small, then the distance between their latent factors is also small. Under this assumption, although time-invariant latent factors are unobserved, individuals with similar latent factors are identified using the pseudo-distance. I match treated units to untreated units with similar covariates and a small pseudo-distance to impute untreated counterfactual outcomes, and then identify the dynamic ATT. Under staggered adoption, this matching strategy is combined with a backward recursive procedure to accommodate dynamic selection. The identification result is new under selection on both time-invariant latent heterogeneity and time-varying covariates and requires only pairwise latent similarity rather than identification of the latent factors or

their distribution.

Based on this identification strategy, I propose a doubly robust estimator for the dynamic ATT. I use Nadaraya-Watson estimators based on both covariates and the pseudo-distance to nonparametrically estimate counterfactual outcomes and propensity scores. Asymptotic properties are established for both the nuisance parameter estimators and the dynamic ATT estimator. When pretreatment histories are sufficiently long, the nuisance parameter estimators achieve the same convergence rates as if time-invariant latent factors were observed.

I further apply a novel double cross-fitting procedure to obtain valid inference under weaker conditions. Unlike standard cross-fitting (Chernozhukov et al., 2018), double cross-fitting constructs the counterfactual outcome and propensity score estimators using separate training samples. This reduces the dependence between their estimation errors and, when combined with undersmoothing, yields a faster convergence rate for the proposed estimator. I establish the root-n consistency, asymptotic normality, and asymptotic unbiasedness of the proposed estimator, and show that these properties hold under weaker dimension restrictions than those required for standard cross-fitting.

Extensive simulation studies evaluate the finite-sample performance of the proposed estimator. The estimator performs well in terms of bias, standard deviation, and coverage across different data-generating processes and sample sizes, even when the number of time periods is moderate. Based on these simulation results, practical guidance on bandwidth selection is provided for applied researchers.

I apply the proposed method to the U.S. family planning programs studied by Bailey (2012) and reestimate their effects on fertility rates. First, I find strong evidence that counties with lower lagged fertility rates are more likely to be treated, which motivates the use of the proposed method. Second, my method estimates statistically significant and persistent reductions in fertility. This finding is consistent with the results in Bailey (2012) and those obtained using DiD. Lastly, the estimated reductions are larger than the DiD estimates, and the gap widens over longer post-treatment horizons. This difference reflects the importance of accounting for dynamic selection based on lagged fertility rates.

Literature review This paper contributes to the large and rapidly growing literature on causal inference with panel data by allowing treatment selection to depend on time-varying covariates. I refer readers to De Chaisemartin and d'Haultfoeuille (2023) and Roth et al. (2023) for recent surveys of the DiD literature and to Arkhangelsky and Imbens (2024) for a broader survey of causal inference with panel data. The implications of treatment selection for the validity of causal panel methods have received increasing attention. Earlier discussions of this issue in DiD include Ashenfelter and Card (1985) and Abadie (2005),

and more recently, Ghanem, Sant'Anna and Wüthrich (2022) and Marx, Tamer and Tang (2024*b*) show that, except in a few special cases, the parallel trends assumption fails when treatment decisions respond to previous outcomes. Arkhangelsky and Imbens (2024) further emphasize that many causal panel results are established under strict exogeneity, which rules out treatment selection based on time-varying outcome shocks. Although Arkhangelsky and Hirshberg (2023) show that this restriction can be relaxed for synthetic control under simultaneous-treatment, their arguments do not directly extend to staggered adoption.

A related strand of the causal panel literature addresses treatment selection on time-varying covariates. For DiD, these methods rely on conditional parallel trends (Abadie, 2005; Wooldridge, 2025), with extensions to doubly robust estimation (Sant'Anna and Zhao, 2020), high-dimensional covariates (Chang, 2020), and treatment-affected covariates (Caetano et al., 2022). However, conditional parallel trends may fail when treatment selection depends jointly on observed covariates and time-invariant latent factors. For the dynamic treatment effect literature, Lewis and Syrgkanis (2020) and Viviano and Bradic (2026) allow selection on high-dimensional covariates but assume that treatment assignment is independent of time-invariant latent factors conditional on these covariates. In addition, Marx, Tamer and Tang (2024*a*) and Botosaru and Liu (2025) focus on dynamic treatment effects under linear panel models. In contrast, the proposed approach uses a long pretreatment panel to accommodate selection on both time-varying covariates and time-invariant latent factors without imposing functional-form restrictions on either the propensity score or the untreated outcome process. Lastly, Gulek and Vives-i Bastida (2025) combine synthetic control and instrumental variables to address unmeasured time-varying confounding, while my analysis focuses on treatment selection based on observed time-varying covariates and does not require external instruments.

Beyond the literature on causal panel methods, this paper connects to the broader econometric literature on panel-data models. This literature has paid close attention to individual heterogeneity and outcome dynamics in both linear settings (e.g., Arellano and Bond, 1991; Blundell and Bond, 1998; Alvarez and Arellano, 2003) and nonlinear settings (e.g., Blundell, Griffith and Windmeijer, 2002; Honoré and Kyriazidou, 2000; Bonhomme, 2012; Bonhomme and Manresa, 2015; Bonhomme, Lamadon and Manresa, 2022), but has primarily focused on estimating parametric models. By contrast, the causal panel literature allows for rich heterogeneity in treatment effects but often abstracts away from dynamic dependence in outcomes (Arkhangelsky and Imbens, 2024). The proposed method bridges these two literatures by allowing for both flexible interactions between time-varying covariates and latent heterogeneity and dynamic dependence in untreated potential outcomes, without imposing a parametric outcome model.

The nonparametric model for untreated potential outcomes connects this paper to the nonparametric panel literature. For example, Chernozhukov et al. (2013), Hoderlein and White (2012), and Chernozhukov et al. (2026) study average structural and causal effects under time-homogeneity conditions. The proposed method allows for rich time heterogeneity in untreated potential outcomes without imposing time-homogeneity conditions. Another strand of literature uses observations from different periods as measurements or proxies for (possibly time-varying) latent confounders (e.g., Hu and Shum, 2012; Sasaki, 2015; Deaner, 2018). These methods only need short panels but rely on completeness conditions. In comparison, although the proposed method requires a long panel, it imposes no completeness conditions and avoids solving an inverse problem.

Methodologically, this paper builds on the recent literature on nonlinear factor models for causal inference with panel data, particularly Deaner, Hsiang and Zeleneev (2025). The pseudo-distance approach to measuring similarity between latent factors was first developed by Zhang, Levina and Zhu (2017) for graphon estimation. Since then, different forms of the pseudo-distance have been used in a range of applications, including nonparametric graphon estimation (Zeleneev, 2020; Nowakowicz, 2024), network imputation (Sun and Zhang, 2026), controlling for unobserved heterogeneity using network data (Auerbach, 2022), nonlinear factor model estimation (Feng, 2023), and causal inference (Feng, 2020; Deaner, Hsiang and Zeleneev, 2025; Athey and Imbens, 2025; Hoshino and Yanagi, 2024). However, the validity of most pseudo-distance methods is established only under pure factor models, which rule out dynamic dependence in panel settings. I extend the pseudo-distance approach to dynamic models in which outcomes depend on both latent factors and lagged outcomes. This extension is new and is of independent interest for the study of dynamic nonseparable panel models beyond causal inference.

This paper contributes to the double/debiased machine learning literature (e.g., Chernozhukov et al., 2018) in causal inference by introducing a novel double cross-fitting procedure. Doubly robust estimation combines Neyman-orthogonal moments with cross-fitting to obtain valid inference on target parameters in the presence of nonparametric nuisance parameters.[1] I apply double cross-fitting (Newey and Robins, 2018; McClean et al., 2026) to dynamic ATT estimation with nonparametric regressions on *latent factors* and show that, when combined with undersmoothing, double cross-fitting achieves a faster convergence rate

[1] In the causal panel literature, doubly robust methods have been developed for DiD and TWFE models (Arkhangelsky and Imbens, 2022; Arkhangelsky et al., 2024; Sant'Anna and Zhao, 2020), dynamic treatment effects (Chernozhukov et al., 2022; Lewis and Syrgkanis, 2020), and linear factor models (Abadie et al., 2024). More recently, Feng (2020) and Deaner, Hsiang and Zeleneev (2025) extend doubly robust methods to nonlinear factor models. The proposed estimator builds directly on the doubly robust framework of Deaner, Hsiang and Zeleneev (2025) and likewise requires only weak smoothness conditions.

and valid inference under weaker dimensionality restrictions than standard cross-fitting. This result may be of independent interest for broader causal inference methods involving non-parametric nuisance parameters.

The rest of the paper is organized as follows. Section 2 presents the basic setup and the selection mechanism. Section 3 provides an overview of the proposed method. Section 4 establishes identification. Section 5 derives the asymptotic distribution of the proposed estimators. Section 6 extends the analysis to staggered adoption. Section 7 provides simulation evidence, and Section 8 applies the method to the setting studied by Bailey (2012) to reexamine the effects of U.S. family planning programs on fertility rates.

2 Selection and Model

This section introduces the basic setup of this paper. I discuss the treatment-selection problem in panel-data causal inference, focusing on settings where treatment assignment depends on both pretreatment information and latent heterogeneity. I then introduce a general outcome model that accommodates individual and time fixed effects and dynamic dependence. To fix ideas, I first focus on block assignment settings where all treated units receive treatment at the same time period. I extend the analysis to staggered adoption in Section 6.

2.1 Data

The panel data (Y_{it}, X_{it}, D_i) are indexed by $i = 1, \ldots, n$ and $t = 0, 1, \ldots, T$. Here, $Y_{it} \in \mathbb{R}$ denotes the outcome, and $X_{it} \in \mathbb{R}^{d_X}$ denotes a vector of time-varying covariates that may affect treatment selection. A leading example is that X_{it} includes lagged outcomes, e.g., $X_{it} := Y_{it-1}$. Treatment status is denoted by $D_i \in \{0, 1\}$. Multidimensional time-invariant latent heterogeneity is captured by $\alpha_i \in \mathcal{A} \subset \mathbb{R}^{d_\alpha}$. I consider a simultaneous-treatment setting in which treatment is assigned only once, at time T_0. Treatment is absorbing, so once a unit is treated at T_0, it remains treated thereafter. I assume that X_{iT_0} is realized before treatment assignment at T_0 and therefore constitutes pretreatment information.

Let $Y_{it}(0)$ and $Y_{it}(1)$ denote the untreated and treated potential outcomes, respectively. I impose no spillover effects across units, so that unit i's potential outcomes depend only on its own treatment status. All treated units receive treatment at period T_0. Then, under the

non-anticipation condition,

$$Y_{it} = \begin{cases} Y_{it}(0), & t = 0, \ldots, T_0 - 1, \\ D_i Y_{it}(1) + (1 - D_i) Y_{it}(0), & t = T_0, \ldots, T. \end{cases}$$

I also allow for an unobserved time factor γ_t, which captures aggregate shocks, such as macroeconomic conditions, that affect all units in the same period. Letting $\Gamma_t := (\ldots, \gamma_{t-1}, \gamma_t)$ denote the realized time-factor path up to time t, I assume that $(Y_{it}, X_{it}, \alpha_i, D_i)$ are i.i.d. across units conditional on fixed effects Γ_T.

Estimand For notational simplicity, define $\mathbb{E}_T(\cdot) := \mathbb{E}(\cdot \mid \Gamma_T)$ and $\mathbb{P}_T(\cdot) := \mathbb{P}(\cdot \mid \Gamma_T)$ as the expectation and probability conditional on the realized time-factor path Γ_T. For each $t = T_0, \ldots, T$, the estimand of interest is the dynamic ATT [2]:

$$\mathrm{ATT}(t) := \mathbb{E}_T \left(Y_{it}(1) - Y_{it}(0) \mid D_i = 1 \right).$$

2.2 Selection

Self-selection is common in economic analyses using observational data, and failing to account for such selection can lead to biased estimates of causal effects. Panel data provide a way to account for selection on unobserved time-invariant characteristics because they contain repeated observations for the same units. The recent and rapidly growing literature on causal panel methods, including DiD, synthetic control, and matrix completion methods, builds on this feature to control for selection on fixed effects.

However, most existing causal panel methods do not accommodate selection on time-varying covariates. For DiD, Ghanem, Sant'Anna and Wüthrich (2022) and Marx, Tamer and Tang (2024*b*) show that the parallel trends assumption fails when treatment decisions respond to both time-invariant characteristics and past outcomes. This concern is not limited to DiD but arises for many other causal panel methods (Arkhangelsky and Imbens, 2024). Although Arkhangelsky and Hirshberg (2023) show that selection on lagged outcomes is less relevant for synthetic control under simultaneous treatment, their argument does not directly extend to staggered adoption.

Ashenfelter's dip To illustrate why selection on past outcomes is problematic for panel causal methods, consider Ashenfelter's dip in the context of DiD (Ashenfelter, 1978; Ashen-

[2] The definition differs from the standard ATT, $\mathbb{E}\left(Y_{it}(1) - Y_{it}(0) \mid D_i = 1\right)$, only because I condition explicitly on the realized time fixed effects. Since these effects are fixed over the sample period, this is purely notational.

felter and Card, 1985). Labor economists find that participants often experience a decline in earnings before joining a job training program, which violates parallel trends. To see this, suppose that untreated earnings are i.i.d. over time. For participants, since the decline in earnings is temporary, their earnings would *recover* on average in the next period even without treatment. Therefore,

$$\mathbb{E}(Y_{iT_0}(0) - Y_{iT_0-1}(0) \mid D_i = 1) > 0.$$

By the same logic, we also have $\mathbb{E}(Y_{iT_0}(0) - Y_{iT_0-1}(0) \mid D_i = 0) < 0$ for nonparticipants. These differences violate the parallel trends assumption and lead to an upward-biased DiD estimator. Therefore, controlling for selection on Y_{iT_0-1} is necessary to disentangle the treatment effect from the dynamic recovery effects.

Conditional parallel trends A common practice in the DiD literature to allow for selection on pretreatment covariates is to impose the conditional parallel trends assumption (Abadie, 2005; Sant'Anna and Zhao, 2020; Chang, 2020; Caetano et al., 2022; Wooldridge, 2025). Assume $X_{it} := Y_{it-1}$ for simplicity. This assumption requires that treated and untreated units with the same lagged outcomes Y_{iT_0-1} have parallel trends, i.e.,

$$\mathbb{E}\left(Y_{iT_0}(0) - Y_{iT_0-1}(0) \mid D_i = 1, Y_{iT_0-1}\right) = \mathbb{E}\left(Y_{iT_0}(0) - Y_{iT_0-1}(0) \mid D_i = 0, Y_{iT_0-1}\right). \tag{2.1}$$

However, (2.1) may still fail when treatment selection depends jointly on α_i and Y_{iT_0-1}, even when untreated potential outcomes admit a *static* TWFE form. To see this, suppose $Y_{it}(0) = \alpha_i + \gamma_t + \epsilon_{it}$, where ϵ_{it} is exogenous and independent across t and $D_i \sim \text{Logit}(\alpha_i - Y_{iT_0-1})$. Cancelling $Y_{iT_0-1}(0)$ from both sides of (2.1) shows that the conditional parallel trends assumption holds if and only if

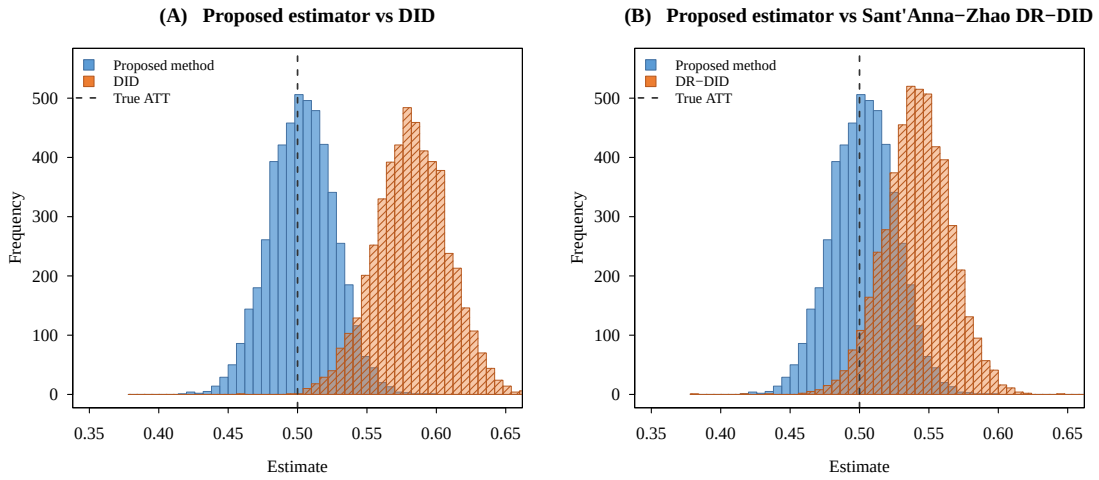
$$\mathbb{E}\left(Y_{iT_0}(0) \mid D_i = 1, Y_{iT_0-1}\right) = \mathbb{E}\left(Y_{iT_0}(0) \mid D_i = 0, Y_{iT_0-1}\right). \tag{2.2}$$

However, the selection mechanism implies that treated units have higher α_i than untreated units with the same Y_{iT_0-1}. Therefore, the LHS of (2.2) is greater than the RHS, violating the conditional parallel trends assumption.

Figure 1 presents a simple simulation to illustrate how treatment selection on fixed effects and lagged outcomes leads to biased estimates from both standard DiD and DiD based on conditional parallel trends, even when $Y_{it}(0)$ follows a static TWFE model and the treatment effect is homogeneous across units. The standard DiD estimator overestimates the treatment effect, as shown in Figure 1(A). In addition, the doubly robust DiD estimator proposed by

Sant'Anna and Zhao (2020), which relies on conditional parallel trends, is also heavily biased, as shown in Figure 1(B). By contrast, the distribution of the proposed estimator is centered on the true treatment effect.

Figure 1: Finite-sample distributions: proposed estimator, standard DiD, and DR DiD



Note: $(Y_{it}(0), Y_{it}(1))$ are generated by $Y_{it}(0) = \alpha_i + \gamma_t + \epsilon_{Y,it}$ and $Y_{it}(1) = Y_{it}(0) + 0.5$. α_i and γ_t are independent random variables drawn from the uniform distribution on $(-\frac{1}{2}, \frac{1}{2})$. $\{\epsilon_{Y,it}\}$ are independent of (α_i, γ_t) and are i.i.d. across both dimensions, uniformly distributed on $(-\frac{1}{2}, \frac{1}{2})$. In addition, $D_i = \mathbf{1}(2\alpha_i - Y_{iT_0-1} + \epsilon_{D,i} \geqslant 0)$, where $\{\epsilon_{D,i}\}$ are i.i.d. logistic random variables independent of $\{\alpha_i, \gamma_t, \epsilon_{Y,it}\}$.

I formalize the selection mechanism as follows:

Assumption 2.1 (Selection) *For each $t = T_0, \ldots, T$,*

$$Y_{it}(0) \perp D_i \mid X_{iT_0}, \alpha_i, \Gamma_T. \tag{2.3}$$

Assumption 2.1 requires that, after conditioning on pretreatment covariates X_{iT_0} and time-invariant heterogeneity α_i, treatment assignment is independent of the untreated potential outcome $Y_{it}(0)$ for each $t = T_0, \ldots, T$. I also explicitly condition on Γ_T to maintain the fixed-effect interpretation of the time factors. The conditional independence is imposed only on untreated potential outcomes, rather than on treated potential outcomes, because the estimand of interest is the ATT and only untreated counterfactual outcomes for treated individuals need to be imputed.

Assumption 2.1 extends the conventional selection-on-fixed-effects assumption by additionally allowing treatment selection to depend on time-varying pretreatment covariates. In addition, this assumption nests the sequential unconfoundedness assumption (Robins, Hernan and Brumback, 2000; Viviano and Bradic, 2026; Marx, Tamer and Tang, 2024*a*),

$Y_{it}(0) \perp D_i \mid X_{iT_0}$, as a special case in which treatment selection does not depend on unobserved fixed effects. Lastly, Assumption 2.1 arises naturally in dynamic economic models (Heckman and Navarro, 2007), in which a forward-looking agent chooses treatment to maximize expected utility based on the state variables (X_{iT_0}, α_i).

Remark. *Since X_{iT_0} is realized before treatment assignment, it cannot include potential outcomes at or after treatment assignment, such as $(Y_{iT_0}(0), Y_{iT_0}(1))$. This timing restriction therefore rules out anticipation effects (see the discussion in Roth et al. (2023)) and Roy-type selection. In addition, Assumption 2.1 rules out unobserved time-varying confounders that jointly affect treatment assignment and untreated potential outcomes. Addressing such confounding typically requires additional information, such as an external instrument. I refer readers to Gulek and Vives-i Bastida (2025), who combine synthetic control and instrumental variables to address such confounding.*

The following assumption imposes a Markov restriction, under which earlier histories have no additional predictive power for future untreated outcomes or treatment assignment when conditioning on (X_{iT_0}, α_i).

Assumption 2.2 (Markov) *Let $H_{it} := \{(Y_{i\tau}, X_{i\tau})\}_{\tau \leqslant t}$ denote the history up to t. Assume*

$$(Y_{it}(0), D_i) \perp H_{iT_0-1} \mid X_{iT_0}, \alpha_i, \Gamma_T, \quad \forall\, t = T_0, \ldots, T.$$

This assumption is stronger than necessary and can be relaxed to conditional moment restrictions on $(Y_{it}(0), D_i)$[3]. I retain the Markov restriction in this paper for ease of exposition.

2.3 Dynamic nonparametric panel

I assume that the untreated outcome $Y_{it}(0)$ is generated according to

$$Y_{it}(0) = f(X_{it}, \alpha_i, \gamma_t) + \epsilon_{it}, \quad \mathbb{E}_T(\epsilon_{it} \mid H_{it-1}, X_{it}, \alpha_i) = 0. \tag{2.4}$$

Here, $f(\cdot)$ is a data-generating function, α_i and γ_t are unobserved individual and time fixed effects, respectively. The exogeneity condition $\mathbb{E}_T(\epsilon_{it} \mid H_{it-1}, X_{it}, \alpha_i) = 0$ corresponds to the weak exogeneity condition in structural panel model literature that allows for feedbacks (Bonhomme, 2025). The data-generating process in (2.4) is flexible. I impose no

[3]Specifically, let $Z_{it} = (Y_{it}(0), \alpha_i)'$, it suffices to require that, for each $t = T_0, \ldots, T$,

$$\mathbb{E}_T(Z_{it} \mid H_{iT_0-1}, X_{iT_0}, \alpha_i) = \mathbb{E}_T(Z_{it} \mid X_{iT_0}, \alpha_i), \text{ and } \mathbb{E}_T(Z_{it}Z_{it}' \mid H_{iT_0-1}, X_{iT_0}, \alpha_i) = \mathbb{E}_T(Z_{it}Z_{it}' \mid X_{iT_0}, \alpha_i).$$

parametric restrictions on the function f, allowing for rich interactions between individual-specific and time-specific heterogeneity. In addition, I do not specify the joint distribution of X_{it} and the latent factors.

The model accommodates dynamic dependence in untreated potential outcomes when $X_{it} := (Y_{it-1}(0), \ldots, Y_{it-p}(0))$. Dynamic dependence is ubiquitous in economics. Aggregate outcomes such as local employment and fertility rates often exhibit dynamic dependence because shocks are persistent over time, while individual outcomes may also exhibit state dependence, i.e., past employment may affect present employment (Heckman, 1981). Such dynamic dependence is particularly relevant in staggered-adoption settings, as it allows researchers to disentangle treatment effects from dependence on lagged outcomes.

Remark. *The specification (2.4) can be motivated by a nonseparable panel model:*

$$Y_{it}(0) = h(X_{it}, \alpha_i, \gamma_t, \omega_{it}),$$

where ω_{it} is an unobserved shock. I allow its distribution to depend on $(X_{it}, \alpha_i, \gamma_t)$, which is more general than the time-stationarity condition imposed in the nonseparable panel literature (e.g., Chernozhukov et al., 2013, 2026). If I further suppose that ω_{it} is independent of the previous history conditional on $(X_{it}, \alpha_i, \Gamma_T)$, i.e., $\omega_{it} \perp H_{it-1} \mid X_{it}, \alpha_i, \Gamma_T$, then $f(\cdot)$ in (2.4) can be regarded as the projection of the untreated outcome $Y_{it}(0)$, and ϵ_{it} is the corresponding projection error.

Model (2.4) incorporates a number of important data-generating processes that are commonly used in the structural and causal panel literature. Here are some examples.

Example 1 (Linear panel with IFE) *Consider the following linear panel model with covariates X_{it} and interactive fixed effects:*

$$Y_{it}(0) = \beta' X_{it} + \gamma_t' \alpha_i + \epsilon_{it}, \quad \mathbb{E}_T(\epsilon_{it} \mid H_{it-1}, X_{it}, \alpha_i) = 0. \tag{2.5}$$

Here, γ_t is a vector of macroeconomic shocks, e.g., technology shocks and financial crises, and α_i captures heterogeneous responses to these shocks. For example, in labor economics, Y_{it} represents the wage rate, X_{it} represents experience, α_i is a vector of unobserved skills, and γ_t contains the time-varying market returns to these skills. Changes in technology or labor-market conditions therefore affect workers differently according to their latent skill composition.

Model (2.5) is studied in Bai (2009) and nests many commonly used untreated-outcome specifications in causal panel methods. For example, setting $\beta = 0$ and imposing an additive factor structure gives $Y_{it}(0) = \alpha_i + \gamma_t + \epsilon_{it}$, which is the untreated-outcome structure in

standard DiD designs. Setting $\beta = 0$ while allowing unrestricted interactive fixed effects gives $Y_{it}(0) = \alpha_i'\gamma_t + \epsilon_{it}$, which is typically assumed to support recently developed synthetic control and related factor-based causal panel methods (Arkhangelsky et al., 2021; Athey et al., 2021; Bai and Ng, 2021; Chernozhukov et al., 2023).

Example 2 (Dynamic linear panel with IFE) *Consider the following dynamic linear panel model with $X_{it} := Y_{it-1}(0)$ and interactive fixed effects:*

$$Y_{it}(0) = \rho Y_{it-1}(0) + \alpha_i'\gamma_t + \epsilon_{it}, \quad \mathbb{E}_T(\epsilon_{it} \mid \{Y_{i\tau}(0)\}_{\tau \leqslant t-1}, \alpha_i) = 0. \tag{2.6}$$

Model (2.6) captures outcome persistence through the lagged outcome $Y_{it-1}(0)$ and allows individuals to respond heterogeneously to time-varying aggregate shocks through the interactive term $\alpha_i'\gamma_t$. This model is studied in Moon and Weidner (2017). A special case with additive fixed effects is $Y_{it}(0) = \rho Y_{it-1}(0) + \alpha_i + \gamma_t + \epsilon_{it}$, which is a benchmark specification in the dynamic panel literature (Arellano and Bond, 1991).

Example 3 (Dynamic binary panel with IFE) *Consider a binary response panel model,*

$$Y_{it}(0) = \mathbf{1}(\beta Y_{it-1}(0) + \alpha_i'\gamma_t - u_{it} \geqslant 0), \tag{2.7}$$

where the exogenous errors $\{u_{it}\}$ are independent (across i and t) draws from a distribution with cumulative distribution function $F(\cdot)$.[4] The binary panel (2.7) allows for heterogeneous responses of units to aggregate shocks and state dependence, e.g., women's labor-force participation decisions may depend on their employment histories, as those who have been out of the labor market for a long time may be less likely to re-enter. When $F(\cdot)$ can stand for the logistic distribution, the data generating process $f(\cdot)$ as in (2.4) takes the form

$$f(Y_{it-1}(0), \alpha_i, \gamma_t) = \frac{\exp\left(\beta Y_{it-1}(0) + \alpha_i'\gamma_t\right)}{1 + \exp\left(\beta Y_{it-1}(0) + \alpha_i'\gamma_t\right)},$$

and ϵ_{it} is a centered Bernoulli error with $\mathbb{E}_T(\epsilon_{it} \mid \{Y_{i\tau}(0)\}_{\tau \leqslant t-1}, \alpha_i) = 0$.

Assumption 2.3 (Potential Outcome) *$Y_{it}(0)$ evolves according to (2.4) and satisfies:*

(i) ***(Time factor exogeneity)*** *conditional on the individual latent factors α_i and the time-factor path Γ_t, $Y_{it}(0)$ is mean independent of future time factors, i.e.,*

$$\mathbb{E}\left(Y_{it}(0) \mid X_{it}, \alpha_i, \Gamma_T\right) = \mathbb{E}\left(Y_{it}(0) \mid X_{it}, \alpha_i, \Gamma_t\right), \quad \forall t = 1, \ldots, T.$$

[4] E.g., $F(\cdot)$ can stand for the logistic or standard normal CDF in Logit and Probit models, respectively.

(ii) ***(Weak dependence)*** $(\ldots, Y_{iT_0-1}(0), Y_{iT_0}(0), \ldots)$ *are conditionally weakly dependent, such that* $\sum_{s=-\infty}^{\infty} |\text{Cov}\left(Y_{it}(0), Y_{is}(0) \mid \alpha_i = \alpha, \Gamma_T = \Gamma\right)| < \infty$ *uniformly for any supported* α*, any possible time path* Γ*, and any* $t \in \mathbb{Z}$*.*

Assumption 2.3(i) requires that future time factors provide no additional information for predicting $Y_{it}(0)$ when conditioning on $(\alpha_i, X_{it}, \Gamma_t)$. This restriction is mild and has natural interpretations. First, it reflects the principle that the future cannot affect the present. This assumption does not preclude forward-looking behavior based on expectations formed using the macroeconomic information set Γ_t. Second, since $(\gamma_{t+1}, \gamma_{t+2}, \ldots)$ capture aggregate conditions, they should not be affected by any single individual's behavior.

Assumption 2.3(ii) states that, conditional on the fixed effects (α_i, Γ_T), the serial dependence of the untreated outcome vanishes sufficiently fast such that the sum of the absolute conditional autocovariances is finite. This condition holds uniformly over any supported individual fixed effects α_i and any possible time path Γ_T. This conditional weak dependence requirement is also mild since persistence captured by the individual and time fixed effects, α_i and Γ_T, is unrestricted.

The conditional weak dependence assumption is crucial to the proposed method. I discuss sufficient conditions under which it holds in Examples 1-3.

Example 1 (Linear panel with IFE - Continued) *It is common in the literature to assume that* X_{it} *also admits a factor structure (Bai, 2009). For simplicity, set* $d_X = 1$*. Then,*

$$Y_{it}(0) = \beta X_{it} + \gamma_{Y,t}'\alpha_{Y,i} + \epsilon_{Y,it}, \quad \mathbb{E}_T(\epsilon_{Y,it} \mid H_{it-1}, X_{it}, \alpha_i) = 0;$$
$$X_{it} = \gamma_{X,t}'\alpha_{X,i} + \epsilon_{X,it}, \quad \mathbb{E}_T(\epsilon_{X,it} \mid H_{it-1}, \alpha_i) = 0.$$

Here, $\alpha_{Y,i}$ *and* $\alpha_{X,i}$ *are (possibly overlapping) subvectors of* α_i*, and* $\gamma_t := (\gamma_{Y,t}', \gamma_{X,t}')'$*. Assumption 2.3(ii) holds if I additionally assume that the idiosyncratic terms* $(\epsilon_{X,it}, \epsilon_{Y,it})$ *are not persistent over time, i.e., for all* $\alpha \in \mathcal{A}$*, all possible time paths* Γ*, and any* $t \in \mathbb{Z}$*,*

$$\sum_{s=-\infty}^{\infty} |\text{Cov}\left(\epsilon_{Y,it}, \epsilon_{Y,is} \mid \alpha_i = \alpha, \Gamma_T = \Gamma\right)| < \infty, \quad \sum_{s=-\infty}^{\infty} |\text{Cov}\left(\epsilon_{X,it}, \epsilon_{X,is} \mid \alpha_i = \alpha, \Gamma_T = \Gamma\right)| < \infty,$$

which are standard in the factor-model literature. The result extends to multidimensional X_{it} *by imposing the same weak-dependence condition on each component of* $\epsilon_{X,it}$*.*

Example 2 (Dynamic linear panel with IFE - Continued) *For the dynamic linear model with IFE, (2.6) in Example 2, it is straightforward to verify that Assumption 2.3(ii) holds if I additionally assume that (i)* $|\rho| < 1$*, and (ii) the idiosyncratic terms* (ϵ_{it}) *are not persistent*

over time, i.e., for all $\alpha \in \mathcal{A}$*, all possible time paths* Γ*, and any* $t \in \mathbb{Z}$*,*

$$\sum_{s=-\infty}^{\infty} |\text{Cov}\left(\epsilon_{it}, \epsilon_{is} \mid \alpha_i = \alpha, \Gamma_T = \Gamma\right)| < \infty.$$

The condition $|\rho| < 1$ *is standard in the time-series literature to ensure that the serial dependence decays geometrically over time. The result extends to higher-order lags as long as all roots of the autoregressive characteristic polynomial lie outside the unit circle.*

Example 3 (Dynamic nonlinear panel - Continued) *For the dynamic model (2.7) in Example 3, since* $\{u_{it}\}_{t\in\mathbb{Z}}$ *is independent across* t*,* $\{Y_{it}(0)\}_{t\in\mathbb{Z}}$ *is a time-inhomogeneous Markov process conditional on the fixed effects* (α_i, Γ_T)*.*

If I further assume that (i) the cumulative distribution function $F(\cdot)$ *is continuous and its derivative satisfies* $F^{(1)}(\cdot) > 0$*, and (ii) the supports of* α_i *and* γ_t *are uniformly bounded across* i *and* t*, then each entry of the transition matrix is bounded away from* 0 *and* 1 *uniformly over all periods. This ensures that Assumption 2.3(ii) holds. The commonly used logit and probit specifications satisfy these restrictions on* $F(\cdot)$*.*

3 Method

This section provides an overview of the proposed method, which uses pretreatment histories to recover information about latent heterogeneity, and introduces a doubly robust estimator for the dynamic ATT. I focus on simultaneous treatment settings and formally extend the analysis to staggered adoption in Section 6.

3.1 Recovering latent heterogeneity

The key technical challenge is that the individual fixed effects α_i are unobserved by econometricians. If α_i were observed, the problem would fall under selection on observables, and one could directly apply existing methods to estimate the dynamic ATT. Therefore, the problem of estimating dynamic ATT reduces to recovering information about the latent factors α_i from the observed information.

To see the basic idea, first consider a pure factor model without dynamics:

$$Y_{it}(0) = f(\alpha_i, \gamma_t) + \epsilon_{it}, \quad \mathbb{E}(\epsilon_{it} \mid \alpha_i, \Gamma_T) = 0.$$

This is a nonlinear factor model in which the untreated outcome depends on an individual latent factor α_i and a time-varying aggregate factor γ_t, and $\{\epsilon_{it}\}$ are weakly dependent across

time. If two individuals i and j have similar latent factors, then, for each t, the smoothness of $f(\cdot)$ implies that $f(\alpha_i, \gamma_t) \approx f(\alpha_j, \gamma_t)$. Thus, individuals with similar latent heterogeneity should have similar untreated histories up to idiosyncratic innovations. This observation provides the intuition for the reverse direction: similarity in pretreatment histories can be informative about similarity in latent factors. This is the logic behind the pseudo-distance approach studied in Feng (2023) and Deaner, Hsiang and Zeleneev (2025).

However, this idea does not directly extend to models with dynamic effects. This is because even if two individuals i and j have similar latent factors, the terms $f(X_{it}, \alpha_i, \gamma_t)$ and $f(X_{jt}, \alpha_j, \gamma_t)$ can still differ when their covariates are different. Despite this complication, I show that, even in the presence of dynamic effects, the latent factors can still be recovered from pretreatment histories as long as the sequence is conditionally weakly dependent as in Assumption 2.3(ii). Let us first consider the linear model in Example 2:

$$Y_{it}(0) = \rho Y_{it-1}(0) + \alpha_i' \gamma_t + \epsilon_{it}, \quad \mathbb{E}(\epsilon_{it} \mid \{Y_{i\tau}(0)\}_{\tau \leqslant t-1}, \alpha_i, \Gamma_T) = 0.$$

I impose $|\rho| < 1$ to ensure that the sequence is conditionally weakly dependent. Iterating the model backward gives, for any lag $L > 0$,

$$Y_{it}(0) = \rho^L Y_{it-L}(0) + \alpha_i' \sum_{\tau=0}^{L-1} \rho^\tau \gamma_{t-\tau} + \sum_{\tau=0}^{L-1} \rho^\tau \epsilon_{it-\tau}.$$

Letting $L \to \infty$, the first term vanishes under $|\rho| < 1$. Hence,

$$Y_{it}(0) = \underbrace{\alpha_i' \sum_{\tau=0}^{\infty} \rho^\tau \gamma_{t-\tau}}_{:= g(\alpha_i, \Gamma_t)} + \underbrace{\sum_{\tau=0}^{\infty} \rho^\tau \epsilon_{it-\tau}}_{\tilde{\epsilon}_{it}} = g(\alpha_i, \Gamma_t) + \tilde{\epsilon}_{it}.$$

Thus, the dynamic linear model admits a transformed pure factor representation. The transformed innovation $\tilde{\epsilon}_{it}$ is weakly dependent over time conditional on (α_i, Γ_T), and satisfies $\mathbb{E}(\tilde{\epsilon}_{it} \mid \alpha_i, \Gamma_T) = 0$ by sequential exogeneity. This implies that the pseudo-distance method based on pure factor models can also be applied to dynamic linear models. The intuition is that the dynamic effect of the initial condition vanishes, i.e., $\lim_{L \to \infty} \rho^L Y_{it-L}(0) = 0$, and the behavior of the path is eventually driven by latent heterogeneity α_i.

I now extend the idea from the linear dynamic model to nonlinear cases. Motivated by the idea of backward iteration, I project $Y_{it}(0)$ onto the space of functions of (α_i, Γ_T):

$$Y_{it}(0) = \underbrace{\mathbb{E}(Y_{it}(0) \mid \alpha_i, \Gamma_T)}_{:= g(\alpha_i, \Gamma_t)} + \underbrace{Y_{it}(0) - \mathbb{E}(Y_{it}(0) \mid \alpha_i, \Gamma_T)}_{\tilde{\epsilon}_{it}} = g(\alpha_i, \Gamma_t) + \tilde{\epsilon}_{it}. \tag{3.1}$$

Here, since future time factors provide no additional information about $Y_{it}(0)$ conditional on (α_i, Γ_t), the transformed pure factor representation $g(\alpha_i, \Gamma_t)$ depends only on time factors up to period t. In addition, the projection error $\tilde{\epsilon}_{it}$ is weakly dependent over time conditional on (α_i, Γ_T) and satisfies $\mathbb{E}(\tilde{\epsilon}_{it} \mid \alpha_i, \Gamma_T) = 0$. Thus, I show that the dynamic nonlinear model also admits a transformed pure factor representation.

Remark. *While the pure factor representation (3.1) provides information about similarities in α_i, it is not sufficient to identify treatment effects because $\tilde{\epsilon}_{it}$ is not mean independent of treatment conditional on α_i and Γ_t alone.*

Example 3 provides a concrete example of this transformed pure factor representation under nonlinear models.

Example 3 (Dynamic nonlinear panel - Continued) *Suppose that $Y_{it}(0)$ evolves as in Example 3 and let $F(\cdot)$ denote the cumulative distribution function of u_{it}. For notational simplicity, define $\rho_{it} := F(\beta + \gamma_t'\alpha_i) - F(\gamma_t'\alpha_i)$ as the time-inhomogeneous discount factor, which satisfies $\sup_{i,t} |\rho_{it}| < 1$. It is straightforward to verify that one-step backward iteration gives $g(\alpha_i, \Gamma_t) = F(\gamma_t'\alpha_i) + \rho_{it} g(\alpha_i, \Gamma_{t-1})$. Then, for any $L > 0$, L-step backward iteration gives*

$$g(\alpha_i, \Gamma_t) = \sum_{\tau=0}^{L-1} \left(\prod_{s=0}^{\tau-1} \rho_{i,t-s} \right) F\left(\gamma_{t-\tau}'\alpha_i\right) + \left(\prod_{s=0}^{L-1} \rho_{i,t-s} \right) g(\alpha_i, \Gamma_{t-L}),$$

where the empty product is equal to one. Letting $L \to \infty$, the last term vanishes as $\sup_{i,t} |\rho_{it}| < 1$. Hence,

$$g(\alpha_i, \Gamma_t) = \sum_{\tau=0}^{\infty} \left(\prod_{s=0}^{\tau-1} \rho_{i,t-s} \right) F\left(\gamma_{t-\tau}'\alpha_i\right).$$

A rigorous treatment is provided in the appendix.

The pure factor representation of the dynamic model allows us to compare individuals through their pretreatment histories. Intuitively, when pretreatment histories are informative about latent factors, closeness between the pretreatment histories of individuals i and j implies similarity between their latent factors, i.e., $\|\alpha_i - \alpha_j\| \approx 0$.

A natural measure of the difference between pretreatment histories is the squared L_2 distance. Specifically, for any i, j,

$$\widehat{d}_{2,ij}^2 := \frac{1}{T_0} \sum_{t=0}^{T_0-1} (Y_{it} - Y_{jt})^2.$$

However, $\widehat{d}^2_{2,ij}$ cannot serve as a good distance for comparing histories, because it captures not only differences in the factor component of the history but also the conditional variance of idiosyncratic innovations (see the discussion in Appendix B.1). As an alternative, I employ the pseudo-distance approach from the nonlinear factor literature (Feng, 2023; Deaner, Hsiang and Zeleneev, 2025). Specifically, for any i, j, define the pseudo-distance as:

$$\widehat{d}_{ij} = \max_{\substack{k_1,k_2=1,\ldots,n,\\ k_1,k_2\neq i,j}} \left| \frac{1}{T_0} \sum_{t=0}^{T_0-1} (Y_{k_1 t} - Y_{k_2 t})(Y_{it} - Y_{jt}) \right|. \tag{3.2}$$

Unlike the L_2 distance, the pseudo-distance effectively filters the idiosyncratic innovations by comparing the history difference between i and j against history differences between other individuals. It therefore provides a cleaner measure of similarity in the denoised histories.

The key requirement for using the pseudo-distance $\widehat{d}_{ij}$ as a proxy for the latent distance $\|\alpha_i - \alpha_j\|$ is *informativeness*: individuals who are close in terms of the pseudo-distance should also be close in their underlying latent factors. This condition is related to completeness conditions in nonparametric identification literature. Under the informativeness condition, matching on latent factors reduces to matching units with small pseudo-distances. Therefore, standard nonparametric methods, such as kernel regression or k-nearest-neighbor methods, can be used to impute missing potential outcomes and estimate propensity scores. I formally define the informativeness condition and discuss sufficient conditions under which it holds in Sections 4 and 5.

Remark. ***(Alternatives to long histories)*** *This paper uses long pretreatment histories to recover time-invariant heterogeneity, consistent with synthetic control and matrix completion methods. However, this is not the only way to do so. For example, when the panel is short but rich cross-sectional variables are available, these variables can instead be used to recover latent heterogeneity. This idea is closely related to factor-augmented regression (Stock and Watson, 2002) and the method of Feng (2020).*

3.2 Estimator

For any $t = T_0, \ldots, T$, define the regression outcome and propensity score as

$$\begin{aligned} m_t(x, \alpha) &:= \mathbb{E}_T(Y_{it}(0) \mid X_{iT_0} = x, \alpha_i = \alpha), \\ \pi(x, \alpha) &:= \mathbb{P}_T(D_i = 1 \mid X_{iT_0} = x, \alpha_i = \alpha). \end{aligned} \tag{3.3}$$

Here, the function $m_t : \mathbb{R}^{d_X} \times \mathbb{R}^{d_\alpha} \mapsto \mathbb{R}$ denotes the conditional expectation of the untreated outcome at time t given pretreatment information[5], and the function $\pi : \mathbb{R}^{d_X} \times \mathbb{R}^{d_\alpha} \mapsto [0,1]$ denotes the propensity score given individuals' pretreatment information.

For notational convenience, let $m_{it} := m_t(X_{iT_0}, \alpha_i)$ and $\pi_i := \pi(X_{iT_0}, \alpha_i)$. Under the selection mechanism in Assumption 2.1, an oracle doubly robust estimator for $\mathrm{ATT}(t)$ reads

$$\widehat{\mathrm{ATT}(t)}^{\text{oracle}} := \frac{1}{n_1} \sum_{i=1}^{n} \left(D_i Y_{it} - \frac{(1 - D_i)\pi_i Y_{it} + (D_i - \pi_i) m_{it}}{1 - \pi_i} \right), \tag{3.4}$$

where $n_1 := \sum_{i=1}^{n} \mathbf{1}(D_i = 1)$. The estimator is oracle in the sense that it uses the true propensity score π_i and outcome regression m_{it}, which are unknown in practice.

Nadaraya-Watson estimator If latent heterogeneity α_i were observable, researchers could apply the Nadaraya-Watson (NW) estimator to nonparametrically estimate the functions $m_t(\cdot,\cdot)$ and $\pi(\cdot,\cdot)$, and thus m_{it} and π_i. Since the pseudo-distance $\widehat{d}_{ij}$ serves as a proxy for the latent distance $\|\alpha_i - \alpha_j\|$, I replace the latent distance with the pseudo-distance $\widehat{d}_{ij}$ and obtain a feasible NW estimator. Specifically, for any $i, j = 1, \ldots, n$, define the kernel weight

$$\widehat{K}_{h,ij} := K\left(\left(X_{jT_0} - X_{iT_0} \right) / h \right) K\left(\widehat{d}_{ij} / h \right).$$

Here, $K(\cdot)$ denotes the (product) kernel function and $h \to 0$ is the bandwidth. The hat notation emphasizes that the kernel weight is constructed using the estimated pseudo-distance $\widehat{d}_{ij}$. The Nadaraya-Watson estimators based on the pseudo-distance are given by

$$\widehat{m}_{it} = \sum_{j=1}^{n} (1 - D_j) Y_{jt} \widehat{K}_{h,ij} \Big/ \sum_{j=1}^{n} (1 - D_j) \widehat{K}_{h,ij}, \quad \widehat{\pi}_i = \sum_{j=1}^{n} D_j \widehat{K}_{h,ij} \Big/ \sum_{j=1}^{n} \widehat{K}_{h,ij}.$$

Plugging these estimates into (3.4) yields a feasible doubly robust estimator:

$$\widehat{\mathrm{ATT}(t)} := \frac{1}{n_1} \sum_{i=1}^{n} \left(D_i Y_{it} - \frac{(1 - D_i)\widehat{\pi}_i Y_{it} + (D_i - \widehat{\pi}_i) \widehat{m}_{it}}{1 - \widehat{\pi}_i} \right). \tag{3.5}$$

This estimator is doubly robust in the sense that it is consistent if either $\widehat{m}_{it}$ or $\widehat{\pi}_i$ is consistent. In the double machine learning literature, doubly robust estimators are combined with standard cross-fitting, which uses separate samples to obtain the nuisance estimates $(\widehat{m}_{it}, \widehat{\pi}_i)$ and evaluate the ATT score, to achieve a faster convergence rate and make the bias arising

[5]It is straightforward to verify that, for contemporaneous outcomes, $m_{T_0}(x, \alpha) = f(x, \alpha, \gamma_{T_0})$.

from nuisance estimation negligible for inference (Chernozhukov et al., 2018).

Double cross-fitting Although standard cross-fitting eliminates overfitting bias, the estimation errors in $\widehat{m}_{it}$ and $\widehat{\pi}_i$ remain correlated because $(\widehat{m}_{it}, \widehat{\pi}_i)$ are estimated using the same subsample. I further employ a *double* cross-fitting procedure (Newey and Robins, 2018; McClean et al., 2026) to estimate the outcome regressions and propensity scores on separate samples, thereby reducing the dependence between their estimation errors.

The double cross-fitting is implemented as follows. I randomly partition the sample into three separate folds, $\{\mathcal{I}_1, \mathcal{I}_2, \mathcal{I}_3\}$. For each unit i, let $\mathcal{I}(i)$ denote the fold containing unit i, and let $\mathcal{I}_m(i)$ and $\mathcal{I}_\pi(i)$ denote the folds used to estimate $\widehat{m}_{it}$ and $\widehat{\pi}_i$ for unit i, respectively[6]. In addition, different bandwidths $\{h_m, h_\pi\}$ are used to estimate m_{it} and π_i, respectively. Therefore, the NW estimators under double cross-fitting are given by

$$\widehat{m}_{it} = \frac{\sum_{j \in \mathcal{I}_m(i)} (1 - D_j) Y_{jt} \widehat{K}_{h_m, ij}}{\sum_{j \in \mathcal{I}_m(i)} (1 - D_j) \widehat{K}_{h_m, ij}}, \quad \widehat{\pi}_i = \frac{\sum_{j \in \mathcal{I}_\pi(i)} D_j \widehat{K}_{h_\pi, ij}}{\sum_{j \in \mathcal{I}_\pi(i)} \widehat{K}_{h_\pi, ij}}.$$

Plugging these estimates into (3.5) yields the proposed estimator for ATT(t). To highlight the main idea of the proposed algorithm, a simplified version that abstracts from cross-validation for selecting the bandwidths $\{h_m, h_\pi\}$ is summarized in Algorithm 1.

I show that, compared with standard cross-fitting, combining double cross-fitting with undersmoothing yields faster convergence and relaxes the dimensionality restrictions associated with nonparametric nuisance estimation. This delivers a root-n consistent, asymptotically normal, and asymptotically unbiased ATT estimator under more general conditions. These results are novel in the literature and contribute to the application of double machine learning to causal inference, particularly in settings with latent confounding.

4 Theory for Identification

This section establishes identification of the dynamic ATT. If α_i were observed or identified, the selection mechanism would reduce to selection on observables, and the dynamic ATT could be identified directly under standard conditions. I show that, although α is unobserved, the pseudo-distance is informative about latent heterogeneity, thereby enabling a matching strategy that identifies the dynamic ATT.

[6] For example, when $i \in \mathcal{I}_1$, I use $\mathcal{I}_m(i) = \mathcal{I}_2$ to estimate the potential-outcome regression and obtain $\widehat{m}_{it}$, and use $\mathcal{I}_\pi(i) = \mathcal{I}_3$ to estimate the propensity score and obtain $\widehat{\pi}_i$. More broadly, if $i \in \mathcal{I}_s$, then $\mathcal{I}(i) = \mathcal{I}_s$, $\mathcal{I}_m(i) = \mathcal{I}_{1+(s \bmod 3)}$, and $\mathcal{I}_\pi(i) = \mathcal{I}_{1+((s+1) \bmod 3)}$.

Algorithm 1 Doubly robust estimator for dynamic ATT (Basic Idea)

Input: Kernel $K(\cdot)$, bandwidths h_m, h_π, and confidence level $1-\alpha$.

Output: Dynamic ATT estimates $\widehat{\mathrm{ATT}}(t)$, standard errors $\widehat{\mathrm{se}}_t$, and level $1-\alpha$ confidence interval.

Step 1: Sample-splitting
Randomly partition the sample into 3 disjoint folds $\{\mathcal{I}_1, \mathcal{I}_2, \mathcal{I}_3\}$ of equal sizes.

Step 2: Calculate similarity
For each $i, j = 1, \ldots, n$, calculate the pseudo-distance using pretreatment outcomes:

$$\widehat{d}_{ij} = \max_{\substack{k_1,k_2=1,\ldots,n,\\ k_1,k_2\neq i,j}} \left| \frac{1}{T_0} \sum_{t=0}^{T_0-1} (Y_{k_1 t} - Y_{k_2 t})(Y_{it} - Y_{jt}) \right|.$$

Step 3: Compute expected outcomes and propensity scores
The Nadaraya-Watson estimators under double cross-fitting are given by

$$\widehat{m}_{it} = \frac{\sum_{j\in\mathcal{I}_m(i)} (1-D_j) Y_{jt} \widehat{K}_{h_m,ij}}{\sum_{j\in\mathcal{I}_m(i)} (1-D_j) \widehat{K}_{h_m,ij}}, \quad \widehat{\pi}_i = \frac{\sum_{j\in\mathcal{I}_\pi(i)} D_j \widehat{K}_{h_\pi,ij}}{\sum_{j\in\mathcal{I}_\pi(i)} \widehat{K}_{h_\pi,ij}}.$$

For any $i, j = 1, \ldots, n$ and $h > 0$, the kernel weight $\widehat{K}_{h,ij}$ is defined as:

$$\widehat{K}_{h,ij} := K\left(\left(X_{jT_0} - X_{iT_0} \right) / h \right) K\left(\widehat{d}_{ij} / h \right).$$

Step 4: Construct $\widehat{\mathrm{ATT}}(t)$, standard error $\widehat{\mathrm{se}}_t$, and confidence interval
Let $n_1 := \sum_{i=1}^n \mathbf{1}(D_i = 1)$, and compute

$$\widehat{\mathrm{ATT}}(t) = \frac{1}{n_1} \sum_{i=1}^n \left(D_i Y_{it} - \frac{(1-D_i)\widehat{\pi}_i Y_{it} + (D_i - \widehat{\pi}_i)\widehat{m}_{it}}{1-\widehat{\pi}_i} \right),$$

$$\widehat{\mathrm{se}}_t = \left(\frac{1}{n_1^2} \sum_{i=1}^n \left(D_i Y_{it} - \frac{(1-D_i)\widehat{\pi}_i Y_{it} + (D_i - \widehat{\pi}_i)\widehat{m}_{it}}{1-\widehat{\pi}_i} - D_i \widehat{\mathrm{ATT}}(t) \right)^2 \right)^{1/2}.$$

The confidence interval is $\mathrm{CI}_t = [\widehat{\mathrm{ATT}}(t) \pm Z_{1-\alpha/2} \cdot \widehat{\mathrm{se}}_t]$.

Pseudo-distance I first discuss conditions under which the sample pseudo-distance $\widehat{d}_{ij}$ serves as a proxy for the latent distance. The argument proceeds in two steps: (i) I show that, as $n, T_0 \to \infty$, $\widehat{d}_{ij}$ converges to a *population* pseudo-distance $d(\alpha_i, \alpha_j)$ that depends only on the latent heterogeneity α_i and α_j, and (ii) I introduce an informativeness condition under which the population pseudo-distance $d(\alpha_i, \alpha_j)$ reveals the latent distance $\|\alpha_i - \alpha_j\|$.

Establishing the population limit of $\widehat{d}_{ij}$ requires the underlying time averages to converge as $T_0 \to \infty$, which in turn requires restrictions on the long-run behavior of the time fixed effects. One way to formulate these restrictions is to condition on the path $\{\gamma_t\}_{t\in\mathbb{Z}}$ and impose restrictions directly on this path. This approach is consistent with the fixed-effect interpretation of γ_t and is common in the factor model literature, but the resulting conditions are typically high-level and difficult to interpret, especially in nonlinear models.

To focus on the main idea and provide more primitive conditions, I instead embed the realized time effects in a superpopulation model. Specifically, Γ_T is viewed as one realization of an underlying stochastic process, and restrictions are imposed on this process to characterize the limiting behavior of the realized path as $T_0 \to \infty$[7]. The superpopulation assumption serves only to provide transparent and easy-to-verify sufficient conditions for the required time-series limits. Importantly, all stochastic-process assumptions imposed below can be replaced by direct restrictions on the realized path that imply the same limiting results.

Assumption 4.1 (Pseudo distance) *I assume that:*

(i) ***(Sampling)*** *Conditional on $\{\gamma_t\}_{t\in\mathbb{Z}}$, the panel $\{(Y_{it}, X_{it}, D_i, \alpha_i)\}_{i=1,\ldots,n,t=1,\ldots,T}$ is i.i.d. across i, and $\{\gamma_t\}_{t\in\mathbb{Z}}$ is stationary and ergodic over time. The panel is large with $n \to \infty$ and the number of pretreatment periods $T_0 \to \infty$.*

(ii) ***(Compact)*** *The support of α is compact.*

(iii) ***(Moment)*** $\mathbb{E}\left(Y_{it}^2(0) \mid \alpha_i = \alpha, \Gamma_t = \Gamma, D_i = d\right) < \infty$ *uniformly over all supported α, all possible paths Γ, $d \in \{0, 1\}$, and $t = 1, \ldots, T$.*

(iv) ***(Envelope)*** *The envelope of the factor structure $g(\alpha, \Gamma)$ (defined in (3.1)) has a finite second moment, i.e., $\mathbb{E}\left(\sup_{\alpha\in\mathcal{A}}\left(g(\alpha, \Gamma)\right)^2\right) < \infty$.*

(v) ***(Smoothness)*** *The factor structure $g(\alpha, \Gamma)$ defined in (3.1) is uniformly continuous in α, i.e., for any $\epsilon > 0$, there exists a $\delta > 0$ such that*

$$\sup_{\Gamma\in\text{supp}(\Gamma_T)} \sup_{\alpha,\alpha'\in\mathcal{A},\|\alpha-\alpha'\|\leqslant\delta} |g(\alpha, \Gamma) - g(\alpha', \Gamma)| \leqslant \epsilon.$$

[7] This formulation does not change the fixed-effect interpretation in the analysis, because once the sample path is realized, the time effects are treated as fixed.

Assumption 4.1(i) requires panel data with $n \to \infty$ individuals and $T_0 \to \infty$ pretreatment periods, and imposes conditional independence across individuals i given the realized time fixed effects Γ_T. The conditional independence condition is weak and does not require time-homogeneity condition (Chernozhukov et al., 2013, 2026). I also assume that $\{\gamma_t\}_{t\in\mathbb{Z}}$ is stationary and ergodic over time, so that $\{\Gamma_t\}_{t\in\mathbb{Z}}$ is also stationary and ergodic. Consequently, the distribution of Γ_t is invariant over t. These restrictions, similar to the stochastic-process restrictions imposed in Feng (2023) and Deaner, Hsiang and Zeleneev (2025), ensure the convergence of the time averages in the sample pseudo-distance. These stationarity and ergodicity assumptions are stronger than necessary and serve as primitive conditions for the identification argument in the main text. Appendix B discusses alternative conditions that allow for nonstationarity.

Assumption 4.1(ii) is a standard compact-support condition on the latent heterogeneity.

Assumption 4.1(iii) imposes moment conditions on untreated potential outcomes. It requires that the conditional second moment $\mathbb{E}\left(Y_{it}^2(0) \mid \alpha_i = \alpha, \Gamma_t = \Gamma, D_i = d\right)$ be uniformly bounded over the support of α, all possible realizations of Γ, $d \in \{0,1\}$, and all time periods $t = 1, \ldots, T$. It is worth noting that, although the presence of unbounded common time trends in $Y_{it}(0)$ violates Assumption 4.1(iii), the analysis remains valid because the common trend cancels out in the cross-sectional differences used to construct the pseudo-distance, as discussed in Section 3.

Assumption 4.1(iv) requires that the envelope of $g^2(\cdot,\cdot)$ is integrable over the time-factor path Γ. The condition is mild in that it requires only a finite second moment of the envelope and does not require $g(\alpha,\Gamma)$ to be uniformly bounded.

Assumption 4.1(v) imposes that $g(\alpha,\Gamma)$ is uniformly continuous in α over all possible paths Γ. This assumption is necessary because it ensures that individuals with similar α behave similarly over a long time horizon.

Proposition 1 (Identification of $d(\alpha_i,\alpha_j)$) *Under Assumptions 2.3 and 4.1, for each $i,j = 1,\ldots,n$, $\widehat{d}_{ij} \xrightarrow{p} d(\alpha_i,\alpha_j)$, as $n, T_0 \to \infty$, where $d(\alpha_i,\alpha_j)$ is given by*

$$d(\alpha_i,\alpha_j) := \sup_{\alpha_1,\alpha_2\in\mathcal{A}} \left| \int (g(\alpha_1,\Gamma) - g(\alpha_2,\Gamma))(g(\alpha_i,\Gamma) - g(\alpha_j,\Gamma)) \mathrm{d}\mathbb{P}(\Gamma) \right|. \tag{4.1}$$

The identification of $d(\alpha_i,\alpha_j)$ follows directly from establishing that $\widehat{d}_{ij} \xrightarrow{p} d(\alpha_i,\alpha_j)$. This constitutes an identification-by-construction approach (see the discussion in Lewbel (2019)). The population pseudo-distance $d(\alpha_i,\alpha_j)$ depends only on (α_i,α_j) and does not depend on the realized outcome paths or the realization of the time fixed effects. In addition, it is finite, symmetric, and uniformly continuous in (α_i,α_j), with $d(\alpha_i,\alpha_j) = 0$ whenever

$\alpha_i = \alpha_j$.

The integral in (4.1) is taken with respect to the stationary distribution of the time-factor path Γ because, under stationarity and ergodicity, the ergodic theorem implies that time averages along the realized path $\{\Gamma_t\}_{t\in\mathbb{Z}}$ converge to expectations under this stationary distribution. The existence of the population pseudo-distance and the identification result, however, do not rely on stationarity or ergodicity. I provide the definition of $d(\alpha_i, \alpha_j)$ under more general conditions in Appendix B.2.

Remark. *Proposition 1 identifies the $n(n-1)/2$ pairwise pseudo-distances $d(\alpha_i, \alpha_j)$ in the observed sample. The proposition does not identify the latent heterogeneity α_i itself or the functional form of $d(\cdot, \cdot)$ over the entire latent space. Nevertheless, I show that this pairwise information is sufficient for the identification of the dynamic ATT, as it allows us to identify and match units with similar latent heterogeneity.*

Remark. ***(Common time trend)*** *The pseudo-distance remains well-defined even when $Y_{it}(0)$ contains a common time trend, in which case the factor representation of the data-generating process takes the form*

$$Y_{it}(0) = \underbrace{\lambda_t + \tilde{g}(\alpha_i, \Gamma_t)}_{=g(\alpha_i,\Gamma_t)} + \tilde{\epsilon}_{it} = g(\alpha_i, \Gamma_t) + \tilde{\epsilon}_{it},$$

where λ_t is common across units and may be nonstationary and unbounded. The pseudo-distance proposed in this paper, unlike the pseudo-distances in Feng (2023), Feng (2020), and Deaner, Hsiang and Zeleneev (2025), accommodates such common trends without modification because these trends cancel in the cross-sectional differences entering construction (3.2).

Identification of ATT(t) To use the identified pseudo-distance to recover latent similarity, I impose the following informativeness condition.

Assumption 4.2 (Informativeness - identification) *Suppose the population pseudo-distance exists and for every $\varepsilon > 0$, there exists a $\delta > 0$, such that*

$$\sup_{\alpha_1,\alpha_2\in\mathcal{A}, d(\alpha_1,\alpha_2)\leqslant\delta} \|\alpha_1 - \alpha_2\| \leqslant \varepsilon.$$

This assumption is key to identifying dynamic ATT, which requires that, if two individuals are close in terms of $d(\alpha_i, \alpha_j)$, then their latent factors must also be close. Assumption 4.2, analogous to the identification assumption in Auerbach (2022, Assumption 3), is mild and

holds if distinct latent factors cannot generate exactly the same factor structure, i.e., for any $\alpha_1 \neq \alpha_2$, $g(\alpha_1, \Gamma)$ and $g(\alpha_2, \Gamma)$ are not almost surely identical (see Lemma B.1 in Appendix B.1 for a formal discussion). In addition, the dynamic panel models in Examples 2 and 3 satisfy the informativeness assumption under general conditions, and the discussion is postponed to Section 5. This assumption is called informativeness because, although α is unobserved, the pretreatment history contains sufficient information to recover similarity in latent heterogeneity through the population pseudo-distance $d(\alpha_i, \alpha_j)$.

Assumption 4.3 (Identification of ATT) *I assume that:*

(i) ***(Moment)*** $\mathbb{E}_T\left(|Y_{it}(0)| \mid D_i = 1\right), \mathbb{E}_T\left(|Y_{it}(1)| \mid D_i = 1\right) < \infty$ *for each* $t = T_0, \ldots, T$.

(ii) ***(Smoothness)*** *For each* $t = T_0, \ldots, T$, *the expected outcome* $m_t(x, \alpha)$ *defined in* (3.3), *is uniformly continuous in* (x, α), *i.e., for any* $\epsilon > 0$, *there exists* $\delta > 0$ *such that for all* $(x, \alpha) \in \operatorname{supp}(X_{iT_0}, \alpha_i)$,

$$\sup_{\|(x', \alpha') - (x, \alpha)\| \leqslant \delta} |m_t(x, \alpha) - m_t(x', \alpha')| \leqslant \epsilon.$$

(iii) ***(Overlap)*** $\mathbb{P}_T(D_i = 1 \mid X_{iT_0}, \alpha_i) \in (0, 1)$ *a.s.*

Assumption 4.3(i) imposes finite first-moment conditions on both untreated and treated potential outcomes. Assumption 4.3(ii) imposes uniform continuity on expected outcome $m_t(\cdot, \cdot)$. Uniform continuity of m_t is a mild and standard condition commonly imposed for nonparametric identification of conditional expectations. Assumption 4.3(iii) imposes the standard overlap condition, which is commonly required in causal inference to identify counterfactual outcomes.

The following theorem presents the identification result for the dynamic ATT. The key idea is that, although α is unobserved to econometricians, the identified population pseudo-distance reveals latent similarity. Therefore, the dynamic ATT is identified by matching units with similar latent heterogeneity and pretreatment history.

Theorem 1 (Identification of ATT) *Suppose that the population pseudo-distance* $d(\alpha_i, \alpha_j)$ *is identified for each* $i, j = 1, \ldots, n$. *Then, under Assumptions 2.1-2.3, 4.2, and 4.3,* $\mathrm{ATT}(t)$ *is identified by*

$$\mathbb{E}_T(Y_{it} \mid D_i = 1) - \mathbb{E}_T\left(\lim_{\delta \downarrow 0} \mathbb{E}_T(Y_{jt} \mid X_{jT_0} = X_{iT_0}, d(\alpha_i, \alpha_j) \leqslant \delta, D_j = 0) \mid D_i = 1\right).$$

Theorem 1 shows that the dynamic ATT is identified using a matching strategy based on the pseudo-distance. Identification of ATT(t) requires imputing the missing untreated counterfactual outcome for treated units, i.e., $\mathbb{E}_T\left(Y_{it}(0) \mid X_{iT_0}, \alpha_i\right)$. I impute this counterfactual using matching. To elaborate, for each treated unit i, I match i to untreated units j with the same pretreatment covariates and a population pseudo-distance $d(\alpha_i, \alpha_j)$ no greater than δ. For a fixed δ, the imputed quantity is

$$\mathbb{E}_T(Y_{jt} \mid X_{jT_0} = X_{iT_0}, d(\alpha_i, \alpha_j) \leqslant \delta, D_j = 0).$$

By the informativeness condition, a small pseudo-distance implies similar latent heterogeneity. Together with the smoothness and selection conditions, this allows δ to approach zero, yielding

$$\mathbb{E}_T\left(Y_{it}(0) \mid X_{iT_0}, \alpha_i\right) = \lim_{\delta \to 0} \mathbb{E}_T(Y_{jt} \mid X_{jT_0} = X_{iT_0}, d(\alpha_i, \alpha_j) \leqslant \delta, D_j = 0).$$

Finally, averaging the imputed counterfactual over the distribution of treated units yields the identification formula in Theorem 1.

My imputation strategy is closely related to the missing-outcome imputation approach in the panel-data causal inference literature, including DiD (Borusyak, Jaravel and Spiess, 2024), synthetic control (Abadie and Gardeazabal, 2003), and factor-model imputation methods (Bai and Ng, 2021). I consider a more general nonlinear dynamic model, and as a result, the latent heterogeneity cannot be eliminated by differencing, as in standard DiD designs, and cannot be directly estimated using linear factor methods. The pseudo-distance identifies only pairwise latent similarity: it reveals whether two units have close latent types, but it does not identify the latent heterogeneity or its distribution. Nevertheless, Theorem 1 shows that even pairwise information is sufficient for identification.

The studies most closely related to this paper are Feng (2023), Athey and Imbens (2025), and Deaner, Hsiang and Zeleneev (2025), which also consider imputation-based methods for nonlinear models. However, their frameworks allow selection to depend only on latent heterogeneity, whereas this paper extends the analysis to settings in which selection also depends on pretreatment covariates.

5 Theory for Estimation and Inference

This section establishes the asymptotic properties of the proposed Nadaraya-Watson estimators using the sample pseudo-distance to impute untreated outcomes and estimate propensity scores. I then provide sufficient conditions under which the doubly robust ATT estimator

with double cross-fitting is $\sqrt{n}$-consistent, asymptotically unbiased, and asymptotically normal, thereby enabling valid inference for the dynamic ATT.

In this section, I focus on the case in which both the latent heterogeneity and the pretreatment covariates are continuously distributed. The analysis extends straightforwardly to settings where either α or X has finite support.

5.1 Asymptotic analysis of pseudo distance

I impose two assumptions to ensure that a sufficiently small pseudo-distance $\widehat{d}_{ij}$ implies that α_i and α_j are close in the latent space. The first assumption characterizes the rate of convergence of $\widehat{d}_{ij}$ to its population counterpart $d(\alpha_i, \alpha_j)$.

Assumption 5.1 (Pseudo-distance approximation) *There exist constants $\lambda_1, \lambda_2 > 0$ such that the following inequality holds wpa1:*

$$\max_{i,j\in\{1,\ldots,n\}} |\widehat{d}_{ij} - d(\alpha_i, \alpha_j)| \leqslant \lambda_1 n^{-1/d_\alpha} \log(n) + \lambda_2 T_0^{-1/2} \sqrt{\log(nT_0)},$$

Assumption 5.1 strengthens the pointwise convergence result used for identification by imposing a uniform convergence rate for $\widehat{d}_{ij}$. This assumption states that the estimation error between $\widehat{d}_{ij}$ and $d(\alpha_i, \alpha_j)$ is uniformly bounded by two terms in an asymptotic sense. The first term, $\lambda_1 n^{-1/d_\alpha}\sqrt{\log n}$, depends only on the dimension of the latent heterogeneity d_α and not on the length of the pretreatment history T_0. It captures the matching discrepancy in the latent characteristics α, and its rate deteriorates as d_α increases. The second term, $\lambda_2 T_0^{-1/2}\sqrt{\log(nT_0)}$, captures the sampling error that arises because the pseudo-distance is constructed from noisy observed outcomes.

The convergence rate in Assumption 5.1 is consistent with the indirect matching rates in Feng (2023, Theorem 4.1) and Deaner, Hsiang and Zeleneev (2025, Lemma A.1) for pure factor models, while the setting considered here is more general in allowing for dynamic dependence. I show that this condition is not restrictive even in the dynamic panel setting and provide sufficient conditions under which it holds (see Lemma B.2 in Appendix B.3). Lastly, for notational simplicity, let $\delta_{n,T_0} := \lambda_1 n^{-1/d_\alpha} \log(n) + \lambda_2 T_0^{-1/2}\sqrt{\log(nT_0)}$ denote a uniform upper bound on these estimation errors.

The second assumption is critical and requires that the population pseudo-distance $d(\alpha_i, \alpha_j)$ be informative to reveal the latent distance $\|\alpha_i - \alpha_j\|$.

Assumption 5.2 (Informativeness-estimation) *There exist constants $0 < \underline{\eta} < \overline{\eta} < \infty$ such that for any $\alpha_1, \alpha_2 \in \mathcal{A}$, $\underline{\eta} d(\alpha_1, \alpha_2) \leqslant \|\alpha_1 - \alpha_2\| \leqslant \overline{\eta} d(\alpha_1, \alpha_2)$.*

Assumption 5.2 strengthens the informativeness condition used for identification (Assumption 4.2) by requiring the latent distance to be linearly bounded by the population pseudo-distance. Assumption 5.2 is analogous to the conditions in Feng (2023, Assumption 4.1) and Deaner, Hsiang and Zeleneev (2025, Assumption 4(vi)). The lower bound condition in Assumption 5.2 is mild and can be directly verified when $g(\alpha, \Gamma)$ is Lipschitz continuous with respect to α, uniformly over Γ. By contrast, verifying the upper bound condition can be challenging, especially in nonlinear models with dynamic effect. I provide sufficient conditions under which many commonly used dynamic panel models, including Examples 1 and 3, satisfy the informativeness assumption.

Example 1 and 2 (Linear model - Continued) *Here I focus on Example 2, since the verification for Example 1 can be obtained by a similar argument. Suppose that* $Y_{it}(0)$ *follows* $Y_{it}(0) = \rho Y_{it-1}(0) + \gamma_t'\alpha_i + \epsilon_{it}$ *with sequentially mean-independent idiosyncratic error* ϵ_{it}*. Then, under certain regularity conditions on* $\mathcal{A}$*, Assumption 5.2 holds if (i)* $|\rho| < 1$*, (ii)* $\{\gamma_t\}_{t\in\mathbb{Z}}$ *are stationary and ergodic over time, and (iii) the eigenvalues of* $\mathbb{E}(\gamma_t\gamma_t')$ *are bounded above and bounded away from zero. Here, the condition* $|\rho| < 1$ *ensures that the effects of past idiosyncratic shocks decay over time. In addition, the positive smallest eigenvalue of* $\mathbb{E}(\gamma_t\gamma_t')$*, which can be viewed as analogous to a strong-factor condition in the factor-model literature, ensures that each dimension of* α *makes an independent contribution to the factor component that cannot be perfectly explained by the remaining dimensions. A formal statement of this result is provided in Lemma B.3 in Appendix B.4.*

Example 3 (Dynamic nonlinear model - Continued) *Suppose that* $Y_{it}(0)$ *evolves according to a dynamic Logit model as in Example 3. Then, under certain regularity conditions on* $\mathcal{A}$*, Assumption 5.2 holds if (i)* $\{\gamma_t\}_{t\in\mathbb{Z}}$ *are i.i.d. over time, (ii) the supports of* α *and* γ_t *are bounded, and (iii) the largest eigenvalue of* $\mathbb{E}(\gamma_t\gamma_t')$ *is finite and its smallest eigenvalue is strictly positive. Similarly, the positive smallest eigenvalue of* $\mathbb{E}(\gamma_t\gamma_t')$ *ensures that each dimension of* α *makes an independent contribution to the factor component that cannot be perfectly explained by the remaining dimensions. The same result also holds for dynamic Probit models. A formal statement is provided in Lemma B.4 in Appendix B.4.*

5.2 Asymptotic theory of $\widehat{\mathrm{ATT}}(t)$

I impose the following regularity conditions to establish the asymptotic theory.

Assumption 5.3 (Estimation) *I assume that*

(i) ***(Sampling)*** *Conditional on* Γ_T*, the panel* $\{(Y_{it}, X_{it}, D_i, \alpha_i)\}_{i=1,\ldots,n,t=1,\ldots,T}$ *is i.i.d. across* i*. The panel is large with* $n \to \infty$ *and* $T_0 \to \infty$*.*

(ii) ***(Moment)*** *Suppose* X_{iT_0} *is uniformly bounded across* i*. In addition, for some* $\sigma > 0$*,* $\mathbb{E}_T\left(\exp\left(\lambda(Y_{it}(0) - m_{it})\right) \mid X_{iT_0}, \alpha_i\right) \leqslant \exp(\sigma^2\lambda^2/2)$ *a.s. for all* $\lambda \in \mathbb{R}$*, all* i*, and all* $t = T_0, \ldots, T$*. Lastly, for some* $\delta > 0$*,* $\mathbb{E}_T\left(|Y_{it}(1)|^{2+\delta}\right) < \infty$ *for all* $t = T_0, \ldots, T$*.*

(iii) ***(Compactness)*** *The support of* α*,* $\mathcal{A}$*, is compact. There exist constants* $\underline{c}_1, \overline{c}_1 > 0$*, such that, for any* $\alpha \in \mathcal{A}$ *and any* $r \in (0,1)$*,* $\underline{c}_1 r^{d_\alpha} \leqslant \mathbb{P}\left(\|\alpha_i - \alpha\| \leqslant r\right) \leqslant \overline{c}_1 r^{d_\alpha}$*. In addition, there exist constants* $0 < \underline{c}_2 \leqslant \overline{c}_2 < \infty$ *such that, for any* $(x, \alpha) \in \operatorname{supp}(X_{iT_0}, \alpha_i)$ *and any* $r \in (0,1)$*,* $\underline{c}_2 r^{d_X + d_\alpha} \leqslant \mathbb{P}_T\left(\|X_{iT_0} - x\| \leqslant r, \|\alpha_i - \alpha\| \leqslant r\right) \leqslant \overline{c}_2 r^{d_X + d_\alpha}$*.*

(iv) ***(Lipschitz)*** $m_t(x, \alpha)$ *and* $\pi(x, \alpha)$ *are Lipschitz continuous on* $\operatorname{supp}(X_{iT_0}, \alpha_i)$ *for each* $t = T_0, \ldots, T$*.*

(v) ***(Common overlap)*** *There exist constants* $\underline{p}, \overline{p} \in (0,1)$ *s.t.* $\underline{p} \leqslant \pi(X_{iT_0}, \alpha_i) \leqslant \overline{p}$ *a.s.*

(vi) ***(Kernel)*** *The kernel* $K : \mathbb{R} \mapsto \mathbb{R}_+$ *is bounded by* $\overline{K} > 0$ *and supported on* $[-1, 1]$*. In addition,* $K(0) > 0$ *and* $K(\cdot)$ *is Lipschitz continuous with constant* $L_K > 0$*.*

Assumption 5.3(i) requires that the panel data be i.i.d. given the time fixed effects Γ_T.

Assumption 5.3(ii) requires that the covariates X_{iT_0} be uniformly bounded over units, which is a standard condition for regressors in nonparametric estimation. This assumption also requires $Y_{it}(0)$ to be uniformly sub-Gaussian conditional on covariates and latent factors. The sub-Gaussian assumption is analogous to Deaner, Hsiang and Zeleneev (2025, Assumption 4(iii)) and facilitates the application of exponential concentration inequalities. The finite $(2+\delta)$-th moment condition on $Y_{it}(1)$ is used to apply the central limit theorem. Note that the uniform boundedness requirement is imposed for technical convenience and is stronger than necessary. It suffices to require that X_{iT_0} be uniformly bounded after removing an additive common time trend, since this trend cancels out under cross-sectional differencing even if it is unbounded.

Assumption 5.3(iii) first imposes compactness on the support of α and requires that the distribution of α be neither locally too sparse nor too concentrated. That is, for any supported α and any small radius r, the probability that α_i falls within a ball of radius r centered at α is of the same order as the volume of that ball. In addition, this assumption imposes the same restriction on the joint distribution of (X_{iT_0}, α) and requires that the joint distribution of (X_{iT_0}, α_i) be neither locally too sparse nor too concentrated. Assumption 5.3(iii) is common in establishing the uniform asymptotic properties of nonparametric kernel estimators and is automatically satisfied if (i) the support of (X_{iT_0}, α) is compact and convex, and (ii) the density of α and the joint density of (X_{iT_0}, α) are uniformly bounded above and bounded away from zero. Lastly, although this assumption is primarily applicable

to continuously distributed α and X, the density condition can be replaced by the corresponding condition on point-mass probabilities to accommodate the case in which either α or X has finite support.

Assumption 5.3(iv) requires that both the expected outcomes and propensity scores be Lipschitz continuous functions of covariates and individual heterogeneity. This condition is mild and does not require differentiability.

Assumption 5.3(v) is a commonly adopted overlap condition in the causal inference literature. Assumption 5.3(vi) imposes standard regularity conditions on the kernel function. These conditions are satisfied by many commonly used kernels, including the Epanechnikov kernel.

Proposition 2 (Convergence rates for counterfactual imputation) *Under Assumptions 2.1-2.3 and Assumptions 5.1-5.3, for any $h \in \{h_\pi, h_m\}$ such that $h \to 0$, $nh^{d_\alpha+d_X}/\log n \to \infty$, and $\delta_{n,T_0}/h \to 0$, the following result holds for each $t = T_0, \ldots, T$,*

$$\max_{i=1,\ldots,n} |\widehat{m}_{it} - m_{it}| = O_P\left(h_m + \left(nh_m^{d_\alpha+d_X}\right)^{-1/2}\sqrt{\log(n)}\right).$$

In addition,

$$\max_{i=1,\ldots,n} |\widehat{\pi}_i - \pi_i| = O_P\left(h_\pi + \left(nh_\pi^{d_\alpha+d_X}\right)^{-1/2}\sqrt{\log(n)}\right).$$

Proposition 2 establishes the uniform convergence rates for counterfactual outcome estimators and propensity score estimators over all individuals. These convergence rates imply that, when the estimation error of the sample pseudo-distance is asymptotically negligible relative to the bandwidth h, the convergence rate of the estimator using $\widehat{d}_{ij}$ is identical (up to a logarithmic term) to that of the Nadaraya-Watson estimator if α were observed.

In standard practice, researchers set $h_m, h_\pi \asymp n^{-1/(d_\alpha+d_X+2)}$ to balance the bias and variance terms. The uniform convergence rates in this case are of order $n^{-1/(d_\alpha+d_X+2)}\log(n)$ and achieve Stone's optimal rate under Lipschitz continuity (up to a logarithmic term). Then, one typically employs standard single cross-fitting and plugs these nuisance-function estimators into the doubly robust estimator of the ATT. However, I show that the bandwidth choice that is rate-optimal for imputing counterfactual outcomes and estimating propensity scores is not necessarily optimal for estimating the ATT. By combining double cross-fitting with a different choice of bandwidths, I obtain a faster convergence rate for the ATT estimator and establish root-n consistency, asymptotic normality, and asymptotic unbiasedness under a broader range of settings. The following Theorem formalizes this idea.

Theorem 2 (Asymptotic theory for ATT estimator) *Under conditions in Proposition 2, for any $h \in \{h_\pi, h_m\}$ such that $h \to 0$, $nh^{d_\alpha+d_X}/\log n \to \infty$, and $\delta_{n,T_0}/h \to 0$, the following result holds for each $t = T_0, \ldots, T$,*

$$\widehat{\mathrm{ATT}(t)} - \mathrm{ATT}(t) = \frac{1}{\sqrt{n}}\mathcal{N}(0, V_t) + O_P\left(h_m\left(h_\pi + (nh_\pi^{d_\alpha+d_X})^{-1/2}\sqrt{\log(n)}\right)\right) + o_P(n^{-1/2}).$$

Here, $\mathcal{N}(\cdot,\cdot)$ denotes the normal distribution, and V_t is the asymptotic variance of the oracle estimator.

Theorem 2 is the main result of this paper. Ignoring logarithmic terms for simplicity, the theorem shows that the difference between the oracle estimator and the proposed doubly robust estimator using double cross-fitting is $O_P\left(h_m\left(h_\pi + (nh_\pi^{d_\alpha+d_X})^{-1/2}\right)\right)$. Therefore, when this remainder term is $o_P(1/\sqrt{n})$, the proposed estimator is asymptotically equivalent to the oracle estimator and achieves root-n asymptotic normality and asymptotic unbiasedness.

Double cross-fitting achieves a faster convergence rate than standard single cross-fitting. Under standard single cross-fitting, the remainder term is the product of nuisance estimation errors, and is of order $O_P\left(\left(h_m + (nh_m^{d_\alpha+d_X})^{-1/2}\right)\left(h_\pi + (nh_\pi^{d_\alpha+d_X})^{-1/2}\right)\right)$. Double cross-fitting eliminates the variance term $(nh_m^{d_\alpha+d_X})^{-1/2}$ from the remainder, which enables researchers to adopt an *undersmoothing* choice of h_m rather than the bandwidth that is optimal for estimating m_{it} and achieves a faster convergence rate for the remainder term.

The intuition behind this improvement is that double cross-fitting reduces the dependence between the outcome regression and propensity score estimators. Under single cross-fitting, $\widehat{m}_{it}$ and $\widehat{\pi}_i$ are constructed from the same sample and hence have correlated estimation errors. As a result, the product error is controlled by the Cauchy-Schwarz inequality, which depends on the convergence rates of both $\widehat{m}_{it}$ and $\widehat{\pi}_i$ in the L_2 norm. Double cross-fitting removes this dependence by estimating the two nuisance functions using separate samples. Therefore, the remainder term depends on the average bias of $\widehat{m}_{it}$ (which is of order h_m) rather than the L_2 norm of the estimation errors, allowing for undersmoothing the outcome regression.

Corollary 1 (Inference) *Under conditions in Proposition 2, set $h_\pi \asymp n^{-1/(d_\alpha+d_X+2)}$ and $h_m \asymp n^{-1/(d_\alpha+d_X)}(\log(n))^{2/(d_\alpha+d_X)}$. Then, for each $t = T_0, \ldots, T$ and $d_\alpha + d_X \leqslant 3$,*

$$\sqrt{n}\left(\widehat{\mathrm{ATT}(t)} - \mathrm{ATT}(t)\right) \xrightarrow{d} \mathcal{N}(0, V_t).$$

Here, V_t is the asymptotic variance of the oracle estimator.

Corollary 1 shows that, when $d_\alpha + d_X \leqslant 3$, combining double cross-fitting with undersmoothing yields a root-n asymptotically normal and asymptotically unbiased ATT estimator, en-

abling valid inference on the dynamic ATT. I choose $h_\pi \asymp n^{-1/(d_\alpha+d_X+2)}$, which is optimal for the MSE of the propensity score estimator, while aggressively undersmoothing h_m because the outcome regression contributes to the remainder through its bias. Corollary 1 expands the range of settings in which root-n inference can be obtained relative to the standard single cross-fitting, which establishes root-n inference for the ATT estimator only when $d_\alpha + d_X = 1$ (see Deaner, Hsiang and Zeleneev (2025)).

Remark. ***(Other choices of bandwidths)*** *Although Corollary 1 requires h_m to be undersmoothed as aggressively as possible to achieve the fastest convergence rate of the remainder term, the bandwidth choice is not unique. It is straightforward to verify that there exists a range of bandwidth choices that make the remainder term $o_P(1/\sqrt{n})$ and therefore yield root-n valid inference. I adopt this most aggressive undersmoothing choice because it leads to a concise theoretical characterization of the remainder term while also providing practical guidance for selecting bandwidths, which will be discussed later.*

Remark. ***(Discrete Covariates)*** *Proposition 2, Theorem 2, and Corollary 1 are primarily applicable to continuously distributed covariates. When X_{it} takes discrete values, for example, in the binary response model as in Example 3, the proposed estimator and the associated analysis remain valid, and the only difference is that d_X does not enter the convergence rate asymptotically. Specifically, one only needs to replace all terms of the form $nh^{d_\alpha+d_X}$ with nh^{d_α} in the previous discussion to accommodate the discrete case. In addition, combining double cross-fitting with undersmoothing yields a root-n asymptotically normal and asymptotically unbiased ATT estimator when $d_\alpha \leqslant 3$.*

Remark. ***(Asymptotics of T_0)*** *Although Corollary 1 does not impose an explicit restriction on T_0, T_0 needs to grow sufficiently fast to achieve root-n valid inference, as it affects the accuracy of $\widehat{d}_{ij}$ through δ_{n,T_0}. A sufficient condition is that $T_0 \gg n^{2/(d_\alpha+d_X)}$.*

Remark. ***(Parametric assumption)*** *The proposed approach is nonparametric and does not impose functional form restrictions on the potential outcome or propensity score. If researchers are willing to impose parametric restrictions, for example, assuming that the potential outcome follows a linear dynamic model with interactive fixed effects or that the propensity score follows a logit model with a linear index, the dimensionality restrictions can be substantially relaxed. In addition, how to incorporate high-dimensional covariates into the outcome regression (see, e.g., Viviano and Bradic (2026)), while controlling for individual fixed effects remains a promising direction for future research.*

5.3 Bandwidth selection in practice

The theoretical bandwidth choice in Corollary 1 provides practical guidance for selecting h_m and h_π using data-driven methods. Since h_π is set to minimize MSE of the propensity score estimator, I propose to use standard leave-one-out cross-validation. To elaborate, given $h > 0$, for each $i = 1, \ldots, n$, compute the within-sample leave-one-out estimator,

$$\widehat{\pi}_i^{(-i)}(h) = \sum_{i' \in \mathcal{I}(i) \setminus \{i\}} D_{i'} \widehat{K}_{h,ii'} \Big/ \sum_{i' \in \mathcal{I}(i) \setminus \{i\}} \widehat{K}_{h,ii'},$$

and select the optimal bandwidth h_π^* as $h_\pi^* = \operatorname{argmin}_{h>0} \frac{1}{n} \sum_{i=1}^n (D_i - \widehat{\pi}_i^{(-i)}(h))^2$.

Since Corollary 1 requires $h_m \asymp n^{-1/(d_\alpha + d_X)} (\log(n))^{2/(d_X + d_\alpha)}$ to be undersmoothed as aggressively as possible, this immediately means that the neighbors for each unit is of order $\sim (\log(n))^2$. Given any bandwidth $h > 0$, for each i, I define the effective sample size as

$$n_{i,\text{eff}}(h) = \Big(\sum_{i' \in \mathcal{I}(i) \setminus \{i\}} (1 - D_{i'}) \widehat{K}_{h,ii'} \Big)^2 \Big/ \sum_{i' \in \mathcal{I}(i) \setminus \{i\}} (1 - D_{i'}) \widehat{K}_{h,ii'}^2,$$

Following the theoretical guidance, a proper choice of h_m should be as small as possible while maintaining a sufficient number of effective matches of order $\sim (\log(n))^2$. Therefore, in practice, I choose predetermined constants $\kappa > 0$ and $\underline{q} \in (0, 1]$ and select the smallest bandwidth that ensures at least a fraction $\underline{q}$ of observations have an effective sample size larger than $(\kappa \log(n))^2$:

$$h_m^* = \min \left\{ h > 0 : \frac{1}{n} \sum_{i=1}^n \mathbf{1} \left(n_{i,\text{eff}}(h) > (\kappa \log(n))^2 \right) \geqslant \underline{q} \right\}.$$

Numerical simulations in Section 7 show that $(\kappa = 0.2, \underline{q} = 0.8)$ performs well under various settings, and I recommend these values for empirical applications.

6 Extension: Staggered Adoption

Section 6 extends the analysis to staggered adoptions, which are ubiquitous in empirical research. I adopt the idea of sequential unconfoundedness from Robins, Hernan and Brumback (2000) and combine it with the pseudo-distance approach to identify and estimate dynamic ATT. The identification, estimation, and inference results developed for simultaneous treatment continue to hold under this extension with appropriate modifications.

6.1 Setup and dynamic selection

Consider the panel data $\{(Y_{it}, X_{it}, D_{it}, \alpha_i)\}$ on n units, indexed by $i = 1, \ldots, n$, over T periods, indexed by $t = 1, \ldots, T$. The outcome variable is $Y_{it} \in \mathbb{R}$. The pretreatment covariates are $X_{it} \in \mathbb{R}^{d_X}$ and can include lagged outcomes, for example, $X_{it} = Y_{it-1}$. The treatment status is denoted by $D_{it} \in \{0, 1\}$. I assume that treatment is *absorbing*, i.e., $D_{it+1} \geqslant D_{it}$ for any t, in the sense that once a unit becomes treated, it remains treated in all subsequent periods. Let G_i denote the first period in which unit i receives the treatment, i.e., $D_{it} = \mathbf{1}\{t \geqslant G_i\}$. For those individuals never treated in the observed time period, we set $G_i = \infty$. α_i is the unobservable individual fixed effect.

To allow for dynamic treatment effects, potential outcomes are indexed by the entire treatment sequence $(d_1, ..., d_T) \in \{0, 1\}^T$, $Y_{it}(d_1, ..., d_T)$. Since treatment is an absorbing state, the potential outcomes can be indexed by the first treatment period G_i only. Here, we define $Y_{it}(g) := Y_{it}(\mathbf{0}_{g-1}, \mathbf{1}_{T-g+1})$, and $Y_{it}(\infty) := Y_{it}(\mathbf{0}_T)$. I follow the standard non-anticipation assumption such that

$$Y_{it} = \begin{cases} Y_{it}(\infty), & t < G_i \\ Y_{it}(G_i), & t \geqslant G_i \end{cases}$$

For each $t \geqslant g \geqslant T_0$, the estimand of interest is the dynamic ATT:

$$\mathrm{ATT}(g, t) := \mathbb{E}_T \left(Y_{it}(g) - Y_{it}(\infty) \mid G_i = g \right).$$

It denotes the ATT at time t for units that first receive the treatment in period g. As discussed in Section 2, although the definition of the ATT here appears to differ from the standard definition, $\mathbb{E}(Y_{it}(g) - Y_{it}(\infty) \mid G_i = g)$, the two definitions are equivalent, and the difference is purely notational.

Assumption 6.1 (Selection) *Assume that for each $T_0 \leqslant g \leqslant t \leqslant T$,*

$$Y_{it}(\infty) \perp \{G_i = g\} \mid X_{ig}, \alpha_i, \Gamma_T, G_i > g - 1.$$

The assumption extends Assumption 2.1 to staggered adoption, and is closely related to the sequential unconfoundedness assumption in the dynamic treatment effect literature. Assumption 6.1 requires that, for individuals who have not yet received the treatment before period g, treatment adoption at period g is independent of future untreated potential outcomes, conditional on pretreatment covariates X_{ig}, individual fixed effects, and time factors. The conditional orthogonality is imposed only on untreated potential outcomes because the estimand of interest is the ATT, and only untreated outcomes need to be imputed.

The following assumption imposes a Markov restriction.

Assumption 6.2 (Markov) *Let* $H_{it} := \{(Y_{i\tau}, X_{i\tau}, D_{i\tau})\}_{\tau \leqslant t}$ *denote the history up to* t*,*

$$(Y_{it}(\infty), D_{it'}) \perp H_{it'-1} \mid X_{it'}, \alpha_i, \Gamma_T, G_i > t' - 1, \quad \forall\, t = T_0, \ldots, T.$$

This assumption extends Assumption 2.2 to staggered adoption. It is stronger than necessary and can be relaxed to conditional moment restrictions on $(Y_{it}(0), D_{it'})$ (see the discussion in Section 2). Nevertheless, I retain the Markov restriction for ease of exposition.

Assumption 6.3 (Potential Outcome) $Y_{it}(\infty)$ *evolves according to* (2.4) *and satisfies:*

(i) ***(Time factor exogeneity)*** *conditional on the individual latent factors* α_i *and the time-factor path* Γ_t*,* $Y_{it}(\infty)$ *is mean independent of future time factors, i.e.,*

$$\mathbb{E}\left(Y_{it}(\infty) \mid X_{it}, \alpha_i, \Gamma_T\right) = \mathbb{E}\left(Y_{it}(\infty) \mid X_{it}, \alpha_i, \Gamma_t\right), \quad \forall t.$$

(ii) ***(Weak dependence)*** $(\ldots, Y_{it-1}(\infty), Y_{it}(\infty), \ldots)$ *are conditionally weakly dependent, such that* $\sum_{s=-\infty}^{\infty} |\mathrm{Cov}\left(Y_{it}(\infty), Y_{is}(\infty) \mid \alpha_i = \alpha, \Gamma_T = \Gamma\right)| < \infty$ *uniformly for any supported* α*, any possible time path* Γ*, and any* $t \in \mathbb{Z}$*.*

Assumption 6.3 is identical to Assumption 2.3, except that $Y_{it}(0)$ is replaced by $Y_{it}(\infty)$ to denote the untreated potential outcomes under staggered adoption.

Similar to (3.3), for each $T_0 \leqslant t' \leqslant t \leqslant T$ and $\alpha \in \mathcal{A}$, define

$$\begin{aligned} m_{t|t'}(x, \alpha) :=& \mathbb{E}_T(Y_{it}(\infty) \mid X_{it'} = x, \alpha_i = \alpha, G_i > t' - 1), \\ \pi_{t'}(x, \alpha) :=& \mathbb{P}(G_i = t' \mid X_{it'} = x, \alpha_i = \alpha, G_i > t' - 1). \end{aligned} \tag{6.1}$$

The function $m_{t|t'}(x, \alpha)$ is the conditional expectation of the untreated outcome at period t for units that remain untreated at time $t' - 1$, given their pretreatment information summarized by $X_{it'} = x$ and latent heterogeneity $\alpha_i = \alpha$. The function $\pi_{t'}(x, \alpha)$ is the propensity score of treatment adoption at time t' for units that have not been treated by time $t' - 1$, given observed covariates $X_{it'} = x$, and individual fixed effects $\alpha_i = \alpha$.

6.2 Identification

The identification of the pseudo-distance and its informativeness under staggered adoption is identical to that under simultaneous treatment (see Assumptions 4.1 and 4.2 and Proposition 1), except that $Y_{it}(0)$ is replaced by $Y_{it}(\infty)$ to accommodate staggered adoption.

Therefore, I omit the discussion of pseudo-distance identification and informativeness and focus directly on the identification of the dynamic ATT.

Assumption 6.4 (Identification of ATT) *I assume that:*

(i) ***(Moment)*** *For each* $t \geqslant g \geqslant T_0$, $\mathbb{E}_T\left(|Y_{it}(\infty)| \mid G_i = g\right), \mathbb{E}_T\left(|Y_{it}(g)| \mid G_i = g\right) < \infty$.

(ii) ***(Smoothness)*** *For each* $T_0 \leqslant t' \leqslant t \leqslant T$, $m_{t|t'}(x, \alpha)$ *is uniformly continuous in* (x, α), *i.e., for any* $\epsilon > 0$, *there exists* $\delta > 0$ *such that for all* $(x, \alpha) \in \operatorname{supp}(X_{it'}, \alpha_i)$

$$\sup_{\|(x',\alpha')-(x,\alpha)\| \leqslant \delta} \left|m_{t|t'}(x, \alpha) - m_{t|t'}(x', \alpha')\right| \leqslant \epsilon.$$

(iii) ***(Overlap)*** $\mathbb{P}_T(G_i = t \mid X_{it}, \alpha_i, G_i > t-1) \in (0, 1)$ *a.s.*

Assumption 6.4(i) is standard and imposes finite first-moment conditions on both untreated and treated potential outcomes. Assumption 6.4(ii) extends Assumption 4.3(ii) by imposing uniform continuity on the expected untreated outcomes $m_{t|t'}(\cdot, \cdot)$. Uniform continuity of $m_{t|t'}(\cdot, \cdot)$ is a mild and standard condition commonly imposed for nonparametric identification of conditional expectations. Assumption 6.4(iii) strengthens Assumption 4.3(iii) and is commonly required in causal inference to identify counterfactual outcomes.

Theorem 3 (Identification) *Suppose that the population pseudo-distance* $d(\alpha_i, \alpha_j)$ *is identified for each* $i, j = 1, \ldots, n$. *Then, under Assumptions 6.1-6.4 and Assumption 4.2, the following identification results hold: (i) For each* $t \geqslant T_0$ *and each unit* i *such that* $G_i > t-1$, *the contemporaneous counterfactual is identified by*

$$m_{t|t}(X_{it}, \alpha_i) = \lim_{\delta \downarrow 0} \mathbb{E}_T(Y_{jt} \mid X_{jt} = X_{it}, d(\alpha_j, \alpha_i) \leqslant \delta, G_j > t);$$

(ii) for any $t' = t-1, \ldots, g$ *and each unit* i *such that* $G_i > t'-1$, *the dynamic counterfactual expectation is identified recursively backward by*

$$m_{t|t'}(X_{it'}, \alpha_i) = \lim_{\delta \downarrow 0} \mathbb{E}_T\left(m_{t|t'+1}(X_{jt'+1}, \alpha_j) \mid X_{jt'} = X_{it'}, d(\alpha_j, \alpha_i) \leqslant \delta, G_j > t'\right);$$

and (iii) for any $t \geqslant g \geqslant T_0$, *the dynamic ATT is identified by*

$$\mathrm{ATT}(g, t) = \mathbb{E}_T\left(Y_{it} \mid G_i = g\right) - \mathbb{E}_T\left(m_{t|g}(X_{ig}, \alpha_i) \mid G_i = g\right).$$

Theorem 3 extends the identification result to staggered-adoption settings and establishes the identification of $\mathrm{ATT}(g, t)$. It uses a pseudo-distance-based matching strategy with a

backward recursive procedure to accommodate dynamic selection. The theorem proceeds in three steps. First, it identifies the contemporaneous counterfactual expectation $m_{t|t}(\cdot,\cdot)$. Second, it identifies dynamic counterfactual expectations $m_{t|t'}(\cdot,\cdot)$ through a backward recursive procedure over $t' = t, t-1, \ldots, g$. Finally, averaging the imputed counterfactual $m_{t|g}(\cdot,\cdot)$ over the distribution of units being treated at g identifies $\mathrm{ATT}(g,t)$.

6.3 Estimator

For notational simplicity, for each $t \geqslant t' \geqslant T_0$ and each i who has not been treated at time $t'-1$, define its expected (counterfactual) potential outcome as $m_{i,t|t'} := m_{t|t'}(X_{it'}, \alpha_i)$. In addition, for each $t = T_0, \ldots, T$ and each i such that $G_i > t-1$, define its propensity score as $\pi_{it} = \pi_t(X_{it}, \alpha_i)$.

An oracle estimator for $\mathrm{ATT}(g,t)$ under staggered adoption takes the following doubly robust form:

$$\widehat{ATT(g,t)}^{\text{oracle}} = \frac{1}{n_g}\sum_{i=1}^{n}\Bigg[\mathbf{1}\left(G_i = g\right)\left(Y_{it} - m_{i,t|g}\right) - \sum_{t'=g}^{t}\mathbf{1}\left(G_i > t'\right)\frac{\pi_{ig}}{1-\pi_{ig}}\left(\prod_{r=g+1}^{t'}\frac{1}{1-\pi_{ir}}\right)\left(m_{i,t|t'+1} - m_{i,t|t'}\right)\Bigg]. \tag{6.2}$$

Here, (i) $n_g := \sum_{i=1}^{n}\mathbf{1}(G_i = g)$ denotes the number of units first treated in period g; (ii) $m_{i,t|t+1} := Y_{it}$ is defined for notational convenience; and (iii) an empty product is defined as one, so that $\prod_{r=g+1}^{t'}(1-\pi_{ir})^{-1} = 1$ when $t' = g$. This is an oracle estimator in the sense that it uses the true propensity scores and outcome regressions, which are unknown in practice. As in the simultaneous treatment case, I use Nadaraya-Watson estimators based on the pseudo-distance to nonparametrically estimate $m_{i,t|t'}$ and $\pi_{it'}$, and then employ a *double* cross-fitting procedure to achieve a faster convergence rate.

The full algorithm, similar to the identification strategy, involves a backward procedure and is tedious to present in the main text. Therefore, a simplified version that abstracts from cross-validation for selecting the bandwidths h_m, h_π is summarized in Algorithm A.2 to highlight the main idea. The cross-validation procedure for bandwidth selection is discussed in the following text, and the full algorithm is presented in Algorithm A.3.

6.4 Estimation

The convergence rate of the pseudo-distance and its informativeness under staggered adoption are identical to those under simultaneous treatment (see Assumptions 5.1 and 5.2), ex-

cept that $Y_{it}(0)$ needs to be replaced by $Y_{it}(\infty)$ to accommodate dynamic selection. Therefore, I omit the discussion of the accuracy of the sampled pseudo-distance and informativeness and focus directly on the inference on the dynamic ATT.

Assumption 6.5 (Estimation) *I assume that*

(i) ***(Sampling)*** *Conditional on $\{\gamma_t\}_{t\in\mathbb{Z}}$, the panel $\{(Y_{it}, X_{it}, D_{it}, \alpha_i)\}_{i=1,\dots,n,t=1,\dots,T}$ is i.i.d. across i. The panel is large with $n \to \infty$ and $T_0 \to \infty$.*

(ii) ***(Moment)*** *Suppose X_{it} is uniformly bounded across i and t. In addition, for some $\sigma > 0$, $\mathbb{E}_T\left(\exp\left(\lambda(Y_{it}(\infty) - m_{i,t|t'})\right) \mid X_{it'}, \alpha_i, G_i > t' - 1\right) \leqslant \exp(\sigma^2\lambda^2/2)$ a.s. for all $\lambda \in \mathbb{R}$, all i, and all t, t' satisfying $T_0 \leqslant t' \leqslant t \leqslant T$. Lastly, for some $\delta > 0$, $\mathbb{E}_T\left(|Y_{it}(g)|^{2+\delta}\right) < \infty$ for all g, t satisfying $T_0 \leqslant g \leqslant t \leqslant T$.*

(iii) ***(Compactness)*** *The support of α, $\mathcal{A}$, is compact. There exist constants $\underline{c}_1, \overline{c}_1 > 0$, such that for any $\alpha \in \mathcal{A}$ and any $r \in (0,1)$, $\underline{c}_1 r^{d_\alpha} \leqslant \mathbb{P}\left(\|\alpha_i - \alpha\| \leqslant r\right) \leqslant \overline{c}_1 r^{d_\alpha}$. In addition, for each $t = T_0, \dots, T$, let $\mathcal{S}_{t-1}$ be the support of (X_{it}, α_i) conditional on $\{G_i > t-1\}$. There exist constants $0 < \underline{c}_2 \leqslant \overline{c}_2 < \infty$ such that, for any $t = T_0, \dots, T$, any $r \in (0,1)$, and any $(x, \alpha) \in \mathcal{S}_{t-1}$,*

$$\underline{c}_2 r^{d_X + d_\alpha} \leqslant \mathbb{P}\left(\|X_{it} - x\| \leqslant r, \|\alpha_i - \alpha\| \leqslant r, G_i > t-1\right) \leqslant \overline{c}_2 r^{d_X + d_\alpha}.$$

(iv) ***(Lipschitz)*** *For each $T_0 \leqslant t' \leqslant t \leqslant T$, let $\mathcal{S}_{t'-1}$ be the support of $(X_{it'}, \alpha_i)$ conditional on $\{G_i > t' - 1\}$. Assume $m_{t|t'}(x, \alpha)$ and $\pi_{t'}(x, \alpha)$ are Lipschitz continuous on $\mathcal{S}_{t'-1}$.*

(v) ***(Common overlap)*** *There exist constants $\underline{p}, \overline{p} \in (0,1)$ such that for any $t = T_0, \dots, T$, $\underline{p} \leqslant \pi_t(X_{it}, \alpha_i) \leqslant \overline{p}$ a.s.*

(vi) ***(Kernel)*** *The kernel $K : \mathbb{R} \mapsto \mathbb{R}_+$ is bounded by $\overline{K}$ and supported on $[-1, 1]$. In addition, $K(0) > 0$ and $K(\cdot)$ is Lipschitz continuous with constant $L_K > 0$.*

Assumption 6.5(i) and Assumption 6.5(vi) are identical to those conditions in the simultaneous treatment case. Assumption 6.5(ii) and (iii) extend Assumption 5.3(ii) and (iii) to staggered adoption, respectively. Assumption 6.5(iv) is similar to Assumption 5.3(iv), but extends continuity condition to the dynamic outcome regression functions $m_{t|t'}(\cdot,\cdot)$ and the propensity scores. Lastly, Assumption 6.5(v) is a common overlap condition imposed for treatment adoption at different periods.

Proposition 3 (Convergence rates for counterfactual imputation) *Under Assumptions 5.1-5.3, 6.1-6.3, and 6.5, for any $h \in \{h_\pi, h_m\}$ such that $h \to 0$, $nh^{d_\alpha+d_X}/\log n \to \infty$, and $\delta_{n,T_0}/h \to 0$, we have that for each $T_0 \leqslant t' \leqslant t \leqslant T$,*

$$\max_{i:G_i>t'-1} \left|\widehat{m}_{i,t|t'} - m_{i,t|t'}\right| = O_p\left(h_m + \left(nh_m^{d_\alpha+d_X}\right)^{-1/2}\sqrt{\log(n)}\right).$$

In addition, for each $t = T_0, \ldots, T$,

$$\max_{i:G_i>t-1} |\widehat{\pi}_{it} - \pi_{it}| = O_p\left(h_\pi + \left(nh_\pi^{d_\alpha+d_X}\right)^{-1/2}\sqrt{\log(n)}\right).$$

Proposition 3 establishes the convergence rates for counterfactual outcome estimators and propensity score estimators. The convergence rates are identical to those in the simultaneous treatment case.

Theorem 4 (Asymptotic normality) *Under conditions in Proposition 3, for each $T_0 \leqslant g \leqslant t \leqslant T$,*

$$\widehat{\mathrm{ATT}}(g,t) - \mathrm{ATT}(g,t) = \frac{1}{\sqrt{n}}\mathcal{N}(0, V_{g,t}) + O_p\left(h_m\left(h_\pi + (nh_\pi^{d_\alpha+d_X})^{-1/2}\sqrt{\log(n)}\right)\right).$$

Here, $\mathcal{N}(\cdot,\cdot)$ denotes the normal distribution, and $V_{g,t}$ is the asymptotic variance of the oracle estimator.

Theorem 4 shows that, similar to the simultaneous treatment case, combining double cross-fitting with undersmoothing of h_m yields a faster convergence rate than standard single cross-fitting.

Corollary 2 (Inference) *Under conditions in Proposition 3, set $h_\pi \asymp n^{-1/(d_\alpha+d_X+2)}$ and $h_m \asymp n^{-1/(d_\alpha+d_X)}(\log(n))^{2/(d_X+d_\alpha)}$. In addition, suppose $\delta_{n,T_0}/h_m \to 0$. Then, for each $T_0 \leqslant g \leqslant t \leqslant T$ and $d_\alpha + d_X \leqslant 3$,*

$$\sqrt{n}\left(\widehat{\mathrm{ATT}(g,t)} - \mathrm{ATT}(g,t)\right) \xrightarrow{d} \mathcal{N}(0, V_{g,t}).$$

Here, $\mathcal{N}(\cdot,\cdot)$ denotes the normal distribution, and $V_{g,t}$ is the asymptotic variance of the oracle estimator.

Corollary 2 shows that, when $d_\alpha + d_X \leqslant 3$, combining double cross-fitting with undersmoothing yields a root-n asymptotically normal and asymptotically unbiased estimator of the dynamic ATT, enabling valid inference.

6.5 Bandwidth selection in practice

The theoretical bandwidth choice in Corollary 2 provides practical guidance for selecting h_m and h_π using data-driven methods. Similar to the simultaneous treatment case, the choice of h_π optimizes the MSE of the propensity score estimator, while h_m is aggressively undersmoothed.

I propose to use standard leave-one-out cross-validation for selecting h_π and the effective number of matched neighbors for selecting h_m. The bandwidth selection procedure is almost identical to that in the simultaneous treatment case, except that it is implemented separately for each period. The full procedure is lengthy and is therefore omitted from the main text. The complete algorithm, including the bandwidth selection procedure, is provided in Algorithm A.3.

7 Simulation

In this section, I illustrate the finite-sample properties of the proposed method under different DGPs, including dynamic models with additive fixed effects, dynamic models with interactive fixed effects, and nonlinear dynamic models.

I report three estimators. (i) **DID**: the canonical difference-in-differences estimator. (ii) **DR2**: the doubly robust estimator based on a standard two-way sample split, with m and π estimated using the same training sample. The bandwidths for m and π are selected by leave-one-out cross-validation to minimize mean squared errors. I also use DR2* to denote the infeasible estimator that uses the true distance between individual latent factors rather than the pseudo-distance. (iii) **DR3**: the proposed estimator using a three-way sample split, with m and π estimated on separate training samples. I also use DR3* to denote the infeasible estimator that uses the true distance between individual latent factors rather than the pseudo-distance. I vary κ to examine how different choices affect the estimator's performance.

7.1 Dynamic panel with additive fixed effects

Consider the following data-generating process:

$$Y_{it} = \begin{cases} \rho Y_{it-1} + \alpha_i + \gamma_t + \epsilon_{Y,it}, & t < T_0 \\ \rho Y_{it-1} + \alpha_i + \gamma_t + 0.5 \cdot D_i + \epsilon_{Y,it}, & t \geqslant T_0 \end{cases}. \tag{7.1}$$

The individual latent factors $\{\alpha_i\}_{i=1}^n$ and the time latent factors $\{\gamma_t\}_{t\leqslant T}$ are independent random variables drawn from the uniform distribution on $[-1/4, 1/4]$. The error terms $\{\epsilon_{Y,it}\}_{i=1,\ldots,n,\ t=0,\ldots,T}$ are independent of the latent factors and are i.i.d. across both dimensions, following the uniform distribution on $[-1/2, 1/2]$. I set $\rho = 0.8$ to capture the dynamic effect. The contemporaneous ATT is 0.5, and for each $t > T_0$, the dynamic treatment effect is $0.5\left(1 + \rho + \ldots + \rho^{t-T_0}\right)$. Treatment assignment depends jointly on latent heterogeneity and the lagged outcome:

$$D_i = \mathbf{1}\left(\alpha_i/2 + Y_{iT_0-1}/2 + \epsilon_{D,i} \geqslant 0\right). \tag{7.2}$$

Here, the error terms $\{\epsilon_{D,i}\}_{i=1}^n$ are i.i.d. across individuals and are independent of the latent factors and $\{\epsilon_{Y,it}\}_{i=1,\ldots,n,\ t=0,\ldots,T}$. In the numerical designs, $\epsilon_{D,i}$ follows a standard logistic distribution. I vary $(N, T_0) \in \{(1000, 20), (1000, 40), (200, 100), (1000, 100), (2000, 100)\}$ and perform 5,000 replications for each design.

Table 1 reports the simulation results for the contemporaneous ATT and shows that the proposed estimator performs well across all designs. Relative to DID and the standard doubly robust estimator, the proposed estimator substantially reduces bias and has coverage close to the nominal 95% level. In addition, the proposed estimator using the pseudo-distance performs similarly to the infeasible DR3* using the true latent distance, indicating that the estimated pseudo-distance provides a good approximation to the true latent distance even when $T_0 = 20$. It is worth noting that, when $N = 200$, DR2 has a smaller standard deviation and coverage closer to the nominal level, although the proposed estimator has a smaller absolute bias. This is because the double cross-fitting procedure leaves relatively few observations in each subsample. Finally, the performance of the proposed estimator is also stable across different values of κ, and I recommend $(\kappa, \underline{q}) = (0.2, 0.8)$ as the default choice.

7.2 Dynamic panel with interactive fixed effects

Consider the following data-generating process:

$$Y_{it} = \begin{cases} \rho Y_{it-1} + \alpha_i \gamma_t + \epsilon_{Y,it}, & t < T_0 \\ \rho Y_{it-1} + \alpha_i \gamma_t + 0.5 \cdot D_i + \epsilon_{Y,it}, & t \geqslant T_0 \end{cases}. \tag{7.3}$$

Table 1: Simulation Results: Dynamic Panel with Additive Fixed Effects

	DID	DR2*	DR2	DR3*			DR3		
				$\kappa = .15$	$\kappa = .2$	$\kappa = .25$	$\kappa = .15$	$\kappa = .2$	$\kappa = .25$
$N = 1000, T_0 = 20$									
Bias	2.16	0.50	0.79	0.13	0.18	0.27	0.43	0.49	0.61
SD	1.94	2.05	2.10	2.47	2.40	2.30	2.53	2.43	2.33
Coverage	79.8	93.9	92.9	94.2	94.2	93.9	93.8	93.7	93.3
$N = 1000, T_0 = 40$									
Bias	2.16	0.53	0.81	0.16	0.21	0.31	0.43	0.52	0.63
SD	1.94	2.05	2.11	2.47	2.39	2.29	2.40	2.40	2.33
Coverage	79.8	93.6	92.8	94.2	93.9	94.0	94.4	94.2	93.7
$N = 200, T_0 = 100$									
Bias	2.16	0.90	1.16	0.47	0.49	0.66	0.81	0.83	0.95
SD	1.94	5.37	5.44	6.75	6.92	6.77	7.06	7.01	6.86
Coverage	79.8	93.9	93.5	93.1	92.8	93.0	93.6	93.5	93.3
$N = 1000, T_0 = 100$									
Bias	2.16	0.50	0.71	0.11	0.17	0.27	0.35	0.44	0.55
SD	1.94	2.05	2.08	2.49	2.39	2.32	2.51	2.41	2.33
Coverage	79.8	94.3	93.7	94.3	94.9	94.3	94.4	94.0	93.8
$N = 2000, T_0 = 100$									
Bias	2.16	0.36	0.55	0.10	0.14	0.19	0.29	0.37	0.42
SD	1.94	1.44	1.46	1.69	1.62	1.57	1.72	1.63	1.59
Coverage	79.8	93.8	92.5	94.0	94.0	94.1	94.2	94.2	93.6

Note: The table reports Monte Carlo results for the contemporaneous ATT based on 5000 replications of the dynamic panel model with additive fixed effects specified in (7.1) and the treatment selection mechanism specified in (7.2). I fix $\underline{q} = 80\%$ and vary $\kappa \in \{.15, .2, .25\}$ when selecting the bandwidth h_m^* according to Algorithm A.1. I report the bias, standard deviation, and coverage rate for the difference-in-differences estimator (DID), the infeasible doubly robust estimator based on a standard two-way sample split and the true latent distance (DR2*), and its feasible counterpart based on the estimated pseudo-distance (DR2). I also report the bias, standard deviation, and coverage rate for the infeasible estimator based on a three-way sample split and the true latent distance (DR3*), as well as the proposed estimator based on a three-way sample split and the estimated pseudo-distance (DR3). Biases and standard deviations are reported in units of **0.01**, and coverage rates are reported in percentage points **1%**.

The individual latent factors $\{\alpha_i\}_{i=1}^n$ and the time latent factors $\{\gamma_t\}_{t \leqslant T}$ are independent

random variables drawn from uniform distributions on $[-1, 1]$ and $[-2, 2]$, respectively. The error terms $\{\epsilon_{Y,it}\}_{i=1,\ldots,n,\ t=0,\ldots,T}$ are independent of the latent factors and are i.i.d. across both dimensions, following a uniform distribution on $[-1/2, 1/2]$. I set $\rho = 0.8$ to capture the dynamic effect. The contemporaneous ATT is 0.5, and for each $t > T_0$, the dynamic treatment effect is $0.5\left(1 + \rho + \ldots + \rho^{t-T_0}\right)$. Treatment assignment depends jointly on latent heterogeneity and the lagged outcome:

$$D_i = \mathbf{1}\left(\alpha_i/2 + Y_{iT_0-1}/2 + \epsilon_{D,i} \geqslant 0\right). \tag{7.4}$$

Here, the error terms $\{\epsilon_{D,i}\}_{i=1}^n$ are i.i.d. across individuals and are independent of the latent factors and $\{\epsilon_{Y,it}\}_{i=1,\ldots,n,\ t=0,\ldots,T}$. In the numerical designs, $\epsilon_{D,i}$ follows a standard logistic distribution. I vary $(N, T_0) \in \{(1000, 20), (1000, 40), (200, 100), (1000, 100), (2000, 100)\}$ and perform 5000 replications for each design.

Table 2 reports the simulation results for the contemporaneous ATT under interactive fixed effects (the DiD estimator is omitted because the DGP does not admit an additive two-way fixed-effects representation). The table shows that the proposed estimator reduces bias and achieves coverage close to the nominal 95% level. Also, the proposed estimator performs similarly to the infeasible DR3* estimator, indicating that the estimated pseudo-distance provides a good approximation to the true latent distance even when $T_0 = 20$ under the interactive fixed effects model. When $N = 200$, the performance of the proposed estimator is exceeded by that of the standard doubly robust estimator because the double cross-fitting procedure leaves relatively few observations in each subsample. Still, I recommend $(\kappa, \underline{q}) = (0.2, 0.8)$ as the default choice.

Table 2: Simulation Results: Dynamic Panel with Interactive Fixed Effects

	DR2*	DR2	DR3*			DR3		
			$\kappa = .15$	$\kappa = .2$	$\kappa = .25$	$\kappa = .15$	$\kappa = .2$	$\kappa = .25$
$N = 1000, T_0 = 20$								
Bias	0.60	0.82	0.24	0.34	0.46	0.46	0.57	0.69
SD	2.27	2.56	2.70	2.61	2.53	3.00	2.90	2.82
Coverage	93.5	92.1	93.6	93.6	93.5	93.1	92.8	92.3
$N = 1000, T_0 = 40$								
Bias	0.63	0.80	0.30	0.38	0.49	0.51	0.59	0.71
SD	2.29	2.49	2.70	2.65	2.56	2.87	2.81	2.76
Coverage	93.5	92.1	93.8	93.6	93.3	93.4	92.9	92.3
$N = 200, T_0 = 100$								
Bias	1.06	1.16	0.86	0.90	1.07	0.91	0.94	1.18
SD	6.11	6.29	8.12	8.09	7.97	8.29	8.25	8.17
Coverage	93.9	93.8	92.9	92.6	92.4	92.8	92.8	92.7
$N = 1000, T_0 = 100$								
Bias	0.62	0.75	0.30	0.36	0.50	0.41	0.50	0.62
SD	2.29	2.35	2.74	2.65	2.58	2.83	2.74	2.68
Coverage	93.3	92.7	93.7	93.5	93.1	93.8	93.4	92.6
$N = 2000, T_0 = 100$								
Bias	0.40	0.52	0.13	0.20	0.27	0.25	0.32	0.40
SD	1.56	1.62	1.83	1.76	1.71	1.92	1.84	1.80
Coverage	93.7	92.6	94.2	93.9	93.7	93.8	93.3	93.2

Note: The table reports Monte Carlo results for the contemporaneous ATT based on 5000 replications of the dynamic panel model with interactive fixed effects specified in (7.3) and the treatment selection mechanism specified in (7.4). I fix $\underline{q} = 80\%$ and vary $\kappa \in \{.15, .2, .25\}$ when selecting the bandwidth h_m^* according to Algorithm A.1. I report the bias, standard deviation, and coverage rate for the infeasible doubly robust estimator based on a standard two-way sample split and the true latent distance (DR2*), and its feasible counterpart based on the estimated pseudo-distance (DR2). I also report the bias, standard deviation, and coverage rate for the infeasible estimator based on a three-way sample split and the true latent distance (DR3*), as well as the proposed estimator based on a three-way sample split and the estimated pseudo-distance (DR3). Biases and standard deviations are reported in units of **0.01**, and coverage rates are reported in percentage points **1%**.

7.3 Nonlinear dynamic panel with fixed effects

Consider the following data-generating process:

$$Y_{it} = \begin{cases} \mathbf{1}\left(-2 + \rho Y_{it-1} + 4\alpha_i + \epsilon_{Y,it} \geqslant 0\right), & t < T_0 \\ \mathbf{1}\left(-2 + \rho Y_{it-1} + 4\alpha_i + D_i + \epsilon_{Y,it} \geqslant 0\right), & t \geqslant T_0 \end{cases}. \tag{7.5}$$

The individual latent factors $\{\alpha_i\}_{i=1}^n$ consist of independent random variables drawn from the uniform distribution on $[-1, 1]$. The error terms $\{\epsilon_{Y,it}\}_{i=1,\ldots,n,\ t=0,\ldots,T}$ are independent of the latent factors and are i.i.d. across both dimensions, following the standard logistic distribution. I set $\rho = 0.5$ to capture dynamic dependence. The contemporaneous ATT, computed using numerical simulation, is approximately 0.12. Dynamic treatment effects for $t > T_0$ can also be calculated numerically. Treatment assignment depends jointly on latent heterogeneity and the lagged outcome:

$$D_i = \mathbf{1}\left(\alpha_i/2 + Y_{iT_0-1}/2 + \epsilon_{D,i} \geqslant 0\right). \tag{7.6}$$

Here, the error terms $\{\epsilon_{D,i}\}_{i=1}^n$ are i.i.d. across individuals and are independent of the latent factors and $\{\epsilon_{Y,it}\}_{i=1,\ldots,n,\ t=0,\ldots,T}$. In the numerical designs, $\epsilon_{D,i}$ follows the standard logistic distribution. I vary the sample sizes $(N, T_0) \in \{(1000, 20), (200, 100), (1000, 100), (2000, 100)\}$ and perform 5000 replications for each design.

Table 3 reports the simulation results for the contemporaneous ATT under the nonlinear dynamic panel model (the DiD estimator is omitted because the DGP does not admit an additive two-way fixed-effects representation). The table shows that both the standard doubly robust estimator and the proposed estimator have small biases and achieve coverage close to the nominal 95% level. The standard doubly robust estimator performs particularly well in this design because Y_{it} is discrete, reducing the effective dimension of the nonparametric regression to one and allowing the estimator to be root-N consistent and asymptotically unbiased (Deaner, Hsiang and Zeleneev, 2025). The proposed estimator performs similarly to the infeasible DR3* estimator, indicating that the estimated pseudo-distance provides a good approximation to the true latent distance even when $T_0 = 20$. Finally, I recommend $(\kappa, \underline{q}) = (0.2, 0.8)$ as the default choice.

Overall, the proposed estimator substantially reduces bias and achieves coverage rates close to the nominal level. It performs well across different DGPs and sample sizes. In particular, its performance remains strong when the number of pretreatment periods is moderate, such as $T_0 = 20$. This result is especially relevant because panels with a very large time dimension are uncommon in empirical applications. The simulations show that

the performance of the proposed estimator is more sensitive to the cross-sectional sample size N and deteriorates when N is small.

Table 3: Simulation Results: Nonlinear Dynamic Panel with Fixed Effects

	DR2*	DR2	DR3*			DR3		
			$\kappa = .15$	$\kappa = .2$	$\kappa = .25$	$\kappa = .15$	$\kappa = .2$	$\kappa = .25$
$N = 1000, T_0 = 20$								
Bias	0.20	1.10	-0.31	-0.24	-0.02	-0.25	-0.05	0.38
SD	2.51	2.67	3.21	3.17	2.97	3.08	3.07	3.04
Coverage	95.1	93.1	94.2	94.3	94.9	94.0	94.1	94.1
$N = 200, T_0 = 100$								
Bias	0.38	0.77	-0.48	-0.47	-0.38	-0.54	-0.50	-0.33
SD	6.72	6.96	9.90	9.89	9.83	10.01	10.01	10.01
Coverage	94.3	94.5	93.6	93.4	93.2	92.4	92.4	92.6
$N = 1000, T_0 = 100$								
Bias	0.21	0.63	-0.31	-0.25	-0.06	-0.09	-0.04	0.06
SD	2.56	2.58	3.19	3.14	2.99	3.08	3.06	3.03
Coverage	95.1	94.5	94.5	94.6	94.2	93.9	94.1	94.1
$N = 2000, T_0 = 100$								
Bias	0.12	0.55	-0.06	-0.06	-0.06	-0.05	-0.04	0.00
SD	1.76	1.77	2.06	2.06	2.06	2.06	2.05	2.05
Coverage	95.1	94.5	94.8	94.8	94.6	94.2	94.3	94.2

Note: The table reports Monte Carlo results for the contemporaneous ATT based on 5000 replications of the nonlinear dynamic panel model with fixed effects specified in (7.5) and the treatment selection mechanism specified in (7.6). I fix $\underline{q} = 80\%$ and vary $\kappa \in \{.15, .2, .25\}$ when selecting the bandwidth h_m^* according to Algorithm A.1. I report the bias, standard deviation, and coverage rate for the infeasible doubly robust estimator based on a standard two-way sample split and the true latent distance (DR2*), and its feasible counterpart based on the estimated pseudo-distance (DR2). I also report the bias, standard deviation, and coverage rate for the infeasible estimator based on a three-way sample split and the true latent distance (DR3*), as well as the proposed estimator based on a three-way sample split and the estimated pseudo-distance (DR3). Biases and standard deviations are reported in units of **0.01**, and coverage rates are reported in percentage points **1%**.

8 Empirical Application

I apply the proposed method to the setting studied by Bailey (2012), estimating the long-run effects of U.S. family planning programs on childbearing, allowing treatment assignment to depend on county fixed effects and pretreatment fertility rates.

The original data is a county-year panel covers $3,037$ counties from 1959 to 1988. The outcome, Y_{it}, is the general fertility rate (GFR), defined as the number of live births per 1000 women aged $15-44$ in county i and year t. Treatment timing is summarized in Table 4 and is measured by the fiscal year in which a county first received a recorded federal family planning grant. Counties were treated sequentially between 1965 and 1973, and those without a recorded grant during this period are coded as untreated. Following Bailey (2012), I pool adjacent treatment years into three groups: $1965-1967$, $1968-1969$, and $1970-1973$[8].

Table 4: Treatment Timing

	Years	No. treated by year	Total
Group 1	$1965-1967$	6, 43, 74	123
Group 2	$1968-1969$	52, 278	330
Group 3	$1970-1973$	63, 75, 53, 10	201

These groups correspond to different stages of policy implementation. The first stage $(1965-67)$ was the initial phase funded by the Community Action Program (CAP) with informal funding procedures. The second stage $(1968-69)$ saw rapid expansion of the program, which the federal government had made a national emphasis program. The third stage $(1970-73)$ saw further expansion and was more institutionalized under the 1970 Title X legislation, which allowed the Department of Health, Education, and Welfare to provide grants directly to local family planning organizations. These stages were heterogeneous in funding priorities and the agencies responsible for awarding grants, leading to differences in selection mechanisms and treatment effects, as discussed below.

Pseudo distance To obtain a longer pretreatment outcome history, I match the Bailey replication sample to the *U.S. County-Level Natality and Mortality Data, 1915–2007*[9] assembled by Bailey et al. (2016). This database provides county-year live births by the mother's

[8]This is because estimating the ATT separately for each $g = 1965, \ldots, 1973$ is infeasible in practice: the number of counties first treated in some years (e.g., 1965 and 1973) is very small, making the corresponding nonparametric estimates unstable and imprecise. In addition, since the analysis focuses on medium- and long-run effects, pooling adjacent cohorts sacrifices little timing variation while preserving the staggered rollout of the program.

[9]The data is publicly available at https://www.icpsr.umich.edu/web/ICPSR/studies/36603.

county of residence and estimates of the female population aged $15-44$ since 1937. This enables me to construct the historical GFR from 1937 to 1958[10]. The final panel contains $3,017$ county units observed from 1937 to 1988 with 28 years of pretreatment outcomes for each county[11].

I use the pretreatment outcomes from $1937-1964$ to calculate pseudo-distance between each pair of counties. Although someone may argue that there are structural changes in the series because the period spans the Second World War and the baby boom, this does not affect the construction when the structural change can be captured by an additive common time trend, as the common trend is differenced out by cross-sectional differencing. The results are similar when I use data from $1946-1964$ to calculate the pseudo-distance.

Selection The selection mechanism of the family planning program consists of two steps: first, local organizations decide whether to apply for the program, and second, conditional on the applicaiton, the federal official decide whether and when to approve the applications. Bailey (2012) documents strong evidence that the selection depends on fixed effects: she finds that treated counties were larger and more urban and, on average, more educated and affluent than unfunded counties, which supports the selection on county fixed effects.

In addition, pretreatment fertility rates may affect treatment assignment through several channels. For example, counties with higher recent fertility rates may have greater demand for subsidized family planning services and, therefore, stronger incentives to apply for federal funding. This channel leads to a positive relationship between pretreatment fertility rates and selection into the program. On the other hand, federal officials may have been more likely to approve applications from counties with lower lagged fertility rates, as this would make it easier to claim the success of the program, generating a negative relationship between pretreatment fertility rates and selection into the program.

I find strong reduced-form empirical evidence that lagged fertility rates affect selection into treatment. Specifically, for each group g and each t in the group, I estimate the following regression for units that are untreated at time t:

$$D_{it} \sim \text{Logit}\left(\beta_g' X_{it} + \alpha_i\right).$$

In Table 5, for each treatment group, I report logistic regressions of selection into treatment on either (1) the one-period lagged outcome or (2) the average of the three most recent

[10]For the overlapping period 1959-1988, I just retain the GFR reported in the Bailey (2012) replication files rather than replacing it with the reconstructed series.

[11]Counties that could not be matched or lacked the required outcome history were excluded from the final sample. For example, Los Alamos was excluded because its outcome data are unavailable before 1950.

lagged outcomes. The fixed effect is unobserved and is therefore proxied by urbanization, education, income, race, and population. Only the coefficient on the lagged outcome is reported. I find strong empirical evidence that lagged outcomes are associated with treatment for counties treated in $1965-67$, with negative and statistically significant coefficients across both specifications. The coefficients for $1968-69$ are not statistically significant, which may reflect the mixed selection channels (I also estimate year-specific regressions, reported in Table A.1, and find a negative and statistically significant coefficient for 1968, indicating negative selection on lagged fertility rates in at least one year within the second group). In addition, I find a positive and statistically significant coefficient for $(1970-73)$ when using the one-period lagged outcome, but the coefficient is not statistically significant when using the three-period average. In summary, the reduced-form empirical evidence supports selection on lagged fertility rates, although the effect is heterogeneous over time.

Table 5: Selection on lagged outcomes

	Treated $1965-67$		Treated $1968-69$		Treated $1970-73$	
	(1)	(2)	(1)	(2)	(1)	(2)
$\widehat{\beta}_g$	-2.08^{***}	-1.76^{***}	0.46	0.15	0.56^{**}	0.46
	(0.56)	(0.56)	(0.34)	(0.35)	(0.27)	(0.37)
AME	-0.03	-0.02	0.02	0.01	0.01	0.01

Note: This table reports the estimated logistic regression coefficients on lagged outcomes for counties first funded in 1965-67, 1968-69, and 1970-73. Columns (1) and (2) use the one-period lag and the average of the three preceding lags, respectively. I control for urbanization (the percentage of the population living in urban areas in 1960), education (the percentage of the population with more than 12 years of education in 1960), income (the percentage of the population with an annual income below 3,000 US dollars), race (the percentage of the population that was nonwhite in 1960), and population (in 1960). Standard errors are reported in parentheses. $***$, $**$, and $*$ denote statistical significance at the 1%, 5%, and 10% levels, respectively.

Dynamic dependence I also find strong reduced-form empirical evidence supporting dynamic dependence in fertility rates. Specifically, I estimate the following linear panel model:

$$Y_{it}(\infty) = \rho Y_{it-1}(\infty) + \alpha_i + \gamma_t + \epsilon_{it}, \quad \forall t = 1965, \ldots, 1988.$$

I report the results using (i) the Arellano–Bond estimator (Arellano and Bond, 1991) with different sets of instruments and (ii) a least-squares estimator with analytical *bias correction*. Table 6 reports the results. Across specifications, I find positive and statistically significant dynamic dependence in fertility rates, even after controlling for unit and time fixed effects.

Table 6: Dynamic dependence

	Arellano–Bond			Dynamic-panel
	IV lag 2	IV lags $2-4$	IV lags $2-6$	
$\widehat{\rho}$	0.21^{***}	0.24^{***}	0.34^{***}	0.50^{***}
	(0.03)	(0.04)	(0.04)	(0.02)

Notes: Standard errors are reported in parentheses. The first three columns report two-step Arellano–Bond estimator using the indicated lag windows. The final column reports the bias-corrected dynamic-panel within estimate. $^{***}p < 0.01$.

Estimation For estimation, I control for unobserved time-invariant heterogeneity using the pseudo-distance and for observed lagged fertility rates. I consider the following specifications[12]. I set $\kappa = 0.2$, as suggested by the simulations. The estimates in Bailey (2012) are based on event-study regressions and are therefore potentially affected by negative weights, so I also report estimates obtained using the DiD estimator of Callaway and Sant'Anna (2021):

	Pretreatment history	X_{it}
Model 1	$1937-1964$	$(Y_{it-1}+Y_{it-2}+Y_{it-3})/3$
Model 2	$1937-1964$	Y_{it-1}
Model 3	$1946-1964$	$(Y_{it-1}+Y_{it-2}+Y_{it-3})/3$

For each treatment group, I report ATT estimates beginning in the year after the latest treatment year in that group. For example, for the $1965-1967$ group, the reported ATT estimates begin in 1968.

Results Figure 2 presents the dynamic ATT estimates and 95% confidence intervals obtained using the proposed method and the DiD estimator of Callaway and Sant'Anna (2021). The first row displays the dynamic ATT estimates from the proposed method. The estimates show that the family planning program has a statistically significant negative effect on fertility. In addition, the long-run effect is statistically significant for the group treated in 1968-69. The DiD estimates in the second row show similar results.

The last row in Figure 2 compares estimates from the proposed method and the DiD estimator over the first 15 post-treatment periods. The proposed method yields larger estimated reductions in fertility for the first two treatment groups, whereas the estimates are similar for the third group. For the first two groups, the gap becomes larger at longer post-treatment

[12]The average is calculated using fertility rates from the three years preceding the earliest treatment year in that group. For example, for the $1965-1967$ group, I use the average fertility rate over $1962-1964$.

horizons. These results highlight the importance of accounting jointly for selection on lagged outcomes and their dynamic effects when evaluating policies.

Table 7 reports treatment effects averaged over three five-year windows and across the three treatment groups. The Callaway-Sant'Anna DiD estimates are similar to those reported by Bailey (2012), but the proposed estimator produces more negative estimates than the Callaway-Sant'Anna DiD estimator in all three windows. The magnitude of the difference increases from 0.32 in years $1-5$ to 0.76 in years $6-10$ and then to 1.61 in years $11-15$.

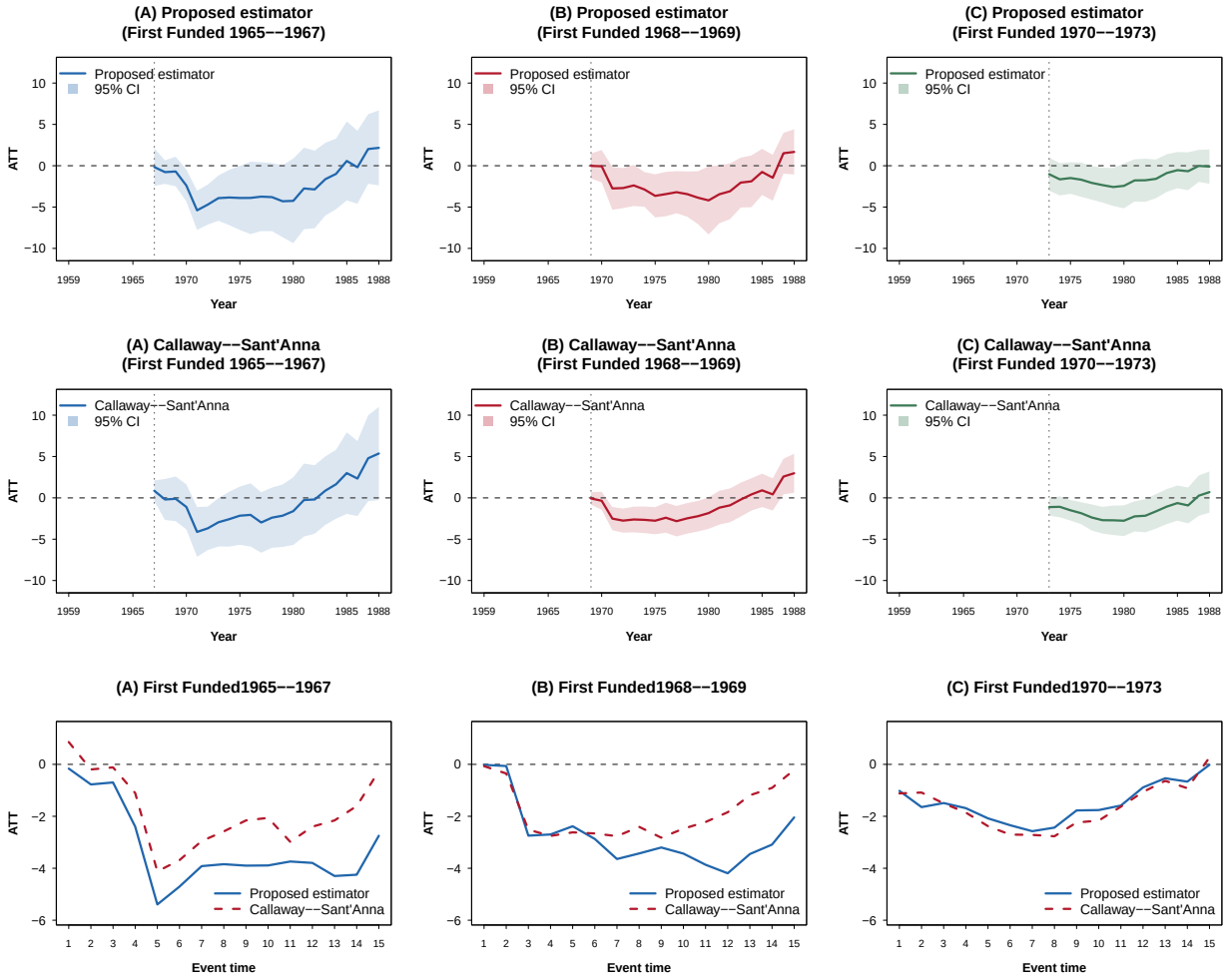


Figure 2: Weighted dynamic ATT (Model 1).

Note: Following Bailey (2012), the results are weighted by the number of women aged $15-44$ in 1970.

Table 7: ATT by Post-Treatment Horizon

	ATT (1 - 5)	ATT (6 - 10)	ATT (11 - 15)
Proposed estimator (Model 1)	-1.70^{**}	-3.39^{***}	-3.03^{**}
	(0.67)	(1.01)	(1.39)
Proposed estimator (Model 2)	-1.74^{***}	-3.47^{***}	-2.92^{**}
	(0.59)	(0.90)	(1.36)
Proposed estimator (Model 3)	-1.92^{***}	-3.56^{***}	-3.06^{**}
	(0.69)	(1.06)	(1.40)
DiD Callaway-Sant'Anna	-1.37^{**}	-2.63^{***}	-1.42
	(0.54)	(0.80)	(0.94)
Difference (Model 1)	-0.32	-0.76	-1.61
	(0.76)	(1.11)	(1.28)
Difference (Model 2)	-0.37	-0.89	-1.51
	(0.60)	(0.94)	(1.22)
Difference (Model 3)	-0.55	-0.93	-1.64
	(0.82)	(1.21)	(1.33)

Note: Following Bailey (2012), the results are weighted by the number of women aged $15-44$ in 1970.

Discussion The differences between the two estimators can be explained by selection on lagged fertility rates and dynamic dependence. For counties first funded in $1965-67$, Table 5 indicates that counties with lower lagged fertility rates are more likely to be treated. Since these negative shocks to fertility rates are temporary, fertility in the treated counties would recover on average in the following period and generate increases in untreated fertility. DiD attributes this recovery effect to the program, generating upward bias and understating the reduction in fertility. In addition, since the recovery occurs gradually, it generates larger upward bias at longer horizons, consistent with the empirical results.

For the second group, although Table 5 does not provide evidence of selection on negative lagged fertility shocks, the regression for 1968 reported in Table A.1 shows a negative coefficient. Thus, the insignificant coefficients in Table 5 may reflect offsetting selection effects across years. Hence, the same recovery mechanism explain the differences between the two estimators for the second group. For the third group, Table 5 shows evidence of selection on positive fertility shocks, while the two estimators produce similar treatment effect estimates, as shown in Figure 2. This similarity may be related to the slowdown in the decline of U.S. fertility rates after 1973.

References

Abadie, Alberto. 2005. "Semiparametric difference-in-differences estimators." *The review of economic studies*, 72(1): 1–19.

Abadie, Alberto, and Javier Gardeazabal. 2003. "The economic costs of conflict: A case study of the Basque Country." *American economic review*, 93(1): 113–132.

Abadie, Alberto, Anish Agarwal, Raaz Dwivedi, and Abhin Shah. 2024. "Doubly robust inference in causal latent factor models." *arXiv preprint arXiv:2402.11652*.

Alvarez, Javier, and Manuel Arellano. 2003. "The time series and cross-section asymptotics of dynamic panel data estimators." *Econometrica*, 71(4): 1121–1159.

Arellano, Manuel, and Stephen Bond. 1991. "Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations." *The review of economic studies*, 58(2): 277–297.

Arkhangelsky, Dmitry, and David Hirshberg. 2023. "Large-sample properties of the synthetic control method under selection on unobservables." *arXiv preprint arXiv:2311.13575*.

Arkhangelsky, Dmitry, and Guido Imbens. 2024. "Causal models for longitudinal and panel data: A survey." *The Econometrics Journal*, 27(3): C1–C61.

Arkhangelsky, Dmitry, and Guido W Imbens. 2022. "Doubly robust identification for causal panel data models." *The Econometrics Journal*, 25(3): 649–674.

Arkhangelsky, Dmitry, Guido W Imbens, Lihua Lei, and Xiaoman Luo. 2024. "Design-robust two-way-fixed-effects regression for panel data." *Quantitative Economics*, 15(4): 999–1034.

Arkhangelsky, Dmitry, Susan Athey, David A Hirshberg, Guido W Imbens, and Stefan Wager. 2021. "Synthetic difference-in-differences." *American economic review*, 111(12): 4088–4118.

Ashenfelter, Orley. 1978. "Estimating the effect of training programs on earnings." *The Review of Economics and Statistics*, 60(1): 47–57.

Ashenfelter, Orley C, and David Card. 1985. "Using the longitudinal structure of earnings to estimate the effect of training programs."

Athey, Susan, and Guido Imbens. 2025. "Identification of average treatment effects in nonparametric panel models." *arXiv preprint arXiv:2503.19873*.

Athey, Susan, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi. 2021. "Matrix completion methods for causal panel data models." *Journal of the American Statistical Association*, 116(536): 1716–1730.

Auerbach, Eric. 2022. "Identification and estimation of a partially linear regression model using network data." *Econometrica*, 90(1): 347–365.

Bai, Jushan. 2009. "Panel data models with interactive fixed effects." *Econometrica*, 77(4): 1229–1279.

Bai, Jushan, and Serena Ng. 2021. "Matrix completion, counterfactuals, and factor analysis of missing data." *Journal of the American Statistical Association*, 116(536): 1746–1763.

Bailey, Martha J. 2012. "Reexamining the impact of family planning programs on US fertility: evidence from the War on Poverty and the early years of Title X." *American Economic Journal: Applied Economics*, 4(2): 62–97.

Bailey, Martha, Karen Clay, Price Fishback, Michael Haines, Shawn Kantor, Edson Severnini, and Anna Wentz. 2016. "U.S. County-Level Natality and Mortality Data, 1915–2007." Version V4.

Blundell, Richard, and Stephen Bond. 1998. "Initial conditions and moment restrictions in dynamic panel data models." *Journal of econometrics*, 87(1): 115–143.

Blundell, Richard, Rachel Griffith, and Frank Windmeijer. 2002. "Individual effects and dynamics in count data models." *Journal of econometrics*, 108(1): 113–131.

Bonhomme, Stéphane. 2012. "Functional differencing." *Econometrica*, 80(4): 1337–1385.

Bonhomme, Stephane. 2025. "Back to feedback: Dynamics and heterogeneity in panel data." *arXiv preprint arXiv:2512.17576*.

Bonhomme, Stéphane, and Elena Manresa. 2015. "Grouped patterns of heterogeneity in panel data." *Econometrica*, 83(3): 1147–1184.

Bonhomme, Stéphane, Thibaut Lamadon, and Elena Manresa. 2022. "Discretizing unobserved heterogeneity." *Econometrica*, 90(2): 625–643.

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess. 2024. "Revisiting event-study designs: robust and efficient estimation." *Review of economic studies*, 91(6): 3253–3285.

Botosaru, Irene, and Laura Liu. 2025. "Time-varying heterogeneous treatment effects in event studies." *arXiv preprint arXiv:2509.13698*.

Caetano, Carolina, Brantly Callaway, Stroud Payne, and Hugo Sant'Anna Rodrigues. 2022. "Difference in differences with time-varying covariates." *arXiv preprint arXiv:2202.02903*.

Callaway, Brantly, and Pedro HC Sant'Anna. 2021. "Difference-in-differences with multiple time periods." *Journal of econometrics*, 225(2): 200–230.

Chang, Neng-Chieh. 2020. "Double/debiased machine learning for difference-in-differences models." *The Econometrics Journal*, 23(2): 177–191.

Chernozhukov, Victor, Ben Deaner, Ying Gao, Jerry Hausman, and Whitney K Newey. 2026. "Linear estimation of structural and causal effects for nonseparable panel data." *arXiv preprint arXiv:2607.28291*.

Chernozhukov, Victor, Christian Hansen, Yuan Liao, and Yinchu Zhu. 2023. "Inference for low-rank models." *The Annals of statistics*, 51(3): 1309–1330.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. "Double/debiased machine learning for treatment and structural parameters." *The Econometric Journal*.

Chernozhukov, Victor, Iván Fernández-Val, Jinyong Hahn, and Whitney Newey. 2013. "Average and quantile effects in nonseparable panel models." *Econometrica*, 81(2): 535–580.

Chernozhukov, Victor, Whitney Newey, Rahul Singh, and Vasilis Syrgkanis. 2022. "Automatic debiased machine learning for dynamic treatment effects and general nested functionals." *arXiv preprint arXiv:2203.13887*.

Deaner, Ben. 2018. "Proxy controls and panel data." *arXiv preprint arXiv:1810.00283*.

Deaner, Ben, Chen-Wei Hsiang, and Andrei Zeleneev. 2025. "Inferring treatment effects in large panels by uncovering latent similarities." *arXiv preprint arXiv:2503.20769*.

De Chaisemartin, Clément, and Xavier d'Haultfoeuille. 2023. "Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey." *The econometrics journal*, 26(3): C1–C30.

Feng, Yingjie. 2020. "Causal inference in possibly nonlinear factor models." *arXiv preprint arXiv:2008.13651*.

Feng, Yingjie. 2023. "Optimal Estimation of Large-Dimensional Nonlinear Factor Models." *arXiv preprint arXiv:2311.07243*.

Ghanem, Dalia, Pedro HC Sant'Anna, and Kaspar Wüthrich. 2022. "Selection and parallel trends." *arXiv preprint arXiv:2203.09001*.

Gulek, Ahmet, and Jaume Vives-i Bastida. 2025. "Synthetic IV Estimation in Panels." Technical report, MIT, 2025. Job Market Paper.

Hayashi, Fumio. 2011. *Econometrics*. Princeton University Press.

Heckman, James J. 1981. "Heterogeneity and state dependence." In *Studies in labor markets*. 91–140. University of Chicago Press.

Heckman, James J, and Salvador Navarro. 2007. "Dynamic discrete choice and dynamic treatment effects." *Journal of Econometrics*, 136(2): 341–396.

Hoderlein, Stefan, and Halbert White. 2012. "Nonparametric identification in nonseparable panel data models with generalized fixed effects." *Journal of Econometrics*, 168(2): 300–314.

Honoré, Bo E, and Ekaterini Kyriazidou. 2000. "Panel data discrete choice models with lagged dependent variables." *Econometrica*, 68(4): 839–874.

Hoshino, Tadao, and Takahide Yanagi. 2024. "Estimating Dyadic Treatment Effects with Unknown Confounders." *arXiv preprint arXiv:2405.16547*.

Hu, Yingyao, and Matthew Shum. 2012. "Nonparametric identification of dynamic models with unobserved state variables." *Journal of Econometrics*, 171(1): 32–44.

Lewbel, Arthur. 2019. "The identification zoo: Meanings of identification in econometrics." *Journal of Economic Literature*, 57(4): 835–903.

Lewis, Greg, and Vasilis Syrgkanis. 2020. "Double/debiased machine learning for dynamic treatment effects via g-estimation." *arXiv preprint arXiv:2002.07285*.

Marx, Philip, Elie Tamer, and Xun Tang. 2024*a*. "Heterogeneous Treatment Effects via Linear Dynamic Panel Data Models." *arXiv preprint arXiv:2410.19060*.

Marx, Philip, Elie Tamer, and Xun Tang. 2024*b*. "Parallel trends and dynamic choices." *Journal of Political Economy Microeconomics*, 2(1): 129–171.

McClean, Alec, Sivaraman Balakrishnan, Edward H Kennedy, and Larry Wasserman. 2026. "Double cross-fit doubly robust estimators: Beyond series regression." *Journal of the Royal Statistical Society Series B: Statistical Methodology*, qkag057.

Merlevède, Florence, Magda Peligrad, and Emmanuel Rio. 2009. "Bernstein inequality and moderate deviations under strong mixing conditions." In *High dimensional probability V: the Luminy volume*. Vol. 5, 273–293. Institute of Mathematical Statistics.

Moon, Hyungsik Roger, and Martin Weidner. 2017. "Dynamic linear panel regression models with interactive fixed effects." *Econometric Theory*, 33(1): 158–195.

Newey, Whitney K, and James R Robins. 2018. "Cross-fitting and fast remainder rates for semiparametric estimation." *arXiv preprint arXiv:1801.09138*.

Nowakowicz, J. 2024. "Nonparametric network bootstrap." *arXiv preprint arXiv:2401.12345*.

Robins, James M, Miguel Angel Hernan, and Babette Brumback. 2000. "Marginal structural models and causal inference in epidemiology." *Epidemiology*, 11(5): 550–560.

Roth, Jonathan, Pedro HC Sant'Anna, Alyssa Bilinski, and John Poe. 2023. "What's trending in difference-in-differences? A synthesis of the recent econometrics literature." *Journal of econometrics*, 235(2): 2218–2244.

Sant'Anna, Pedro HC, and Jun Zhao. 2020. "Doubly robust difference-in-differences estimators." *Journal of econometrics*, 219(1): 101–122.

Sasaki, Yuya. 2015. "Heterogeneity and selection in dynamic panel data." *Journal of Econometrics*, 188(1): 236–249.

Stock, James H, and Mark W Watson. 2002. "Forecasting using principal components from a large number of predictors." *Journal of the American statistical association*, 97(460): 1167–1179.

Sun, Ge, and Weisheng Zhang. 2026. "Flexible Imputation of Incomplete Network Data." *arXiv preprint arXiv:2604.03171*.

Viviano, Davide, and Jelena Bradic. 2026. "Dynamic covariate balancing: estimating treatment effects over time with potential local projections." *Biometrika*, 113(2): asag016.

Wainwright, Martin J. 2019. *High-dimensional statistics: A non-asymptotic viewpoint.* Vol. 48, Cambridge university press.

Wooldridge, Jeffrey M. 2025. "Two-way fixed effects, the two-way mundlak regression, and difference-in-differences estimators." *Empirical Economics*, 69(5): 2545–2587.

Zeleneev, Andrei. 2020. "Identification and estimation of network models with nonparametric unobserved heterogeneity." *Department of Economics, Princeton University*.

Zhang, Yuan, Elizaveta Levina, and Ji Zhu. 2017. "Estimating network edge probabilities by neighbourhood smoothing." *Biometrika*, 104(4): 771–783.

A Additional Empirical Results, Algorithm, and Simulations

For each $t = 1965, \ldots, 1973$, I estimate the following regression for units that are untreated at time t:

$$D_{it} \sim \text{Logit}\left(\beta_t' X_{it} + \alpha_i\right).$$

In Table A.1, for each t, I report logistic regressions of selection into treatment on either (1) the one-period lagged outcome or (2) the average of the three most recent lagged outcomes. The fixed effect is unobserved and is therefore proxied by urbanization, education, income, race, and population. Only the coefficient on the lagged outcome is reported.

Table A.1: Selection on lagged outcomes - suupplement

Year	Treated	Model (1)		Model (2)	
Initial CAP programs					
1965	6	−5.01**	(1.99)	−4.19*	(2.14)
1966	43	−3.37	(2.94)	−2.41	(2.82)
1967	74	2.45	(1.56)	2.34	(1.61)
National-emphasis expansion					
1968	52	−2.45	(1.74)	−3.82*	(2.14)
1969	278	0.48	(1.33)	1.27	(1.21)
Title X expansion					
1970	63	1.81	(1.34)	1.54	(1.42)
1971	75	1.99*	(1.20)	3.01**	(1.42)
1972	53	0.73	(2.45)	0.76	(2.81)
1973	10	−1.11	(4.56)	−1.92	(4.85)

Notes: Each row reports a separate weighted Logit regression. Coefficients and standard errors, reported in parentheses, are multiplied by 100. The regressions are weighted by the number of women aged 15-44 in 1970.

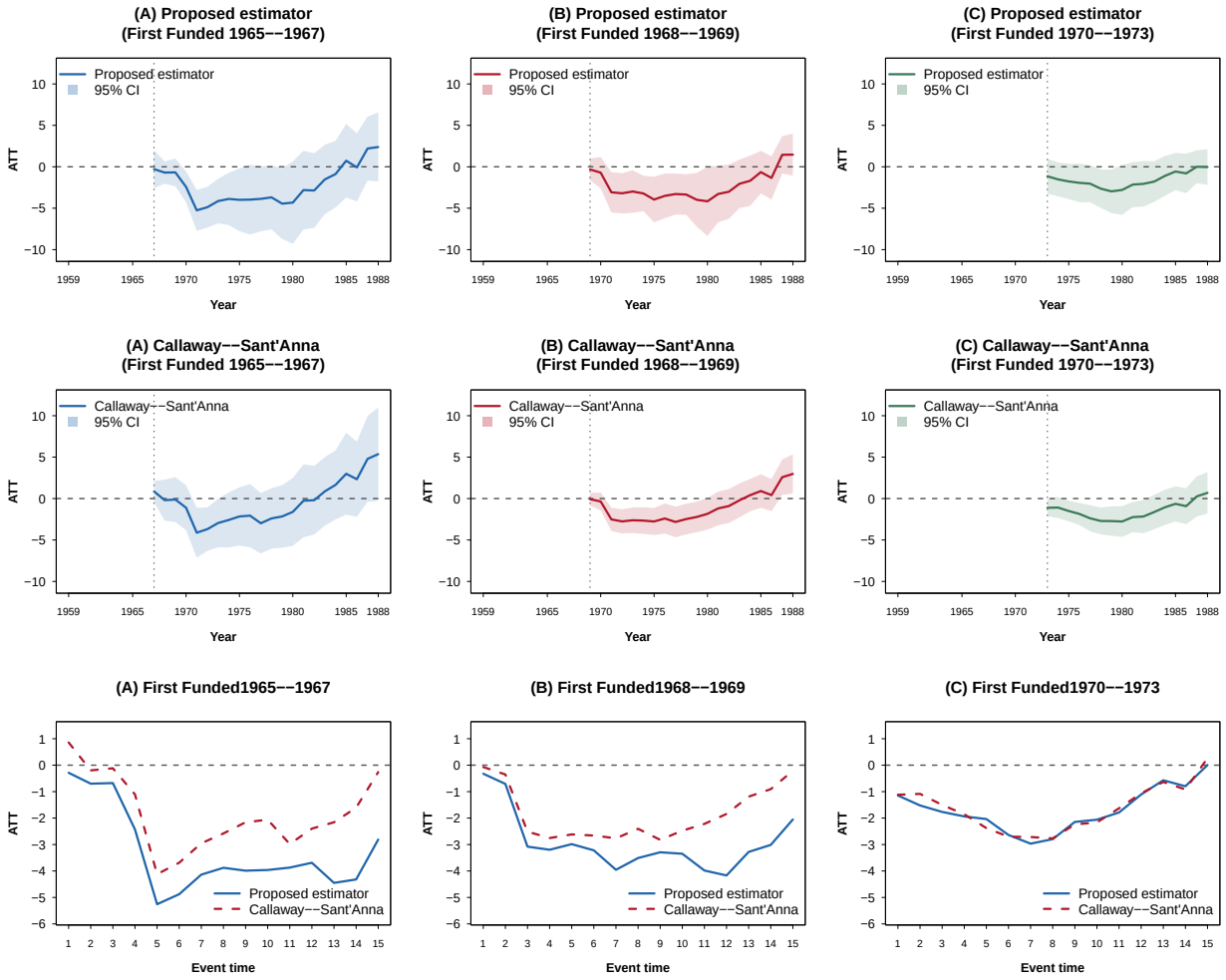


Figure 3: Weighted dynamic ATT (Model 2).

Note: Following Bailey (2012), the results are weighted by the number of women aged $15-44$ in 1970.

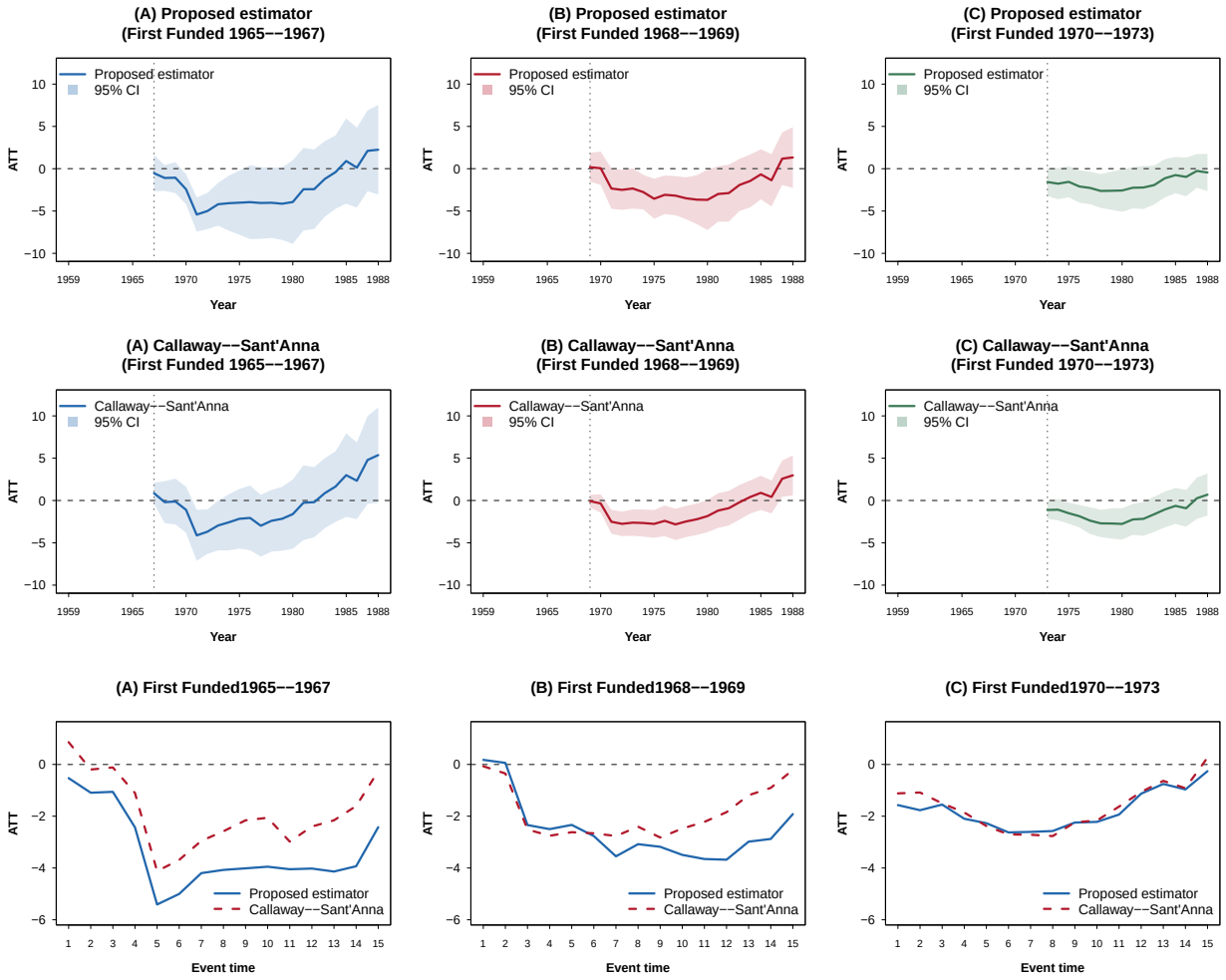


Figure 4: Weighted dynamic ATT (Model 3).

Note: Following Bailey (2012), the results are weighted by the number of women aged $15-44$ in 1970.

Algorithm A.1 Doubly robust dynamic ATT estimator - simultaneous treatment

Input: Kernel $K(\cdot)$, two lists of bandwidths, $\{h_{m1}, \ldots, h_{mb}\}$ and $\{h_{\pi 1}, \ldots, h_{\pi b}\}$, tuning parameters $\kappa > 0$ and $\underline{q} \in (0, 1]$, and confidence level $1 - \alpha$.

Output: Dynamic ATT estimates $\widehat{\text{ATT}}(t)$, stand errors $\widehat{\text{se}}_t$, $1 - \alpha$ confidence interval.

Step 1: Sample-splitting Randomly partition samples into disjoint $\{\mathcal{I}_1, \mathcal{I}_2, \mathcal{I}_3\}$.

Step 2: Calculate similarity For each $i, j = 1, \ldots, n$, calculate the pseudo-distance:

$$\widehat{d}_{ij} = \max_{k_1, k_2 = 1, \ldots, n, k_1, k_2 \neq i, j} \left| \frac{1}{T_0} \sum_{t=0}^{T_0 - 1} (Y_{k_1 t} - Y_{k_2 t})(Y_{it} - Y_{jt}) \right|.$$

Step 3: Cross validation

For each $i, j = 1, \ldots, n$, any $h > 0$, $\widehat{K}_{h,ij} := K\left(\left(X_{jT_0} - X_{iT_0} \right) / h \right) K(\widehat{d}_{ij}/h)$.

(i) Leave-one-out cross validation for π. For each i, compute the leave-one-out predictors for each $h \in \{h_{\pi 1}, \ldots, h_{\pi b}\}$,

$$\widetilde{\pi}_i^{(-i)}(h) = \sum_{i' \in \mathcal{I}(i) \setminus \{i\}} D_{i'} \widehat{K}_{h,ii'} \Big/ \sum_{i' \in \mathcal{I}(i) \setminus \{i\}} \widehat{K}_{h,ii'},$$

and select the optimal bandwidth h_π^* as

$$h_\pi^* \in \operatorname*{argmin}_{h \in \{h_{\pi 1}, \ldots, h_{\pi b}\}} \frac{1}{n} \sum_{i=1}^{n} (D_i - \widetilde{\pi}_i^{(-i)}(h))^2.$$

(ii) Cross validation for m. For i, compute effective sample size for each $h \in \{h_{m1}, \ldots, h_{mb}\}$,

$$n_{i,\text{eff}}(h) = \left(\sum_{i' \in \mathcal{I}(i) \setminus \{i\}} (1 - D_{i'}) \widehat{K}_{h,ii'} \right)^2 \Big/ \sum_{i' \in \mathcal{I}(i) \setminus \{i\}} (1 - D_{i'}) \widehat{K}_{h,ii'}^2,$$

and select h_m^* as

$$h_m^* \in \min \left\{ h \in \{h_{m1}, \ldots, h_{mb}\} : \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\left(n_{i,\text{eff}}(h) > (\kappa \log(n))^2 \right) \geqslant \underline{q} \right\}.$$

Step 4: Compute $\widehat{m}_{it}$ and $\widehat{\pi}_i$ For each $i = 1, \ldots, n$, compute

$$\widehat{m}_{it} = \frac{\sum_{j \in \mathcal{I}_m(i)} (1 - D_j) \widehat{K}_{h_m^*,ij} Y_{jt}}{\sum_{j \in \mathcal{I}_m(i)} (1 - D_j) \widehat{K}_{h_m^*,ij}}, \quad \widehat{\pi}_i = \frac{\sum_{j \in \mathcal{I}_\pi(i)} D_j \widehat{K}_{h_\pi^*,ij}}{\sum_{j \in \mathcal{I}_\pi(i)} \widehat{K}_{h_\pi^*,ij}}$$

Step 5: Construct $\widehat{\text{ATT}}(t)$ and its CI

Let $n_1 := \sum_{i=1}^{n} \mathbf{1}(D_i = 1)$, compute

$$\widehat{\text{ATT}}(t) = \frac{1}{n_1} \sum_{i}^{n} \left(D_i Y_{it} - \frac{(1 - D_i)\widehat{\pi}_i Y_{it} + (D_i - \widehat{\pi}_i)\widehat{m}_{it}}{1 - \widehat{\pi}_i} \right),$$

$$\widehat{\text{se}}_t = \left(\frac{1}{n_1^2} \sum_{i=1}^{n} \left(D_i Y_{it} - \frac{(1 - D_i)\widehat{\pi}_i Y_{it} + (D_i - \widehat{\pi}_i)\widehat{m}_{it}}{1 - \widehat{\pi}_i} - D_i \widehat{\text{ATT}}(t) \right)^2 \right)^{1/2}.$$

The confidence interval is $\text{CI}_t = [\widehat{\text{ATT}}(t) \pm Z_{1-\alpha/2} \cdot \widehat{\text{se}}_t]$.

Algorithm A.2 Doubly robust dynamic ATT estimator - staggered adoption (Basic Idea)

Input: Kernel $K(\cdot)$, bandwidths h_m, h_π, and confidence level $1-\alpha$.

Output: Dynamic ATT estimates $\widehat{\text{ATT}(g,t)}$, stand errors $\widehat{\text{se}}_t$, $1-\alpha$ confidence interval.

Step 1: Sample-splitting Randomly partition samples into disjoint $\{\mathcal{I}_1, \mathcal{I}_2, \mathcal{I}_3\}$.

Step 2: Calculate similarity For each $i, j = 1, \ldots, n$, calculate the pseudo-distance:

$$\widehat{d}_{ij} = \max_{k_1,k_2=1,\ldots,n, k_1,k_2 \neq i,j} \left| \frac{1}{T_0} \sum_{t=0}^{T_0-1} (Y_{k_1 t} - Y_{k_2 t})(Y_{it} - Y_{jt}) \right|.$$

Step 3: Compute expected outcomes and propensity scores

For each $i, j = 1, \ldots, n$, any $h > 0$, and each t', $\widehat{K}^{(t')}_{h,ij} := K\left(\left(X_{jt'} - X_{it'}\right)/h\right) K(\widehat{d}_{ij}/h)$.

Initialization. For each $j = 1, \ldots, n$, set $\widetilde{m}_{j,t|t+1} := Y_{jt}$.

for $t' = t, t-1, \ldots, g$ **do**

For each $i = 1, \ldots, n$, compute the Nadaraya-Watson estimators

$$\widehat{m}_{i,t|t'} = \frac{\sum_{j \in \mathcal{I}_m(i)} \mathbf{1}\left(G_j > t'\right) \widetilde{m}_{j,t|t'+1} \widehat{K}^{(t')}_{h_m,ij}}{\sum_{j \in \mathcal{I}_m(i)} \mathbf{1}\left(G_j > t'\right) \widehat{K}^{(t')}_{h_m,ij}}, \quad \widehat{\pi}_{it'} = \frac{\sum_{j \in \mathcal{I}_\pi(i)} \mathbf{1}\left(G_j = t'\right) \widehat{K}^{(t')}_{h_\pi,ij}}{\sum_{j \in \mathcal{I}_\pi(i)} \mathbf{1}\left(G_j > t'-1\right) \widehat{K}^{(t')}_{h_\pi,ij}},$$

wherefor each $j = 1, \ldots, n$, the within-sample estimators are:

$$\widetilde{m}_{j,t|t'} = \sum_{j' \in \mathcal{I}(j)} \mathbf{1}\left(G_{j'} > t'\right) \widetilde{m}_{j',t|t'+1} \widehat{K}^{(t')}_{h_m,jj'} \Big/ \sum_{j' \in \mathcal{I}(j)} \mathbf{1}\left(G_{j'} > t'\right) \widehat{K}^{(t')}_{h_m,jj'}.$$

end for

Step 4: Construct $\widehat{\text{ATT}(t)}$, standard error $\widehat{\text{se}}_t$, and confidence interval

Let $n_g := \sum_{i=1}^n \mathbf{1}(G_i = g)$, compute

$$\widehat{ATT(g,t)} = \frac{1}{n_g} \sum_{i=1}^n \Bigg[\mathbf{1}\left(G_i = g\right) \left(Y_{it} - \widehat{m}_{i,t|g}\right) - \sum_{t'=g}^{t} \mathbf{1}\left(G_i > t'\right) \frac{\widehat{\pi}_{ig}}{1 - \widehat{\pi}_{ig}} \left(\prod_{r=g+1}^{t'} \frac{1}{1 - \widehat{\pi}_{ir}} \right) \left(\widehat{m}_{i,t|t'+1} - \widehat{m}_{i,t|t'}\right) \Bigg],$$

where an empty product is defined as one. Then, calculate

$$\widehat{\text{se}}_t = \left(\frac{1}{n_g^2} \sum_{i=1}^n \Bigg[\mathbf{1}\left(G_i = g\right) \left(Y_{it} - \widehat{m}_{i,t|g} - \widehat{ATT(g,t)}\right) - \sum_{t'=g}^{t} \mathbf{1}\left(G_i > t'\right) \frac{\widehat{\pi}_{ig}}{1 - \widehat{\pi}_{ig}} \left(\prod_{r=g+1}^{t'} \frac{1}{1 - \widehat{\pi}_{ir}} \right) \left(\widehat{m}_{i,t|t'+1} - \widehat{m}_{i,t|t'}\right) \Bigg]^2 \right)^{1/2}.$$

The confidence interval is $\text{CI}_t = [\widehat{\text{ATT}(t)} \pm Z_{1-\alpha/2} \cdot \widehat{\text{se}}_t]$.

Algorithm A.3 Doubly robust estimator for dynamic ATT under staggered adoption

Input: Kernel $K(\cdot)$, two lists of bandwidths, $\boldsymbol{h}_m := \{h_{m1}, \ldots, h_{mb}\}$ and $\boldsymbol{h}_\pi := \{h_{\pi 1}, \ldots, h_{\pi b}\}$, tuning parameters $\kappa > 0$ and $\underline{q} \in (0, 1]$, and confidence level $1 - \alpha$.

Output: Dynamic ATT estimates $\widehat{\mathrm{ATT}(g,t)}$, stand errors $\widehat{\mathrm{se}}_t$, $1-\alpha$ confidence interval.

Step 1: Sample-splitting Randomly partition samples into disjoint $\{\mathcal{I}_1, \mathcal{I}_2, \mathcal{I}_3\}$.

Step 2: Calculate similarity For each $i, j = 1, \ldots, n$, calculate the pseudo-distance:

$$\widehat{d}_{ij} = \max_{k_1,k_2=1,\ldots,n,k_1,k_2\neq i,j} \left| \frac{1}{T_0} \sum_{t=0}^{T_0-1} (Y_{k_1 t} - Y_{k_2 t})(Y_{it} - Y_{jt}) \right|.$$

Step 3: Compute expected outcomes and propensity scores

Given $h > 0$, for each i, j and each t', define $\widehat{K}^{(t')}_{h,ij} := K\left(\left(X_{jt'} - X_{it'}\right)/h\right) K\left(\widehat{d}_{ij}/h\right)$.

Initialization. For each $j = 1, \ldots, n$, set $\widetilde{m}_{j,t|t+1}(h) := Y_{jt}$ for any $h > 0$,

for $t' = t, t-1, \ldots, g$ **do**

Cross-validation For each $j = 1, \ldots, n$, each $h_m \in \boldsymbol{h}_m$, and each $h_\pi \in \boldsymbol{h}_\pi$, compute the leave-one-out within-sample estimators and effective sample sizes

$$\widetilde{\pi}^{(-i)}_{it'}(h_\pi) = \frac{\sum_{i' \in \mathcal{I}(i)\setminus\{i\}} \mathbf{1}\left(G_{i'} = t'\right) \widehat{K}^{(t')}_{h_\pi, ii'}}{\sum_{i' \in \mathcal{I}(i)\setminus\{i\}} \mathbf{1}\left(G_{i'} > t' - 1\right) \widehat{K}^{(t')}_{h_\pi, ii'}}, \quad n_{it',\mathrm{eff}}(h_m) = \frac{\left(\sum_{i' \in \mathcal{I}(i)\setminus\{i\}} \mathbf{1}(G_{i'} > t') \widehat{K}^{(t')}_{h_m, ii'}\right)^2}{\sum_{i' \in \mathcal{I}(i)\setminus\{i\}} \mathbf{1}(G_{i'} > t') \widehat{K}^{(t')2}_{h_m, ii'}}.$$

Then, $h^*_\pi(t') \in \mathrm{argmin}_{h_\pi \in \boldsymbol{h}_\pi} \sum_{i=1}^n \mathbf{1}(G_i > t' - 1)(\mathbf{1}(G_i = t') - \widetilde{\pi}^{(-i)}_{it'}(h_\pi))^2$ and $h^*_m(t') = \min\left\{h_m \in \boldsymbol{h}_\pi : \frac{1}{\sum_{i=1}^n \mathbf{1}(G_i > t'-1)} \sum_{i=1}^n \mathbf{1}\left(G_i > t' - 1, n_{it',\mathrm{eff}}(h_m) > (\kappa \log(n))^2\right) \geqslant \underline{q}\right\}$.

Calculate $\widehat{m}_{i,t|t'}$ and $\widehat{\pi}_{it'}$ For each $i, j = 1, \ldots, n$, calculate $\widehat{m}_{i,t|t'}$, $\widehat{\pi}_{it'}$, and $\widetilde{m}_{j,t|t'}$:

$$\widehat{m}_{i,t|t'} = \frac{\sum_{j \in \mathcal{I}_m(i)} \mathbf{1}\left(G_j > t'\right) \widetilde{m}_{j,t|t'+1} \widehat{K}^{(t')}_{h^*_m(t'),ij}}{\sum_{j \in \mathcal{I}_m(i)} \mathbf{1}\left(G_j > t'\right) \widehat{K}^{(t')}_{h^*_m(t'),ij}}, \quad \widehat{\pi}_{it'} = \frac{\sum_{j \in \mathcal{I}_\pi(i)} \mathbf{1}\left(G_j = t'\right) \widehat{K}^{(t')}_{h^*_\pi(t'),ij}}{\sum_{j \in \mathcal{I}_\pi(i)} \mathbf{1}\left(G_j > t' - 1\right) \widehat{K}^{(t')}_{h^*_\pi(t'),ij}},$$

$$\widetilde{m}_{j,t|t'} = \sum_{j' \in \mathcal{I}(j)} \mathbf{1}\left(G_{j'} > t'\right) \widetilde{m}_{j',t|t'+1} \widehat{K}^{(t')}_{h^*_m(t'),jj'} \Big/ \sum_{j' \in \mathcal{I}(j)} \mathbf{1}\left(G_{j'} > t'\right) \widehat{K}^{(t')}_{h^*_m(t'),jj'}.$$

end for

Step 3: Construct $\widehat{\mathrm{ATT}(g,t)}$, $\widehat{\mathrm{se}}_t$, and CI. $\widehat{ATT(g,t)}$ is obtained via

$$\frac{1}{n_g} \sum_{i=1}^n \left[\mathbf{1}\left(G_i = g\right) \left(Y_{it} - \widehat{m}_{i,t|g}\right) - \sum_{t'=g}^{t} \mathbf{1}\left(G_i > t'\right) \frac{\widehat{\pi}_{ig}}{1 - \widehat{\pi}_{ig}} \left(\prod_{r=g+1}^{t'} \frac{1}{1 - \widehat{\pi}_{ir}} \right) \left(\widehat{m}_{i,t|t'+1} - \widehat{m}_{i,t|t'}\right) \right],$$

where $n_g := \sum_{i=1}^n \mathbf{1}(G_i = g)$, and an empty product is defined as one. $\widehat{\mathrm{se}}_t$ is

$$\widehat{\mathrm{se}}_t = \left(\frac{1}{n_g^2} \sum_{i=1}^n \left[\mathbf{1}\left(G_i = g\right) \left(Y_{it} - \widehat{m}_{i,t|g} - \widehat{ATT(g,t)}\right) \right.\right.$$
$$\left.\left. - \sum_{t'=g}^{t} \mathbf{1}\left(G_i > t'\right) \frac{\widehat{\pi}_{ig}}{1 - \widehat{\pi}_{ig}} \left(\prod_{r=g+1}^{t'} \frac{1}{1 - \widehat{\pi}_{ir}} \right) \left(\widehat{m}_{i,t|t'+1} - \widehat{m}_{i,t|t'}\right) \right]^2 \right)^{1/2}.$$

The confidence interval is $\mathrm{CI}_t = [\widehat{\mathrm{ATT}(t)} \pm Z_{1-\alpha/2} \cdot \widehat{\mathrm{se}}_t]$.

B Discussion on Pseudo Distance and Informativeness Condition

In this section, I (i) discuss the relationships among the population pseudo-distance $d(\alpha_i, \alpha_j)$, the sample squared L_2 distance $\widehat{d}^2_{2,ij}$, and the population squared L_2 distance $d^2_2(\alpha_i, \alpha_j)$, as well as the sufficient conditions for the informativeness used in identification (Assumption 4.2), (ii) give the definition of $d(\alpha_i, \alpha_j)$ when $\{\gamma_t\}_{t\in\mathbb{Z}}$ is neither stationary nor ergodic over time, (iii) provide a set of primitive sufficient conditions for Assumption 5.1, and (iv) discuss the informativeness condition used in estimation (Assumption 5.2) and verify it under Example 2 and 3.

B.1 Pseudo distance and L_2 distance

Recall the definition of the sample squared L_2 distance $\widehat{d}^2_{2,ij}$:

$$\widehat{d}^2_{2,ij} = \frac{1}{T_0}\sum_{t=0}^{T_0-1}(Y_{it} - Y_{jt})^2.$$

Under the regularity conditions in Assumption 4.1, define the population squared L_2 distance $d^2_2(\alpha_i, \alpha_j)$ as

$$d^2_2(\alpha_i, \alpha_j) = \int \left(g(\alpha_i, \Gamma) - g(\alpha_j, \Gamma_t)\right)^2 \mathrm{d}\mathbb{P}(\Gamma).$$

Although the population squared L_2 distance $d^2_2(\alpha_i, \alpha_j)$ serves as a natural measure of the distance between α_i and α_j in the function space, it is infeasible to estimate directly from the sample squared L_2 distance $\widehat{d}^2_{2,ij}$. To see this, note that under the regularity conditions in Assumption 4.1,

$$\widehat{d}^2_{2,ij} = d^2_2(\alpha_i, \alpha_j) + \mathbb{E}_T\left(\widetilde{\epsilon}^2_{it} \mid \alpha_i\right) + \mathbb{E}_T\left(\widetilde{\epsilon}^2_{jt} \mid \alpha_j\right) + o_P(1).$$

The result follows from an argument similar to that used to establish the consistency of $\widehat{d}_{ij}$ in Appendix C, and the proof is therefore omitted. The additional variance terms, $\mathbb{E}_T\left(\widetilde{\epsilon}^2_{it} \mid \alpha_i\right)$ and $\mathbb{E}_T\left(\widetilde{\epsilon}^2_{jt} \mid \alpha_j\right)$, do not vanish asymptotically and therefore contaminate the estimation. In general, these conditional variance terms depend on latent characteristics and cannot be removed, and only under homoskedasticity can one recover $d^2_2(\alpha_i, \alpha_j)$ up to an additive constant.

By definition of pseudo distance, it is straightforward for verify that the population

pseudo distance provides an upper bound for $d_2^2(\alpha_i, \alpha_j)$:

$$d(\alpha_i, \alpha_j) := \sup_{\alpha_1, \alpha_2 \in \mathcal{A}} \left| \int (g(\alpha_1, \Gamma) - g(\alpha_2, \Gamma))(g(\alpha_i, \Gamma) - g(\alpha_j, \Gamma)) \mathrm{d}\mathbb{P}(\Gamma) \right| \geqslant d_2^2(\alpha_i, \alpha_j).$$

This relationship is useful for proving the following lemma:

Lemma B.1 (Sufficient conditions for Assumption 4.2) *Under Assumption 4.1, suppose further that for any $\alpha_1 \neq \alpha_2$, $g(\alpha_1, \Gamma)$ and $g(\alpha_2, \Gamma)$ are not almost surely identical. Then Assumption 4.2 holds.*

Proof of Lemma B.1. For each pair $\alpha_1, \alpha_2 \in \mathcal{A}$ such that $\alpha_1 \neq \alpha_2$, since $g(\alpha_1, \Gamma)$ and $g(\alpha_2, \Gamma)$ are not almost surely identical, we have

$$d(\alpha_1, \alpha_2) \geqslant d_2^2(\alpha_1, \alpha_2) > 0.$$

By the continuity of $g(\cdot, \cdot)$ and envelope condition (see Assumption 4.1(iv) and (v)), $d(\cdot, \cdot)$ is also continuous. In addition, for any $\epsilon > 0$, since the set $\{(\alpha_1, \alpha_2) \in \mathcal{A}^2 \mid \|\alpha_1 - \alpha_2\| \geqslant \epsilon\}$ is compact. It follows that for any $\epsilon > 0$, there exists a constant $\delta_\epsilon > 0$ such that

$$\inf_{(\alpha_1, \alpha_2) \in \mathcal{A}^2, \|\alpha_1 - \alpha_2\| \geqslant \epsilon} d(\alpha_1, \alpha_2) \geqslant \delta_\epsilon.$$

It immediately implies that for any $\epsilon > 0$, as long as $d(\alpha_1, \alpha_2) \leqslant \delta_\epsilon/2$, $\|\alpha_1 - \alpha_2\| \leqslant \epsilon$. This completes the proof. ■

B.2 Pseudo distance without stationarity and ergodicity

When stationarity and ergodicity are not imposed, define $d(\alpha_i, \alpha_j)$ as

$$d(\alpha_i, \alpha_j) := \sup_{\alpha_1, \alpha_2 \in \mathcal{A}} \left| \lim_{T_0 \to \infty} \frac{1}{T_0} \sum_{t=1}^{T_0} \mathbb{E}\left[(g(\alpha_1, \Gamma_t) - g(\alpha_2, \Gamma_t))(g(\alpha_i, \Gamma_t) - g(\alpha_j, \Gamma_t)) \right] \right| \tag{B.1}$$

It is straightforward to see that, if the limit in (B.1) exists, then $d(\alpha_i, \alpha_j)$ does not depend on the realization of $\{\gamma_t\}_{t \in \mathbb{Z}}$. The existence of the limit in (B.1) requires restrictions on the limiting distribution of γ_t, and a discussion of such restrictions is therefore omitted.

In addition, it is straightforward to verify that, under Assumption 4.1(ii)-(v) and the following conditions: (i) $\{\gamma_t\}$ is weakly dependent, and (ii) information from the distant past

becomes asymptotically negligible, i.e.,

$$\sup_{i,t} |g(\alpha_i, \Gamma_t) - \mathbb{E}(g(\alpha_i, \Gamma_t) \mid \gamma_t, \ldots, \gamma_{t-m})| \to 0, \quad \text{as } m \to \infty,$$

whenever the limit in (B.1) exists, we have $\widehat{d}_{ij} \xrightarrow{p} d(\alpha_i, \alpha_j)$.

B.3 Sufficient conditions for Assumption 5.1

I introduce the following assumptions:

Assumption B.1 *I assume that*

(i) ***(Mixing)*** *Conditional on $\{\gamma_t\}_{t\in\mathbb{Z}}$, the panel $\{(Y_{it}, X_{it}, D_{it}, \alpha_i)\}_{i=1,\ldots,n,t=1,\ldots,T}$ is i.i.d. across i. In addition, For any $\alpha \in \mathcal{A}$ and any $\Gamma \in \text{supp}(\Gamma_T)$, conditional on $(\alpha_i, \Gamma_T) = (\alpha, \Gamma)$, $\{Y_{it}(\infty)\}_{t\in\mathbb{Z}}$ are α-mixing with mixing coefficient $\phi_i(\tau; \alpha, \Gamma) \to 0$, where*

$$\phi_i(\tau; \alpha, \Gamma) = \sup_{t\in\mathbb{Z}} \sup_{A\in\mathcal{A}_{it}, B\in\mathcal{B}_{it+\tau}} \Big|\mathbb{P}(B \cap A \mid \alpha_i = \alpha, \Gamma_T = \Gamma) - \mathbb{P}(A \mid \alpha_i = \alpha, \Gamma_T = \Gamma)\mathbb{P}_T(B \mid \alpha_i = \alpha, \Gamma_T = \Gamma)\Big|.$$

Here, $\mathcal{A}_{it}$ is the sigma-field generated by $\{Y_{is}(\infty)\}_{s=-\infty}^{t-1}$, and $\mathcal{B}_{it+\tau}$ is the sigma-field generated by $\{Y_{it+\tau}(\infty), \ldots, Y_{iT}(\infty)\}$. $\phi_i(\tau; \alpha, \Gamma)$ exhibits a uniformly exponential decay rate: there exists $\zeta_0 > 0$ such that $\max_{i=1,\ldots,n} \phi(\tau; \alpha, \Gamma) \leqslant e^{-\zeta_0 \tau}$ uniformly over $\alpha \in \mathcal{A}$ and $\Gamma \in \text{supp}(\Gamma_T)$.

(ii) ***(Convergence)*** *$\{\gamma_t\}_{t\in\mathbb{Z}}$ is stationary and ergodic over time. In addition, there exists a constant $C > 0$ such that*

$$\sup_{\alpha_1,\alpha_2,\alpha_i,\alpha_j\in\mathcal{A}} \Big| \frac{1}{T_0} \sum_{t=0}^{T_0-1} (g(\alpha_1, \Gamma_t) - g(\alpha_2, \Gamma_t))(g(\alpha_i, \Gamma_t) - g(\alpha_j, \Gamma_t)) - \int (g(\alpha_1, \Gamma) - g(\alpha_2, \Gamma))(g(\alpha_i, \Gamma) - g(\alpha_j, \Gamma)) \mathrm{d}\mathbb{P}(\Gamma) \Big| \leqslant C \sqrt{\frac{\log(T_0)}{T_0}}$$

with probability approaching to 1.

(iii) ***(Boundedness)*** *$g(\cdot, \cdot)$ is uniformly bounded on the relavent supports. In addition, for some $\sigma > 0$, $\mathbb{E}\left(\exp\left(\lambda(Y_{it}(\infty) - g(\alpha_i, \Gamma_t))\right) \mid \alpha_i, \Gamma_T\right) \leqslant \exp(\sigma^2\lambda^2/2)$ a.s. for all $\lambda \in \mathbb{R}$, all i and t.*

(iv) ***(Compactness)*** *The support of α, $\mathcal{A}$, is compact. There exist constants $\underline{c}_1, \bar{c}_1 > 0$, such that for any $\alpha \in \mathcal{A}$ and any $r \in (0, 1)$, $\underline{c}_1 r^{d_\alpha} \leqslant \mathbb{P}\left(\|\alpha_i - \alpha\| \leqslant r\right) \leqslant \bar{c}_1 r^{d_\alpha}$.*

(v) ***(Lipschitz)*** *There exists a constant L_g such that for any $\alpha_1, \alpha_2 \in \mathcal{A}$,*

$$\sup_{\Gamma \in \text{supp}(\Gamma_T)} |g(\alpha_1, \Gamma) - g(\alpha_2, \Gamma)| \leqslant L_g \|\alpha_1 - \alpha_2\|_\infty.$$

Assumption B.1(i) strengthens the weak-dependence condition in Assumption 6.3(ii) by imposing conditional α-mixing with mixing coefficients that decay uniformly at an exponential rate. Assumption B.1(ii) strengthens the ergodic convergence theorem by imposing an explicit uniform convergence rate. It is satisfied when the influence of distant time factors on $g(\alpha, \Gamma_t)$ decays sufficiently fast. Assumption B.1(v) imposes Lipschitz continuity, which is stronger than the uniform continuity condition in Assumption 4.1(v).

Lemma B.2 (Sufficient conditions for Assumption 5.1) *Assumption B.1 implies Assumption 5.1.*

B.4 Discussion on Assumption 5.2

Verifying the lower bound $\underline{\eta} d(\alpha_1, \alpha_2) \leqslant \|\alpha_1 - \alpha_2\|$ is straightforward. When g is Lipschitz as in Assumption B.1(v),

$$\begin{aligned} d(\alpha_i, \alpha_j) &:= \sup_{\alpha_1, \alpha_2 \in \mathcal{A}} \left| \int (g(\alpha_1, \Gamma) - g(\alpha_2, \Gamma))(g(\alpha_i, \Gamma) - g(\alpha_j, \Gamma)) \mathrm{d}\mathbb{P}(\Gamma) \right| \\ &\overset{\text{(i)}}{\leqslant} \sup_{\alpha_1, \alpha_2 \in \mathcal{A}} \left(\int (g(\alpha_1, \Gamma) - g(\alpha_2, \Gamma))^2 \mathrm{d}\mathbb{P}(\Gamma) \right)^{1/2} L_g \sqrt{d_\alpha} \|\alpha_i - \alpha_j\|, \end{aligned}$$

where (i) follows from the Cauchy-Schwarz inequality and the Lipschitz continuity of g. Thus, the lower bound holds if the second moment of g is uniformly bounded over $\mathcal{A}$, i.e.,

$$\sup_{\alpha \in \mathcal{A}} \int g^2(\alpha, \Gamma) \mathrm{d}\mathbb{P}(\Gamma) < \infty.$$

To verify the upper bound, I introduce the following non-redundancy condition:

Definition B.1 (Nonredundancy) *We say that $\mathcal{A} \subseteq \mathbb{R}^d$ is* effective *if there does not exist a nontrivial partition of coordinates $\mathcal{I}_1, \mathcal{I}_2 \subseteq \{1, \ldots, d\}$ with $\mathcal{I}_1 \cap \mathcal{I}_2 = \varnothing$, $\mathcal{I}_2 \neq \varnothing$, and a measurable function $q : \mathbb{R}^{|\mathcal{I}_1|} \to \mathbb{R}^{|\mathcal{I}_2|}$ such that*

$$\alpha_{\mathcal{I}_2} = q(\alpha_{\mathcal{I}_1}), \quad \forall \alpha \in \mathcal{A}.$$

That is, no subset of coordinates of α can be written as a deterministic function of the remaining coordinates on $\mathcal{A}$, or equivalently, $\mathcal{A}$ is not contained in a lower-dimensional

manifold. By definition, nonredundancy immediately implies that affine$(\mathcal{A}) = \mathbb{R}^{d_\alpha}$.

Lemma B.3 (Linear dynamic panel with interactive fixed effects) *Suppose*

$$Y_{it}(\infty) = \rho Y_{it-1}(\infty) + \gamma_t'\alpha_i + \epsilon_{it},$$
$$\mathbb{E}\left(\epsilon_{it} \mid Y_{it-1}, \alpha_i, \Gamma_T\right) = 0, \quad \mathbb{E}\left(\epsilon_{it}^2 \mid Y_{it-1}, \alpha_i, \Gamma_T\right) < \infty.$$

Then, the informativeness condition (Assumption 5.2) holds if (i) $\mathcal{A}$ is nonredundant, compact, and convex, (ii) $|\rho| < 1$, (iii) γ_t is ergodic and stationary, and (iv) $0 < s_{\min}(\mathbb{E}(\gamma_t\gamma_t')) \leqslant s_{\max}(\mathbb{E}(\gamma_t\gamma_t')) < \infty$, where $s_{\min}(\cdot)$ and $s_{\max}(\cdot)$ denote the smallest and largest eigenvalue, respectively.

The condition that γ_t is ergodic and stationary seems restrictive, but it is imposed to simplify the technical discussion. It is straightforward to show that, Lemma B.3 still holds if I replace conditions (iii) and (iv) with

$$\frac{1}{T_0}\sum_{t=0}^{T_0-1}\gamma_t\gamma_t' \text{ exists, and } 0 < s_{\min}\left(\mathbb{E}\left(\frac{1}{T_0}\sum_{t=0}^{T_0-1}\gamma_t\gamma_t'\right)\right) \leqslant s_{\max}\left(\mathbb{E}\left(\frac{1}{T_0}\sum_{t=0}^{T_0-1}\gamma_t\gamma_t'\right)\right) < \infty.$$

For nonlinear dynamic panel models discussed in Example 3, recall that the outcome is generated according to

$$Y_{it}(\infty) = \mathbf{1}\left(\beta Y_{it-1}(\infty) + \gamma_t'\alpha_i - u_{it} \geqslant 0\right),$$

and let $F(\cdot)$ denote the cumulative distribution function of u_{it}. Still, we assume that u_{it} is independent of past histories conditional on all fixed effects. Note that the outcome process is also written as

$$Y_{it}(\infty) = F\left(\beta Y_{it-1}(\infty) + \gamma_t'\alpha_i\right) + \epsilon_{it}.$$

For notational simplicity, define $g_t(\alpha_i) := g(\alpha_i, \Gamma_t) = \mathbb{E}\left(Y_{it}(\infty) \mid \alpha_i, \Gamma_t\right)$. In addition, let $\rho_t(\alpha_i) := F\left(\beta + \gamma_t'\alpha_i\right) - F\left(\gamma_t'\alpha_i\right)$ denote the stochastic discount factor. Then,

$$g_t(\alpha_i) = (1 - g_{t-1}(\alpha_i))F\left(\gamma_t'\alpha_i\right) + g_{t-1}(\alpha_i)F\left(\beta + \gamma_t'\alpha_i\right) = F\left(\gamma_t'\alpha_i\right) + \rho_t(\alpha_i)g_{t-1}(\alpha_i).$$

Let $f_t := \mathbb{E}\left(Y_{it} \mid \alpha_i, \Gamma\right)$. Let $\rho(\gamma_t'\alpha_i)$ denote $\Gamma\left(1+\gamma_t'\alpha_i\right) - \Gamma\left(\gamma_t'\alpha_i\right)$. Then,

$$\begin{aligned} g_t(\alpha_i) =& F\left(\gamma_t'\alpha_i\right) + \rho_t(\alpha_i) g_{t-1}(\alpha_i) \\ =& F\left(\gamma_t'\alpha_i\right) + \rho_t(\alpha_i)\left(F\left(\gamma_{t-1}'\alpha_i\right) + \rho_{t-1}(\alpha_i) g_{t-2}(\alpha_i)\right) \\ =& F\left(\gamma_t'\alpha_i\right) + \rho_t(\alpha_i) F\left(\gamma_{t-1}'\alpha_i\right) + \rho_t(\alpha_i)\rho_{t-1}(\alpha_i) F_{t-2}(\alpha_i) \\ =& \sum_{\tau=0}^{\infty}\left(\prod_{s=0}^{\tau-1}\rho_{t-s}(\alpha_i)\right) F\left(\gamma_{t-\tau}'\alpha_i\right) \end{aligned} \tag{B.2}$$

Lemma B.4 (Nonlinear dynamic panel with interactive fixed effects) *Suppose*

$$Y_{it}(\infty) = \mathbf{1}\left(\beta Y_{it-1}(\infty) + \gamma_t'\alpha_i - u_{it} \geqslant 0\right).$$

The error terms $\{u_{it}\}_{i=1,\ldots,n,\, t\in\mathbb{Z}}$ *are independent of* $\{Y_{i\tau}(\infty)\}_{i=1,\ldots,n,\, \tau\leqslant t-1}$ *and the fixed effects. In addition, they are i.i.d. across time and units. Let* $F(\cdot)$ *denote the cumulative distribution function of* u_{it}*.*

Then, the informativeness condition (Assumption 5.2) holds if (i) $F(\cdot)$ *is twice continuously differentiable and* $F^{(1)}(\cdot)$ *on the index space; (ii)* $\mathcal{A}$ *is nonredundant, compact, and convex; (iii)* $\{\gamma_t\}_{t\in\mathbb{Z}}$ *are i.i.d. over time, with bounded* $\operatorname{supp}(\gamma_t)$*; and (iv)* $s_{\min}\left(\mathbb{E}(\gamma_t\gamma_t')\right) > 0$*, where* $s_{\min}(\cdot)$ *denotes the smallest eigenvalue.*

C Proofs of Identification

We provide only the proof of Theorem 3, as it extends Theorem 1 by incorporating staggered adoptions.

Proof of Theorem 3. We first prove the identification of the population pseudo-distance in large-n, large-T_0 panel data by showing that the sample pseudo-distance $\widehat{d}_{ij}$ converges in probability to $d(\alpha_i, \alpha_j)$. We then establish the identification of the contemporaneous ATT. Finally, we show how to identify the dynamic ATT using a recursive argument.

Identification of $d(\alpha_i, \alpha_j)$ For notationally simplicity, for any $\alpha_1, \alpha_2 \in \mathcal{A}$ and any $\Gamma \in \text{supp}(\Gamma_T)$, define $\Delta(\alpha_1, \alpha_2, \Gamma) := (g(\alpha_1, \Gamma) - g(\alpha_2, \Gamma))$. Note that for any $i, j \in 1, \ldots, n$,

$$
\begin{aligned}
&\left|\widehat{d}_{ij} - d(\alpha_i, \alpha_j)\right| \\
\leqslant &\underbrace{\max_{k_1,k_2 \in \{1,\ldots,n\}\setminus\{i,j\}} \left| \frac{1}{T_0} \sum_{t=0}^{T_0-1} \Delta(\alpha_{k_1}, \alpha_{k_2}, \Gamma_t)\Delta(\alpha_i, \alpha_j, \Gamma_t) - \mathbb{E}\left(\Delta(\alpha_{k_1}, \alpha_{k_2}, \Gamma)\Delta(\alpha_i, \alpha_j, \Gamma) \mid \alpha_{k_1}, \alpha_{k_2}, \alpha_i, \alpha_j\right) \right|}_{A_1} \\
&+ \underbrace{\left| \max_{k_1,k_2 \in \{1,\ldots,n\}\setminus\{i,j\}} \mathbb{E}\left(\Delta(\alpha_{k_1}, \alpha_{k_2}, \Gamma)\Delta(\alpha_i, \alpha_j, \Gamma) \mid \alpha_{k_1}, \alpha_{k_2}, \alpha_i, \alpha_j\right) - d(\alpha_i, \alpha_j) \right|}_{A_2} \\
&+ \underbrace{\max_{k_1,k_2 \in \{1,\ldots,n\}\setminus\{i,j\}} \left| \frac{1}{T_0} \sum_{t=0}^{T_0-1} \Delta(\alpha_i, \alpha_j, \Gamma_t)(\widetilde{\epsilon}_{k_1 t} - \widetilde{\epsilon}_{k_2 t}) \right|}_{A_3} + \underbrace{\max_{k_1,k_2 \in \{1,\ldots,n\}\setminus\{i,j\}} \left| \frac{1}{T_0} \sum_{t=0}^{T_0-1} \Delta(\alpha_{k_1}, \alpha_{k_2}, \Gamma_t)(\widetilde{\epsilon}_{it} - \widetilde{\epsilon}_{jt}) \right|}_{A_4} \\
&+ \underbrace{\max_{k_1,k_2 \in \{1,\ldots,n\}\setminus\{i,j\}} \left| \frac{1}{T_0} \sum_{t=0}^{T_0-1} (\widetilde{\epsilon}_{it} - \widetilde{\epsilon}_{jt})(\widetilde{\epsilon}_{k_1 t} - \widetilde{\epsilon}_{k_2 t}) \right|}_{A_5}
\end{aligned}
$$

For the first term A_1, since (i) $\mathcal{A}$ is compact, (ii) fixed $\alpha_1, \alpha_2 \in \mathcal{A}$, $\Delta(\alpha_1, \alpha_2, \Gamma_t)$ is stationary and ergodic over time by Assumption 4.1(i), (iii) $\Delta(\alpha_1, \alpha_2, \Gamma_t)$ is continuous in (α_1, α_2) for all Γ_t by Assumption 4.1(v), and (iv) the envelope condition in Assumption 4.1(iv),

$$
\mathbb{E}\left(\sup_{\alpha_1,\alpha_2} |\Delta(\alpha_{k_1}, \alpha_{k_2}, \Gamma)\Delta(\alpha_i, \alpha_j, \Gamma)| \right) \leqslant 4\mathbb{E}\left(\sup_{\alpha} |g(\alpha, \Gamma)|^2 \right) < \infty,
$$

we employ the uniform law of large numbers of ergodic and stationary process (Hayashi (2011, Lemma 7.2)) to show that $A_1 = o_P(1)$.

For the second term, since $\mathbb{E}\left(\Delta(\alpha_{k_1}, \alpha_{k_2}, \Gamma)\Delta(\alpha_i, \alpha_j, \Gamma) \mid \alpha_{k_1}, \alpha_{k_2}, \alpha_i, \alpha_j\right)$ is continuous in $(\alpha_{k_1}, \alpha_{k_2})$ when fixing (α_i, α_j) (see Hayashi (2011, Lemma 7.2)), then, by the compactness

of $\mathcal{A}$ in Assumption 4.1(ii), it immediately follows that

$$\max_{k_1,k_2\in\{1,\ldots,n\}\setminus\{i,j\}} \mathbb{E}\left(\Delta(\alpha_{k_1},\alpha_{k_2},\Gamma)\Delta(\alpha_i,\alpha_j,\Gamma) \mid \alpha_{k_1},\alpha_{k_2},\alpha_i,\alpha_j\right) \xrightarrow{p} \sup_{\alpha_1,\alpha_2\in\mathcal{A}} \int \Delta(\alpha_1,\alpha_2,\Gamma)\Delta(\alpha_i,\alpha_j,\Gamma)\mathrm{d}\mathbb{P}(\Gamma).$$

In other words, the sample maximum over $(\alpha_{k_1},\alpha_{k_2})$ converges in probability to the population supremum over $(\alpha_1,\alpha_2)\in\mathcal{A}^2$. Therefore, $A_2 = o_P(1)$.

For A_3, Since $\mathbb{E}\left(\widetilde{\epsilon}_{k_1t} \mid \{\alpha_i\}_{i=1}^n,\Gamma_T\right) = \mathbb{E}\left(\widetilde{\epsilon}_{k_2t} \mid \{\alpha_i\}_{i=1}^n,\Gamma_T\right) = 0$, we have

$$\begin{aligned}
&\mathrm{Var}\left(\frac{1}{T_0}\sum_{t=0}^{T_0-1}\Delta(\alpha_i,\alpha_j,\Gamma_t)(\widetilde{\epsilon}_{k_1t}-\widetilde{\epsilon}_{k_2t}) \mid \{\alpha_i\}_{i=1}^n,\Gamma_T\right)\\
=&\frac{1}{T_0^2}\sum_{s,t=1,\ldots,T_0}\Delta(\alpha_i,\alpha_j,\Gamma_s)\Delta(\alpha_i,\alpha_j,\Gamma_t)\mathbb{E}\left(\widetilde{\epsilon}_{k_1t}\widetilde{\epsilon}_{k_1s}+\widetilde{\epsilon}_{k_2t}\widetilde{\epsilon}_{k_2s} \mid \{\alpha_i\}_{i=1}^n,\Gamma_T\right)\\
\leqslant&\frac{1}{T_0^2}\sum_{s,t=1,\ldots,T_0}\left|\Delta(\alpha_i,\alpha_j,\Gamma_s)\Delta(\alpha_i,\alpha_j,\Gamma_t)\right|\left|\mathbb{E}\left(\widetilde{\epsilon}_{k_1t}\widetilde{\epsilon}_{k_1s}+\widetilde{\epsilon}_{k_2t}\widetilde{\epsilon}_{k_2s} \mid \{\alpha_i\}_{i=1}^n,\Gamma_T\right)\right|\\
\leqslant&\left(2\max_{t=1,\ldots,T_0}\Delta^2(\alpha_i,\alpha_j,\Gamma_t)\right)\frac{1}{T_0^2}\sum_{t=1}^{T_0}\sum_{s=1}^{T_0}\left(\left|\mathbb{E}\left(\widetilde{\epsilon}_{k_1t}\widetilde{\epsilon}_{k_1s} \mid \alpha_{k_1},\Gamma_T\right)\right|+\left|\mathbb{E}\left(\widetilde{\epsilon}_{k_2t}\widetilde{\epsilon}_{k_2s} \mid \alpha_{k_2},\Gamma_T\right)\right|\right),
\end{aligned}$$

The weak dependence condition in Assumption 6.3(ii) requires that

$$\frac{1}{T_0}\sum_{t=1}^{T_0}\sum_{s=1}^{T_0}\left|\mathbb{E}\left(\widetilde{\epsilon}_{k_1t}\widetilde{\epsilon}_{k_1s} \mid \alpha_{k_1},\Gamma_T\right)\right|+\left|\mathbb{E}\left(\widetilde{\epsilon}_{k_2t}\widetilde{\epsilon}_{k_2s} \mid \alpha_{k_2},\Gamma_T\right)\right| < \infty$$

In addition, it is straightforward to verify that, by moment restriction in Assumption 4.1(iii),

$$\max_{t=1,\ldots,T_0}\Delta^2(\alpha_i,\alpha_j,\Gamma_t) <\leqslant \max_{t=1,\ldots,T_0}\mathbb{E}\left(Y_{it}^2(\infty) \mid \alpha_i,\Gamma_T\right) < \infty. \tag{C.1}$$

Then, it follows that there exists a constant C_1 (irrelevant of the realization of $(\{\alpha_i\}_{i=1}^n,\Gamma_T)$) such that

$$\mathrm{Var}\left(\frac{1}{T_0}\sum_{t=0}^{T_0-1}\Delta(\alpha_i,\alpha_j,\Gamma_t)(\widetilde{\epsilon}_{k_1t}-\widetilde{\epsilon}_{k_2t}) \mid \{\alpha_i\}_{i=1}^n,\Gamma_T\right) \leqslant \frac{C_1}{T_0}$$

Then, by Markov inequality, for any $\delta > 0$,

$$\begin{aligned}\mathbb{P}\left(|A_3| \geqslant \delta \mid \{\alpha_i\}_{i=1}^n, \Gamma_T\right) =& \mathbb{P}\left(\bigcup_{k_1,k_2}\left\{\left|\frac{1}{T_0}\sum_{t=0}^{T_0-1}\Delta(\alpha_i,\alpha_j,\Gamma_t)(\widetilde{\epsilon}_{k_1t}-\widetilde{\epsilon}_{k_2t})\right| \geqslant \delta\right\} \mid \{\alpha_i\}_{i=1}^n, \Gamma_T\right)\\ \leqslant& \sum_{k_1,k_2=1,\ldots,n}\mathbb{P}\left(\left|\frac{1}{T_0}\sum_{t=0}^{T_0-1}\Delta(\alpha_i,\alpha_j,\Gamma_t)(\widetilde{\epsilon}_{k_1t}-\widetilde{\epsilon}_{k_2t})\right| \geqslant \delta \mid \{\alpha_i\}_{i=1}^n, \Gamma_T\right)\\ \leqslant& n^2\frac{C_1}{T_0\delta^2}.\end{aligned}$$

Since the inequalities above does not depend on the realization of Γ_T, we integrate them out to obtain

$$\mathbb{P}\left(|A_3| \geqslant \delta\right) \leqslant n^2\frac{C_1}{T_0\delta^2}. \tag{C.2}$$

The bound above implies that $A_3 = o_P(1)$ whenever $n^2/T_0 \to 0$. This relative-rate restriction, however, is not essential for identification. Indeed, even when $n^2/T_0 \nrightarrow 0$, it suffices to construct the sample maximum using a subsample of n_{T_0} individuals such that $n_{T_0} \to \infty$ and $n_{T_0} = o(\sqrt{T_0})$. The sample maximum over this subsample still converges to the population supremum, which implies that $A_3 = o_P(1)$. Applying the same argument on A_4 yields $A_4 = o_P(1)$.

For the last term A_5, note that $\mathbb{E}\left((\widetilde{\epsilon}_{it}-\widetilde{\epsilon}_{jt})(\widetilde{\epsilon}_{k_1t}-\widetilde{\epsilon}_{k_2t}) \mid \{\alpha_i\}_{i=1}^n, \Gamma_T\right) = 0$, and

$$\begin{aligned}&\mathrm{Var}\left(\frac{1}{T_0}\sum_{t=0}^{T_0-1}(\widetilde{\epsilon}_{it}-\widetilde{\epsilon}_{jt})(\widetilde{\epsilon}_{k_1t}-\widetilde{\epsilon}_{k_2t}) \mid \{\alpha_i\}_{i=1}^n, \Gamma_T\right)\\ =&\frac{1}{T_0^2}\sum_{s,t=1,\ldots,T_0}\mathbb{E}\left(\widetilde{\epsilon}_{it}\widetilde{\epsilon}_{is}+\widetilde{\epsilon}_{jt}\widetilde{\epsilon}_{js} \mid \{\alpha_i\}_{i=1}^n, \Gamma_T\right)\mathbb{E}\left(\widetilde{\epsilon}_{k_1t}\widetilde{\epsilon}_{k_1s}+\widetilde{\epsilon}_{k_2t}\widetilde{\epsilon}_{k_2s} \mid \{\alpha_i\}_{i=1}^n, \Gamma_T\right)\\ \leqslant& \max_{t=1\ldots,T_0}\left(\mathbb{E}\left(\widetilde{\epsilon}_{it}\widetilde{\epsilon}_{is}+\widetilde{\epsilon}_{jt}\widetilde{\epsilon}_{js} \mid \{\alpha_i\}_{i=1}^n, \Gamma_T\right)\right)\frac{1}{T_0^2}\sum_{t=1}^{T_0}\sum_{s=1}^{T_0}\left(\left|\mathbb{E}\left(\widetilde{\epsilon}_{k_1t}\widetilde{\epsilon}_{k_1s} \mid \alpha_{k_1}, \Gamma_T\right)\right|+\left|\mathbb{E}\left(\widetilde{\epsilon}_{k_2t}\widetilde{\epsilon}_{k_2s} \mid \alpha_{k_2}, \Gamma_T\right)\right|\right)\end{aligned}$$

Since for any realization of $(\{\alpha_i\}_{i=1}^n, \Gamma_T)$, (i) $\max_{t=1\ldots,T_0}\left(\mathbb{E}\left(\widetilde{\epsilon}_{it}\widetilde{\epsilon}_{is}+\widetilde{\epsilon}_{jt}\widetilde{\epsilon}_{js} \mid \{\alpha_i\}_{i=1}^n, \Gamma_T\right)\right) < \infty$ by Assumption 4.1(iii), and $\frac{1}{T_0}\sum_{t=1}^{T_0}\sum_{s=1}^{T_0}\left(\left|\mathbb{E}\left(\widetilde{\epsilon}_{k_1t}\widetilde{\epsilon}_{k_1s} \mid \alpha_{k_1}, \Gamma_T\right)\right|+\left|\mathbb{E}\left(\widetilde{\epsilon}_{k_2t}\widetilde{\epsilon}_{k_2s} \mid \alpha_{k_2}, \Gamma_T\right)\right|\right) < \infty$ by the weak dependence condition in Assumption 6.3(ii). There exists a constant C_2 (irrelevant of the realization of $(\{\alpha_i\}_{i=1}^n, \Gamma_T)$) such that

$$\mathrm{Var}\left(\frac{1}{T_0}\sum_{t=0}^{T_0-1}(\widetilde{\epsilon}_{it}-\widetilde{\epsilon}_{jt})(\widetilde{\epsilon}_{k_1t}-\widetilde{\epsilon}_{k_2t}) \mid \{\alpha_i\}_{i=1}^n, \Gamma_T\right) \leqslant \frac{C_2}{T_0}$$

Then, by Markov inequality, for any $\delta > 0$,

$$\begin{aligned}\mathbb{P}\left(|A_5| \geqslant \delta \mid \{\alpha_i\}_{i=1}^n, \Gamma_T\right) =& \mathbb{P}\left(\bigcup_{k_1,k_2}\left\{\left|\frac{1}{T_0}\sum_{t=0}^{T_0-1}(\widetilde{\epsilon}_{it}-\widetilde{\epsilon}_{jt})(\widetilde{\epsilon}_{k_1t}-\widetilde{\epsilon}_{k_2t})\right| \geqslant \delta\right\} \mid \{\alpha_i\}_{i=1}^n, \Gamma_T\right) \\ \leqslant & \sum_{k_1,k_2=1,\ldots,n}\mathbb{P}\left(\left|\frac{1}{T_0}\sum_{t=0}^{T_0-1}(\widetilde{\epsilon}_{it}-\widetilde{\epsilon}_{jt})(\widetilde{\epsilon}_{k_1t}-\widetilde{\epsilon}_{k_2t})\right| \geqslant \delta \mid \{\alpha_i\}_{i=1}^n, \Gamma_T\right) \\ \leqslant & n^2\frac{C_2}{T_0\delta^2}.\end{aligned}$$

Since the inequalities above does not depend on the realization of $(\{\alpha_i\}_{i=1}^n, \Gamma_T)$, we integrate them out to obtain

$$\mathbb{P}\left(|A_5| \geqslant \delta\right) \leqslant n^2\frac{C_1}{T_0\widetilde{\epsilon}^2}. \tag{C.3}$$

The bound above implies that $A_5 = o_P(1)$ whenever $n^2/T_0 \to 0$. This relative-rate restriction, however, as we discussed previously, is not essential for identification. Because even when $n^2/T_0 \nrightarrow 0$, it suffices to construct the sample maximum using a subsample of n_{T_0} individuals such that $n_{T_0} \to \infty$ and $n_{T_0} = o(\sqrt{T_0})$. The sample maximum over this subsample still converges to the population supremum, which implies that $A_3 = o_P(1)$.

Therefore, combining the probability bounds for A_1–A_5, we have that, for any $i, j = 1, \ldots, n$

$$\widehat{d}_{ij} - d(\alpha_i, \alpha_j) = o_P(1),$$

which immediately implies that $d(\alpha_i, \alpha_j)$ is identified by the construction of $\widehat{d}_{ij}$.

Identification of $\mathrm{ATT}(t,t)$ For any $t \geqslant T_0$, we have

$$\mathrm{ATT}(t,t) = \mathbb{E}(Y_{it}(t) \mid G_i = t) - \mathbb{E}(Y_{it}(\infty) \mid G_i = t)$$

Since $\mathbb{E}(Y_{it}(t) \mid G_i = t)$ is identified from the sample, we only need to identify counterfactual $\mathbb{E}(Y_{it}(\infty) \mid G_i = t)$. Note that by iterated expectation,

$$\begin{aligned}\mathbb{E}\left(Y_{it}(\infty) \mid G_i = t\right) =& \mathbb{E}\left(\mathbb{E}\left(Y_{it}(\infty) \mid X_{it}, \alpha_i, G_i > t-1, G_i = t\right) \mid G_i = t\right) \\ \overset{\text{(i)}}{=}& \mathbb{E}\left(\mathbb{E}\left(Y_{it}(\infty) \mid X_{it}, \alpha_i, G_i > t\right) \mid G_i = t\right) \\ =& \mathbb{E}\left(m_{t|t}\left(X_{it}, \alpha_i\right) \mid G_i = t\right),\end{aligned}$$

where (i) follows from Assumption 6.1. Let $\mathcal{S}_{t-1}$ be the support of (X_{it}, α_i) conditional on $\{G_i > t-1\}$. It is straightforward to verify that, by the overlap condition in Assumption 6.4(iii), $\mathcal{S}_{t-1}$ is also the support of (X_{it}, α_i) conditional on $\{G_i > t\}$ and conditional on $\{G_i = t\}$. Then, by the uniform continuity of $m_{t|t}(\cdot,\cdot)$ (Assumption 6.4(ii)), for any $\epsilon > 0$, there exists a $\widetilde{\delta} > 0$ such that

$$\sup_{(x,\alpha)\in\mathcal{S}_{t-1}} \left| \mathbb{E}\left(Y_{jt}(\infty) \mid X_{jt} = x, \|\alpha_j, \alpha\| \leqslant \widetilde{\delta}, G_j > t \right) - m_{t|t}(x,\alpha) \right| \leqslant \epsilon.$$

Therefore, combining the inequality above and the informativeness condition in Assumption 4.2 implies that there exist a $\delta > 0$ such that

$$\sup_{(x,\alpha)\in\mathcal{S}_{t-1}} \left| \mathbb{E}\left(Y_{jt}(\infty) \mid X_{jt} = x, d(\alpha_j, \alpha) \leqslant \delta, G_j > t \right) - m_{t|t}(x,\alpha) \right| \leqslant \epsilon.$$

It immediately follows that for any i such that $G_i > t-1$, the inequality below holds wpa1,

$$\left| \mathbb{E}\left(Y_{jt}(\infty) \mid X_{jt} = X_{it}, d(\alpha_j, \alpha_i) \leqslant \delta, G_j > t \right) - m_{t|t}(X_{it}, \alpha_i) \right| \leqslant \epsilon$$

Since $d(\alpha_j, \alpha_i)$ is identified, and $\mathbb{E}\left(Y_{jt}(\infty) \mid X_{jt} = X_{it}, d(\alpha_j, \alpha_i) \leqslant \delta, G_j > t \right)$ is directly identified by $\mathbb{E}\left(Y_{jt} \mid X_{jt} = X_{it}, d(\alpha_j, \alpha_i) \leqslant \delta, G_j > t \right)$, for any i such that $G_i > t-1$, the contemporaneous counterfactual expectation is identified by

$$m_{t|t}(X_{it}, \alpha_i) = \lim_{\delta \downarrow 0} \mathbb{E}\left(Y_{jt} \mid X_{jt} = X_{it}, d(\alpha_j, \alpha_i) \leqslant \delta, G_j > t \right). \tag{C.4}$$

This completes the proof of part (i) in Theorem 3.

Averaging the imputed counterfactual over the distribution of treated units in (C.4) yields

$$\begin{aligned}
&\left| \mathbb{E}\left(Y_{it}(\infty) \mid G_i = t \right) - \mathbb{E}\left(\mathbb{E}\left(Y_{jt}(\infty) \mid X_{jt} = X_{it}, d(\alpha_j, \alpha_i) \leqslant \delta, G_j > t \right) \mid G_i = t \right) \right| \leqslant \epsilon \\
\Rightarrow\ &\mathbb{E}\left(Y_{it}(\infty) \mid G_i = t \right) = \mathbb{E}\left(\lim_{\delta \downarrow 0} \mathbb{E}\left(Y_{jt}(\infty) \mid X_{jt} = X_{it}, d(\alpha_j, \alpha_i) \leqslant \delta, G_j > t \right) \mid G_i = t \right) \\
\Rightarrow\ &\mathbb{E}\left(Y_{it}(\infty) \mid G_i = t \right) = \mathbb{E}\left(\lim_{\delta \downarrow 0} \mathbb{E}\left(Y_{jt} \mid X_{jt} = X_{it}, d(\alpha_j, \alpha_i) \leqslant \delta, G_j > t \right) \mid G_i = t \right)
\end{aligned}$$

This give the contemporaneous ATT identification formula in Theorem 1.

Identification of $\mathrm{ATT}(g,t)$ The proof follows a backward induction. For any $t > t' \geqslant g$, we show that when $m_{t|t'+1}(X_{it'+1}, \alpha_i)$ is identified for any i such that $G_i > t'$, then $m_{t|t'}(X_{it'}, \alpha_i)$ is also identified for any i such that $G_i > t'-1$. Now, assume that $m_{t|t'+1}(X_{it'+1}, \alpha_i)$

is identified for any i such that $G_i > t'$. Note that for any $(x, \alpha) \in \mathcal{S}_{t'-1}$,

$$\begin{aligned}
m_{t|t'}(x,\alpha) =& \mathbb{E}\left(Y_{jt}(\infty) \mid X_{jt'} = x, \alpha_j = \alpha, G_j > t'\right) \\
\overset{\text{(i)}}{=}& \mathbb{E}\left(\mathbb{E}\left(Y_{jt}(\infty) \mid X_{jt'+1}, X_{jt'} = x, \alpha_j, G_j > t'\right) \mid X_{jt'} = x, \alpha_j = \alpha, G_j > t'\right) \\
\overset{\text{(ii)}}{=}& \mathbb{E}\left(\mathbb{E}\left(Y_{jt}(\infty) \mid X_{jt'+1}, \alpha_j, G_j > t'\right) \mid X_{jt'} = x, \alpha_j = \alpha, G_j > t'\right) \\
\overset{\text{(ii)}}{=}& \mathbb{E}\left(\mathbb{E}\left(Y_{jt}(\infty) \mid X_{jt'+1}, \alpha_j, G_j > t'+1\right) \mid X_{jt'} = x, \alpha_j = \alpha, G_j > t'\right) \\
=& \mathbb{E}\left(m_{t|t'+1}(X_{jt'+1}, \alpha_j) \mid X_{jt'} = x, \alpha_j = \alpha, G_j > t'\right)
\end{aligned}$$

Here, (i) comes from the law of iterated expectation, (ii) follows from mean-independence assumption in Assumption 6.2, and (iii) holds because of selection mechanism in Assumption 6.1. And $m_{t|t'+1}(X_{jt'+1}, \alpha_j)$ is identified for each j such that $G_i > t'$ by assumption.

Then, by the informativeness condition in Assumption 4.2 and the uniform continuity of $m_{t|t'}$ (Assumption 6.4(ii)), for any $\epsilon > 0$, there exists $\delta > 0$ such that for all $(x, \alpha) \in \mathcal{S}_{t'-1}$,

$$\left|\mathbb{E}\left(m_{t|t'+1}(X_{jt'+1}, \alpha_j) \mid X_{jt'} = x, d(\alpha_j, \alpha) \leqslant \delta, G_j > t'\right) - m_{t|t'}(x, \alpha)\right| \leqslant \epsilon,$$

which implies that the following inequality holds with probability 1 for any i such that $G_i > t' - 1$:

$$\left|\mathbb{E}\left(m_{t|t'+1}(X_{jt'+1}, \alpha_j) \mid X_{jt'} = X_{it'+1}, d(\alpha_j, \alpha_i) \leqslant \delta, G_j > t'\right) - m_{t|t'}(X_{it'}, \alpha_i)\right| \leqslant \epsilon$$

Then, for any i such that $G_i > t' - 1$, $m_{t|t'}(X_{it'}, \alpha_i)$ is identified by

$$m_{t|t'}(X_{it'}, \alpha_i) = \lim_{\delta \downarrow 0} \mathbb{E}\left(m_{t|t'+1}(X_{it'+1}, \alpha_j) \mid X_{jt'} = X_{it'}, d(\alpha_j, \alpha_i) \leqslant \delta, G_j > t'\right).$$

Since $m_{t|t}(X_{it}, \alpha_i)$ is identified for any i such that $G_i > t - 1$, we iterate the argument backward from t to g and obtain that $m_{t|g}(X_{ig}, \alpha_i)$ is identified for any i such that $G_i > g-1$. This completes the proof for the second part in Theorem 3. Lastly, averaging the $m_{t|g}(X_{ig}, \alpha_i)$ over the distribution of units with $G_i = g$ yields

$$\mathbb{E}\left(Y_{it}(\infty) \mid G_i = g\right) = \mathbb{E}\left(m_{t|g}(X_{ig}, \alpha_i) \mid G_i = g\right)$$

$\mathrm{ATT}(g, t)$ is identified by

$$\mathrm{ATT}(g,t) = \mathbb{E}\left(Y_{it} \mid G_i = g\right) - \mathbb{E}\left(m_{t|g}(X_{ig}, \alpha_i) \mid G_i = g\right).$$

This completes the proof. ■

D Proofs of Estimation

For notational simplicity, for each $i = 1, \ldots, n$, let $\mathcal{I}(i)$ denote the subsample containing unit i, and let $\mathcal{I}_m(i)$ and $\mathcal{I}_\pi(i)$ denote the subsamples used to estimate the potential-outcome regression and the propensity score for unit i, respectively. More precisely, if $i \in \mathcal{I}_s$, then $\mathcal{I}(i) = \mathcal{I}_s$, $\mathcal{I}_m(i) = \mathcal{I}_{1+(s \bmod 3)}$, and $\mathcal{I}_\pi(i) = \mathcal{I}_{1+((s+1) \bmod 3)}$.

For notational simplicity, for any $i, j = 1, \ldots, n$, any bandwidth $h > 0$, and any $t \geqslant t' - 1 \geqslant T_0 - 1$, define

$$\widehat{W}_{h,ij}^{(t,t')} = \mathbf{1}(G_j > t)\widehat{K}_{h,ij}^{(t')}$$

To simplify the notation, we also assume, without loss of generality, that the kernel function is bounded by $\overline{K} < 1$, so that the product kernel satisfies $|\widehat{W}_{h,ij}^{(t,t')}| < \overline{K}$. Then, for each $i \in \mathcal{I}_s$, the estimated propensity score $\widehat{\pi}_{it}$ can be rewritten as

$$\widehat{\pi}_{it} = \frac{\sum_{j \in \mathcal{I}_{s_p}} \mathbf{1}(G_j = t)\widehat{W}_{h_\pi,ij}^{(t-1,t)}}{\sum_{j \in \mathcal{I}_{s_p}} \widehat{W}_{h_\pi,ij}^{(t-1,t)}}.$$

In addition, for any $i \in \mathcal{I}_s$, $j \in \mathcal{I}_{s_m}$, and for any $t > g \geqslant T_0$, we can rewrite $\widehat{m}_{j,t|t'}$ in a recursive way:

$$\begin{aligned}
\widetilde{m}_{j,t|t} &= \sum_{j' \in \mathcal{I}_{s_m}} Y_{j't}\widehat{W}_{h_m,jj'}^{(t,t)} \Big/ \sum_{j' \in \mathcal{I}_{s_m}} \widehat{W}_{h_m,jj'}^{(t,t)}, \\
\widetilde{m}_{j,t|t-1} &= \sum_{j' \in \mathcal{I}_{s_m}} \widehat{m}_{j,t|t}\widehat{W}_{h_m,jj'}^{(t-1,t-1)} \Big/ \sum_{j' \in \mathcal{I}_{s_m}} \widehat{W}_{h_m,jj'}^{(t-1,t-1)}, \\
&\vdots \\
\widetilde{m}_{j,t|g+1} &= \sum_{j' \in \mathcal{I}_{s_m}} \widehat{m}_{j,t|g+2}\widehat{W}_{h_m,jj'}^{(g+1,g+1)} \Big/ \sum_{j' \in \mathcal{I}_{s_m}} \widehat{W}_{h_m,jj'}^{(g+1,g+1)},
\end{aligned}$$

and

$$\begin{aligned}
\widehat{m}_{i,t|t} &= \sum_{j \in \mathcal{I}_{s_m}} Y_{jt}\widehat{W}_{h_m,ij}^{(t,t)} \Big/ \sum_{j \in \mathcal{I}_{s_m}} \widehat{W}_{h_m,ij}^{(t,t)}, \\
\widehat{m}_{i,t|t-1} &= \sum_{j \in \mathcal{I}_{s_m}} \widetilde{m}_{j,t|t}\widehat{W}_{h_m,ij}^{(t-1,t-1)} \Big/ \sum_{j \in \mathcal{I}_{s_m}} \widehat{W}_{h_m,ij}^{(t-1,t-1)}, \\
&\vdots \\
\widehat{m}_{i,t|g} &= \sum_{j \in \mathcal{I}_{s_m}} \widetilde{m}_{j,t|g+1}\widehat{W}_{h_m,ij}^{(g,g)} \Big/ \sum_{j \in \mathcal{I}_{s_m}} \widehat{W}_{h_m,ij}^{(g,g)}
\end{aligned}$$

Also, for notational simplicity, we use $\mathcal{A}$ to denote the event

$$\mathcal{A} := \left\{ \max_{i,j=1,\ldots,n} |\widehat{d}_{ij} - d(\alpha_i, \alpha_j)| \leqslant \delta_{n,T_0} \right\}. \tag{D.1}$$

By Assumption 5.1, $\lim_{n,T_0\to\infty} \mathbb{P}(\mathcal{A}) = 1$. Let $\phi := \{\{\alpha_i\}_{i=1,\ldots,N}, \Gamma_t\}$ denote the set of factors. Then, It is straightforward to verify that $\mathcal{A}$ belongs to the σ-field generated by ϕ and pre-treatment histories $\{(X_{it}, Y_{it})\}_{i=1,\ldots,N,t=0,\ldots,T_0-1}$. Let $\mathcal{F}_t$ denote the collection of fixed effects and information available at period t, i.e.,

$$\mathcal{F}_t := \{\phi, \{(Y_{i\tau}, X_{i\tau}, D_{i\tau})\}_{i=1,\ldots,n,\tau=1,\ldots,t}\}.$$

Lastly, for notational simplicity, I write

$$\begin{aligned} Y_{it}(\infty) =& m_{t|t'}(X_{it'}, \alpha_i) + u_{i,t|t'}, && \forall\ T_0 \leqslant t' \leqslant t \leqslant T, \text{ and } G_i > t' - 1, \\ D_{it'} =& \pi_{t'}(X_{it'}, \alpha_i) + e_{it'}, && \forall\ T_0 \leqslant t' \leqslant T, \text{ and } G_i > t' - 1. \end{aligned}$$

Here, $u_{i,t|t'}$ and $e_{it'}$ are projection errors satisfying $\mathbb{E}_T(u_{i,t|t'} \mid X_{it'}, \alpha_i, G_i > t' - 1) = \mathbb{E}_T(e_{it'} \mid X_{it'}, \alpha_i, G_i > t' - 1) = 0$. In addition, when $t = t'$, we use $u_{it} := u_{i,t|t'}$.

Lemma D.1 *Under conditions in Theorem 4, let* $\delta_{n,T_0} := \lambda_1 n^{-1/d_\alpha} \log(n) + \lambda_2 T_0^{-1/2} \sqrt{\log(nT_0)}$ *be the estimation error of* $\widehat{d}_{ij}$ *stated as in Assumption 5.1. Then, For any* $h > 0$ *such that* $\delta_{n,T_0}/h \to 0$*, and for all* $i, j = 1, \ldots, n$*, if* $\widehat{d}_{ij} \leqslant h$*, then*

$$\|\alpha_i - \alpha_j\| \leqslant 2\overline{\eta} h$$

holds with probability approaching 1*, where* η *is defined as in Assumption 5.2.*

Proof of Lemma D.1. Note that for all $i, j = 1, \ldots, n$, since $\delta_{n,T_0}/h \to 0$, the following inequality holds with probability approach to 1,

$$\left\{ \widehat{d}_{ij} \leqslant h \right\} \subseteq \{ d(\alpha_i, \alpha_j) \leqslant h + \delta_{n,T_0} \} \subseteq \{ d(\alpha_i, \alpha_j) \leqslant 2h \}.$$

Also, by informativeness condition (see Assumption 5.2), we have

$$\{ d(\alpha_i, \alpha_j) \leqslant 2h \} \subseteq \{ \|\alpha_i - \alpha_j\| \leqslant 2\overline{\eta} h \}.$$

This completes the proof. ■

D.1 Proofs for Proposition 2 and Proposition 3

It sufices to prove Proposition 3 because Proposition 2 is a special case of Proposition 3.

Proof of Proposition 3. We establish the proposition in three steps. We first derive the uniform convergence rate of the contemporaneous outcome-regression estimator, then establish the corresponding result for the propensity-score estimator, and finally extend the outcome-regression result to the recursive dynamic estimator by backward induction.

Bounds for $\max_{i=1,\dots,n}\left|\widehat{m}_{i,t|t}-m_{i,t|t}\right|$ For each $i=1,\dots,n$, consider the following decomposition:

$$\widehat{m}_{i,t|t}-m_{i,t|t}=\underbrace{\sum_{j\in\mathcal{I}_m(i)}(m_{j,t|t}-m_{i,t|t})\widehat{W}^{(t,t)}_{h_m,ij}\Big/\sum_{j\in\mathcal{I}_m(i)}\widehat{W}^{(t,t)}_{h_m,ij}}_{:=B_{h_m,it}}+\underbrace{\sum_{j\in\mathcal{I}_m(i)}u_{jt}\widehat{W}^{(t,t)}_{h_m,ij}\Big/\sum_{j\in\mathcal{I}_m(i)}\widehat{W}^{(t,t)}_{h_m,ij}}_{:=V_{h_m,it}},$$

and the estimation errors of counterfactual potential outcomes are bounded by:

$$\max_{i=1,\dots,n}\left|\widehat{m}_{i,t|t}-m_{i,t|t}\right|\leqslant\max_{i=1,\dots,n}\left|B_{h_m,it}\right|+\max_{i=1,\dots,n}\left|V_{h_m,it}\right|.$$

To establish the bound for $\max_{i=1,\dots,n}|B_{h_m,it}|$, note that the following bound holds wpa1:

$$\begin{aligned}
&\max_{i=1,\dots,n}\left|B_{h_m,it}\right|\\
&\overset{\text{(i)}}{=}\max_{i=1,\dots,n}\left|\sum_{j\in\mathcal{I}_m(i)}\mathbf{1}\left(\|X_{it}-X_{jt}\|_\infty,\widehat{d}_{ij}\leqslant h\right)(m_{j,t|t}-m_{i,t|t})\widehat{W}^{(t,t)}_{h_m,ij}\Big/\sum_{j\in\mathcal{I}_m(i)}\widehat{W}^{(t,t)}_{h_m,ij}\right|\\
&\overset{\text{(ii)}}{\leqslant}\max_{i=1,\dots,n}\left|\sum_{j\in\mathcal{I}_m(i)}\mathbf{1}\left(\|X_{it}-X_{jt}\|_\infty\leqslant h,\|\alpha_i-\alpha_j\|\leqslant\frac{2}{\overline{\eta}}h\right)(m_{j,t|t}-m_{i,t|t})\widehat{W}^{(t,t)}_{h_m,ij}\Big/\sum_{j\in\mathcal{I}_m(i)}\widehat{W}^{(t,t)}_{h_m,ij}\right|\\
&\overset{\text{(ii)}}{\leqslant}L_m(d_\alpha+d_X)(1+2\overline{\eta})h.
\end{aligned}\tag{D.2}$$

Here, (i) follows from Assumption 6.5(vi), which requires that the kernel function be supported on $[-1,1]$. Inequality (ii) follows from Lemma D.1, and (iii) follows from the Lipschitz continuity of $m_{t|t'}(\cdot)$ imposed in Assumption 6.5(iv), where L_m denotes its Lipschitz constant. To establish the bound for $\max_{i=1,\dots,n}|V_{h_m,it}|$, note that the following inequality holds wpa1

by Lemma D.2:

$$\begin{aligned}\max_{i=1,\dots,n}|V_{h_m,it}| &\leqslant \left(\min_{i=1,\dots,n}\sum_{j\in\mathcal{I}_m(i)}\widehat{W}^{(t,t)}_{h_m,ij}\right)^{-1}\max_{i=1,\dots,n}\left|\sum_{j\in\mathcal{I}_m(i)}u_{jt}\widehat{W}^{(t,t)}_{h_m,ij}\right| \\ &\leqslant \frac{\max_{i=1,\dots,n}\left|\sum_{j\in\mathcal{I}_m(i)}u_{jt}\widehat{W}^{(t,t)}_{h_m,ij}\right|}{\underline{C}(t,t)nh^{d_\alpha+d_X}}.\end{aligned}$$

In addition, since the event $\mathcal{A}$ (defined as in (D.1)) and $\{\widehat{d}_{ij}\}_{i,j}$ belong to the σ-field generated by the set of factors ϕ and pretreatment histories $\{(X_{it},Y_{it})\}_{i=1,\dots,n,t=0,\dots,T_0-1}$, it is straightforward to verify that (i) $\mathbb{E}\left(u_{jt}\mid\{X_{it}\}_{i=1}^n,\mathcal{F}_{t-1},\{\widehat{W}^{(t,t)}_{h_m,ij}\}_{i,j\in\mathcal{I}_m(i)}\right)=0$ by Assumption 6.1 and 6.2, and (ii) u_{jt} is independent across j conditional on $(\{X_{it}\}_{i=1}^n,\mathcal{F}_{t-1},\{\widehat{W}^{(t,t)}_{h_m,ij}\}_{i,j\in\mathcal{I}_m(i)})$ because of Assumption 6.1(i), (iii) given $(\{X_{it}\}_{i=1}^n,\mathcal{F}_{t-1},\{\widehat{W}^{(t,t)}_{h_m,ij}\}_{i,j\in\mathcal{I}_m(i)})$, u_{jt} is conditionally sub-Gaussian by Markov and conditional sub-Gaussian assumptions (Assumption 6.2 and 6.5(ii)). Therefore, we employ the Hoeffding's inequality (Lemma F.2) to show that for each i and for any $\epsilon>0$,

$$\mathbb{P}\left(\left|\sum_{j\in\mathcal{I}_m(i)}u_{jt}\widehat{W}^{(t,t)}_{h_m,ij}\right|\geqslant\epsilon\mid\{X_{it}\}_{i=1}^n,\mathcal{F}_{t-1},\{\widehat{W}^{(t,t)}_{h_m,ij}\}_{i,j\in\mathcal{I}_m(i)}\right)\leqslant 2\exp\left(\frac{-\epsilon^2}{2\sigma^2V_n}\right),$$

where $V_n:=\max_{i=1,\dots,n}\sum_{j\in\mathcal{I}_m(i)}\left(\widehat{W}^{(t,t)}_{h_m,ij}\right)^2$. Therefore, we obtain

$$\begin{aligned}&\mathbb{P}\left(\max_{i=1,\dots,n}\left|\sum_{j\in\mathcal{I}_m(i)}u_{jt}\widehat{W}^{(t,t)}_{h_m,ij}\right|\geqslant\epsilon\mid\{X_{it}\}_{i=1}^n,\mathcal{F}_{t-1},\{\widehat{W}^{(t,t)}_{h_m,ij}\}_{i,j\in\mathcal{I}_m(i)}\right)\\ =&\mathbb{P}\left(\bigcup_{i=1,\dots,n}\left\{\left|\sum_{j\in\mathcal{I}_m(i)}u_{jt}\widehat{W}^{(t,t)}_{h_m,ij}\right|\geqslant\epsilon\right\}\mid\{X_{it}\}_{i=1}^n,\mathcal{F}_{t-1},\{\widehat{W}^{(t,t)}_{h_m,ij}\}_{i,j\in\mathcal{I}_m(i)}\right)\\ \leqslant&\sum_{i=1,\dots,n}\mathbb{P}\left(\left|\sum_{j\in\mathcal{I}_m(i)}u_{jt}\widehat{W}^{(t,t)}_{h_m,ij}\right|\geqslant\epsilon\mid\{X_{it}\}_{i=1}^n,\mathcal{F}_{t-1},\{\widehat{W}^{(t,t)}_{h_m,ij}\}_{i,j\in\mathcal{I}_m(i)}\right)\\ \overset{\text{(i)}}{\leqslant}&2n\exp\left(\frac{-\epsilon^2}{2\sigma^2V_n}\right),\end{aligned}$$

where (i) follows from Hoeffding's inequality discussed above. Also, note that V_n is bounded following inequality wpa1:

$$V_n \overset{\text{(i)}}{\leqslant} \overline{K} \max_{i=1,\dots,n} \sum_{j \in \mathcal{I}_m(i)} \widehat{W}^{(t,t)}_{h_m,ij} \overset{\text{(ii)}}{\leqslant} \overline{KC}(t,t) n h^{d_\alpha + d_X}.$$

Here, (i) follows from Assumption 6.5(vi) and (ii) follows from Lemma D.2. Then, there exists a sufficiently large constant $C_1 > 0$ such that when $\epsilon = C_1 \sqrt{n h^{d_\alpha + d_X} \log(n)}$, [13]

$$\lim_{n \to \infty} \mathbb{P}\left(\max_{i=1,\dots,n} \left| \sum_{j \in \mathcal{I}_m(i)} u_{jt} \widehat{W}^{(t,t)}_{h_m,ij} \right| \geqslant C_1 \sqrt{n h^{d_\alpha + d_X} \log(n)} \right) = 0.$$

Therefore, the following inequality holds wpa1:

$$\max_{i=1,\dots,n} |V_{h_m,it}| \leqslant \frac{C_1}{\underline{C}(t,t)} \left(n h_m^{d_\alpha + d_X} \right)^{-1/2} \sqrt{\log(n)}. \tag{D.3}$$

Combining (D.2) and (D.3) yields

$$\max_{i=1,\dots,n} \left| \widehat{m}_{i,t|t} - m_{i,t|t} \right| = O_P\left(h_m + \left(n h_m^{d_\alpha + d_X} \right)^{-1/2} \sqrt{\log(n)} \right) \tag{D.4}$$

By similar argument, we can also show that for each $j = 1, \dots, n$, the within-sample estimator $\widetilde{m}_{j,t|t}$ admits a similar error bound,

$$\max_{j=1,\dots,n} \left| \widetilde{m}_{j,t|t} - m_{j,t|t} \right| = O_P\left(h_m + \left(n h_m^{d_\alpha + d_X} \right)^{-1/2} \sqrt{\log(n)} \right), \tag{D.5}$$

which is crucial in obtaining the error bound for recursive estimator in the following analysis.

Bounds for $\max_{i \in \mathcal{I}_s} |\widehat{\pi}_{it'} - \pi_{it'}|$ The uniform error bound for $\widehat{\pi}_{it'}$ can be established similarly, and the proof is omitted.

Bounds for $\max_{i \in \mathcal{I}_s} \left| \widehat{m}_{i,t|g} - m_{i,t|g} \right|$ We establish the result by backward induction. The induction statement is that, fixing $t' < t$, the within-sample estimator satisfies

$$\max_{j=1,\dots,n} \left| \widetilde{m}_{j,t|t'+1} - m_{j,t|t'+1} \right| = O_P\left(h_m + \left(n h_m^{d_\alpha + d_X} \right)^{-1/2} \sqrt{\log(n)} \right).$$

[13] Here we integrate $(\{X_{it}\}_{i=1}^n, \mathcal{F}_{t-1}, \{\widehat{W}^{(t,t)}_{h_m,ij}\}_{i,j \in \mathcal{I}_m(i)})$ to obtain unconditional probability.

For the base case $t' = t-1$, this has been established in (D.5). Then, for any $i = 1, \ldots, n$ and fixing $t' < t$, consider the following decomposition:

$$\widehat{m}_{i,t|t'} - m_{i,t|t'} = \frac{\sum_{j\in\mathcal{I}_m(i)} \left(\widetilde{m}_{j,t|t'+1} - m_{j,t|t'+1}\right) \widehat{W}^{(t',t')}_{h_m,ij}}{\sum_{j\in\mathcal{I}_m(i)} \widehat{W}^{(t',t')}_{h_m,ij}} + \frac{\sum_{j\in\mathcal{I}_m(i)} \left(m_{j,t|t'+1} - m_{j,t|t'}\right) \widehat{W}^{(t',t')}_{h_m,ij}}{\sum_{j\in\mathcal{I}_m(i)} \widehat{W}^{(t',t')}_{h_m,ij}} + \frac{\sum_{j\in\mathcal{I}_m(i)} \left(m_{j,t|t'} - m_{i,t|t'}\right) \widehat{W}^{(t',t')}_{h_m,ij}}{\sum_{j\in\mathcal{I}_m(i)} \widehat{W}^{(t',t')}_{h_m,ij}}.$$

By the induction hypothesis, $\max_{j=1,\ldots,n} \left|\widetilde{m}_{j,t|t'+1} - m_{j,t|t'+1}\right| = O_P\left(h_m + \left(nh_m^{d_\alpha+d_X}\right)^{-1/2}\sqrt{\log(n)}\right)$, it follows that the first term is bounded by

$$\max_{i=1,\ldots,n} \left|\frac{\sum_{j\in\mathcal{I}_m(i)} \left(\widetilde{m}_{j,t|t'+1} - m_{j,t|t'+1}\right) \widehat{W}^{(t',t')}_{h_m,ij}}{\sum_{j\in\mathcal{I}_m(i)} \widehat{W}^{(t',t')}_{h_m,ij}}\right| = O_P\left(h_m + \left(nh_m^{d_\alpha+d_X}\right)^{-1/2}\sqrt{\log(n)}\right) \tag{D.6}$$

For the second term, note that since $\mathbb{E}\left(m_{j,t|t'+1} - m_{j,t|t'} \mid \{X_{it'}\}_{i=1}^n, \phi, \mathcal{F}_{t'-1}\right) = 0$ (by Assumption 6.1 and 6.2), we apply the same argument used to establish (D.3) to obtain that the second term is bounded by

$$\max_{i=1,\ldots,n} \left|\frac{\sum_{j\in\mathcal{I}_m(i)} \left(m_{j,t|t'+1} - m_{j,t|t'}\right) \widehat{W}^{(t',t')}_{h_m,ij}}{\sum_{j\in\mathcal{I}_m(i)} \widehat{W}^{(t',t')}_{h_m,ij}}\right| = O_P\left(\left(nh_m^{d_\alpha+d_X}\right)^{-1/2}\sqrt{\log(n)}\right) \tag{D.7}$$

Last, applying the same argument used to establish (D.2), we obtain that the third term is bounded by

$$\max_{i=1,\ldots,n} \left|\frac{\sum_{j\in\mathcal{I}_m(i)} \left(m_{j,t|t'} - m_{i,t|t'}\right) \widehat{W}^{(t',t')}_{h_m,ij}}{\sum_{j\in\mathcal{I}_m(i)} \widehat{W}^{(t',t')}_{h_m,ij}}\right| = O_P\left(h_m\right) \tag{D.8}$$

Combining (D.6), (D.7), and (D.8) yields

$$\max_{i=1,\ldots,n} \left|\widehat{m}_{i,t|t'} - m_{i,t|t'}\right| = O_P\left(h_m + \left(nh_m^{d_\alpha+d_X}\right)^{-1/2}\sqrt{\log(n)}\right).$$

By similar argument, we can also show that for each $j = 1, \ldots, n$, the in-sample estimator $\widehat{m}_{j,t|t'}$ admits a similar error bound,

$$\max_{j=1,\ldots,n} \left|\widetilde{m}_{j,t|t'} - m_{j,t|t'}\right| = O_P\left(h_m + \left(nh_m^{d_\alpha+d_X}\right)^{-1/2}\sqrt{\log(n)}\right).$$

This establishes the induction step. Since the result holds for the base case $t' = t-1$, backward induction implies that it holds for every $t' < t$. This completes the proof. ■

D.2 Proofs for Theorem 4 and Theorem

Proof of Theorem 4. We first establish the result for the contemporaneous ATT and then extend it to the dynamic ATT under staggered adoption.

Contemporaneous ATT Without loss of generality, we assume no unit is treated before t, i.e., $t = T_0$ and $\{G_i > t-1\} = \{1, \dots, n\}$. Consider the following decomposition of contemporaneous ATT estimator $\widehat{\mathrm{ATT}(t,t)}$:

$$\begin{aligned}\widehat{\mathrm{ATT}(t,t)} = &\underbrace{\frac{1}{n_t}\sum_{i=1}^{n}\left(\mathbf{1}(G_i=t)Y_{it} - \frac{\mathbf{1}(G_i>t)Y_{it}\pi_{it} + (\mathbf{1}(G_i=t)-\pi_{it})m_{i,t|t}}{1-\pi_{it}}\right)}_{:=Q_1}\\ &\underbrace{-\frac{1}{n_t}\sum_{i=1}^{n}\frac{\mathbf{1}(G_i>t)(\widehat{\pi}_{it}-\pi_{it})}{(1-\pi_{it})(1-\widehat{\pi}_{it})}u_{it}}_{:=Q_2} \underbrace{-\frac{1}{n_t}\sum_{i=1}^{n}\frac{\widehat{m}_{i,t|t}-m_{i,t|t}}{1-\widehat{\pi}_{it}}e_{it}}_{:=Q_3}\\ &\underbrace{+\frac{1}{n_t}\sum_{i=1}^{n}\frac{\widehat{\pi}_{it}-\pi_{it}}{1-\widehat{\pi}_{it}}(\widehat{m}_{i,t|t}-m_{i,t|t})}_{:=Q_4}.\end{aligned}$$

Here, $n_t = \sum_{i=1}^{n}\mathbf{1}(G_i=t)$, Q_1 stands for the oracle estimator, and Q_2, Q_3, Q_4 are estimation errors introduced by nonparametric estimation. Each of Q_2, Q_3, and Q_4 is the sum of contributions from the three folds. Since the same argument applies to each fold, it suffices to establish the bounds for one fold. Therefore, we focus on $i \in \mathcal{I}_1$ and always assume that unit 1 belongs to $\mathcal{I}_1$ in the following derivations whenever this simplifies the analysis.

Step 1. Bound for Q_2 It is straightforward to verify that, for each $i \in \mathcal{I}_1$,

$$\mathbb{E}\left(u_{it} \mid \{X_{it}\}_{i=1}^{n}, \mathcal{F}_{t-1}, \left\{\frac{\mathbf{1}(G_i>t)(\widehat{\pi}_{it}-\pi_{it})}{(1-\pi_{it})(1-\widehat{\pi}_{it})}\right\}_{i\in\mathcal{I}_1}\right) = 0$$

by sample-splitting, Assumption 6.1, and 6.2. In addition,

$$\begin{aligned}&\mathrm{Var}\left(Q_2 \mid \{X_{it}\}_{i=1}^{n}, \mathcal{F}_{t-1}, \left\{\frac{\mathbf{1}(G_i>t)(\widehat{\pi}_{it}-\pi_{it})}{(1-\pi_{it})(1-\widehat{\pi}_{it})}\right\}_{i\in\mathcal{I}_1}\right)\\ &\overset{\text{(i)}}{\leqslant}\frac{9\sigma^2}{n^2}\sum_{i\in\mathcal{I}_1}\frac{\mathbf{1}(G_i>t)(\widehat{\pi}_{it}-\pi_{it})^2}{(1-\pi_{it})^2(1-\widehat{\pi}_{it})^2} \overset{\text{(ii)}}{=} o_P(1/n).\end{aligned}$$

Here, (i) holds because u_{it} is conditional sub-Gaussian (Assumption 6.5(ii)) and independent across i conditional on $\left(\{X_{it}\}_{i=1}^n, \mathcal{F}_{t-1}, \left\{\frac{\mathbf{1}(G_i>t)(\widehat{\pi}_{it}-\pi_{it})}{(1-\pi_{it})(1-\widehat{\pi}_{it})}\right\}_{i\in\mathcal{I}_1}\right)$ (see Assumption 6.1(i)), (ii) follows from the common overlap condition (Assumption 6.5(v)) and the consistency of $\widehat{\pi}_{it}$ (Proposition 3). Therefore, we can conclude that

$$Q_2 = o_P\left(1/\sqrt{n}\right). \tag{D.9}$$

Step 3. Bound for Q_3 We can use a similar argument in obtaining (D.9) to establish the bound for Q_3. Specifically,

$$Q_3 = o_P\left(1/\sqrt{n}\right). \tag{D.10}$$

The proof is omitted for simplicity.

Step 3. Bound for Q_4 Before deriving the probability bound for Q_4, consider the following decomposition of $\widehat{m}_{i,t|t}$ and $\widehat{\pi}_{it}$:

$$\widehat{m}_{i,t|t} - m_{i,t|t} = \underbrace{\frac{\sum_{j\in\mathcal{I}_m(i)}(m_{j,t|t}-m_{i,t|t})\widehat{W}^{(t,t)}_{h_m,ij}}{\sum_{j\in\mathcal{I}_m(i)}\widehat{W}^{(t,t)}_{h_m,ij}}}_{:=B_{m,it}} + \underbrace{\frac{\sum_{j\in\mathcal{I}_m(i)}u_{jt}\widehat{W}^{(t,t)}_{h_m,ij}}{\sum_{j\in\mathcal{I}_m(i)}\widehat{W}^{(t,t)}_{h_m,ij}}}_{:=V_{m,it}}, \tag{D.11}$$

and

$$\widehat{\pi}_{it} - \pi_{it} = \underbrace{\frac{\sum_{j\in\mathcal{I}_\pi(i)}(\pi_{jt}-\pi_{it})\widehat{W}^{(t-1,t)}_{h_\pi,ij}}{\sum_{j\in\mathcal{I}_\pi(i)}\widehat{W}^{(t-1,t)}_{h_\pi,ij}}}_{:=B_{\pi,it}} + \underbrace{\frac{\sum_{j\in\mathcal{I}_\pi(i)}e_{jt}\widehat{W}^{(t-1,t)}_{h_\pi,ij}}{\sum_{j\in\mathcal{I}_\pi(i)}\widehat{W}^{(t-1,t)}_{h_\pi,ij}}}_{:=V_{\pi,it}}. \tag{D.12}$$

The second-order series expansion of Q_4 reads[14]:

$$Q_4 = \underbrace{\frac{1}{n_t}\sum_{i=1}^n \frac{\widehat{\pi}_{it}-\pi_{it}}{1-\pi_{it}}(\widehat{m}_{i,t|t}-m_{i,t|t})}_{:=S_1} + \underbrace{\frac{1}{n_t}\sum_{i=1}^n \frac{(\widehat{\pi}_{it}-\pi_{it})^2}{(1-\pi_{it})^2}(\widehat{m}_{i,t|t}-m_{i,t|t})}_{:=S_2} + o_P(S_1) + o_P(S_2)$$

To establish the error bound for S_1, note that,

$$S_1 = \frac{1}{n_t}\left(\sum_{i=1}^n \frac{B_{m,it}B_{\pi,it}}{1-\pi_{it}} + \sum_{i=1}^n \frac{B_{m,it}V_{\pi,it}}{1-\pi_{it}} + \sum_{i=1}^n \frac{V_{m,it}B_{\pi,it}}{1-\pi_{it}} + \sum_{i=1}^n \frac{V_{m,it}V_{\pi,it}}{1-\pi_{it}}\right)$$

[14] It is straightforward to verify that higher-order expansions are asymptotically negligible under conditions in Theorem 4.

For the first term, using the same argument in obtaining (D.2), we have

$$\frac{1}{n_t}\sum_{i=1}^{n}\frac{B_{m,it}B_{\pi,it}}{1-\pi_{it}} = O_P\left(h_m h_\pi .\right) \tag{D.13}$$

For the second term (here, it suffices to focus on $i \in \mathcal{I}_1$ where $1 \in \mathcal{I}_1$ to simplify the analysis), since (i) $\mathbb{E}\left(e_{jt} \mid \{X_{it}\}_{i=1}^n, \mathcal{F}_{t-1}, \{B_{m,it}\}_{i\in\mathcal{I}_1}, \{\widehat{W}^{(t-1,t)}_{h_\pi,ij}\}_{i\in\mathcal{I}_1, j\in\mathcal{I}_\pi(i)}\right) = 0$ by sample-splitting, Assumption 6.1, and 6.2, (ii) conditional on $(\{X_{it}\}_{i=1}^n, \mathcal{F}_{t-1}, \{B_{m,it}\}_{i\in\mathcal{I}_1}, \{\widehat{W}^{(t-1,t)}_{h_\pi,ij}\}_{i\in\mathcal{I}_1, j\in\mathcal{I}_\pi(i)})$, e_{jt} is independent across j because of Assumption 6.1(i), and (iii) e_{jt} is conditionally sub-Gaussian given $\left(\{X_{it}\}_{i=1}^n, \mathcal{F}_{t-1}, \{B_{m,it}\}_{i\in\mathcal{I}_1}, \{\widehat{W}^{(t-1,t)}_{h_\pi,ij}\}_{i\in\mathcal{I}_1, j\in\mathcal{I}_\pi(i)}\right)$ by Markov and sub-Gaussian assumptions (Assumption 6.2 and 6.5(ii)), we then have

$$\begin{aligned}
&\mathrm{Var}\left(\frac{1}{n_t}\sum_{i\in\mathcal{I}_1}\frac{B_{m,it}V_{\pi,it}}{1-\pi_{it}} \mid \{X_{it}\}_{i=1}^n, \mathcal{F}_{t-1}, \{B_{m,it}\}_{i\in\mathcal{I}_1}, \{\widehat{W}^{(t-1,t)}_{h_\pi,ij}\}_{i\in\mathcal{I}_1, j\in\mathcal{I}_\pi(i)}\right)\\
\leqslant&\frac{1}{n_t^2}\sum_{j\in\mathcal{I}_\pi(1)}\left(\sum_{i\in\mathcal{I}_1}\frac{B_{m,it}\widehat{W}^{(t-1,t)}_{h_\pi,ij}}{(1-\pi_{it})\sum_{\ell\in\mathcal{I}_\pi(i)}\widehat{W}^{(t-1,t)}_{h_\pi,i\ell}}\right)^2\\
\overset{\text{(i)}}{\leqslant}&\frac{\max_{i=1,\ldots,n}|B_{m,it}|^2}{\left((1-\overline{p})\underline{C}(t-1,t)nh_\pi^{d_\alpha+d_X}\right)^2}\frac{1}{n_t^2}\sum_{j\in\mathcal{I}_\pi(1)}\left(\sum_{i\in\mathcal{I}_1}\widehat{W}^{(t-1,t)}_{h_\pi,ij}\right)^2.
\end{aligned}$$

Here, (i) follows from Lemma D.2. We then employ the same argument in obtaining (D.2) to show that $\max_{i=1,\ldots,n} B_{m,it} = O_P(h_m)$, and together with Lemma D.2, we have

$$\mathrm{Var}\left(\frac{1}{n_t}\sum_{i=1}^{n}\frac{B_{m,it}V_{\pi,it}}{1-\pi_{it}}\right) = O_P\left(\frac{h_m^2}{n}\right) = o_P(1/n) \Rightarrow \frac{1}{n_t}\sum_{i=1}^{n}\frac{B_{m,it}V_{\pi,it}}{1-\pi_{it}} = o_P(1/\sqrt{n}). \tag{D.14}$$

The third term can be bounded similarly:

$$\frac{1}{n_t}\sum_{i=1}^{n}\frac{V_{m,it}B_{\pi,it}}{1-\pi_{it}} = O_P\left(\frac{h_\pi}{\sqrt{n}}\right) = o_P(1/\sqrt{n}). \tag{D.15}$$

The argument is identical to that used for the second term and is therefore omitted.

For the last term, since (i) $\mathbb{E}\left(e_{jt} \mid \{X_{it}\}_{i=1}^n, \mathcal{F}_{t-1}, \{V_{m,it}\}_{i\in\mathcal{I}_1}, \{\widehat{W}^{(t-1,t)}_{h_\pi,ij}\}_{i\in\mathcal{I}_1, j\in\mathcal{I}_\pi(i)}\right) = 0$ by sample-splitting, Assumption 6.1, and 6.2, (ii) e_{jt} is independent across j conditional on $(\{X_{it}\}_{i=1}^n, \mathcal{F}_{t-1}, \{V_{m,it}\}_{i\in\mathcal{I}_1}, \{\widehat{W}^{(t-1,t)}_{h_\pi,ij}\}_{i\in\mathcal{I}_1, j\in\mathcal{I}_\pi(i)})$ because of Assumption 6.1(i), and (iii) e_{jt} is conditionally sub-Gaussian given $\left(\{X_{it}\}_{i=1}^n, \mathcal{F}_{t-1}, \{V_{m,it}\}_{i\in\mathcal{I}_1}, \{\widehat{W}^{(t-1,t)}_{h_\pi,ij}\}_{i\in\mathcal{I}_1, j\in\mathcal{I}_\pi(i)}\right)$, by

Markov and sub-Gaussian assumptions (Assumption 6.2 and 6.5(ii)), we then have

$$\begin{aligned}
&\mathrm{Var}\left(\frac{1}{n_t}\sum_{i\in\mathcal{I}_1}\frac{V_{m,it}V_{\pi,it}}{1-\pi_{it}} \mid \{X_{it}\}_{i=1}^n, \mathcal{F}_{t-1}, \{V_{m,it}\}_{i\in\mathcal{I}_1}, \{\widehat{W}_{h_\pi,ij}^{(t-1,t)}\}_{i\in\mathcal{I}_1, j\in\mathcal{I}_\pi(i)}\right)\\
\leqslant&\frac{1}{n_t^2}\sum_{j\in\mathcal{I}_\pi(1)}\left(\sum_{i\in\mathcal{I}_1}\frac{V_{m,it}\widehat{W}_{h_\pi,ij}^{(t-1,t)}}{(1-\pi_{it})\sum_{\ell\in\mathcal{I}_\pi(i)}\widehat{W}_{h_\pi,i\ell}^{(t-1,t)}}\right)^2\\
\overset{\text{(i)}}{\leqslant}&\frac{\max_{i=1,\ldots,n}|V_{m,it}|^2}{\left((1-\overline{p})\underline{C}(t-1,t)nh_\pi^{d_\alpha+d_X}\right)^2}\frac{1}{n_t^2}\sum_{j\in\mathcal{I}_\pi(1)}\left(\sum_{i\in\mathcal{I}_1}\widehat{W}_{h_\pi,ij}^{(t-1,t)}\right)^2.
\end{aligned}$$

Here, (i) follows from Lemma D.2. We have already shown that (see inequality (D.3)) to show that $\max_{i=1,\ldots,n}|V_{m,it}| = O_P\left(\left(nh_m^{d_\alpha+d_X}\right)^{-1/2}\sqrt{\log(n)}\right)$. Combining it with Lemma D.2 yields

$$\mathrm{Var}\left(\frac{1}{n_t}\sum_{i=1}^n\frac{V_{m,it}V_{p,it}}{1-\pi_{it}}\right) = O_P\left(\frac{\sqrt{\log(n)}}{\sqrt{n}nh_m^{d_\alpha+d_X}}\right) \Rightarrow \frac{1}{n_t}\sum_{i=1}^n\frac{V_{m,it}V_{p,it}}{1-\pi_{it}} = o_P(1\sqrt{n}). \tag{D.16}$$

Combining (D.13), (D.14), (D.15), and (D.16) gives

$$S_1 = O_P\left(h_m h_\pi\right) + o_P\left(1\sqrt{n}\right). \tag{D.17}$$

To establish the error bound for S_2, note that by Proposition 3, we have

$$\max_{i:G_i>t-1}|\widehat{\pi}_{it}-\pi_{it}|^2 = O_P\left(h_\pi^2 + \left(nh_\pi^{d_\alpha+d_X}\right)^{-1}\log(n)\right).$$

It follows that, by the same argument in deriving (D.14),

$$\begin{aligned}
S_2 =&\frac{1}{n_t}\sum_{i=1}^n\frac{(\widehat{\pi}_{it}-\pi_{it})^2}{(1-\pi_{it})^2}B_{m,it} + \frac{1}{n_t}\sum_{i=1}^n\frac{(\widehat{\pi}_{it}-\pi_{it})^2}{(1-\pi_{it})^2}V_{m,it}\\
=&O_P\left(h_m h_\pi^2 + h_m\left(nh_\pi^{d_\alpha+d_X}\right)^{-1}\log(n)\right) + o_P\left(1/\sqrt{n}\right).
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
Q_4 =&O_P\left(h_m h_\pi + h_m h_\pi^2 + h_m\left(nh_\pi^{d_\alpha+d_X}\right)^{-1}\log(n)\right) + o_P(1/\sqrt{n})\\
=&O_P\left(h_m h_\pi + h_m\left(nh_\pi^{d_\alpha+d_X}\right)^{-1}\log(n)\right) + o_P(1/\sqrt{n})
\end{aligned}$$

Lastly, since $h_m \left(nh_\pi^{d_\alpha+d_X}\right)^{-1}\log(n) = o_P(1)$, we obtain a looser by more interpretable bound

$$Q_4 = O_P\left(h_m h_\pi + h_m\left(nh_\pi^{d_\alpha+d_X}\right)^{-1/2}\sqrt{\log(n)}\right) + o_P(1/\sqrt{n}). \tag{D.18}$$

Combining it with (D.9) and (D.10) completes the proof for contemporaneous ATT.

Dynamic ATT We focus on ATT$(t-1,t)$ for brevity. The extension to treatment effects over longer horizons follows from exactly the same argument and is omitted to avoid unnecessary repetition.

Decomposition of $\widehat{m}_{i,t|t-1}$ and $\widehat{\pi}_{it-1}$ $\widehat{m}_{i,t|t-1}$ admits the following decomposition

$$\widehat{m}_{i,t|t-1} - m_{i,t|t-1} = B^{(1)}_{m,it-1} + B^{(2)}_{m,it-1} + V^{(1)}_{m,it-1} + V^{(2)}_{m,it-1}, \tag{D.19}$$

where

$$\begin{aligned}
B^{(1)}_{m,it-1} &:= \sum_{j\in\mathcal{I}_m(i)} \widehat{W}^{(t-1,t-1)}_{h_m,ij} \frac{\sum_{j'\in\mathcal{I}_m(i)} \widehat{W}^{(t,t)}_{m,jj'}\left(m_{j',t|t} - m_{j,t|t}\right)}{\sum_{\ell'\in\mathcal{I}_m(i)} \widehat{W}^{(t,t)}_{m,j\ell'}} \Big/ \sum_{\ell\in\mathcal{I}_m(i)} \widehat{W}^{(t-1,t-1)}_{h_m,i\ell},\\
B^{(2)}_{m,it-1} &:= \sum_{j\in\mathcal{I}_m(i)} \widehat{W}^{(t-1,t-1)}_{h_m,ij} \left(m_{j,t|t-1} - m_{i,t|t-1}\right) \Big/ \sum_{\ell\in\mathcal{I}_m(i)} \widehat{W}^{(t-1,t-1)}_{h_m,i\ell},\\
V^{(1)}_{m,it-1} &:= \sum_{j\in\mathcal{I}_m(i)} \widehat{W}^{(t-1,t-1)}_{h_m,ij} \left(m_{j,t|t} - m_{j,t|t-1}\right) \Big/ \sum_{\ell\in\mathcal{I}_m(i)} \widehat{W}^{(t-1,t-1)}_{h_m,i\ell},\\
V^{(2)}_{m,it-1} &:= \sum_{j\in\mathcal{I}_m(i)} \widehat{W}^{(t-1,t-1)}_{h_m,ij} \frac{\sum_{j'\in\mathcal{I}_m(i)} \widehat{W}^{(t,t)}_{m,jj'} u_{j't}}{\sum_{\ell'\in\mathcal{I}_m(i)} \widehat{W}^{(t,t)}_{m,j\ell'}} \Big/ \sum_{\ell\in\mathcal{I}_m(i)} \widehat{W}^{(t-1,t-1)}_{h_m,i\ell}.
\end{aligned}$$

Similarly, $\widehat{\pi}_{it-1}$ admits the following decomposition:

$$\widehat{\pi}_{it-1} - \pi_{it-1} = \underbrace{\frac{\sum_{k\in\mathcal{I}_\pi(i)} \widehat{W}^{(t-2,t-1)}_{h_\pi,ik}\left(\pi_{kt-1} - \pi_{it-1}\right)}{\sum_{q\in\mathcal{I}_\pi(i)} \widehat{W}^{(t-2,t-1)}_{h_\pi,iq}}}_{:=B_{\pi,it-1}} + \underbrace{\frac{\sum_{k\in\mathcal{I}_\pi(i)} \widehat{W}^{(t-2,t-1)}_{h_\pi,ik} e_{kt-1}}{\sum_{q\in\mathcal{I}_\pi(i)} \widehat{W}^{(t-2,t-1)}_{h_\pi,iq}}}_{:=V_{\pi,it-1}}. \tag{D.20}$$

Decomposition of $\widehat{\mathrm{ATT}}(t-1,t)$ Let $n_{t-1} := \sum_{i=1}^n \mathbf{1}(G_i = t-1)$, the decomposition of $\widehat{\mathrm{ATT}}(t-1,t)$ reads:

$$
\begin{aligned}
&\widehat{\mathrm{ATT}}(t-1,t) - \widehat{\mathrm{ATT}}(t-1,t)^{\text{oracle}} \\
&= \underbrace{-\frac{1}{n_{t-1}}\sum_{i=1}^n \frac{\mathbf{1}(G_i > t-2)}{1-\pi_{it-1}}\left(\widehat{m}_{i,t|t-1} - m_{i,t|t-1}\right)e_{it-1}}_{:=Q_1} \underbrace{-\frac{1}{n_{t-1}}\sum_{i=1}^n \frac{\mathbf{1}(G_i > t-1)\pi_{it-1}}{(1-\pi_{it-1})(1-\pi_{it})}\left(\widehat{m}_{i,t|t} - m_{i,t|t}\right)e_{it}}_{:=Q_2} \\
&\quad \underbrace{-\frac{1}{n_{t-1}}\sum_{i=1}^n \mathbf{1}(G_i > t)\left(\frac{\widehat{\pi}_{it-1}}{1-\widehat{\pi}_{it-1}}\frac{1}{1-\widehat{\pi}_{it}} - \frac{\pi_{it-1}}{1-\pi_{it-1}}\frac{1}{1-\pi_{it}}\right)u_{it}}_{:=Q_3} \\
&\quad \underbrace{-\frac{1}{n_{t-1}}\sum_{i=1}^n \frac{\mathbf{1}(G_i > t-1)(\widehat{\pi}_{it-1} - \pi_{it-1})}{(1-\pi_{it-1})(1-\widehat{\pi}_{it-1})}(m_{i,t|t} - m_{i,t|t-1})}_{:=Q_4} \\
&\quad \underbrace{-\frac{1}{n_{t-1}}\sum_{i=1}^n \frac{\mathbf{1}(G_i > t-1)}{1-\pi_{it-1}}\frac{\widehat{\pi}_{it-1} - \pi_{it-1}}{1-\widehat{\pi}_{it-1}}\left(\widehat{m}_{i,t|t} - m_{i,t|t}\right)}_{:=Q_5} \\
&\quad \underbrace{+\frac{1}{n_{t-1}}\sum_{i=1}^n \frac{\mathbf{1}(G_i > t-1)}{1-\pi_{it-1}}\frac{\widehat{\pi}_{it-1} - \pi_{it-1}}{1-\widehat{\pi}_{it-1}}\left(\widehat{m}_{i,t|t-1} - m_{i,t|t-1}\right)}_{:=Q_6} \\
&\quad \underbrace{+\frac{1}{n_{t-1}}\sum_{i=1}^n \mathbf{1}(G_i > t)\left(\frac{\widehat{\pi}_{it-1}}{1-\widehat{\pi}_{it-1}}\frac{1}{1-\widehat{\pi}_{it}} - \frac{\pi_{it-1}}{1-\pi_{it-1}}\frac{1}{1-\pi_{it}}\right)\left(\widehat{m}_{i,t|t} - m_{i,t|t}\right)}_{:=Q_7}.
\end{aligned}
$$

Applying the same argument used to derive (D.9) and (D.10), we conclude that

$$Q_1, Q_2, Q_3 = o_P\left(1\sqrt{n}\right). \tag{D.21}$$

For Q_4, by selection mechanism in Assumption 6.1 and mean-independence condition in Assumption 6.2, we have $\mathbb{E}\left(m_{i,t|t} - m_{i,t|t-1} \mid \{X_{it-1}\}_{i=1}^n, \mathcal{F}_{t-2}, \left\{\frac{\mathbf{1}(G_i>t-1)(\widehat{\pi}_{it-1}-\pi_{it-1})}{(1-\pi_{it-1})(1-\widehat{\pi}_{it-1})}\right\}_{G_i>t-1}\right) = 0$. Therefore, Applying the same argument used to derive (D.9) and (D.10), we conclude that

$$Q_4 = o_P\left(1\sqrt{n}\right). \tag{D.22}$$

To derive the bound for Q_5, Q_6, Q_7, applying the same argument used to derive (D.9)

and (D.10), we conclude that

$$Q_5, Q_6, Q_7 = O_P\left(h_m h_\pi + h_m \left(n h_\pi^{d_\alpha + d_X}\right)^{-1/2} \sqrt{\log(n)}\right) + o_P(1/\sqrt{n}). \tag{D.23}$$

Combining (D.21)-(D.23) completes the proof for dynamic ATT. ■

Lemma D.2 *Under the conditions in Theorem 4, for any $t \geqslant t' \geqslant T_0$, $h = h_m, h_p$, and for any $s \in \{1,2,3\}$, there exist constants (depending on t and t') $0 < \underline{C}(t,t'+1) < \overline{C}(t,t'+1) < \infty$ such that the following inequalities*

$$\underline{C}(t,t'+1) n h^{d_\alpha + d_X} \leqslant \min_{i=1,\ldots,n} \sum_{j \in \mathcal{I}_s} \widehat{W}_{h,ij}^{(t,t'+1)} \leqslant \max_{i=1,\ldots,n} \sum_{j \in \mathcal{I}_s} \widehat{W}_{h,ij}^{(t,t'+1)} \leqslant \bar{C}(t,t'+1) n h^{d_\alpha + d_X} \tag{D.24}$$

hold with probability approaching to 1.

Proof of Lemma D.2. Define $q_{j,t|t'+1} := \mathbb{P}_T\left(G_j > t \mid H_{jt'}, X_{jt'+1}, \alpha_j, G_j > t'\right)$. We write

$$\begin{aligned}
\sum_{j \in \mathcal{I}_s} \widehat{W}_{h,ij}^{(t,t'+1)} = & \underbrace{\sum_{j \in \mathcal{I}_s} q_{j,t|t'+1} \mathbf{1}(G_j > t') K_{h,ij}^{(t'+1)}}_{:= A_{1i}} \\
& + \underbrace{\sum_{j \in \mathcal{I}_s} q_{j,t|t'+1} \mathbf{1}(G_j > t') (\widehat{K}_{h,ij}^{(t'+1)} - K_{h,ij}^{(t'+1)})}_{:= A_{2i}} \\
& + \underbrace{\sum_{j \in \mathcal{I}_s} \left(\mathbf{1}(G_j > t) - q_{j,t|t'+1}\right) \mathbf{1}(G_j > t') \widehat{K}_{h,ij}^{(t'+1)}}_{A_{3i}}
\end{aligned} \tag{D.25}$$

Bound for A_{1i} Fix $(\Gamma_T, \alpha_i, X_{it'+1})$. Then, by the compact condition in Assumption 6.5(iii) and the kernel function is bounded and compact supported (see Assumption 6.5(vi)) it is straightforward to verify that

$$\max_{i=1,\ldots,n, j \in \mathcal{I}_s} \left| q_{j,t|t'+1} \mathbf{1}(G_j > t') K_{h,ij}^{(t'+1)} \right| \leqslant \overline{K},$$

and there exists a constant $v > 0$ such that

$$\max_{i=1,\ldots,n, j \in \mathcal{I}_s} \mathbb{E}\left(q_{j,t|t'+1}^2 \mathbf{1}(G_j > t') \left(K_{h,ij}^{(t'+1)}\right)^2 \mid \Gamma_T, \alpha_i, X_{it'+1} \right) \leqslant v h^{d_\alpha + d_X}.$$

Since individuals are independent conditional on Γ_T by Assumption 6.5(i), for each i, we apply Bernstein's inequality for independent random variables (Lemma F.3) to show that

there exists a constant $M > 0$ (independent of i) such that, for any $\epsilon > 0$,

$$\mathbb{P}\left(|A_{1i} - \mathbb{E}\left(A_{1i} \mid \Gamma_T, \alpha_i, X_{it'+1}\right)| \geqslant \epsilon \mid \Gamma_T, \alpha_i, X_{it'+1}\right) \leqslant 2\exp\left(\frac{-\epsilon^2}{M^2 n h^{d_\alpha + d_X} + \frac{1}{3}M\epsilon}\right).$$

Since the inequality above does not depend on realization of $(\alpha_i, X_{it'+1})$, we integrate over $(\alpha_i, X_{it'+1})$ and obtain

$$\mathbb{P}\left(|A_{1i} - \mathbb{E}\left(A_{1i} \mid \Gamma_T, \alpha_i, X_{it'+1}\right)| \geqslant \epsilon \mid \Gamma_T\right) \leqslant 2\exp\left(\frac{-\epsilon^2}{M^2 n h^{d_\alpha + d_X} + \frac{1}{3}M\epsilon}\right).$$

Therefore, we obtain

$$\begin{aligned}
\mathbb{P}\left(\max_{i=1,\ldots,n} |A_{1i} - \mathbb{E}\left(A_{1i} \mid \Gamma_T, \alpha_i, X_{it'+1}\right)| \geqslant \epsilon \mid \Gamma_T\right) =& \mathbb{P}\left(\bigcup_{i=1}^{n} \{|A_{1i} - \mathbb{E}\left(A_{1i} \mid \Gamma_T, \alpha_i, X_{it'+1}\right)| \geqslant \epsilon\} \mid \Gamma_T\right) \\
\leqslant& \sum_{i=1,\ldots,n} \mathbb{P}\left(|A_{1i} - \mathbb{E}\left(A_{1i} \mid \Gamma_T, \alpha_i, X_{it'+1}\right)| \geqslant \epsilon \mid \Gamma_T\right) \\
\overset{\text{(i)}}{=}& \sum_{i=1}^{n} 2\exp\left(\frac{-\epsilon^2}{M^2 n h^{d_\alpha + d_X} + \frac{1}{3}M\epsilon}\right) \\
=& 2n\exp\left(\frac{-\epsilon^2}{M^2 n h^{d_\alpha + d_X} + \frac{1}{3}M\epsilon}\right),
\end{aligned}$$

where (i) follows from Bernstein's inequality. It follows that there exists a sufficiently large constant $D_1 > 0$ such that when $\epsilon = D_1\sqrt{n h^{d_\alpha + d_X}\log(n)}$,

$$\lim_{n\to\infty} \mathbb{P}\left(\max_{i=1,\ldots,n} |A_{1i} - \mathbb{E}\left(A_{1i} \mid \Gamma_T, \alpha_i, X_{it'+1}\right)| \geqslant D_1\sqrt{n h^{d_\alpha + d_X}\log(n)} \mid \Gamma_T\right) = 0 \qquad \text{(D.26)}$$

For the bounds on $\min_{i=1,\ldots,n} \mathbb{E}\left(A_{1i} \mid \Gamma_T, \alpha_i, X_{it'+1}\right)$ and $\max_{i=1,\ldots,n} \mathbb{E}\left(A_{1i} \mid \Gamma_T, \alpha_i, X_{it'+1}\right)$. Since $|\mathcal{I}_s| = n/3$, we have

$$\mathbb{E}\left(A_{1i} \mid \Gamma_T, \alpha_i, X_{it'+1}\right) = \frac{n}{3}\mathbb{E}\left(q_{j,t|t'+1}\mathbf{1}(G_j > t')K_{h,ij}^{(t'+1)} \mid \Gamma_T, \alpha_i, X_{it'+1}\right). \qquad \text{(D.27)}$$

Note that

$$\begin{aligned}
&\min_{i=1,\dots,n} \mathbb{E}\left(q_{j,t|t'+1}\mathbf{1}(G_j > t')K_{h,ij}^{(t'+1)} \mid \Gamma_T, \alpha_i, X_{it'+1}\right) \\
&\overset{\text{(i)}}{\geqslant}(1-\overline{p})^{t-t'} \min_{i=1,\dots,n} \mathbb{E}\left(\mathbf{1}(G_j > t')K_{h,ij}^{(t'+1)} \mid \Gamma_T, \alpha_i, X_{it'+1}\right) \\
&=(1-\overline{p})^{t-t'} \min_{i=1,\dots,n} \mathbb{E}\left(K_{h,ij}^{(t'+1)} \mid \Gamma_T, \alpha_i, X_{it'+1}, G_j > t'\right) \mathbb{P}\left(G_j > t' \mid \Gamma_T, \alpha_i, X_{it'+1}\right) \\
&\overset{\text{(ii)}}{\geqslant}(1-\overline{p})^{t-T_0} \min_{i=1,\dots,n} \mathbb{E}\left(K_{h,ij}^{(t'+1)} \mid \Gamma_T, \alpha_i, X_{it'+1}, G_j > t'\right),
\end{aligned} \tag{D.28}$$

where inequalities (i) and (ii) follow from the common overlap condition (Assumption 6.5(v)). In addition, we have

$$\begin{aligned}
&\min_{i=1,\dots,n} \mathbb{E}\left(K_{h,ij}^{(t'+1)} \mid \Gamma_T, \alpha_i, X_{it'+1}, D_j > t'\right) \\
&= \min_{i=1,\dots,n} \int K\left(\frac{X_{jt'+1} - X_{it'+1}}{h}\right) K\left(\frac{d(\alpha_i,\alpha_j)}{h}\right) \mathrm{d}\mathbb{P}((\alpha_j, X_{jt'+1}) \mid \Gamma_T, G_j > t') \\
&\overset{\text{(i)}}{=} \min_{i=1,\dots,n} \int_{\substack{\|X_{jt'+1}-X_{it'+1}\|_\infty, \\ d(\alpha_j-\alpha_i)\leqslant h}} K\left(\frac{X_{jt'+1} - X_{it'+1}}{h}\right) K\left(\frac{d(\alpha_i,\alpha_j)}{h}\right) \mathrm{d}\mathbb{P}((\alpha_j, X_{it'+1}) \mid \Gamma_T, G_j > t') \\
&\overset{\text{(ii)}}{\geqslant} \min_{i=1,\dots,n} \int_{\substack{\|X_{jt'+1}-X_{it'+1}\|_\infty, \\ d(\alpha_j-\alpha_i)\leqslant \frac{K(0)}{2L_K}h}} K\left(\frac{X_{jt'+1} - X_{it'+1}}{h}\right) K\left(\frac{d(\alpha_i,\alpha_j)}{h}\right) \mathrm{d}\mathbb{P}((\alpha_j, X_{jt'+1}) \mid \Gamma_T, G_j > t') \\
&\overset{\text{(iii)}}{\geqslant} \underbrace{\left(\min_{|u|\leqslant \frac{K(0)}{2L_K}} K(u)\right)^{p+1}}_{:=\tilde{K}} \min_{i=1,\dots,n} \mathbb{P}\left(\|X_{jt'+1} - X_{it'+1}\|_\infty, d(\alpha_j-\alpha_i) \leqslant \frac{K(0)}{2L_K}h \mid \Gamma_T, \alpha_i, X_{it'+1}, G_j > t'\right) \\
&\overset{\text{(iv)}}{\geqslant} \tilde{K} \min_{i=1,\dots,n} \mathbb{P}\left(\|X_{jt'+1} - X_{it'+1}\|_\infty \leqslant \frac{K(0)}{2L_K}h, \|\alpha_j-\alpha_i\| \leqslant \frac{\underline{\eta}K(0)}{2L_K}h \mid \Gamma_T, \alpha_i, X_{it'+1}, G_j > t'\right) \\
&\overset{\text{(v)}}{\geqslant} \underbrace{\underline{c}_1\underline{c}_2\tilde{K}\left(\frac{\min\{1,\underline{\eta}\}K(0)}{2L_K}\right)^{d_\alpha+d_X}}_{:=C_1} h^{d_\alpha+d_X}.
\end{aligned} \tag{D.29}$$

Here, (i) holds because $K(\cdot)$ is supported on $[-1,1]$ by Assumption 6.5(vi), (ii) and (iii) follow from $K(0) > 0$ and the Lipschitz continuity of $K(\cdot)$ (Assumption 6.5(vi)), which together ensure that $K(\cdot)$ is strictly positive on $[-K(0)/2L_K, K(0)/2L_K]$. Inequality (iv) follows from the informativeness condition (see Assumption 5.2). Finally, (v) follows from Assumption 6.5(iii). Then, by (D.27), (D.28), and (D.29), we have

$$\min_{i=1,\dots,n} \mathbb{E}\left(A_{1i} \mid \Gamma_T, \alpha_i, X_{it'+1}\right) \geqslant \underbrace{\frac{C_1}{3}(1-\overline{p})^{t-T_0}}_{:=C_1(t,t'+1)} nh^{d_\alpha+d_X} = C_1(t,t'+1)nh^{d_\alpha+d_X}. \tag{D.30}$$

To establish the upper bound $\max_{i\in\mathcal{I}_{s_1}} \mathbb{E}\left(A_{1i} \mid \Gamma_T, \alpha_i, X_{it'+1}\right)$, note that

$$\begin{aligned}
&\max_{i=1,\ldots,n} \mathbb{E}\left(q_{j,t|t'+1}\mathbf{1}(G_j > t')K_{h,ij}^{(t'+1)} \mid \Gamma_T, \alpha_i, X_{it'+1}\right) \\
\overset{\text{(i)}}{\leqslant}&(1-\underline{p})^{t-t'} \max_{i=1,\ldots,n} \mathbb{E}\left(\mathbf{1}(G_j > t')K_{h,ij}^{(t'+1)} \mid \Gamma_T, \alpha_i, X_{it'+1}\right) \\
=&(1-\underline{p})^{t-t'} \max_{i=1,\ldots,n} \mathbb{E}\left(K_{h,ij}^{(t'+1)} \mid \Gamma_T, \alpha_i, X_{it'+1}, G_j > t'\right) \mathbb{P}\left(G_j > t' \mid \Gamma_T, \alpha_i, X_{it'+1}\right) \\
\overset{\text{(ii)}}{\leqslant}&(1-\underline{p})^{t-T_0} \max_{i=1,\ldots,n} \mathbb{E}\left(K_{h,ij}^{(t'+1)} \mid \Gamma_T, \alpha_i, X_{it'+1}, G_j > t'\right),
\end{aligned} \tag{D.31}$$

where inequalities (i) and (ii) follow from the common overlap condition (Assumption 6.5(v)). In addition, we have

$$\begin{aligned}
&\max_{i=1,\ldots,n} \mathbb{E}\left(K_{h,ij}^{(t'+1)} \mid \Gamma_T, \alpha_i, X_{it'+1}, D_j > t'\right) \\
=&\max_{i=1,\ldots,n} \int K\left(\frac{X_{jt'+1} - X_{it'+1}}{h}\right) K\left(\frac{d(\alpha_i,\alpha_j)}{h}\right) \mathrm{d}\mathbb{P}((\alpha_j, X_{jt'+1}) \mid \Gamma_T, G_j > t') \\
\overset{\text{(i)}}{=}&\max_{i=1,\ldots,n} \int_{\substack{\|X_{jt'+1}-X_{it'+1}\|_\infty \leqslant h, \\ d(\alpha_j-\alpha_i)\leqslant h}} K\left(\frac{X_{jt'+1} - X_{it'+1}}{h}\right) K\left(\frac{d(\alpha_i,\alpha_j)}{h}\right) \mathrm{d}\mathbb{P}((\alpha_j, X_{jt'+1}) \mid \Gamma_T, G_j > t') \\
\overset{\text{(ii)}}{\leqslant}&\overline{K}^{p+1} \max_{i=1,\ldots,n} \int_{\|X_{jt'+1}-X_{it'+1}\|_\infty, d(\alpha_j-\alpha_i)\leqslant h} \mathrm{d}\mathbb{P}((\alpha_j, X_{jt'+1}) \mid \Gamma_T, G_j > t') \\
\overset{\text{(iii)}}{\leqslant}&\overline{K}^{p+1} \max_{i=1,\ldots,n} \mathbb{P}\left(\|X_{jt'+1} - X_{it'+1}\|_\infty \leqslant h, \|\alpha_j - \alpha_i\| \leqslant \overline{\eta} h \mid \Gamma_T, \alpha_i, X_{it'+1}, G_j > t'\right) \\
\overset{\text{(iv)}}{\leqslant}&\underbrace{\bar{c}_1\bar{c}_2\overline{K}^{p+1}(1+\overline{\eta})^{d_\alpha+d_X}}_{:=C_2} h^{d_\alpha+d_X}.
\end{aligned} \tag{D.32}$$

Here, (i) holds because $K(\cdot)$ is supported on $[-1,1]$ by Assumption 6.5(vi), (ii) follows from the boundedness kernel function (Assumption 6.5(vi)), and (iii) follows from the information condition (see Assumption 5.2). Finally, inequality (iv) holds because of Assumption 6.5(iii). Then, by (D.27), (D.31), and (D.32), we have

$$\max_{i=1,\ldots,n} \mathbb{E}\left(A_{1i} \mid \Gamma_T, \alpha_i, Y_{it'}^{(p)}\right) \leqslant \underbrace{\frac{C_2}{3}(1-\underline{p})^{t-T_0}}_{:=C_2(t,t'+1)} nh^{d_\alpha+d_X} = C_2(t,t'+1)nh^{d_\alpha+d_X}. \tag{D.33}$$

Therefore, by (D.26), (D.30), and (D.33), we have that

$$\lim_{n\to\infty}\mathbb{P}\Bigg(C_1(t,t'+1)nh^{d_\alpha+d_X}-D_1\sqrt{nh^{d_\alpha+d_X}\log(n)}\leqslant\min_{i=1,\ldots,n}A_{1i}\leqslant\max_{i=1,\ldots,n}A_{1i}\\ \leqslant C_2(t,t'+1)nh^{d_\alpha+d_X}+D_1\sqrt{nh^{d_\alpha+d_X}\log(n)}\mid\Gamma_T\Bigg)=0$$

Since $nh^{d_\alpha+d_X}\log(n)\to\infty$ as required in Theorem 4, we have

$$\lim_{n\to\infty}\mathbb{P}\left(\frac{C_1(t,t'+1)}{2}nh^{d_\alpha+d_X}\leqslant\min_{ii=1,\ldots,n}A_{1i}\leqslant\max_{i=1,\ldots,n}A_{1i}\leqslant 2C_2(t,t'+1)nh^{d_\alpha+d_X}\right)=0 \tag{D.34}$$

Bound for A_{2i} Note that $\max_{i=1,\ldots,n}|A_{2i}|$ is bounded by

$$\begin{aligned}&\max_{i=1,\ldots,n}\sum_{j\in\mathcal{I}_s}\mathbf{1}(G_j>t')K\left(\frac{X_{jt'+1}-X_{it'+1}}{h}\right)\left|K\left(\frac{\widehat{d}_{ij}}{h}\right)-K\left(\frac{d(\alpha_i,\alpha_j)}{h}\right)\right|\\ &\overset{\text{(i)}}{\leqslant}\max_{i=1,\ldots,n}\sum_{j\in\mathcal{I}_s}\mathbf{1}(G_j>t')\mathbf{1}\left(\widehat{d}_{ij}\leqslant h\vee d(\alpha_i,\alpha_j)\leqslant h\right)K\left(\frac{X_{jt'+1}-X_{it'+1}}{h}\right)L_K\frac{|\widehat{d}_{ij}-d(\alpha_i,\alpha_j)|}{h},\end{aligned}$$

where (i) follows from that the kernel function is Lipschitz continuous and has bounded support (Assumption 6.5(vi)). Then, by informativeness condition (Assumption 5.2), we have that, wpa1,

$$\begin{aligned}\max_{i=1,\ldots,n}|A_{2i}|\leqslant&\frac{\delta_{n,T_0}}{h}\max_{i=1,\ldots,n}\sum_{j\in\mathcal{I}_s}\mathbf{1}(G_j>t')\mathbf{1}\left(\widehat{d}_{ij}\leqslant h\vee d(\alpha_i,\alpha_j)\leqslant h\right)K\left(\frac{X_{jt'+1}-X_{it'+1}}{h}\right)\\ \overset{\text{(i)}}{\leqslant}&\frac{\delta_{n,T_0}}{h}\max_{i=1,\ldots,n}\underbrace{\sum_{j\in\mathcal{I}_s}\mathbf{1}(G_j>t')\mathbf{1}\left(\|\alpha_i-\alpha_j\|\leqslant 2\overline{\eta}h\right)K\left(\frac{X_{jt'+1}-X_{it'+1}}{h}\right)}_{:=B_{2i}},\end{aligned} \tag{D.35}$$

where (i) follows from informativeness condition (Assumption 5.2) and Lemma D.1. We employ the same argument used to prove (D.26) to show that there exists a sufficiently large constant $D_2>0$ such that

$$\lim_{n\to\infty}\mathbb{P}\left(\max_{i=1,\ldots,n}|B_{2i}-\mathbb{E}\left(B_{2i}\mid\Gamma_T,\alpha_i,X_{it'+1}\right)|\geqslant D_2\sqrt{nh^{d_\alpha+d_X}\log(n)}\mid\Gamma_T\right)=0. \tag{D.36}$$

Also, the same argument used to prove (D.33), we have

$$\max_{i=1,\ldots,n} \mathbb{E}\left(B_{2i} \mid \Gamma_T, \alpha_i, X_{it'+1}\right) \leqslant \underbrace{\frac{1}{3}(1-\overline{p})^{t'-T_0}\overline{c}_1\overline{c}_2\overline{K}^{p+1}\left(1+2\overline{\eta}\right)^{d_\alpha+d_X}}_{:=C_3(t,t'+1)} nh^{d_\alpha+d_X}. \tag{D.37}$$

Therefore, combining (D.35), (D.36), (D.37), and $nh^{d_\alpha+d_X}/\log(n) \to \infty$ as required in Theorem 4, we have

$$\lim_{n,T_0\to\infty} \mathbb{P}\left(\max_{i=1,\ldots,n} |A_{2i}| \leqslant 2C_3(t,t'+1)nh^{d_\alpha+d_X}\frac{\delta_{n,T_0}}{h}\right) = 0 \tag{D.38}$$

Bound for A_{3i} Conditional on $\{X_{it'+1}\}_{i=1}^n$ and $\mathcal{F}_{t'}$, it is easy to check that $\widehat{K}_{h,ij}^{(t'+1)}$ is deterministic. In addition, by Assumption 6.1, we obtain

$$\mathbb{E}\left(\left(\mathbf{1}(G_j > t) - q_{j,t|t'+1}\right)\mathbf{1}(G_j > t')\widehat{K}_{h,ij}^{(t'+1)} \mid \{X_{it'+1}\}_{i=1}^n, \mathcal{F}_{t'}\right) = 0.$$

By the compact condition (see Assumption 6.5(iii)) and the kernel function is bounded and compact supported (see Assumption 6.5(vi)) it is straightforward to verify that

$$\max_{i=1,\ldots,n, j\in\mathcal{I}_s} \left|\left(\mathbf{1}(G_j > t) - q_{j,t|t'+1}\right)\mathbf{1}(G_j > t')\widehat{K}_{h,ij}^{(t'+1)}\right| \leqslant \overline{K},$$

and there exists a constant $v > 0$ such that

$$\max_{\substack{i=1,\ldots,n,\\ j\in\mathcal{I}_s}} \mathbb{E}\left(\left(\mathbf{1}(G_j > t) - q_{j,t|t'+1}\right)^2\mathbf{1}(G_j > t')\left(\widehat{K}_{h,ij}^{(t'+1)}\right)^2 \mid \{X_{it'+1}\}_{i=1}^n, \mathcal{F}_{t'}\right) \leqslant vh^{d_\alpha+d_X}.$$

Since individuals are independent conditional on $\{X_{it'+1}\}_{i=1}^n$ and $\mathcal{F}_{t'}$ by Assumption 6.5(i), for each i, we apply Bernstein's inequality for independent random variables (Lemma F.3) to show that there exists a constant $M > 0$ (independent of i) such that, for any $\epsilon > 0$,

$$\mathbb{P}\left(|A_{3i}| \geqslant \epsilon \mid \{X_{it'+1}\}_{i=1}^n, \mathcal{F}_{t'}\right) \leqslant 2\exp\left(\frac{-\epsilon^2}{M^2nh^{d_\alpha+d_X} + \frac{1}{3}M\epsilon}\right).$$

Since the inequality above does not depend on realization of $\{\{\alpha_i\}_{i,1\ldots,n}, \{Y_{i\tau}, X_{i\tau}\}_{i=1,\ldots,n,\tau\leqslant t'}\}$ and $\{X_{it'+1}\}_{i=1}^n$, we integrate over $\{\{\alpha_i\}_{i,1\ldots,n}, \{Y_{i\tau}, X_{i\tau}\}_{i=1,\ldots,n,\tau\leqslant t'}\}$ and $\{X_{it'+1}\}_{i=1}^n$ to obtain

$$\mathbb{P}\left(|A_{3i}| \geqslant \epsilon \mid \Gamma_T\right) \leqslant 2\exp\left(\frac{-\epsilon^2}{M^2nh^{d_\alpha+d_X} + \frac{1}{3}M\epsilon}\right).$$

Therefore, we obtain

$$\begin{aligned}\mathbb{P}\left(\max_{i=1,\ldots,n}|A_{3i}|\geqslant\epsilon\mid\Gamma_T\right)=&\mathbb{P}\left(\bigcup_{i=1,\ldots,n}\{|A_{3i}|\geqslant\epsilon\}\mid\Gamma_T\right)\\&\leqslant\sum_{i=1}^{n}\mathbb{P}\left(|A_{3i}|\geqslant\epsilon\mid\Gamma_T\right)\\&\overset{\text{(i)}}{=}\sum_{i=1}^{n}2\exp\left(\frac{-\epsilon^2}{M^2nh^{d_\alpha+d_X}+\frac{1}{3}M\epsilon}\right)\\&=2n\exp\left(\frac{-\epsilon^2}{M^2nh^{d_\alpha+d_X}+\frac{1}{3}M\epsilon}\right),\end{aligned}$$

where (i) follows from Bernstein's inequality. It follows that there exists a sufficiently large constant $D_3>0$ such that when $\epsilon=D_3\sqrt{nh^{d_\alpha+d_X}\log(n)}$,

$$\lim_{n\to\infty}\mathbb{P}\left(\max_{i=1,\ldots,n}|A_{3i}|\geqslant D_3\sqrt{nh^{d_\alpha+d_X}\log(n)}\mid\Gamma_T\right)=0 \tag{D.39}$$

Therefore, combining (D.34), (D.38), and (D.39),

$$\begin{aligned}\lim_{n,T_0\to\infty}\mathbb{P}\Bigg(&\frac{C_1(t,t'+1)}{2}nh^{d_\alpha+d_X}-2C_3(t,t'+1)nh^{d_\alpha+d_X}\frac{\delta_{n,T_0}}{h}-D_3\sqrt{nh^{d_\alpha+d_X}\log(n)}\\&\leqslant\min_{i=1,\ldots,n}\sum_{j\in\mathcal{I}_s}\widehat{W}_{h,ij}^{(t,t'+1)}\leqslant\max_{i=1,\ldots,n}\sum_{j\in\mathcal{I}_s}\widehat{W}_{h,ij}^{(t,t'+1)}\leqslant\\&2C_2(t,t'+1)nh^{d_\alpha+d_X}+2C_3(t,t'+1)nh^{d_\alpha+d_X}\frac{\delta_{n,T_0}}{h}+D_3\sqrt{nh^{d_\alpha+d_X}\log(n)}\mid\Gamma_T\Bigg)=0\end{aligned}$$

Since $nh^{d_\alpha+d_X}/\log(n)\to\infty$ and $h/\delta_{n,T_0}\to\infty$, we have

$$\begin{aligned}\lim_{n,T_0\to\infty}\mathbb{P}\Bigg(&\frac{C_1(t,t'+1)}{4}nh^{d_\alpha+d_X}\leqslant\min_{i=1,\ldots,n}\sum_{j\in\mathcal{I}_s}\widehat{W}_{h,ij}^{(t,t'+1)}\\&\leqslant\max_{i=1,\ldots,n}\sum_{j\in\mathcal{I}_s}\widehat{W}_{h,ij}^{(t,t'+1)}\leqslant4C_2(t,t'+1)nh^{d_\alpha+d_X}\mid\Gamma_T\Bigg)=0\end{aligned}$$

Let $\underline{C}(t,t'+1):=\frac{1}{4}C_1(t,t'+1)$ and $\overline{C}(t,t'+1):=4C_2(t,t'+1)$. We complete the proof. ■

E Proof for Pseudo distance and Informativeness

E.1 Proof for Lemma B.2

Proof of Lemma B.2. Let $\rho_g = \sup_{\alpha,\Gamma} |g(\alpha, \Gamma)|$. It is straightforward to see that $\rho_g < \infty$ by Assumpion B.1(iii). For notationally simplicity, for any $\alpha_1, \alpha_2 \in \mathcal{A}$ and any $\Gamma \in \text{supp}(\Gamma_T)$, define $\Delta(\alpha_1, \alpha_2, \Gamma) := (g(\alpha_1, \Gamma) - g(\alpha_2, \Gamma))$ and

$$\varphi(\alpha_1, \alpha_2, \alpha_3, \alpha_4) := \int \Delta(\alpha_1, \alpha_2, \Gamma)\Delta(\alpha_3, \alpha_4, \Gamma)\mathrm{d}\mathbb{P}(\Gamma).$$

Recall that

$$\widehat{d}_{ij} = \max_{k_1,k_2 \in \{1,\ldots,n\}\setminus\{i,j\}} \frac{1}{T_0} \sum_{t=0}^{T_0-1} (Y_{k_1 t} - Y_{k_2 t})(Y_{it} - Y_{jt}),$$

and

$$d(\alpha_i, \alpha_j) = d(\alpha_i, \alpha_j) := \sup_{\alpha_1,\alpha_2 \in \mathcal{A}} \left| \int \Delta(\alpha_1, \alpha_2, \Gamma)\Delta(\alpha_i, \alpha_j, \Gamma)\mathrm{d}\mathbb{P}(\Gamma) \right|$$

In addition, for notational simplicity, define

$$\bar{d}_{ij} := \max_{k_1,k_2 \in \{1,\ldots,n\}\setminus\{i,j\}} \frac{1}{T_0} \sum_{t=0}^{T_0-1} \Delta(\alpha_{k_1}, \alpha_{k_2}, \Gamma_t)\Delta(\alpha_i, \alpha_j, \Gamma_t).$$

Then

$$\max_{i,j \in \{1,\ldots,n\}} |\widehat{d}_{ij} - d(\alpha_i, \alpha_j)| \leqslant \max_{i,j \in \{1,\ldots,n\}} |\widehat{d}_{ij} - \bar{d}_{ij}| + \max_{i,j \in \{1,\ldots,n\}} |\bar{d}_{ij} - d(\alpha_i, \alpha_j)|.$$

Bound for $\max_{i,j\in\{1,\dots,n\}} |\widehat{d}_{ij} - \bar{d}_{ij}|$ Note that

$$\begin{aligned}\max_{i,j\in\{1,\dots,n\}} \left|\widehat{d}_{ij} - \bar{d}_{ij}\right| = & \underbrace{\max_{i,j\in\{1,\dots,n\}} \max_{k_1,k_2\in\{1,\dots,n\}\setminus\{i,j\}} \left|\frac{1}{T_0}\sum_{t=0}^{T_0-1} \Delta(\alpha_{k_1},\alpha_{k_2},\Gamma_t)(\tilde{\epsilon}_{it}-\tilde{\epsilon}_{jt})\right|}_{A_1} \\ & + \underbrace{\max_{i,j\in\{1,\dots,n\}} \max_{k_1,k_2\in\{1,\dots,n\}\setminus\{i,j\}} \left|\frac{1}{T_0}\sum_{t=0}^{T_0-1} \Delta(\alpha_{i},\alpha_{j},\Gamma_t)(\tilde{\epsilon}_{k_1t}-\tilde{\epsilon}_{k_1t})\right|}_{A_2} \\ & + \underbrace{\max_{i,j\in\{1,\dots,n\}} \max_{k_1,k_2\in\{1,\dots,n\}\setminus\{i,j\}} \left|\frac{1}{T_0}\sum_{t=0}^{T_0-1} (\tilde{\epsilon}_{k_1t}-\tilde{\epsilon}_{k_1t})(\tilde{\epsilon}_{it}-\tilde{\epsilon}_{jt})\right|}_{A_3}\end{aligned}$$

It is straightforward to verify that for any constant $\delta > 0$,

$$\begin{aligned}&\mathbb{P}\left(A_1 \geqslant 2\rho_g\delta \mid \{\alpha_i\}_{i=1}^n, \Gamma_T\right) \\ =&\mathbb{P}\left(\bigcup_{i,j\in\{1,\dots,n\}}\bigcup_{k_1,k_2\in\{1,\dots,n\}\setminus\{i,j\}} \left\{\left|\frac{1}{T_0}\sum_{t=0}^{T_0-1} \Delta(\alpha_{k_1},\alpha_{k_2},\Gamma_t)(\tilde{\epsilon}_{it}-\tilde{\epsilon}_{jt})\right| \geqslant 2\rho_g\delta\right\} \mid \{\alpha_i\}_{i=1}^n, \Gamma_T\right) \\ \leqslant& \sum_{i,j\in\{1,\dots,n\}}\sum_{k_1,k_2\in\{1,\dots,n\}\setminus\{i,j\}} \mathbb{P}\left(\left|\frac{1}{T_0}\sum_{t=0}^{T_0-1} \Delta(\alpha_{k_1},\alpha_{k_2},\Gamma_t)(\tilde{\epsilon}_{it}-\tilde{\epsilon}_{jt})\right| \geqslant 2\rho_g\delta \mid \{\alpha_i\}_{i=1}^n, \Gamma_T\right) \\ \overset{\text{(i)}}{\leqslant}& n^4 \max_{\substack{i,j\in\{1,\dots,n\},\\ k_1,k_2\in\{1,\dots,n\}\setminus\{i,j\}}} \mathbb{P}\left(\left|\frac{1}{T_0}\sum_{t=0}^{T_0-1} (\tilde{\epsilon}_{it}-\tilde{\epsilon}_{jt})\right| \geqslant \delta \mid \{\alpha_i\}_{i=1}^n, \Gamma_T\right),\end{aligned}$$

where (i) follows from $|\Delta(\alpha_{k_1},\alpha_{k_2},\Gamma_t)| \leqslant 2\rho_g$.

Then, since $\mathbb{E}\left(\Delta(\alpha_{k_1},\alpha_{k_2},\Gamma_t)(\tilde{\epsilon}_{it}-\tilde{\epsilon}_{jt}) \mid \{\alpha_i\}_{i=1}^n, \Gamma_T\right) = 0$, by Assumption 5.1(i) and (iii), we employ Lemma F.1 to show that there exist positive constants C_1, C_2, C_3, C_4 that do not depend on (n, T_0), such that

$$\begin{aligned}&\max_{\substack{i,j\in\{1,\dots,n\},\\ k_1,k_2\in\{1,\dots,n\}\setminus\{i,j\}}} \mathbb{P}\left(\left|\frac{1}{T_0}\sum_{t=0}^{T_0-1} (\tilde{\epsilon}_{it}-\tilde{\epsilon}_{jt})\right| \geqslant \delta \mid \{\alpha_i\}_{i=1}^n, \Gamma_T\right) \\ \leqslant& T_0 \exp\left(-\frac{T_0^{2/3}\delta^{2/3}}{C_1}\right) + \exp\left(-\frac{T_0\delta^2}{C_2}\right) + \exp\left(-\frac{T_0\delta^2}{C_3}\exp\left(\frac{T_0^{2/9}\delta^{2/9}}{C_4\left(\log T_0\right)^{2/3}}\right)\right).\end{aligned}$$

Therefore, there exists a sufficiently large constant $M_1 > 0$ that is independent of n, T_0 such

that when $\delta = M_1\sqrt{\frac{\log(nT_0)}{T_0}}$, we have

$$\mathbb{P}\left(A_1 \geqslant M_1\sqrt{\frac{\log(nT_0)}{T_0}} \mid \{\alpha_i\}_{i=1}^n, \Gamma_T\right) \leqslant 1/\sqrt{T_0}.$$

Since the inequality above does not depend on the realization $(\{\alpha_i\}_{i=1}^n, \Gamma_T)$, we integrate over $(\{\alpha_i\}_{i=1}^n, \Gamma_T)$ to obtain

$$\mathbb{P}\left(A_1 \geqslant M_1\sqrt{\frac{\log(nT_0)}{T_0}}\right) \leqslant 1/\sqrt{T_0}.$$

By the similar argument, we can also show that there exist constants $M_2, M_3 > 0$ such that

$$\mathbb{P}\left(A_2 \geqslant M_2\sqrt{\frac{\log(nT_0)}{T_0}}\right) \leqslant 1/\sqrt{T_0}, \quad \text{and } \mathbb{P}\left(A_3 \geqslant M_3\sqrt{\frac{\log(nT_0)}{T_0}}\right) \leqslant 1/\sqrt{T_0}.$$

Therefore, we obtain

$$\mathbb{P}\left(\max_{i,j\in\{1,\ldots,n\}} |\widehat{d}_{ij} - \bar{d}_{ij}| \geqslant (M_1 + M_2 + M_3)\sqrt{\frac{\log(nT_0)}{T_0}}\right) \leqslant 3/\sqrt{T_0}.$$

Bound for $\max_{i,j\in\{1,\ldots,n\}} |\bar{d}_{ij} - d(\alpha_i, \alpha_j)|$ Note that

$$\begin{aligned}
&|\bar{d}_{ij} - d(\alpha_i, \alpha_j)| \\
\leqslant &\max_{k_1,k_2\in\{1,\ldots,n\}\setminus\{i,j\}} \left|\frac{1}{T_0}\sum_{t=0}^{T_0-1} \Delta(\alpha_{k_1}, \alpha_{k_2}, \Gamma_t)\Delta(\alpha_i, \alpha_j, \Gamma_t) - \varphi(\alpha_{k_1}, \alpha_{k_2}, \alpha_i, \alpha_j)\right| \\
&+ \sup_{\alpha_1,\alpha_2\in\mathcal{A}} |\varphi(\alpha_1, \alpha_2, \alpha_i, \alpha_j)| - \max_{k_1,k_2\in\{1,\ldots,n\}\setminus\{i,j\}} |\varphi(\alpha_{k_1}, \alpha_{k_2}, \alpha_i, \alpha_j)|.
\end{aligned}$$

It follows that $\max_{i,j\in\{1,\ldots,n\}} |\bar{d}_{ij} - d(\alpha_i, \alpha_j)|$ is bounded by

$$\underbrace{\max_{i,j\in\{1,\ldots,n\}} \max_{k_1,k_2\in\{1,\ldots,n\}\setminus\{i,j\}} \left|\frac{1}{T_0}\sum_{t=0}^{T_0-1} \Delta(\alpha_{k_1}, \alpha_{k_2}, \Gamma_t)\Delta(\alpha_i, \alpha_j, \Gamma_t) - \varphi(\alpha_{k_1}, \alpha_{k_2}, \alpha_i, \alpha_j)\right|}_{A_4}$$

$$+ \underbrace{\max_{i,j\in\{1,\ldots,n\}} \left(\sup_{\alpha_1,\alpha_2\in\mathcal{A}} \varphi(\alpha_1, \alpha_2, \alpha_i, \alpha_j) - \max_{k_1,k_2\in\{1,\ldots,n\}\setminus\{i,j\}} \varphi(\alpha_{k_1}, \alpha_{k_2}, \alpha_i, \alpha_j)\right)}_{A_5}.$$

By Assumption B.1(ii), there exists a constant $M_4 > 0$ such that $A_4 \leqslant M_4\sqrt{\frac{\log(T_0)}{T_0}}$ wpa1. To establish an error bound for A_5, let (α_1^*, α_2^*) be the maximizer of $d(\alpha_i, \alpha_j)$ such that

$$d(\alpha_i, \alpha_j) = \varphi\left(\alpha_1^*, \alpha_2^*, \alpha_i, \alpha_j\right).$$

Then, we have

$$\begin{aligned}
A_5 \leqslant & \max_{i,j\in\{1,\ldots,n\}} \left(\varphi\left(\alpha_1^*, \alpha_2^*, \alpha_i, \alpha_j\right) - \max_{k_1,k_2\in\{1,\ldots,n\}\setminus\{i,j\}} \varphi\left(\alpha_{k_1}, \alpha_{k_2}, \alpha_i, \alpha_j\right)\right) \\
= & \max_{i,j\in\{1,\ldots,n\}} \min_{k_1,k_2\in\{1,\ldots,n\}\setminus\{i,j\}} \left(\varphi\left(\alpha_1^*, \alpha_2^*, \alpha_i, \alpha_j\right) - \varphi\left(\alpha_{k_1}, \alpha_{k_2}, \alpha_i, \alpha_j\right)\right) \\
\leqslant & \max_{i,j\in\{1,\ldots,n\}} \min_{k_1,k_2\in\{1,\ldots,n\}\setminus\{i,j\}} 2\rho_g L_g \left(\|\alpha_{k_1} - \alpha_1^*\|_\infty + \|\alpha_{k_2} - \alpha_2^*\|_\infty\right) \\
\leqslant & \max_{i,j\in\{1,\ldots,n\}} \min_{k_1,k_2\in\{1,\ldots,n\}\setminus\{i,j\}} 2\rho_g L_g d_\alpha^{1/2} \left(\|\alpha_{k_1} - \alpha_1^*\| + \|\alpha_{k_2} - \alpha_2^*\|\right) \\
\leqslant & 4\rho_g L_g d_\alpha^{1/2} \max_{i,j\in\{1,\ldots,n\}} \sup_{\alpha\in\mathcal{A}} \min_{k\in\{1,\ldots,n\}\setminus\{i,j\}} \|\alpha_k - \alpha\|
\end{aligned}$$

By Assumption B.1(iv), there exists a δ-covering of the support of α consisting of $\mathcal{J} \leqslant 4^{d_\alpha} d_\alpha^{d_\alpha/2} \delta^{-d_\alpha}$ open balls with radius δ, denoted by $\{\bar{\alpha}_\ell\}_{\ell=1}^{\mathcal{J}}$. We obtain

$$\sup_{\alpha\in\mathcal{A}} \min_{k\in\{1,\ldots,n\}\setminus\{i,j\}} \|\alpha_k - \alpha\| \leqslant \max_{\ell=1,\ldots,\mathcal{J}} \min_{k\in\{1,\ldots,n\}\setminus\{i,j\}} \|\alpha_k - \bar{\alpha}_\ell\| + \delta.$$

It follows that, for any $\delta > 0$,

$$
\begin{aligned}
&\mathbb{P}\left(\max_{i,j\in\{1,\ldots,n\}} \sup_{\alpha\in\mathcal{A}} \min_{k\in\{1,\ldots,n\}\setminus\{i,j\}} \|\alpha_k - \alpha\| \geqslant 2\delta\right) \\
=&\mathbb{P}\left(\bigcup_{i,j\in\{1,\ldots,n\}} \left\{\sup_{\alpha\in\mathcal{A}} \min_{k\in\{1,\ldots,n\}\setminus\{i,j\}} \|\alpha_k - \alpha\| \geqslant 2\delta\right\}\right) \\
\leqslant& \sum_{i,j\in\{1,\ldots,n\}} \mathbb{P}\left(\sup_{\alpha\in\mathcal{A}} \min_{k\in\{1,\ldots,n\}\setminus\{i,j\}} \|\alpha_k - \alpha\| \geqslant 2\delta\right) \\
\leqslant& \sum_{i,j\in\{1,\ldots,n\}} \mathbb{P}\left(\max_{\ell=1,\ldots,\mathcal{J}} \min_{k\in\{1,\ldots,n\}\setminus\{i,j\}} \|\alpha_k - \bar{\alpha}_\ell\| \geqslant \delta\right) \\
=& \sum_{i,j\in\{1,\ldots,n\}} \mathbb{P}\left(\bigcup_{\ell=1,\ldots,\mathcal{J}} \left\{\min_{k\in\{1,\ldots,n\}\setminus\{i,j\}} \|\alpha_k - \bar{\alpha}_\ell\| \geqslant \delta\right\}\right) \\
\leqslant& \sum_{i,j\in\{1,\ldots,n\}} \sum_{\ell=1,\ldots,\mathcal{J}} \mathbb{P}\left(\min_{k\in\{1,\ldots,n\}\setminus\{i,j\}} \|\alpha_k - \bar{\alpha}_\ell\| \geqslant \delta\right) \\
\leqslant& \sum_{i,j\in\{1,\ldots,n\}} \sum_{\ell=1,\ldots,\mathcal{J}} \left(\prod_{k\in\{1,\ldots,n\}\setminus\{i,j\}} \mathbb{P}\left(\|\alpha_k - \bar{\alpha}_\ell\| \geqslant \delta\right)\right) \\
\overset{\text{(i)}}{\leqslant}& n^2\mathcal{J}\left(1 - \underline{c}_1\delta^{d_\alpha}\right)^{n-2} \\
\leqslant& 4^{d_\alpha} d_\alpha^{d_\alpha/2} n^2 \delta^{-d_\alpha}\left(1 - \underline{c}\delta^{d_\alpha}\right)^{n-2}
\end{aligned}
$$

Therefore, there exists a sufficiently large constant $M_5 > 0$ such that, when $\delta = M_5\frac{\log(n)}{n^{1/d_\alpha}}$,

$$\mathbb{P}\left(A_5 \geqslant M_5\frac{\log(n)}{n^{1/d_\alpha}}\right) \leqslant 1/\sqrt{n}.$$

Therefore, let $\lambda_1 = M_5$ and $\lambda_2 = M_1 + M_2 + M_3 + M_4$, we have that the following inequality wpa1:

$$\max_{i,j\in\{1,\ldots,n\}} |\widehat{d}_{ij} - d(\alpha_i, \alpha_j)| \leqslant \gamma_1\frac{\log(n)}{n^{1/d_\alpha}} + \gamma_2\sqrt{\frac{\log(nT_0)}{T_0}}.$$

This completes the proof. ■

E.2 Proof for informativeness condition

Proof of Lemma B.3. By backward iteration,

$$g(\alpha, \Gamma_t) = \alpha' v_t, \qquad v_t := \sum_{\tau=0}^{\infty} \rho^\tau \gamma_{t-\tau}.$$

The infinite series converges in mean square because $|\rho| < 1$ and $s_{\max}(\mathbb{E}(\gamma_t \gamma_t')) < \infty$. Moreover, $\{v_t\}_{t\in\mathbb{Z}}$ is stationary and ergodic. Let $\Omega := \mathbb{E}(v_t v_t')$.

We first establish upper and lower bounds for $\mathbb{E}(v_t v_t')$. For any $q \in \mathbb{R}^{d_\alpha}$, the Cauchy–Schwarz inequality and stationarity imply that

$$\begin{aligned} q'\mathbb{E}(v_t v_t')q &= \mathbb{E}\left(\left(\sum_{\tau=0}^{\infty} \rho^\tau q'\gamma_{t-\tau}\right)^2\right) \leqslant \left(\sum_{\tau=0}^{\infty} |\rho|^\tau\right) \sum_{\tau=0}^{\infty} |\rho|^\tau \mathbb{E}\left((q'\gamma_{t-\tau})^2\right) \\ &= (1-|\rho|)^{-2} q'\mathbb{E}(\gamma_t\gamma_t')q. \end{aligned}$$

To obtain the lower bound, note that $\gamma_t = v_t - \rho v_{t-1}$. Hence, again by the Cauchy–Schwarz inequality and stationarity,

$$\begin{aligned} q'\mathbb{E}(\gamma_t\gamma_t')q &= \mathbb{E}\left((q'v_t - \rho q' v_{t-1})^2\right) \\ &= (1+\rho^2) q'\mathbb{E}(v_t v_t')q - 2\rho\mathbb{E}\left((q'v_t)(q'v_{t-1})\right) \\ &\leqslant (1+\rho^2) q'\mathbb{E}(v_t v_t')q + 2|\rho|\sqrt{\mathbb{E}\left((q'v_t)^2\right)\mathbb{E}\left((q'v_{t-1})^2\right)} \\ &= (1+|\rho|)^2 q'\mathbb{E}(v_t v_t')q. \end{aligned}$$

Since these inequalities hold for every $q \in \mathbb{R}^{d_\alpha}$,

$$(1+|\rho|)^{-2}\mathbb{E}(\gamma_t\gamma_t') \preceq \mathbb{E}(v_t v_t') \preceq (1-|\rho|)^{-2}\mathbb{E}(\gamma_t\gamma_t').$$

It follows that $\mathbb{E}(v_t v_t')$ is positive definite and has a finite largest eigenvalue. Then, I substitute the factor representation into the definition of the population pseudo-distance:

$$d(\alpha_i, \alpha_j) = \sup_{\alpha_1,\alpha_2\in\mathcal{A}} \left|\mathbb{E}\left((\alpha_1-\alpha_2)' v_t v_t'(\alpha_i - \alpha_j)\right)\right| = \sup_{\alpha_1,\alpha_2\in\mathcal{A}} \left|(\alpha_1-\alpha_2)'\mathbb{E}(v_t v_t')(\alpha_i-\alpha_j)\right|.$$

This quantity is finite because $\mathcal{A}$ is compact and $s_{\max}(\mathbb{E}(v_t v_t')) < \infty$.

Fix any $\alpha_0 \in \mathcal{A}$. Since $\mathbb{E}(v_t v_t')$ is positive definite, Lemma E.1 implies that there exists a

constant $c > 0$ such that, for all $\alpha_i, \alpha_j \in \mathcal{A}$,

$$d(\alpha_i, \alpha_j) \geqslant \sup_{\alpha \in \mathcal{A}} |(\alpha - \alpha_0)' \mathbb{E}(v_t v_t')(\alpha_i - \alpha_j)| \geqslant c\|\alpha_i - \alpha_j\|.$$

Thus, letting $\overline{\eta} = c^{-1}$ yields $\|\alpha_i - \alpha_j\| \leqslant \overline{\eta} d(\alpha_i, \alpha_j)$, which completes the proof. ■

Proof of Lemma B.4. Recall

$$Y_{it}(\infty) = \mathbf{1}\left(\beta Y_{it-1}(\infty) + \gamma_t' \alpha_i - u_{it} \geqslant 0\right),$$

and let $F(\cdot)$ denote the cumulative distribution function of u_{it}.

It is straightforward to verify that, since $s_{\min}(\mathbb{E}(\gamma_t \gamma_t'))$ is positive definite,

$$\int (g(\alpha_1, \Gamma) - g(\alpha_1, \Gamma))^2 \mathrm{d}\mathbb{P}(\Gamma) > 0, \quad \forall \alpha_1, \alpha_2 \in \mathcal{A}, \text{ and } \alpha_1 \neq \alpha_2. \tag{E.1}$$

Since $\mathcal{A}$ is compact and g is continuous, for any $\delta_1 > 0$ and any $\alpha_1, \alpha_2 \in \mathcal{A}$ such that, there exists a constant $\eta(\delta_1) > 0$ such that

$$\inf_{\alpha_1, \alpha_2 \in \mathcal{A}, \|\alpha_1 - \alpha_2\| \geqslant \delta_1} \int (g(\alpha_1, \Gamma) - g(\alpha_1, \Gamma))^2 \mathrm{d}\mathbb{P}(\Gamma) \geqslant \eta(\delta_1).$$

It follows that

$$\inf_{\alpha_1, \alpha_2 \in \mathcal{A}, \|\alpha_1 - \alpha_2\| \geqslant \delta_1} d(\alpha_1, \alpha_2) \geqslant \inf_{\alpha_1, \alpha_2 \in \mathcal{A}, \|\alpha_1 - \alpha_2\| \geqslant \delta_1} \int (g(\alpha_1, \Gamma) - g(\alpha_1, \Gamma))^2 \mathrm{d}\mathbb{P}(\Gamma) \geqslant \eta(\delta_1). \tag{E.2}$$

For notational simplicity, define $g_t(\alpha_i) := g(\alpha_i, \Gamma_t) = \mathbb{E}\left(Y_{it}(\infty) \mid \alpha_i, \Gamma_t\right)$. In addition, let $\rho_t(\alpha_i) := F\left(\beta + \gamma_t' \alpha_i\right) - F\left(\gamma_t' \alpha_i\right)$ denote the stochastic discount factor. Then,

$$g_t(\alpha_i) = (1 - g_{t-1}(\alpha_i)) F\left(\gamma_t' \alpha_i\right) + g_{t-1}(\alpha_i) F\left(\beta + \gamma_t' \alpha_i\right) = F\left(\gamma_t' \alpha_i\right) + \rho_t(\alpha_i) g_{t-1}(\alpha_i).$$

Taking first derivative with respect to α_i on both sides yields

$$\nabla_\alpha g_t(\alpha_i) = F^{(1)}(\gamma_t' \alpha_i)\gamma_t + g_{t-1}(\alpha_i)\rho_t^{(1)}(\alpha_i)\gamma_t + \rho_t(\alpha_i)\nabla_\alpha g_{t-1}(\alpha_i).$$

Step 2: Positive definiteness of $\mathbb{E}\left(\nabla_\alpha g_t(\alpha_i) \nabla_\alpha g_t(\alpha_i)'\right)$ The critical step for the proof is to show that $\mathbb{E}\left(\nabla_\alpha g_t(\alpha_i) \nabla_\alpha g_t(\alpha_i)'\right)$ is positive definite. The expectation is taken over the

distribution of path Γ_t. First, note that for any $u \in \mathbb{R}^{d_\alpha}$ with $\|u\| = 1$, we have

$$u'\mathbb{E}\left(\nabla_\alpha g_t(\alpha_i)\nabla_\alpha g_t(\alpha_i)'\right)u = \mathbb{E}\left((A+B)^2\right),$$

where

$$A := \left(F^{(1)}(\gamma_t'\alpha_i) + g_{t-1}(\alpha_i)\rho_t^{(1)}(\alpha_i)\right)\gamma_t' u, \quad B := \rho_t(\alpha_i)\nabla_\alpha g_{t-1}(\alpha_i)'u.$$

Then for any $\mu \in (0,1)$, since $2|AB| \leqslant \mu A^2 + 1/\mu B^2$, [15] it follows that

$$u'\mathbb{E}\left(\nabla_\alpha g_t(\alpha_i) g_t(\alpha_i)'\right)u \geqslant (1-\epsilon)\mathbb{E}A^2 - \left(\frac{1}{\epsilon} - 1\right)\mathbb{E}B^2. \tag{E.3}$$

Also, since $\{\gamma_t\}_{t\in\mathbb{Z}}$ is i.i.d. across time, we have

$$\begin{aligned}\mathbb{E}\left(B^2\right) &\overset{\text{(i)}}{=} \mathbb{E}(\rho_t^2(\alpha_i))u'\mathbb{E}\left(\nabla_\alpha g_{t-1}(\alpha_i)\nabla_\alpha g_{t-1}(\alpha_i)'\right)u \\ &\overset{\text{(ii)}}{=} \mathbb{E}(\rho_t^2(\alpha_i))u'\mathbb{E}\left(\nabla_\alpha g_t(\alpha_i)\nabla_\alpha g_t(\alpha_i)'\right)u.\end{aligned} \tag{E.4}$$

Here, (i) follows from the fact that γ_t is independent of Γ_{t-1}, and (ii) follows because Γ_t and Γ_{t-1} are identically distributed. Combining (E.3) and (E.4) gives

$$\begin{aligned}\left(1 + \frac{\epsilon}{1-\epsilon}\mathbb{E}(\rho_t^2(\alpha_i))\right)u'\mathbb{E}\left(\nabla_\alpha g_t(\alpha_i)\nabla_\alpha g_t(\alpha_i)'\right)u \geqslant& (1-\epsilon)\mathbb{E}A^2 \\ \Rightarrow u'\mathbb{E}\left(\nabla_\alpha g_t(\alpha_i)\nabla_\alpha g_t(\alpha_i)'\right)u \geqslant& \frac{1-\epsilon}{1+\frac{\epsilon}{1-\epsilon}\mathbb{E}(\rho_t^2(\alpha_i))}\mathbb{E}(A^2) \\ \geqslant& C_1 u'\mathbb{E}(\gamma_t\gamma_t')u\end{aligned} \tag{E.5}$$

Since A can be rewritten as

$$\begin{aligned}A &= \left(F^{(1)}(\gamma_t'\alpha_i) + g_{t-1}(\alpha_i)\left(F^{(1)}\left(\beta + \gamma_t'\alpha_i\right) - F^{(1)}\left(\gamma_t'\alpha_i\right)\right)\right)\gamma_t' u \\ &= \left((1 - g_{t-1}(\alpha_i))F^{(1)}(\gamma_t'\alpha_i) + g_{t-1}(\alpha_i)F^{(1)}\left(\beta + \gamma_t'\alpha_i\right)\right)\gamma_t' u,\end{aligned}$$

15

$$\left(\sqrt{\mu}A - \frac{1}{\sqrt{\mu}}B\right)^2 = \mu A^2 + \frac{1}{\mu}B^2 - 2AB \geqslant 0 \quad \Rightarrow 2AB \leqslant \mu A^2 + \frac{1}{\mu}B^2$$

$$\left(\sqrt{\mu}A + \frac{1}{\sqrt{\mu}}B\right)^2 = \mu A^2 + \frac{1}{\mu}B^2 + 2AB \geqslant 0 \quad \Rightarrow 2AB \geqslant -\mu A^2 - \frac{1}{\mu}B^2$$

Thus, $2|AB| \leqslant \mu A^2 + 1/\mu B^2$.

where the first part is a weighted average of Since (i) α_i, γ_t are uniformly bounded over α_i and γ_t, and (ii) $F^{(1)}$ is a positive continuous function on its support, then $F^{(1)}(\gamma_t'\alpha_i)$ and $F^{(1)}(\beta + \gamma_t'\alpha_i)$ are bounded above and uniformly bounded away from 0 uniformly over α_i and γ_t. It immediately follows that, by definition, $|\rho_t(\alpha_i)| < 1$ uniformly over α_i and γ_t, and we let $C_1 > 0$ denote such upper bound. Last, by the representation (B.2), we obtain that $g_t(\alpha_i)$ is bounded above and uniformly bounded away from 0 uniformly over i and t. There, there exists a constant $C_2 > 0$, that is independent of n, T_0, such that α_i and γ_t,

$$\left((1 - g_{t-1}(\alpha_i))F^{(1)}(\gamma_t'\alpha_i) + g_{t-1}(\alpha_i)F^{(1)}(\beta + \gamma_t'\alpha_i)\right) \geqslant C_2.$$

Letting $\epsilon = 1/2$. Then, combining the lower bound with (E.5) yields that, for any α_i and γ_t,

$$\begin{aligned} u'\mathbb{E}\left(\nabla_\alpha g_t(\alpha_i)\nabla_\alpha g_t(\alpha_i)'\right)u &\geqslant \frac{1-\epsilon}{1+\frac{\epsilon}{1-\epsilon}C_1^2}C_2^2 u'\mathbb{E}\left(\gamma_t\gamma_t'\right)u \\ &\geqslant \underbrace{\frac{C_2^2}{2+2C_1^2}s_{\min}\left(\mathbb{E}\left(\gamma_t\gamma_t'\right)\right)}_{C_3}. \end{aligned} \tag{E.6}$$

Step 2: Taylor expansion Define $\rho_{g'} := \sup_{\alpha\in\mathcal{A},\Gamma\in\text{supp}(\Gamma_t)} \|\nabla_\alpha g(\alpha,\Gamma)\|$, and define

$$\rho_{g''} := \sup_{\alpha\in\mathcal{A},\Gamma\in\text{supp}(\Gamma_t)} s_{\max}\left(\nabla^2_{\alpha\alpha'}g(\alpha,\Gamma)\right)$$

as the upper bound of g's largest eigenvalue of the second-order derivative matrix. It is straightforward to verify that $\rho_{g'}, \rho_{g''} < \infty$, because the compactness of the support of (α_i, γ_t), the smoothness of F, and (B.2). For any $\alpha_i, \alpha_j \in \mathcal{A}$ and for any $\delta - 2 > 0$, consider the

second order Taylor expansions of $g(\alpha,\Gamma)$ and $g(\alpha_j,\Gamma)$ around α_i:

$$
\begin{aligned}
d(\alpha_i,\alpha_j) \geqslant & \sup_{\alpha\in\mathcal{A}} \left| \int (g(\alpha,\Gamma)-g(\alpha_i,\Gamma))(g(\alpha_i,\Gamma)-g(\alpha_j,\Gamma)) \mathrm{d}\mathbb{P}(\Gamma) \right| \\
\geqslant & \sup_{\alpha\in\mathcal{A}} \left\{ \left| (\alpha-\alpha_i)' \int \nabla_\alpha g(\alpha_i,\Gamma) \nabla_\alpha g(\alpha_i,\Gamma)' \mathrm{d}\mathbb{P}(\Gamma)(\alpha_i-\alpha_j) \right| \right. \\
& \left. - \frac{1}{2}\rho_{g'}\rho_{g''} \left(\|\alpha-\alpha_i\|^2\|\alpha_i-\alpha_j\| + \|\alpha-\alpha_i\|\|\alpha_i-\alpha_j\|^2 \right) - \frac{1}{4}\rho_{g''}^2 \|\alpha-\alpha_i\|^2\|\alpha_i-\alpha_j\|^2 \right\} \\
\geqslant & \sup_{\alpha\in\mathcal{A}\cap\mathcal{B}(\alpha_i,\delta_2)} \left\{ \left| (\alpha-\alpha_i)'' \mathbb{E}\left(\nabla_\alpha g_t(\alpha_i) \nabla_\alpha g_t(\alpha_i)' \right)(\alpha_i-\alpha_j) \right| \right. \\
& \left. - \frac{1}{2}\rho_{g'}\rho_{g''} \left(\|\alpha-\alpha_i\|^2\|\alpha_i-\alpha_j\| + \|\alpha-\alpha_i\|\|\alpha_i-\alpha_j\|^2 \right) - \frac{1}{4}\rho_{g''}^2 \|\alpha-\alpha_i\|^2\|\alpha_i-\alpha_j\|^2 \right\} \\
\overset{\text{(i)}}{\geqslant} & C_3 \min\left\{ \frac{1}{2}, \frac{\delta_2}{2D} \right\} \overline{\rho}\|\alpha_1-\alpha_2\| - \frac{1}{2}\rho_{g'}\rho_{g''}\|\alpha_i-\alpha_j\|^2\delta_2 \\
& - \left(\frac{1}{2}\rho_{g'}\rho_{g''}\|\alpha_i-\alpha_j\| + \frac{1}{4}\rho_{g''}^2\|\alpha_i-\alpha_j\|^2 \right) \delta_2^2.
\end{aligned}
$$

Here, (i) follows from (E.6) and Lemma E.2, where $D := \sup_{\alpha_1,\alpha_2\in\mathcal{A}} \|\alpha_1-\alpha_2\| < \infty$, and $\overline{\rho} := \sup_{\alpha\in\mathcal{A}} \{\rho \geqslant 0 : \ B(\alpha,\rho) \subset \mathcal{A}\}$. One can choose a sufficiently small $\delta_2^* > 0$ such that

$$
C_3 \min\left\{ \frac{1}{2}, \frac{\delta_2^*}{2D} \right\} \overline{\rho}\|\alpha_1-\alpha_2\| \geqslant \left(\rho_{g'}\rho_{g''} + \frac{1}{2}\rho_{g''}^2 D \right) \|\alpha_i-\alpha_j\|\delta_2^{*2},
$$

and consequently,

$$
d(\alpha_i,\alpha_j) \geqslant \frac{1}{2} C_3 \min\left\{ \frac{1}{2}, \frac{\delta_2^*}{2D} \right\} \overline{\rho}\|\alpha_1-\alpha_2\| - \frac{1}{2}\rho_{g'}\rho_{g''}\|\alpha_i-\alpha_j\|^2\delta_2^*.
$$

Also, we choose a sufficiently small $\delta_1^* > 0$ such that $\delta_1^* \leqslant \frac{C_3 \min\left\{ \frac{1}{2}, \frac{\delta_2^*}{2D} \right\} \overline{\rho}}{2\rho_{g'}\rho_{g''}\delta_2^*}$. It is straightforward to verify that

$$
d(\alpha_i,\alpha_j) \geqslant \frac{1}{4} C_3 \min\left\{ \frac{1}{2}, \frac{\delta_2^*}{2D} \right\} \overline{\rho}\|\alpha_1-\alpha_2\|, \quad \forall \alpha_1,\alpha_2 \in \mathcal{A}, \text{ and } \|\alpha_1-\alpha_2\| \leqslant \delta_1^*. \tag{E.7}
$$

Combining (E.2) and (E.7), we have that, for any $\alpha_1,\alpha_2 \in \mathcal{A}$,

$$
d(\alpha_i,\alpha_j) \geqslant \min\left\{ \frac{\eta(\delta_1^*)}{D}, \frac{1}{4} C_3 \min\left\{ \frac{1}{2}, \frac{\delta_2^*}{2D} \right\} \overline{\rho} \right\} \|\alpha_1-\alpha_2\|.
$$

This completes the proof. ■

Lemma E.1 *Suppose $\mathcal{A} \subset \mathbb{R}^{d_\alpha}$ is nonredundant, compact, and convex. Given a symmetric positive definite matrix $\Omega \in \mathbb{R}^{d_\alpha \times d_\alpha}$. Then, there exists a constant $c > 0$ such that for any $\alpha_1, \alpha_2, \alpha_3 \in \mathcal{A}$,*

$$\sup_{\alpha \in \mathcal{A}} |(\alpha - \alpha_3)' \Omega (\alpha_2 - \alpha_1)| \geqslant c \|\alpha_1 - \alpha_2\|$$

Proof of Lemma E.1. By the nonredundancy of $\mathcal{A}$, there exist $\{\tilde{\alpha}_0, \tilde{\alpha}_1, \ldots, \tilde{\alpha}_d\} \subset \mathcal{A}$ such that $\{\tilde{\alpha}_1 - \tilde{\alpha}_0, \ldots, \tilde{\alpha}_d - \tilde{\alpha}_0\}$ are linear independent. We collect $\{\tilde{\alpha}_1 - \tilde{\alpha}_0, \ldots, \tilde{\alpha}_d - \tilde{\alpha}_0\}$ into matrix

$$A = \begin{pmatrix} \tilde{\alpha}_1' - \tilde{\alpha}_0' \\ \vdots \\ \tilde{\alpha}_d' - \tilde{\alpha}_0' \end{pmatrix},$$

and it is straightforward to verify that A is invertible. In addition,

$$\begin{aligned}
&\sup_{\alpha \in \mathcal{A}} |(\alpha - \alpha_3)' \Omega (\alpha_2 - \alpha_1)| \\
=& \max\{\sup_{\alpha \in \mathcal{A}} \alpha' \Omega (\alpha_2 - \alpha_1) - \alpha_3' \Omega (\alpha_2 - \alpha_1), \alpha_3' \Omega (\alpha_2 - \alpha_1) - \inf_{\alpha \in \mathcal{A}} \alpha' \Omega (\alpha_2 - \alpha_1)\} \\
\overset{\text{(i)}}{\geqslant}& \frac{1}{2} \left(\sup_{\alpha \in \mathcal{A}} \alpha' \Omega (\alpha_2 - \alpha_1) - \inf_{\alpha \in \mathcal{A}} \alpha' \Omega (\alpha_2 - \alpha_1) \right) \\
\geqslant& \frac{1}{2} \left(\max_{\alpha \in \{\tilde{\alpha}_0, \tilde{\alpha}_1, \ldots, \tilde{\alpha}_d\}} \alpha' \Omega (\alpha_2 - \alpha_1) - \min_{\alpha \in \{\tilde{\alpha}_0, \tilde{\alpha}_1, \ldots, \tilde{\alpha}_d\}} \alpha' \Omega (\alpha_2 - \alpha_1) \right) \\
=& \frac{1}{2} \left(\max_{\alpha \in \{\tilde{\alpha}_0, \tilde{\alpha}_1, \ldots, \tilde{\alpha}_d\}} (\alpha - \tilde{\alpha}_0)' \Omega (\alpha_2 - \alpha_1) - \min_{\alpha \in \{\tilde{\alpha}_0, \tilde{\alpha}_1, \ldots, \tilde{\alpha}_d\}} (\alpha - \tilde{\alpha}_0)' \Omega (\alpha_2 - \alpha_1) \right) \\
=& \frac{1}{2} \max\{0, \tilde{\alpha}_1' \Omega (\alpha_2 - \alpha_1), \ldots, \tilde{\alpha}_d' \Omega (\alpha_2 - \alpha_1)\} - \frac{1}{2} \min\{0, \tilde{\alpha}_1' \Omega (\alpha_2 - \alpha_1), \ldots, \tilde{\alpha}_d' \Omega (\alpha_2 - \alpha_1)\} \\
\geqslant& \frac{1}{2} \|A \Omega (\alpha_2 - \alpha_1)\|_\infty.
\end{aligned}$$

Here, (i) holds because $\max\{a, b\} \geqslant \frac{a+b}{2}$. In addition, $\frac{1}{2}\|A\Omega(\alpha_2 - \alpha_1)\|_\infty \geqslant \frac{1}{2\sqrt{d}}\|A\Omega(\alpha_2 - \alpha_1)\| \geqslant \frac{\sigma_{\min}(A\Omega)}{2\sqrt{d}}\|(\alpha_2 - \alpha_1)\|$. Let $c = \frac{\sigma_{\min}(A\Omega)}{2\sqrt{d}}$. Since Ω is positive definite and A is invertible, $c > 0$. This completes the proof. ■

Lemma E.2 *Suppose $\mathcal{A} \subset \mathbb{R}^{d_\alpha}$ is nonredundant, compact, and convex. Given a symmetric positive definite matrix $\Omega \in \mathbb{R}^{d_\alpha \times d_\alpha}$ such that the smallest eigenvalue of Ω satisfies $s_{\min}(\Omega) \geqslant \underline{s} > 0$. Then, for any given $\delta > 0$ and any $\alpha_1, \alpha_2 \in \mathcal{A}$, there exists a constant $c(\delta) > 0$ such*

that

$$\sup_{\alpha\in\mathcal{A}\cap\mathcal{B}(\alpha_1,\delta)} |(\alpha-\alpha_1)'\Omega(\alpha_1-\alpha_2)| \geqslant \min\left\{\frac{1}{2}, \frac{\delta}{2D}\right\}\overline{\rho}\underline{s}\|\alpha_1-\alpha_2\|,$$

where $D := \sup_{\alpha_1,\alpha_2\in\mathcal{A}}\|\alpha_1-\alpha_2\| < \infty$, *and* $\overline{\rho} := \sup_{\alpha\in\mathcal{A}}\{\rho \geqslant 0 : \ B(\alpha,\rho)\subset\mathcal{A}\}$.

Proof of Lemma E.2. By definition, $\overline{\rho}$ is the largest radius of a Euclidean ball that can be placed entirely inside $\mathcal{A}$. Since $\mathcal{A}$ is compact, convex, and affine(supp) $= \mathbb{R}^{d_\alpha}$, such balls exist and $0 < \overline{\rho} < \infty$. Pick one of these balls and suppose its center is α_0, then $\mathcal{B}(\alpha_0,\overline{\rho})\subset\mathcal{A}$. Define $\ell := \min\{1/2, \delta/2D\}$ and $\tilde{\alpha}_1 = (1-\ell)\alpha_1 + \ell\alpha_0$. In addition, define a new ball as

$$\mathcal{B}(\tilde{\alpha}_1, \ell\overline{\rho}) := \{(1-\ell)\alpha_1 + \ell\alpha \mid \alpha_0 \in \mathcal{B}(\alpha_0,\overline{\rho})\}.$$

It is straightforward to verify that $\mathcal{B}(\tilde{\alpha}_1,\ell\overline{\rho})\subset\mathcal{B}(\alpha_1,\delta)$, and by the convexity of $\mathcal{A}$, $\mathcal{B}(\tilde{\alpha}_1,\ell\overline{\rho})\subset\mathcal{A}$ as well. Then,

$$\begin{aligned}
&\sup_{\alpha\in\mathcal{A}\cap\mathcal{B}(\alpha_1,\delta)} |(\alpha-\alpha_1)'\Omega(\alpha_1-\alpha_2)| \\
\geqslant &\sup_{\alpha\in\mathcal{B}(\tilde{\alpha}_1,\ell\overline{\rho})} |(\alpha-\alpha_1)'\Omega(\alpha_1-\alpha_2)| \\
\geqslant &\max\left\{\left|\ell(\alpha_1-\alpha_0)'\Omega(\alpha_2-\alpha_1) + \frac{\ell\overline{\rho}}{\|\alpha_1-\alpha_2\|}(\alpha-\alpha_1)'\Omega(\alpha_1-\alpha_2)\right|,\right. \\
&\qquad\left.\left|\ell(\alpha_1-\alpha_0)'(\alpha_2-\alpha_1) - \frac{\ell\overline{\rho}}{\|\alpha_1-\alpha_2\|}(\alpha-\alpha_1)'\Omega(\alpha_1-\alpha_2)\right|\right\} \\
\geqslant &\ell\overline{\rho}\underline{s}\|\alpha_2-\alpha_1\| \\
\overset{\text{(i)}}{\geqslant} &\min\left\{\frac{1}{2}, \frac{\delta}{2D}\right\}\overline{\rho}\underline{s}\|\alpha_1-\alpha_2\|.
\end{aligned}$$

Here, (i) follows from the fact that $\max\{|a+b|, |a-b|\} \geqslant |b|$. This completes the proof. ∎

F Supplementary Proofs

Lemma F.1 (Based on Theorem 1 of Merlevède, Peligrad and Rio (2009)) *When (i) $(Y_t)_{t=1}^T$ is mean-zero and sub-Gaussian, i.e., there exist constant $\sigma > 0$ such that $\max_{t=1,\ldots,T} \mathbb{E}(\exp(\lambda Y_t)) \leqslant \exp(\frac{\lambda^2\sigma^2}{2})$, and (ii) $(Y_t)_{t=1}^T$ is α-mixing, and for a constant $c > 0$, $\sup_{\tau \geqslant 1} \alpha(\tau) \leqslant \exp(-c\tau)$. Then, for any $\epsilon > 0$, there exists constants $C_1, C_2, C_3, C_4 > 0$ such that*

$$\begin{aligned}\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T} Y_t\right| \geqslant \epsilon\right) \leqslant & T\exp\left(-\frac{T^{2/3}\epsilon^{2/3}}{C_1}\right) + \exp\left(-\frac{T\epsilon^2}{C_2}\right) \\ & + \exp\left(-\frac{T\epsilon^2}{C_3}\exp\left(\frac{T^{2/9}\epsilon^{2/9}}{C_4\left(\log T\right)^{2/3}}\right)\right).\end{aligned} \tag{F.1}$$

Lemma F.2 (Hoeffding inequality) *Suppose that the $\{X_i\}_{i=1}^n$ are conditionally independent across i given $\{Z_i\}_{i=1}^n$. Assume that $\mathbb{E}(X_i \mid \{Z_i\}_{i=1}^n) = 0$ for each i, and that, for some constant $\sigma > 0$,*

$$\mathbb{E}\left(\exp(\lambda X_i) \mid \{Z_i\}_{i=1}^n\right) \leqslant \exp\left(\frac{\sigma^2\lambda^2}{2}\right),$$

for all $\lambda \in \mathbb{R}$ and all i. Then, for every $\epsilon > 0$,

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} Z_i X_i\right| \geqslant \epsilon \mid \{Z_i\}_{i=1}^n\right) \leqslant 2\exp\left(-\frac{\epsilon^2}{2\sigma^2\sum_{i=1}^n Z_i^2}\right).$$

Lemma F.3 (Bernstein inequality) *Suppose that, conditional on random variables $(Z_i)_{i=1}^n$, $\{X_i\}_{i=1}^n$ are independent and satisfy $\mathbb{E}(X_i \mid (Z_i)_{i=1}^n) = 0$ for each $i = 1, \ldots, n$. Define*

$$B_n := \max_{i=1,\ldots,n} |Z_i X_i|, \quad V_n := \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left(Z_i^2 X_i^2 \mid (Z_i)_{i=1}^n\right)$$

Then, for any $\epsilon > 0$,

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} Z_i X_i\right| \geqslant \epsilon \mid (Z_i)_{i=1}^n\right) \leqslant 2\exp\left(-\frac{\epsilon^2}{2\left(nV_n + \frac{B_n}{3}\epsilon\right)}\right). \tag{F.2}$$

In addition, when there exist constants $b, v \in (0, \infty)$ such that $B_n \leqslant b$ and $V_n \leqslant v$ with probability approaching to 1, then

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} Z_i X_i\right| \geqslant \epsilon\right) \leqslant 2\exp\left(-\frac{\epsilon^2}{2\left(nv + \frac{b}{3}\epsilon\right)}\right) + r_n, \tag{F.3}$$

with $r_n \to 0$ *as* $n \to \infty$.

Proof of Lemma F.3. Conditional on $(Z_i)_{i=1}^n$, the variables $(Z_i X_i)_{i=1}^n$ are independent, have conditional mean zero, and satisfy the stated boundedness and conditional second moment restrictions. Therefore, the Bernstein inequality (Wainwright (2019, Proposition 2.14)) implies the conditional bound (F.2). Define $\mathcal{A} := \{B_n \leqslant b, V_n \leqslant v\}$. Taking expectations with respect to $(Z_i)_{i=1}^n$ (on $\mathcal{A}$ and $\mathcal{A}^c$ separately) yields the stated unconditional bound (F.3). ■